Lecture Notes on:

Broadband Circuits for Optical Fiber Communication

Eduard Säckinger

Agere Systems
formerly
Bell Laboratories
Lucent Techologies
Holmdel, NJ

Contents

Preface

Purpose. These lecture notes have been written as a supplement to my lecture on "Broadband Circuits for Optical Fiber Communication" which I have taught on several occasions (VLSI Symposium, June 2000; MEAD Microelectronics, three times in 2001). Since the lecture time is usually limited to $1-3$ hours, it is impossible to treat the whole subject in a thorough way. These notes provide the interested student with more explanations and more details on the subject.

Scope. We will discuss five types of broadband circuits: transimpedance amplifiers, limiting amplifiers, automatic gain control (AGC) amplifiers, laser drivers, and modulator drivers. Some background information about optical fiber, receiver theory, photodetectors, lasers, and modulators is provided because it is important to understand the environment in which the circuits are used. The most important specifications for all five circuit types are explained and illustrated with example numerical values. In many IC design projects a significant amount of time is spent figuring out the right specs for the new design. It is therefore important to understand how these specs relate to the system performance. Then, circuit concepts for all five circuit types are discussed and illustrated with practical implementations taken from the literature. A broad range of implementations in MESFET, HFET, BJT, HBT, BiCMOS, and CMOS technologies are covered. Finally, a brief overview of product examples and current research topics is given.

Emphasis is on circuits for digital continuous-mode transmission which are used for example in SONET, SDH, Gigabit-Ethernet, and 10 Gb/s Ethernet applications. Furthermore, we concentrate on high-speed circuits in the range $2.5\,\text{Gb/s}-40\,\text{Gb/s}$, typically used in back-bone networks. We will have a quick look at circuits for burst-mode transmission which are used for example in access networks such as Passive Optical Networks (PON). These circuits typically operate at a lower speed in the range $50\,\text{Mb/s}-1.25\,\text{Gb/s}$.

It is assumed that the reader is familiar with basic analog IC design as presented for example in Gray and Meyer: *Analysis and Design of Analog Integrated Circuits* or a similar book [GM77, AH87, JM97].

Style. My aim has been to present an overview of the field, with emphasis on an intuitive and qualitative understanding without attempting any real mathematical rigor, but I hope that there are no gross omissions or errors which will seriously offend the specialist. Many references to the literature are made throughout this text to guide the interested reader to a more complete and in-depth treatment of the various topics. In general, the mind-set and notation used are those of an Electrical Engineer rather than that of a Physicist. I

hope this text may be useful to students or professionals who might wish for some survey of this subject without becoming embroiled in too much technical detail.

Acknowledgments. I am deeply indebted to the many reviewers who have given freely of their time to read through the text, in part or in full. In particular, I am most grateful to Hercules Avramopoulos, Kamran Azadet, Renuka Jindal, Helen H. Kim, Patrik Larsson, Marc Loinaz, Nicolas Nodenot, Yusuke Ota, Joe H. Othmer, Hans Ransijn, Behzad Razavi, Fadi Saibi, Leilei Song, Jim Yoder, [**your name**].

Despite the effort made, there are undoubtedly many errors, omissions, and awkward explanations left in this text. If you have any corrections or suggestions, please e-mail them to `edi@agere.com`. Thank you!

Chapter 1

Introduction

Optical Transceiver. Figure 1.1 shows a block diagram of a typical optical transceiver front-end.[1] The optical signal from the fiber is received by a *Photodetector* (PD) which produces a small output current proportional to the optical signal. This current is amplified and converted to a voltage by the *Transimpedance Amplifier* (TIA or TZA). The voltage signal is further amplified by either a *Limiting Amplifier* (LA) or an *Automatic Gain Control Amplifier* (AGC). The LA and AGC amplifier are collectively known as *Main Amplifiers* (MA) or *Post Amplifiers*. The resulting signal, which is now several 100 mV strong, is fed into a *Clock and Data Recovery Circuit* (CDR) which extracts the clock signal and retimes the data signal. In high-speed receivers, a *Demultiplexer* (DMUX) converts the fast serial data stream into n parallel lower-speed data streams which can be processed conveniently by the digital logic block. Some CDR designs (parallel sampling architectures) perform the DMUX task as part of their functionality and an explicit DMUX is not needed in this case [HG93]. The digital logic block descrambles the bits, performs error checks, extracts the payload data from the framing information, synchronizes to another clock domain, etc.

On the transmitter side the same process happens in reverse order. The parallel data from the digital logic block is merged into a single high-speed data stream using a *Multiplexer* (MUX). In order to control the select lines of the MUX, a bit-rate (or half-rate) clock must be synthesized from the slower word clock. This task is performed by the *Clock Synthesizer*. Finally, a *Laser Driver* or *Modulator Driver* drives the optical transmitter. A laser driver modulates the current of a *Laser Diode*, whereas a modulator driver modulates the voltage across a *Modulator* which in turn modulates the light from a *Continuous Wave* (CW) laser. Some laser/modulator drivers also require a bit-rate (or half-rate) clock which is used to retime the data bits. This clock is supplied by the clock synthesizer.

In this text we will discuss the transimpedance amplifier, limiting amplifier, automatic gain control amplifier, laser driver, and modulator driver in more detail.

[1] The term transceiver is a contraction of the words "*trans*mitter" and "re*ceiver*" and we use it here to mean any combination of transmitter and receiver. However, some authors make a distinction between *Transceiver* (only PD, TIA, LA/AGC, Driver, and LD of Fig. 1.1) and *Transponder* (transceiver plus CDR, DMUX, Clock Synthesis, and MUX).

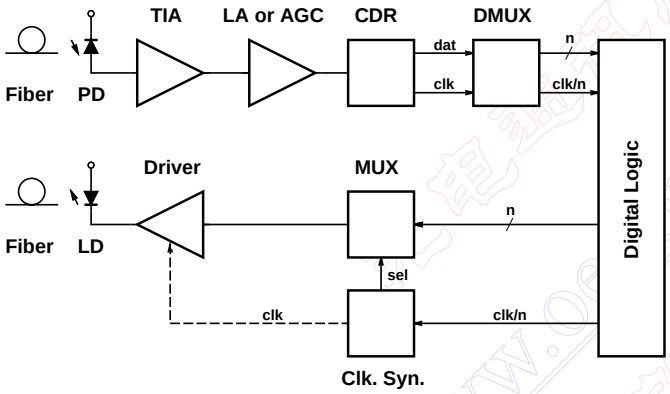

Figure 1.1: Block diagram of an optical transceiver front-end.

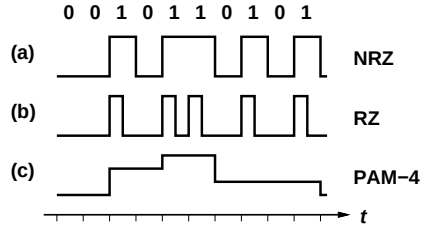

Figure 1.2: Modulation schemes: (a) NRZ, (b) RZ, and (c) PAM-4.

Modulation Schemes. The most commonly used modulation format in optical communication is the *Non-Return to Zero* (NRZ) format shown in Fig. 1.2(a). This format is a form of *On-Off Keying* (OOK): The signal is *on* to transmit a one and *off* to transmit a zero. When the signal (i.e., the light source) is on, it stays on for the entire bit period. For example transmitting the periodic sequence "010101010…" at 10 Gb/s in NRZ format results in a 5 GHz square wave with 50 % duty cycle.[2]

In high-speed and long-distance transmission (e.g., fiber links between two continents) the *Return to Zero* (RZ) format, shown in Fig. 1.2(b), is preferred. The shorter pulses, in contrast to the NRZ format, occupy only a fraction (e.g., 50 %) of the bit period. The advantages are that the required signal-to-noise ratio is lower[3] and that more pulse distortion and spreading can be tolerated for this narrow pulse format without disturbing the adjacent bits. The latter makes the RZ format more immune to effects of fiber non-linearity and polarization-mode dispersion. On the downside, an RZ modulated signal occupies more bandwidth than the corresponding NRZ signal at the same bit rate. Therefore RZ-modulated wavelengths can be packed less densely in a DWDM system and faster, more expensive transmission equipment (laser/modulator, detector, front-end electronics, etc.) is required.

In *Community-Antenna Television* (CATV) systems the TV signal is often first transported *optically* from the distribution center to the neighborhood, then it is distributed to the individual homes on coax cable. This combination is called *Hybrid Fiber-Coax* (HFC) and has the advantage over an all-coax system that it saves many electronic amplifiers (the loss in a fiber is much lower than the loss in a coax cable) and provides better signal quality. In the optical part of the system, the laser light is modulated with multiple RF carriers corresponding to the TV channels. This is called *Sub-Carrier Multiplexing* (SCM): The laser light is the main carrier and the RF signals are the subcarriers. In turn, each subcarrier is modulated using, for example, *Amplitude Modulation with Vestigial Sideband* (AM-VSB) for analog TV or *Quadrature Amplitude Modulation* (QAM) for digital TV. A QAM-signal can be decomposed into two independent *Pulse Amplitude Modulation* (PAM) signals. Fig. 1.2(c) shows an example for 4-level PAM modulation in which groups of two successive bits are encoded with a particular signal level. Compared to the NRZ signal, this PAM-4 signal requires a higher signal-to-noise ratio for reliable reception but occupies a narrower bandwidth due to its reduced symbol rate.

In the remainder of this text we will always assume that NRZ modulation is used, unless stated otherwise.

Line Codes. Before data is modulated onto the optical carrier it is usually preconditioned with a a so-called *Line Code*. Line coding provides the transmitted data stream with the following desirable properties: *DC Balance*, short *Run Lengths*, and a

[2] In some standards the *Non-Return to Zero change on Ones* (NRZ1 or NRZI) format is used (e.g., in Fast Ethernet, FDDI, etc.). This format is just a combination of a simple line code with NRZ modulation. The NRZ1 line-code changes its (binary) value when the bit to be transmitted is a one and leaves the output value unchanged when the bit is a zero.

[3] To receive data at a bit-error rate of 10^{-12}, we need a signal-to-noise ratio of 16.9 dB for NRZ modulation, 15.7 dB for 50 %-RZ modulation, and 23.9 dB for PAM-4 modulation assuming additive, Gaussian noise.

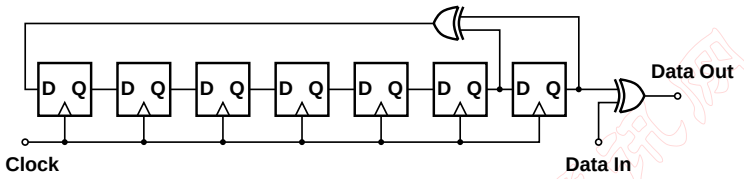

Figure 1.3: Implementation of a SONET scrambler.

high *Transition Density*. A DC balanced data stream contains the same number of zeros and ones on average. This is equivalent to saying that the average *Mark Density* (number of one-bits divided by all bits) is 50 %. A DC-balanced signal has the nice property that its average value (DC component) is always centered halfway between the zero and one levels (1/2 of the peak-to-peak value). This property often permits the use of AC coupling between circuit blocks, simplifying their design. Furthermore, it is desirable to keep the number of successive zeros and ones, the run length, to a small value. This provision reduces the low-frequency content of the transmitted signal and avoids the associated *DC Wander* problem. Finally, a high transition density is desirable to simplify the clock recovery process.

In practice line coding is implemented by either *Scrambling*, *Block Coding*, or a combination of the two:

- *Scrambling*. In this case a *Pseudo-Random Bit Sequence* (PRBS) is generated with a feedback shift register and xor'ed with the data stream (see Fig. 1.3). Scrambling provides DC balance without adding overhead bits to the data stream thus preserving the bit rate. On the down side, the maximum run length (successive runs of zeros or ones) is not strictly limited, i.e., there is a small chance for very long runs of zeros or ones which can be hazardous. In practice it is often assumed that runs are limited to 72 bits. The scrambling method is used in the U.S. telecommunication system described in the SONET (Synchronous Optical NETwork) standard [Bel95, Bel98] and the almost identical SDH (Synchronous Digital Hierarchy) standard used in Europe and Japan.

- *Block Coding*. In this case groups of bits (a block) are replaced by another slightly larger group of bits such that the average mark density becomes 50 % and DC balance is guaranteed. For example, in the 8B10B Code 8-bit groups are replaced with 10-bit patterns. The 8B10B code increases the bit rate by 25 %, however, the maximum run length is strictly limited to 5 zeros or ones in a row. The 8B10B code is used in the *Gigabit Ethernet* (GbE, 1000Base-SX, 1000Base-LX) and *Fiber Channel* data communication system.[4]

- *Combination*. In the serial *10-Gigabit Ethernet* (10GbE) system DC balance is achieved by first scrambling the signal and then applying a block-code (64B66B Code) to it. This combination has the benefit of a low overhead ($\approx 3\,\%$ increase in bit rate) and a strictly limited run length.

[4]The 4B5B code used in Fast Ethernet, FDDI, etc. does not achieve perfect DC balance. The worst-case unbalance is 10 % [RS98].

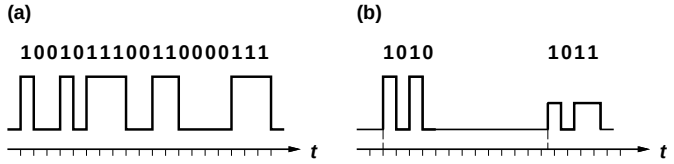

Figure 1.4: (a) Continuous-mode vs. (b) burst-mode signals.

Continuous Mode vs. Burst Mode. It is important to distinguish two types of transmission modes because they call for different circuit designs: *Continuous Mode* and *Burst Mode*. The signals corresponding to these two modes are shown schematically in Fig. 1.4.

In *Continuous-Mode Transmission* a continuous, uninterrupted stream of bits is transmitted as shown in Fig. 1.4(a). The transmitted signal is *DC Balanced* using one of the line codes described earlier. The resulting properties permit the use of AC coupled circuits.

In *Burst-Mode Transmission* data is transmitted in short bursts; in between bursts the transmitter remains silent. See Fig. 1.4(b) for an example, but note that real bursts are much longer than those shown in the figure, typically longer than 400 bits.

Bursts can be fixed or variable in length. Bursts which contain ATM (Asynchronous Transfer Mode) cells have a *fixed length*, they always contain 53 bytes plus some overhead (e.g., 3 bytes). Bursts which contain *Ethernet* frames have a *variable length* (70 − 1524 bytes). In either case, the bursts start out with a preamble (a.k.a. overhead) followed by the payload. The burst-mode receiver uses the preamble to synchronize its clock and establish the logic threshold level (slice level). In PON systems (see below) bursts arrive *asynchronously* and with strongly *varying power levels* (up to 30 dB), therefore the clock must be synchronized and the slice level adjusted for every single burst.

The average value (DC component) of a burst-mode signal is varying with time, depending on the burst activity. If the activity is high, it may be close to the halfway point between the zero and one levels, as in continuous mode systems; if the activity is low, the average comes arbitrarily close to the zero level. This means that the signal is *not DC balanced* and in general we cannot use AC coupling, in order to avoid DC wander. (Note that the mark density *within a burst* may well be 50 %, but the overall signal is still not DC balanced.) This lack of DC balance and the fact that bursts arrive with varying amplitudes necessitate specialized amplifier and driver circuits for burst-mode applications. Furthermore, the asynchronous arrival of the bursts requires specialized fast-locking CDRs.

In the remainder of this text we will always assume that we are dealing with DC-balanced, continuous-mode signals except if stated otherwise like in the sections on burst-mode circuits.

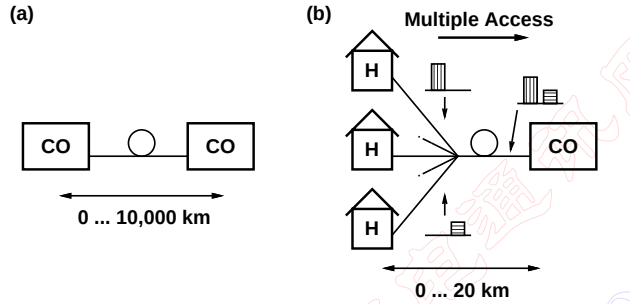

Figure 1.5: Example of (a) point-to-point link and (b) point-to-multipoint network.

Point-to-Point vs. Point-to-Multipoint Networks. There are two important types of optical networks: the simple *Point-to-Point Connection* and the *Point-to-Multipoint Network*.

An optical *Point-to-Point Connection* between two *Central Offices* (CO) is illustrated schematically in Fig. 1.5(a). An example for this is a SONET OC-192 link operating at 10 Gb/s (9.953 28 Gb/s to be precise) carrying about 130,000 voice calls. Point-to-point links are used over a wide range of bit rates and distances, from short computer-to-computer links to intercontinental undersea lightwave systems.

Point-to-point connections can be assembled into more complex structures such as *Ring Networks* and *Active Star Networks*. Examples for ring networks are provided by SONET/SDH rings and FDDI token rings. An active star is formed for example by Gigabit Ethernet or 10-Gigabit Ethernet links converging into a hub. It is important to realize that each individual optical connection of the star has an electrical receiver/transmitter at each end and therefore forms an optical point-to-point link. This is in contrast to a *Passive Star Network* or an optical *Point-to-Multipoint Network* where multiple optical fibers are coupled with a passive optical device.

Continuous-mode transmission is used on most point-to-point connections. In particular this is the case for unidirectional point-to-point connections as well as for bidirectional connections with two fibers – one for each direction, so-called *Space Division Multiplexing* (SDM) systems. Continuous-mode transmission is further used for bidirectional point-to-point connections employing two wavelengths for the two directions, so-called *Wavelength Division Multiplexing* (WDM) systems. However, burst-mode transmission is required for bidirectional point-to-point connections over a single fiber and a single wavelength. In this case, bidirectional communication is achieved by periodically reversing the direction of traffic in a ping-pong fashion, so-called *Time Compression Multiplexing* (TCM). For efficiency reasons, the TCM method is limited to short links such as in home networks.

An optical *Point-to-Multipoint Network* is illustrated schematically in Figure 1.5(b). This network is also known as a *Passive Optical Network* (PON) because it consists only of passive components such as the fibers and an optical power splitter/combiner

shown at the center. Since a PON does not require outside power supplies for switches, amplifiers, repeaters, etc. it is low in cost, easy to maintain, and reliable. PON is an attractive solution for *Fiber-To-The-Home Systems* (FTTH) where the fiber connection reaches all the way from the *Central Office* (CO) to the *Homes* (H). PON FTTx networks are limited to relatively small distances (< 20 km) and are currently operated at modest bit rates ($50 - 622$ Mb/s). Note that a PON constitutes a *shared* optical medium very different from a collection of point-to-point connections.

The most popular PON systems are ATM-PON (Asynchronous Transfer Mode - Passive Optical Network) [Kil96] defined in the FSAN (Full Service Access Network) standard [FSA01, IT98] and EPON (Ethernet - Passive Optical Network) currently being studied by an IEEE task force [IEE01]. In a typical ATM-PON FTTH scenario, $16 - 32$ homes share a bit rate of 155 Mb/s, giving each subscriber an average speed of $5 - 10$ Mb/s. This is sufficient for fast Internet access, telephone, and TV service.

In these PON systems (ATM-PON or EPON) bidirectional transmission is implemented with two different wavelengths, which is known as *Wavelength Division Multiplexing* (WDM). For *Downstream Communication* the CO broadcasts data to all subscribers (homes) in sequential order and each subscriber selects the information with the appropriate address. This method is called *Time Division Multipexing* (TDM) and continuous-mode transmission is used for the downstream direction. For *Upstream Communication* all subscribers transmit to the CO on the same wavelength through a shared medium making *Data Collisions* a concern. To avoid such collisions only one subscriber at a time is allowed to transmit a *burst* of data. Therefore, *Burst-Mode Transmission* is required for the upstream direction. Furthermore, a protocol coordinating who is allowed to send a burst at which time is needed. In PON systems the *Time Division Multiple Access* (TDMA) protocol is used for this purpose.

Another type of PON system, WDM-PON, assigns a different wavelength to each subscriber. In this way data collisions are avoided without the need for burst-mode transmission. However, the optical WDM components required for such a system are currently too expensive making WDM-PON uneconomical.

Further Reading. Useful information about optical fibers, lasers, detectors, optical amplifiers, as well as optical networks can be found in [Agr97, RS98]. The book [Wan99] contains an interesting collection of papers on high-speed circuits for lightwave communication. Finally, I recommend [Hec99] for an entertaining and informative historical account on fiber optics.

Chapter 2

Optical Fiber

It is useful to have a basic understanding of the fiber over which the optical signal is transmitted. However, it is well beyond the scope of this text to treat this subject in depth. Therefore, we will concentrate in this chapter just on what is needed for the remainder of this text, such as the relationship between pulse spreading and transmitter linewidth used in Chapter 7. For a full treatment of this subject there are many excellent books such as [Agr97, RS98].

2.1 Loss and Bandwidth

Loss. As the optical signal travels through a long stretch of fiber it is attenuated because of scattering, material impurities, and other effects. The attenuation is measured in dBs ($10 \cdot \log$ of power ratio) and is proportional to the length of the fiber. *Fiber Attenuation* or *Fiber Loss* is therefore specified in dB/km.

As shown in Fig. 2.1, *Silica Glass* has two low-absorption windows, one around the wavelength $\lambda = 1.3\,\mu$m and one around $\lambda = 1.55\,\mu$m, which are both used for optical fiber communication.[1] The popular single-mode fiber has a loss of about 0.25 dB/km at the $1.55\,\mu$m wavelength and 0.4 dB/km at the $1.3\,\mu$m wavelength. Because the loss is lower at $1.55\,\mu$m, this wavelength is preferred for long-distance communication. A third range of wavelengths around $\lambda = 0.85\,\mu$m where the loss is fairly high, about 2.5 dB/km, is used for short-distance (data) communication applications, mostly because low-cost optical sources and detectors are available at this wavelength.

The loss of modern silica-glass fiber is phenomenally low compared to that of an RF coax cable. A high-performance RF coax cable operating at 10 GHz has an attenuation of about 500 dB/km. Compare this to 0.25 dB/km for a fiber! On a historical note, it is interesting to know that low-loss fiber has not always been available. In 1965 the best glass fiber had a loss of around 1000 dB/km. It was estimated that for a fiber to be useful for optical communication its loss must be reduced to 20 dB/km or less, i.e., an improvement by *98 orders of magnitude* was required!! It is therefore understandable that in '65 most researchers thought that using glass fiber for optical communication was

[1] Note that these wavelengths, and all wavelengths we will refer to later, are defined in the *vacuum*. Thus an optical signal with $\lambda = 1.55\,\mu$m has a wavelength of about $1.1\,\mu$m in the fiber!

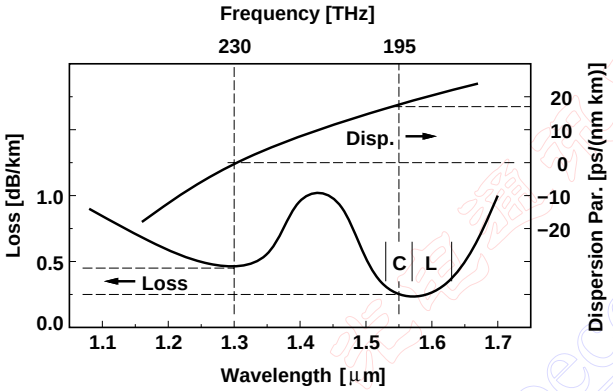

Figure 2.1: Loss of and dispersion parameter D of a standard single-mode fiber.

a hopelessly crazy idea. They spent their time working on "reasonable" approaches to optical communication such as metal pipes which contain periodically spaced lenses (so-called *Confocal Waveguides*) or pipes heated in such a way that the air in them formed *Gas Lenses*. Nevertheless, in 1970 a team of researchers at the *Corning Glass Works* managed to reduce the fiber loss below 20 dB/km by using ultra-pure silica glass rather than the ordinary compound glass. [Hec99] So, next time your circuit parameters are off by 98 orders or magnitude, don't give up ...

Plastic Optical Fiber (POF) is very low in cost and also permits the use of low-cost connectors (its core size is almost 1 mm in diameter). However, it has a huge loss of about 180 dB/km even when operated in the "low-loss" window at 0.65 μm (visible red). It is therefore restricted to very-short distance applications such as home networks and consumer electronics.

Although the loss in silica glass fiber is very low, it is still not low enough for ultra long-haul (e.g., intercontinental) communication. What can we do to further reduce the loss? First, we must operate the fiber at the 1.55 μm wavelength where loss is the lowest. Then, we can use *Optical In-Line Amplifiers* to periodically boost the signal. Two types of fiber amplifiers are in use: (i) the *Erbium-Doped Fiber Amplifier* (EDFA) which provides gain in the 1.55 μm band and (ii) the *Raman Amplifier* which provides distributed gain in the transmission fiber itself at a selectable wavelength (13 THz below the pump frequency).

Bandwidth. In addition to the very low loss, optical fiber also has a huge bandwidth. The low-loss window around the 1.55 μm wavelength is subdivided into two bands (C band for "conventional" and L band for "long-wavelength") and together they have a bandwidth of more than 10 THz (see Fig. 2.1). This means, for example, that a FTTH system with 100 users sharing a single fiber could be upgraded to about 100 Gb/s per user! This should be enough for the next few years and that's why FTTH advocates tout

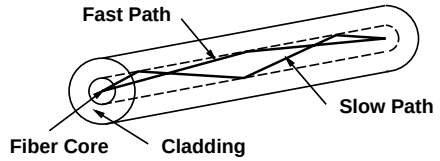

Figure 2.2: Modal dispersion in a multi-mode fiber.

their system as *future proof*.

Given that we have more than 10 THz of bandwidth, could we take a 1.55 μm laser, modulate it with a 10 Tb/s NRZ data stream and use this arrangement for optical transmission, at least in theory? No this would not work well! The received signal would be totally distorted already after a very short distance. The transmitted optical signal in our hypothetical system has a very large spectral width, filling all of the C and L band. Although each spectral component is in the low-loss window and arrives intact at the other end of the fiber, each component is delayed by a *different* amount and the superposition of all components, the received signal, is severely distorted. The dependence of delay on wavelength is known as chromatic dispersion and will be discussed in the next section.

It is therefore important to distinguish between two types of *Fiber Bandwidths*: The *bandwidth for the optical carrier*, which is very wide ($>$ 10 THz) and the *bandwidth for the modulation signal*, which is limited by dispersion and is much, much smaller. For example the modulation-signal bandwidth for 1 km of standard single-mode fiber at 1.55 μm is only a few 10 GHz. For a multi-mode fiber it is as small as 500 MHz! More on this bandwidth in Section 2.4.

Does this mean that we cannot really use the huge bandwidth that the fiber offers? Yes we can, if we use multiple optical carriers each one modulated at a modest bit rate. For example instead of one carrier modulated at 10 Tb/s we could use 1000 carriers, each one modulated at 10 Gb/s. This approach is known as *Dense Wavelength Division Multiplexing* (DWDM).

2.2 Dispersion

Modal Dispersion. An optical fiber consist of a core surrounded by a cladding where the cladding has a slightly lower refractive index than the core to guide the light beam by total internal reflection (see Fig. 2.2). In principle, air which has a lower refractive index than glass could act as the cladding. However, the fiber surface would then be extremely sensitive to dirt and scratches and two fibers touching each other would leak light. The invention of the *Clad Fiber* was a major breakthrough on the way to a practical optical fiber. [Hec99]

Depending on the size of the fiber core, there is only one or several pathways (modes) for the light beam to propagate through the fiber. If there are multiple pathways they have different propagation delays (see Fig. 2.2) and thus produce a distorted (spread out)

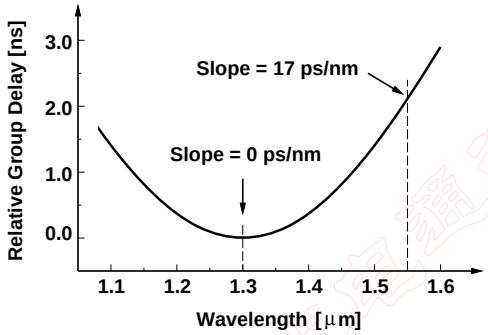

Figure 2.3: Relative group delay as a function of wavelength for 1 km of standard SMF.

signal at the receiver end. This effect is known as *Modal Dispersion*.

The core of a *Multi-Mode Fiber* (MMF) is large enough ($50-100\,\mu$m) for the light to take multiple pathways from the transmitter to the receiver (see Fig. 2.2). This leads to severe signal distortion even on short fiber links. For example at 10 Gb/s the fiber length is limited to about $100-300$ m. The core of a *Single-Mode Fiber* (SMF) is much smaller ($8-10\,\mu$m) and permits only one pathway (a single mode) of light propagation from the transmitter to the receiver and thus distortions due to modal dispersion are suppressed.[2]

SMF is preferred in telecommunication applications (long-haul, metro, and access) where distance matters. MMF is mostly used within buildings for data communication (computer interconnects), and in consumer electronics. Because the MMF has a larger core size, alignment of the fiber with another fiber or a laser chip is less critical. A transverse alignment error between a laser and a SMF of just $0.5\,\mu$m causes a power penalty of about 1 dB, whereas the laser-to-MMF alignment is about 5× less critical [Shu88]. Thus components interfacing to MMF are generally lower in cost.

Chromatic Dispersion. *Chromatic Dispersion* is another source of signal distortions and is caused by different wavelengths (colors) traveling at different speeds through the fiber. This situation is illustrated in Fig. 2.3, where we see how the group delay varies with wavelength for 1 km of standard SMF. We recognize that the change in group delay is very large at $1.55\,\mu$m while it is small at $1.3\,\mu$m. In practice chromatic dispersion is specified by the *change* in group delay per nm wavelength and km length:

$$D = \frac{1}{L}\cdot\frac{\partial\tau}{\partial\lambda} \tag{2.1}$$

[2]The reader may wonder why a $8-10\,\mu$m core is small enough to ensure single-mode propagation of light that has a wavelength of $1.3-1.55\,\mu$m. The condition for single-mode propagation is that the core diameter must be smaller than $d < \lambda\cdot 0.766/\sqrt{n_{\text{cor}}^2 - n_{\text{clad}}^2}$, where n_{cor} and n_{clad} are the refractive indices of the core and cladding, respectively. Because the difference between n_{cor} and n_{clad} is small (less than 1 %), the core can be made larger than the wavelength λ, simplifying the light coupling into the fiber. Another advantage of the clad fiber!

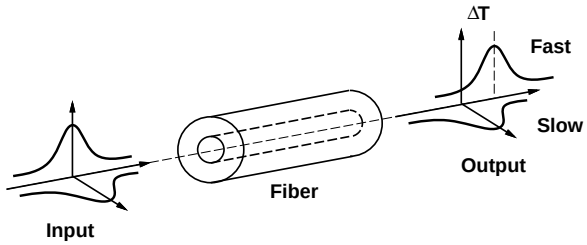

Figure 2.4: Polarization-mode dispersion in a short piece of fiber.

where D is known as the *Dispersion Parameter*, L is the fiber length, τ is the group delay, and λ is the wavelength. A standard SMF at $1.55\,\mu$m has $D = 17\,\text{ps}/(\text{nm}\cdot\text{km})$ which means that a change in wavelength of 1 nm will change the group delay by 17 ps in a 1 km long piece of fiber (cf. Fig. 2.3). The dependence of D on wavelength is plotted together with the fiber loss in Fig. 2.1.

How much distortion chromatic dispersion is causing depends on the spectral linewidth of the transmitter. If the transmitter operates at precisely a single wavelength, chromatic dispersion doesn't matter, but if the transmitter operates over a range of wavelengths, as it usually does, chromatic dispersion causes pulse distortions. This is rather important when designing a transmitter and we will discuss the pulse spreading caused by chromatic dispersion in more detail in Section 2.4.

What can we do to reduce the chromatic dispersion parameter D? Since dispersion is a linear phenomenon it can be reversed by applying an equal amount of *negative* dispersion. This method is called dispersion compensation. For example, so-called *Dispersion Compensating Fiber* (DCF) with a large negative value of D such as $-200\,\text{ps}/(\text{nm}\cdot\text{km})$ can be appended to the standard SMF to compensate for its dispersion. Alternatively, we can transmit at the $1.3\,\mu$m wavelength, where the dispersion parameter D of a standard SMF is much smaller than at $1.55\,\mu$m (see Fig. 2.1). But, as we know, at the $1.3\,\mu$m wavelength the loss is higher. To resolve this dilemma, fiber manufacturers have come up with a so-called *Dispersion-Shifted Fiber* (DSF), which has a value of D close to zero at the $1.55\,\mu$m wavelength while preserving the low loss. This fiber, however, has a disadvantage in WDM systems that we will discuss in Section 2.3 on nonlinearities.

Polarization-Mode Dispersion. Another source of distortions is *Polarization-Mode Dispersion* (PMD) which is caused by different polarization modes traveling at different speeds as shown schematically in Fig. 2.4. This effect occurs in fibers with a slightly elliptic core or asymmetrical mechanical stress. As a result of PMD the receiver sees time-shifted copies of the transmitted sequence superimposed on top of each other. For a long stretch of fiber the situation is complicated by the fact that the fiber's polarization properties change randomly along its *length*. The averaged delay (over many fibers) between the two principal states of polarization is proportional to the square root of the fiber length

L and can be written:

$$\Delta T = D_{PMD} \cdot \sqrt{L}. \tag{2.2}$$

In addition to the statistical uncertainty, PMD also varies slowly over *time* making it more difficult to compensate. As a rule of thumb, we should keep ΔT less than 10 % of the bit interval ($0.1/B$) to keep the power penalty due to PMD below 1 dB for most of the time.

What can we do against PMD? New fiber has a very low PMD parameter around $D_{PMD} = 0.1\,\mathrm{ps}/\sqrt{\mathrm{km}}$. This means after 100 km of fiber the average delay is only 1 ps, no problem even for 40 Gb/s. Older fiber, which is widely deployed and has a slightly elliptic cross section of the fiber core due to manufacturing tolerances, has a much larger PMD parameter around $D_{PMD} = 2\,\mathrm{ps}/\sqrt{\mathrm{km}}$. In this case an optical PMD compensator or an electrical adaptive equalizer (see Section 6.1) can be used.

2.3 Nonlinearities

Attenuation and dispersion are known as *linear* effects because they can be described by a linear relationship between the electrical field and its induced polarization. Apart from these linear effects, the fiber suffers from a number of *nonlinear* effects which may distort, attenuate, or produce crosstalk between optical signals. The most important ones are *Self-Phase Modulation* (SPM), *Cross-Phase Modulation* (CPM or XPM), *Stimulated Raman Scattering* (SRS), *Stimulated Brillouin Scattering* (SBS), and *Four-Wave Mixing* (FWM). These effects become important for long optical fibers operated at high optical power levels. While SPM, SRS, and SBS cause pulse distortions and attenuation in single wavelength systems, CPM, SRS, SBS, and FWM are of particular concern when transmitting multiple bit streams at different wavelengths over a single fiber, i.e., when using *Wavelength Division Multiplexing* (WDM).

In a WDM system, the bits in different channels interact with each other through nonlinear effects resulting in a change of pulse shape and amplitude (cf. Section 6.2.7). The longer the interacting bits stay together, the stronger the crosstalk distortions. For this reason it is advantageous if the different wavelength channels propagate at slightly different speeds, i.e., if there is a small amount of chromatic dispersion. A special fiber called *Non-Zero Dispersion-Shifted Fiber* (NZ-DSF) has been created which has a small value of $|D| = 1 - 6\,\mathrm{ps}/(\mathrm{nm}\cdot\mathrm{km})$, large enough to create a "walk-off" between the bit streams limiting nonlinear interactions, but small enough to limit the amount of dispersion compensation needed or to avoid it altogether.

2.4 Pulse Spreading due to Chromatic Dispersion

In the following we want to investigate the pulse distortions caused by chromatic dispersion in greater detail. An understanding of this effect will be important in Chapter 7 when we discuss transmitter design.

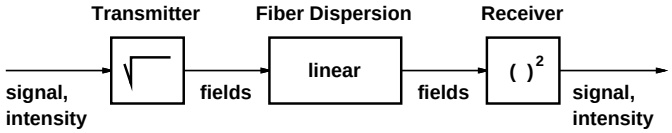

Figure 2.5: Communication channel with intensity modulation, fiber dispersion, and intensity detection.

Nonlinear Character of Optical Fiber Communication. To transmit an optical pulse we modulate the *intensity* of a light source and to receive the pulse we detect the *intensity* of the light.[3] Let's assume the fiber in between the transmitter and the receiver exhibits dispersion. Dispersion is a linear phenomenon, but linear in the *fields* not the *intensity*! So, we end up with the system shown in Fig. 2.5. The signal is converted to a proportional intensity; the intensity is carried by an electromagnetic field which is proportional to the square-root of the intensity; the field disperses linearly in the fiber; the resulting field is characterized by an intensity which is proportional to the square of the field; finally, the intensity is detected and a signal proportional to it is generated. This is a *nonlinear* system!

Now we understand that, in general, we cannot apply linear system theory to analyze the pulse distortions caused by fiber dispersion, making this a rather nasty problem. However, there is an approximation which we are going to use in the following. If we use a light source with a bandwidth much greater than the signal bandwidth, we can approximately describe the channel with a linear response [Ben83]. If we further assume that the source spectrum is Gaussian we find the impulse response of the channel:

$$h(t) = h(0) \cdot \exp\left(-\frac{t^2}{2 \cdot (\Delta T/2)^2}\right) \tag{2.3}$$

where

$$\Delta T = |D| \cdot L \cdot \Delta\lambda \tag{2.4}$$

and $\Delta\lambda$ is the 2σ-linewidth of the source. In other words, a Dirac impulse will spread out into a Gaussian pulse as it propagates along the fiber. The 2σ-width of the spreading pulse is the ΔT given in Eq. (2.4). For example, a very narrow pulse launched into a standard SMF will spread out to 17 ps after 1 km given a source width of 1 nm at 1.55 μm. We have just discovered a new interpretation for the dispersion parameter D: it tells us how fast a narrow pulse is spreading out!

Time-Domain Analysis. Now that we have a linear model, we are on familiar territory and we can calculate how a regular data pulse spreads out. The math is easiest if we assume that the transmitted pulse is Gaussian. The convolution of the (Gaussian) input pulse with the (Gaussian) impulse response produces a Gaussian output pulse! The

[3] This method is known as *Direct Detection*. An alternative is *Coherent Detection*, but this subject is beyond the scope of this text.

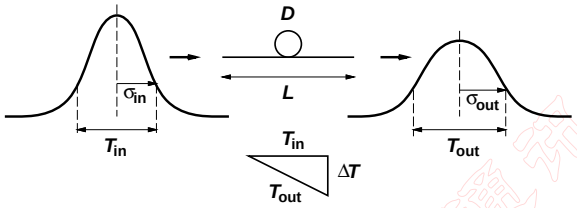

Figure 2.6: Pulse spreading due to chromatic dispersion.

relationship between the 2σ-width of the input pulse, T_{in}, and the 2σ-width of the output pulse T_{out} turns out to be:

$$T_{\text{out}} = \sqrt{T_{\text{in}}^2 + \Delta T^2}. \tag{2.5}$$

This situation is illustrated in Fig. 2.6. For example, on 1 km of standard SMF with a transmitter linewidth of 1 nm at 1.55 μm, a 100 ps pulse will broaden to $\sqrt{(100\,\text{ps})^2 + (17\,\text{ps})^2} = 101.4\,\text{ps}$.

The maximum amount of spreading, ΔT, that can normally be tolerated in practical systems is half a bit period.[4] In mathematical terms, this limit is:

$$\Delta T < \frac{1}{2B}. \tag{2.6}$$

At that point the pulse width increased by $\sqrt{1^2 + 0.5^2} = 1.12$ or about 12 %. This amount of spreading causes a power penalty of approximately 1 dB [HLG88].

As we can see from Eq. (2.4), the linewidth of the transmitter, $\Delta\lambda$, is of critical importance in determining the pulse spreading and thus the maximum transmission distance trough a dispersive fiber. We will discuss in Chapter 7 how this linewidth is related to the type of source (FP or DFB laser), the type of modulation (direct or external), and the bit rate used.

As we have already pointed out, Eqs. (2.3) – (2.5) are strictly valid only for sources with a wide linewidth $\Delta\lambda$. They work approximately for some narrow-linewidth sources, but they don't work at all for pulses with *negative Chirp* and for *Solitons*. Pulses with negative chirp[5] are characterized by a temporary decrease in optical frequency (red shift) during the leading edge and an increase in frequency (blue shift) during the trailing edge. Such pulses get *compressed* up to a certain distance and then undergo broadening just like regular pulses in a dispersive medium. Solitons are short ($\approx 10\,\text{ps}$) and powerful optical pulses of a certain shape. They do not broaden at all because chromatic dispersion is counterbalanced by the nonlinear fiber effect *Self-Phase Modulation*.

Frequency-Domain Analysis. Given the expression for the impulse response of a dispersive fiber, we can easily transform it into the frequency domain and discuss the corresponding channel bandwidth. Transforming the Gaussian impulse response in Eq. (2.3) yields the Gaussian frequency response:

$$H(f) = H(0) \cdot \exp\left(-\frac{(2\pi f)^2(\Delta T/2)^2}{2}\right). \tag{2.7}$$

The 3-dB bandwidth can be found by setting this equation equal to $1/2 \cdot H_0$ and solving for f. Together with Eq. (2.4) we find:[6]

$$BW_{\text{3dB}} = \frac{0.375}{\Delta T} = \frac{0.375}{|D| \cdot L \cdot \Delta\lambda}. \tag{2.8}$$

This is the fiber bandwidth for the modulation signal introduced in Section 2.1. Its value reduces as we increase the fiber length L or increase the dispersion parameter D. For example, 1 km of standard SMF with a transmitter linewidth of 1 nm at 1.55 μm has a bandwidth of just 22 GHz. If we replace the SMF with a NZ-DSF that has a dispersion parameter of only $D = 5\,\text{ps}/(\text{nm} \cdot \text{km})$, the bandwidth increases to 75 GHz.

What is the interpretation of the spreading limit, Eq. (2.6), in the frequency domain? Inserting Eq. (2.6) into Eq. (2.8) we find:

$$BW_{\text{3dB}} > 0.75 \cdot B. \tag{2.9}$$

The fiber bandwidth must be larger than 3/4 of the bit rate to avoid excessive distortions. Given this bandwidth, the attenuation at 1/2 the bit rate, where most of the energy of the "10101010 ..." sequence is located, is about 1 dB and this, in fact, is how the limit in Eq. (2.6) was derived in [HLG88].

2.5 Summary

Optical fiber is characterized by a very low loss of about 0.25 dB/km and a huge bandwidth of more than 10 THz when operated in the 1.55 μm wavelength band.

On the down side, dispersion causes the optical pulses to spread out in time and interfere with each other. There are several types of dispersion:

- Modal dispersion which only occurs in multi-mode fibers.
- Chromatic dispersion which is small at 1.3 μm but presents a significant limitation at the 1.55 μm wavelength in standard single-mode fibers. The impact of chromatic dispersion on pulse spreading can be ameliorated by using narrow-linewidth transmitters.
- Polarization-mode dispersion which occurs in high-speed, long-haul transmission over older types of fiber and is slowly varying in time.

Furthermore, at elevated power levels nonlinear effects can cause attenuation, pulse distortions, and crosstalk in WDM systems.

[4] This is equal to a rms impulse spread of a quarter bit period [HLG88].

[5] There is no consensus on the definition of positive or negative chirp. However, in this text we use the term "positive chirp" to describe a leading edge with blue shift.

[6] In the electrical domain, this bandwidth is the 6-dB bandwidth, because 3 optical dBs convert to 6 electrical dBs (cf. Section 3.1)!

2.6 Problems

2.1 Transmission System at 1310 nm. A 1.31 μm transmitter with a 3-nm linewidth launches a 2.5 Gb/s NRZ signal with 1 mW into a standard SMF. (a) How long can we make the fiber until the power is attenuated to -24.3 dBm? (b) How long can we make the fiber before chromatic dispersion causes too much pulse spreading? Assume $D = 0.5\,\mathrm{ps}/(\mathrm{nm}\cdot\mathrm{km})$.

2.2 Transmission System at 1550 nm. Now we use a 1.55 μm transmitter with the same linewidth, bit rate, and launch power as in Problem 2.1. How does the situation change?

2.3 Transmitter Linewidth. (a) In which system, problem 2.1 or 2.2, would it make sense to use a narrow-linewidth transmitter? How far could we go if we reduce the linewidth to 0.02 nm?

2.4 Fiber PMD. We are using "old" fiber with $D_{PMD} = 2\,\mathrm{ps}/\sqrt{\mathrm{km}}$. Do we have to be concerned about PMD in one of the above transmission systems?

Chapter 3

Photodetectors

The first element in an optical receiver chain is the photodetector. It is important to understand the main characteristics of this device, namely responsivity and noise, in order to be able to discuss and calculate the receiver's performance. The photodetector together with the transimpedance amplifier largely determine the receiver's sensitivity. There are three types of photodetectors which are commonly used: the p-i-n detector, the APD detector, and the optically preamplified p-i-n detector; we will discuss them in this order. More information on photodetectors can be found in [Agr97, Sze81, RS98].

3.1 p-i-n Photodetector

The simplest detector is the p-i-n photodiode shown in Fig. 3.1. A p-i-n photodetector consists of a p-n junction with a layer of intrinsic (undoped) material sandwiched in between the p- and the n-doped material. The junction is reverse biased to create a strong electric field in the intrinsic material. Photons hitting the i-layer create electron-hole pairs which get separated quickly by the electric field and produce an electrical current.

The width W of the i-layer controls the trade-off between efficiency and speed of the detector. The fraction of photons converted to electron-hole pairs is called *Quantum Efficiency* and designated by η. The wider W is made the better the chances that a

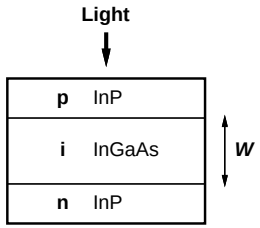

Figure 3.1: p-i-n Photodetector (schematically).

photon is absorbed in the detector and thus the higher the quantum efficiency. But, the wider W is made the longer it takes for the electrons and holes to traverse the i-layer at a given reverse voltage, making the photodiode response slower. To escape this dilemma, in very fast photodetectors the i-layer is illuminated from the side, so-called *Waveguide Photodetectors*. In these detectors the speed is still controlled by W, but the quantum efficiency is now controlled by the orthogonal dimension.

Most semiconductor materials are transparent, i.e., don't absorb photons at the wavelengths $1.3\,\mu$m and $1.55\,\mu$m commonly used in fiber optics. For example, silicon only absorbs photons with $\lambda < 1.06\,\mu$m, GaAs only with $\lambda < 0.87\,\mu$m, and InP only with $\lambda < 0.92\,\mu$m. Therefore a special compound with a narrow bandgap, InGaAs, is used for the i-layer. InGaAs p-i-n photodiodes are sensitive in the range $1.0 - 1.65\,\mu$m. The p- and n-layers of the photodiode are made from InP material which is transparent at the wavelength in question and thus no absorption takes place. Detectors for the $0.85\,\mu$m wavelength, used in data communication, can be made from silicon.

Electron-hole pairs created outside the drift field in the i-layer cause a slow response component because these carriers propagate to the electrodes very slowly by diffusion (e.g., $4\,\text{ns}/\mu\text{m}$). As a result, an undesired "tail current" follows the intended current pulse corresponding to the optical signal. In burst-mode receivers this tail current can cause problems when a very strong burst signal is followed by a very weak one. Tail currents can be minimized by using transparent materials for the p- and n-layers and precisely aligning the fiber to the active part of the i-layer.

Responsivity. We can write down the light-current relationship for a p-i-n diode knowing that the fraction η of all photons creates electrons. Each photon has the energy hc/λ and each electron carries the charge q, thus the electrical current (I) produced for a given amount of optical power (P) illuminating the photodiode is:

$$I_{PIN} = \eta \cdot \frac{\lambda q}{hc} \cdot P. \tag{3.1}$$

To derive this equation, remember that current is "electron charge per time" and optical power is "photon energy per time". To simplify matters we call the constant relating I and P the *Responsivity* of the photodiode and use the symbol $\mathcal{R}$ for it:

$$I_{PIN} = \mathcal{R} \cdot P \quad \text{with} \quad \mathcal{R} = \eta \cdot \frac{\lambda q}{hc}. \tag{3.2}$$

Let's make a numerical example to get a feeling for practical values. For the commonly used wavelength $\lambda = 1.55\,\mu$m and the quantum efficiency $\eta = 0.6$ we get a responsivity $\mathcal{R} = 0.75\,\text{A/W}$. This means for every milli-Watt of optical power incident onto the photodiode we obtain 0.75 mA of current.

A Two-for-One Special. Let's examine this relationship in more detail. Equation (3.2) means that if we double the power we get twice as much current. Now this is very odd! We are used to situations where the power grows with the *square* of the current and not linearly with the current. For instance in a wireless receiver, if we double the RF power we get $\sqrt{2}$ more current from the antenna. Or if we double the current flowing

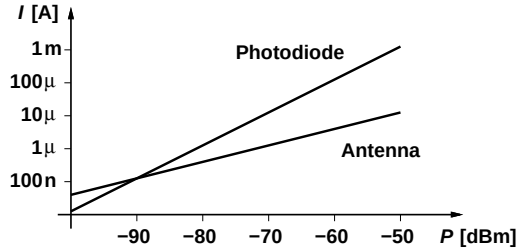

Figure 3.2: Comparison of antenna and photodiode at 1 GHz.

into a resistor, 4× as much power is dissipated. This square relationship is the reason why we use "10 log" to calculate power dBs and "20 log" to calculate current or voltage dBs. By using this convention a 3 dB increase in RF power translates into a 3 dB increase in current from the antenna. Or a 3 dB increase in current results in a 3 dB increase in power dissipation in the resistor. For a photodiode, however, a 3 dB increase of optical power translates into a 6 dB increase in current. What a bargain!

Wireless Receiver with a Photodiode? In contrast to optical receivers, wireless receivers are using antennas to detect the RF photons. The rms current that is produced by an antenna under matched conditions is:

$$i_{\text{ANT}}^{rms} = \sqrt{P/R_{\text{ANT}}} \tag{3.3}$$

where P is the received power[1] and R_{ANT} is the antenna resistance. For example for a $-50\,\text{dBm}$ signal ($10^{-8}\,\text{W}$) we get about $14\,\mu$A rms from an antenna with $R_{\text{ANT}} = 50\,\Omega$.

What if we would use a hypothetical hyper-infrared photodiode instead? These detectors are made sensitive to low-energy RF photons ($4\,\mu$eV) through advanced bandgap engineering. They are cooled down to a few milli-Kelvins to prevent currents caused by thermal electron-hole generation. Using Eq. (3.2) we can calculate the responsivity of this device to be an impressive 120 kA/W at 1 GHz with a quantum efficiency of $\eta = 0.5$. So, at the same received power level of $-50\,\text{dBm}$ we get a current of 1.2 mA. About 86× more than that of the old fashioned antenna!

But don't launch your start-up company to market this idea just yet! What happens if we reduce the received power? After all, this is where the detector's responsivity matters the most. The signal from the photodiode decreases *linearly*, while the signal from the antenna decreases more slowly following the *square-root* law. Once we are down to $-90\,\text{dBm}$ ($10^{-12}\,\text{W}$) we get about $0.14\,\mu$A from the antenna and $0.12\,\mu$A from the photodiode (see Fig. 3.2). About the same! Unfortunately, our invention turns out to be a disappointment ...

[1] More precisely, P is the power incident upon the effective aperture of the antenna [Kra88].

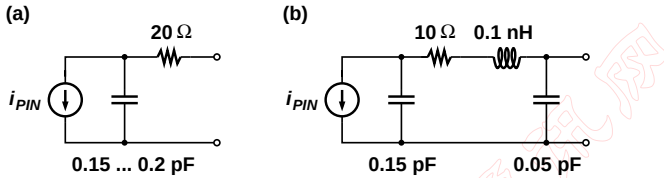

Figure 3.3: Equivalent AC circuits for 10 Gb/s p-i-n photodetectors: (a) bare photodiode [Ran01] and (b) photodiode with packaging parasitics [Gre01].

Bandwidth. Let's get back to more productive work! Figure 3.3 shows two equivalent AC circuits for 10 Gb/s p-i-n photodetectors, one for a bare detector with just the internal R and C and one for a detector with packaging parasitics. The current source in both models represents the current generated by the light and has the value $i_{PIN}(t) = \mathcal{R} \cdot P(t)$. We can see that the output impedance of the p-i-n detector is mostly capacitive.

The bandwidth of the photodiode is determined by two time constants: (i) the transit time, i.e., the time it takes the carriers to travel through the depletion region and (ii) the RC time constant given by the internal parallel capacitance C_{PD} and the series resistance R_{PD}. The bandwidth of a p-i-n photodiode can be written as [Agr97]:

$$BW = \frac{1}{2\pi} \cdot \frac{1}{W/v_n + R_{PD}C_{PD}} \tag{3.4}$$

where W is the width of the depletion region and v_n is the carrier velocity. For high-speed operation, the reverse voltage must be chosen large enough such that the carrier velocity v_n saturates at its maximum value and the transit time is minimized. Typically, a reverse voltage of about $5 - 10$ V is required. Photodiodes with bandwidths well in excess of 100 GHz have been demonstrated.

Shot Noise. The p-i-n photodiode not only produces the signal current I_{PIN} but also a noise current, the so-called *Shot Noise*. This noise current is due to the fact that the photocurrent is not continuous but a collection of random pulses corresponding to the electron/hole pairs created by the photons. If we approximate these pulses with Dirac pulses, the shot-noise spectrum is white and its mean-square value is:

$$\overline{i^2_{n,PIN}} = 2qI_{PIN} \cdot BW \tag{3.5}$$

where I_{PIN} is the signal current and BW is the bandwidth in which we measure the noise current. For example, a received optical power of 1 mW generates an average current of 0.75 mA (assuming $\mathcal{R} = 0.75$ A/W) and a shot-noise current of about 1.6 μA rms in a 10-GHz bandwidth. The signal-to-noise ratio comes out as $10\log(0.75\,\text{mA}/1.6\,\mu\text{A})^2 = 53.4$ dB.

As we can see from Eq. (3.5), the shot-noise current is *signal dependent*. If the received optical power is increased, the noise increases too. But fortunately the rms-noise grows only with the *square root* of the signal amplitude, so we still gain in signal-to-noise ratio.

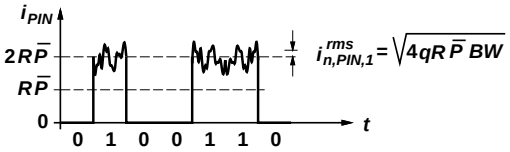

Figure 3.4: Output current of a p-i-n photodiode.

If we double the power in our previous example to 2 mW, we get an average current of 1.5 mA and a shot-noise current of 2.2 μA and the signal-to-noise ratio improved by 3 dB to 56.7 dB. Conversely, If the received optical power is reduced, the noise reduces too. For example, if we reduce the optical power by 3 dB, the signal current is reduced by 6 dB, but this time we are lucky, and the signal-to-noise ratio degrades only by 3 dB.

If we receive an NRZ signal with a p-i-n photodiode the noise on the "one" bits is much larger than that on the "zero" bits. In fact, if we turn our transmitter (laser) completely off during the transmission of a zero (infinite extinction ratio), there will be no signal and therefore no noise for the zeros. Let's assume that we receive the average power $\overline{P}$ and that the signal is an NRZ signal with 50 % mark density. It follows that the optical power for ones is $P_1 = 2\overline{P}$ and that for zeros is $P_0 \approx 0$. Then the noise currents for zeros and ones are:

$$\overline{i^2_{n,PIN,0}} \approx 0, \tag{3.6}$$

$$\overline{i^2_{n,PIN,1}} = 4q\mathcal{R}\overline{P} \cdot BW. \tag{3.7}$$

Figure 3.4 illustrates the signal and noise currents produced by a p-i-n photodiode in response to an optical NRZ signal. Signal and noise magnitudes are expressed in terms of average received power $\overline{P}$.

Dark Current. The p-i-n photodiode produces a very small current I_{DK} even when it is in total darkness. This so-called *Dark Current* depends on temperature and processing but is usually less than 20 nA for an InGaAs or silicon photodiode [Agr97]. The dark current and its associated shot-noise current interfere with the received signal. However, in Gb/s p-i-n receivers this effect is negligible. To demonstrate this let's calculate the optical power for which the worst-case dark current reaches 10 % of the signal current. As long as our received optical power is larger than this, we are fine:

$$\overline{P} > 10 \cdot \frac{I_{DK}(\text{max})}{\mathcal{R}}. \tag{3.8}$$

With the values $\mathcal{R} = 0.75$ A/W and $I_{DK}(\text{max}) = 20$ nA we find $\overline{P} > -35.7$ dBm. We will see later that Gb/s p-i-n receivers require much more signal power than this to work at an acceptable bit-error rate and therefore we don't need to worry about dark current in such receivers. However, in high-sensitivity receivers (at low speed and/or with APD detector) dark current can be an important limitation. In Section 4.5 we will formulate the impact of the dark current in a more precise way.

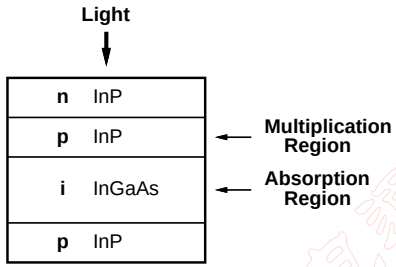

Figure 3.5: Avalanche photodetector (schematically).

3.2 Avalanche Photodetector

The basic structure of the avalanche photodetector is shown in Fig. 3.5. Like the p-i-n detector, the avalanche photodetector is a reverse biased diode. However, in contrast to the p-i-n diode it features an additional layer, the *Multiplication Region*. This layer provides gain through avalanche multiplication of the electron-hole pairs generated in the i-layer, a.k.a. the *Absorption Region*. For the avalanche process to set in, the APD must be operated at a fairly high reverse bias voltage of about $40-60\,\mathrm{V}$. In comparison a p-i-n photodiode operates at about $5-10\,\mathrm{V}$.

To make the APD detector sensitive in the $1.0-1.65\,\mu\mathrm{m}$ wavelength range, InGaAs is used for the absorption region, just like in the case of the p-i-n detector. The multiplication region, however, is typically made from InP material which can sustain a higher electric field.

Responsivity. The APD gain is called *Avalanche Gain* or *Multiplication Factor* and is designated by the letter M. A typical value for an InGaAs APD is $M = 10$. Light power is therefore converted to electrical current as:

$$I_{APD} = M \cdot \mathcal{R} P \tag{3.9}$$

where $\mathcal{R}$ is the responsivity of the APD detector without avalanche gain, similar to that of a p-i-n diode. Assuming that $\mathcal{R} = 0.75\,\mathrm{A/W}$ as for the p-i-n detector and $M = 10$, the APD detector generates $7.5\,\mathrm{A/W}$. Therefore we can also say that the APD detector has a responsivity $\mathcal{R}_{APD} = 7.5\,\mathrm{A/W}$, but we have to be careful to avoid confusion with the responsivity $\mathcal{R}$ in Eq. (3.9) which does not include the avalanche gain.

The avalanche gain M is a sensitive function of the reverse bias voltage (see Fig. 3.6) and therefore this voltage must be well controlled. Furthermore, the avalanche gain is temperature dependent and a temperature compensated bias voltage source is needed to keep the gain constant. The circuit in Fig. 3.7 uses a thermistor (ThR) to measure the APD temperature and a control loop to adjust the reverse bias voltage V_{APD} at a rate of $0.2\,\%/°\mathrm{C}$ [Luc99]. The dependence of the avalanche gain on the bias voltage can

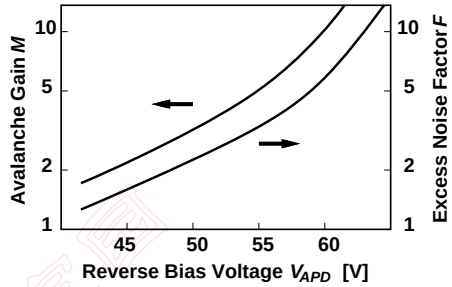

Figure 3.6: Avalanche gain and excess noise factor as a function of reverse voltage for a typical InGaAs APD.

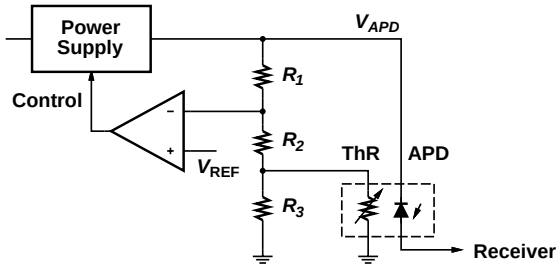

Figure 3.7: Temperature compensated APD bias circuit.

also be exploited to implement a gain-control mechanism (AGC) increasing the receiver's dynamic range.

Avalanche Noise. Unfortunately, the APD does not only provide more signal but also more noise, in fact, *more* noise than simply the amplified shot noise. More precisely, each primary carrier is multiplied by a random gain factor: one photon creates 9 electron/hole pairs, the next one 13, and so on. The gain M introduced before is really the *average* avalanche gain. Taking the random nature of the gain process into account, we can express the mean-square noise current of the APD:

$$\overline{i^2_{n,APD}} = F \cdot M^2 \cdot 2qI_{PIN} \cdot BW \tag{3.10}$$

where F is the so-called *Excess Noise Factor* and I_{PIN} is the current produced by an equivalent p-i-n diode with responsivity $\mathcal{R}$ that receives the same amount of light as the APD diode under question. An excess noise factor $F = 1$ is the ideal case where there is only amplified shot noise and no randomness in the gain process. In reality this factor is more typically around $F = 6$.

The excess noise factor increases with the avalanche gain M and thus is also a sensitive function of the reverse voltage (cf. Fig. 3.6). In mathematical terms the relationship between F and M can be written:

$$F = k_A \cdot M + (1 - k_A) \cdot \left(2 - \frac{1}{M}\right) \tag{3.11}$$

where k_A is the *Ionization-Coefficient Ratio*. If only one type of carriers, usually electrons, participate in the avalanche process, $k_A = 0$ and the excess noise factor is minimized. However, if electrons and holes are both participating, $k_A > 0$ and more excess noise is the result. For an InGaAs APD, $k_A = 0.5 - 0.7$ and the excess noise factor is almost proportional to M as illustrated in Fig. 3.6; for a silicon APD, $k_A = 0.02 - 0.05$ and the excess noise factor increases much more slowly [Agr97]. Researchers are currently working on reducing k_A for long-wavelength APDs by experimenting with new materials (InAlAs) and structures in the multiplication layer.

Since the APD gain can only be increased at the expense of generating more detector noise, as given by Eq. (3.11), there is an *Optimum APD Gain* which depends on the APD material (k_A), the amplifier noise, and the received power level. In Section 4.3 we will derive the APD gain that maximizes the receiver sensitivity.

Just like in the case of the p-i-n detector, the APD noise is signal dependendent and hence the noise currents for zeros and ones are different:

$$\overline{i^2_{n,APD,0}} \approx 0, \tag{3.12}$$

$$\overline{i^2_{n,APD,1}} = F \cdot M^2 \cdot 4q\mathcal{R}\overline{P} \cdot BW. \tag{3.13}$$

Dark Current. Just like the p-i-n photodiode the APD diode also suffers from a dark current. The so-called *Primary Dark Current* is less than 5 nA for InGaAs and silicon photodiodes [Agr97]. This dark current, just like a signal current, gets amplified to $M \cdot I_{DK}$ and the associated avalanche noise is $F \cdot M^2 \cdot 2qI_{DK} \cdot BW$. We can again use

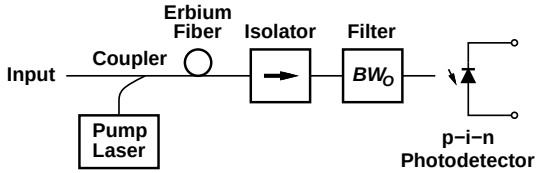

Figure 3.8: p-i-n photodetector with erbium-doped fiber preamplifier (schematically).

Eq. (3.8) to judge if this dark current is harmful. With the values $\mathcal{R} = 0.75\,\mathrm{A/W}$, and $I_{DK}(\max) = 5\,\mathrm{nA}$ we find that we are o.k. if $\overline{P} > -41.8\,\mathrm{dBm}$. Most Gb/s APD receivers require more signal power than this to work at an acceptable bit-error rate and dark current is not a big worry.

Bandwidth. Increasing the APD gain not only adds more noise but also reduces the *bandwidth*. Like in a single-stage amplifier, the product of gain and bandwidth remains approximately constant and is therefore used to quantify the speed of an APD. The gain-bandwidth product of a typical high-speed APD is in the range $100 - 150\,\mathrm{GHz}$. The equivalent AC circuit for an APD detector is similar to those shown in Fig. 3.3. However, the current source now has the value $i_{APD}(t) = M \cdot \mathcal{R}P(t)$ and the parasitic capacitances are usually somewhat larger.

APDs are in widespread use for receivers up to and including 2.5 Gb/s. However, it is challenging to fabricate APDs with a high enough gain-bandwidth product to be useful at 10 Gb/s and above. For this reason high sensitivity 10 Gb/s+ receivers are using optically preamplified p-i-n detectors instead of APD detectors. Optically preamplified p-i-n detectors are more expensive but feature superior sensitivity and speed.

3.3 p-i-n Detector with Optical Preamplifier

An alternative to the APD is the p-i-n detector with optical preamplifier. The p-i-n detector provides high speed while the optical preamplifier provides high gain over a wide bandwidth, eliminating the gain-bandwidth trade-off known from APDs. Furthermore, the optically preamplified p-i-n detector has superior noise characteristics when compared to an APD. However, the cost of a high-quality optical preamplifier (EDFA) is very high.

The optical preamplifier can be implemented as a *Semiconductor Optical Amplifier* (SOA) which is small and can be integrated together with the p-i-n detector on the same InP substrate. However, for best performance the *Erbium-Doped Fiber Amplifier* (EDFA) which has high gain and low noise in the important $1.55\,\mu$m band is a popular choice. See Fig. 3.8 for the operating principle of an EDFA-preamplified p-i-n detector. The received optical signal is combined with the light from a continuous-wave pump laser, typically providing a power of a few 10 mW. The pump wavelength for an EDFA can be $0.98\,\mu$m or $1.48\,\mu$m, but $\lambda = 0.98\,\mu$m is preferred for low-noise preamplifiers. The signal and the pump power are sent through an Erbium doped fiber of about 10 m length where the

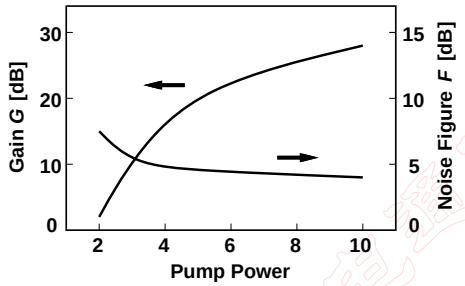

Figure 3.9: EDFA gain and noise figure as a function of the pump power.

amplification takes place by means of stimulated emission. An optical isolator prevents reflection of the optical signal back into the amplifier, which would cause instability. An optical filter with bandwidth BW_O reduces the noise power before the optical signal is converted to an electrical signal in the p-i-n photodiode. Noise is generated in the EDFA due to a process called *Amplified Spontaneous Emission* (ASE). The power spectral density of the ASE noise is designated with S_{ASE} and is nearly white.[2] Thus the optical noise power which reaches the detector is $P_{ASE} = S_{ASE} \cdot BW_O$. To keep P_{ASE} low we want to use a narrow optical filter.

Responsivity. The optical amplifier is characterized by a power gain, designated with G. The gain value of an EDFA depends on the length of the Erbium-doped fiber and increases with pump power as shown in Fig. 3.9. A typical value is $G = 100$ corresponding to 20 dB. The current produced by the p-i-n photodiode, I_{OA}, expressed as a function of the optical power at the input of the preamplifier, P is:

$$I_{OA} = G \cdot \mathcal{R}P \tag{3.14}$$

where $\mathcal{R}$ is the responsivity of the p-i-n diode.

Because of the strong dependence of the gain on pump power, EDFAs usually contain a microcontroller adjusting the pump laser in response to a small light sample split off from the amplified output signal. In this way an automatic gain control (AGC) mechanism is provided which increases the receiver's dynamic range [FJ97].

While the APD detector gave us about one order of magnitude of gain ($M = 10$), the optically preamplified p-i-n detector gives us about two orders of magnitude of gain ($G = 100$) relative to a regular p-i-n detector. So, the total responsivity of the combined preamplifier and p-i-n detector is 75 A/W given $\mathcal{R} = 0.75$ A/W and $G = 100$.

ASE Noise. We have already mentioned the ASE noise generated in the EDFA. How is this optical noise converted to electrical noise in the photodiode? If you thought that it

[2] In the following, S_{ASE} always refers to the spectral density in *both* polarization modes, i.e., $S_{ASE} = 2 \cdot S'_{ASE}$ where S'_{ASE} is the spectral density in one polarization mode.

was odd that optical signal *power* gets converted to electrical signal *amplitude*, wait until you hear this! Because the optical detector responds to intensity which is proportional to the *square* of the fields (cf. Fig. 2.5), the optical noise causes multiple electrical beat noise terms. The two most important ones are [Agr97]:

$$\overline{i^2_{n,ASE}} = \mathcal{R}^2 \cdot (2P_S S_{ASE} + S^2_{ASE} \cdot BW_O) \cdot BW. \tag{3.15}$$

The first term is called *Signal-Spontaneous Beat Noise* and is usually the dominant term. This noise component is proportional to the signal power P_S at the output of the EDFA. So, a constant, signal-independent optical noise density S_{ASE} generates a *signal dependent* noise term in the electrical domain! Furthermore, this noise term is *not* affected by the optical filter bandwidth BW_O, but the electrical bandwidth BW does have an influence. The second term is called *Spontaneous-Spontaneous Beat Noise* and may be closer to your expectations![3] Just like with the signal, the electrical amplitude of this noise component is proportional to the optical power. And the optical filter bandwidth does have an influence on the spontaneous-spontaneous beat noise component. In addition to the ASE noise terms in Eq (3.15), the p-i-n diode also produces the well-known signal-dependent shot noise. But the latter noise is so small that it is usually neglected.

By now you have probably developed a healthy respect for the unusual ways optical quantities translate to the electrical domain. Now let's see what happens to the *Signal-to-Noise Ratio* (SNR)! For a continuous-wave signal with optical power P_S and the signal power in the electrical domain is $\overline{i^2_S} = \mathcal{R}^2 P_S^2$, the noise power is given by Eq. (3.15). The ratio of these two expressions is:

$$SNR = \frac{(P_S/P_{ASE})^2}{P_S/P_{ASE} + 1/2} \cdot \frac{BW_O}{2BW}. \tag{3.16}$$

Now P_S/P_{ASE} is also known as the *Optical Signal-to-Noise Ratio* (OSNR) measured in the optical bandwidth BW_O. If the OSNR is much larger than 1/2 we can neglect the contribution from spontaneous-spontaneous beat noise (this is where the 1/2 in the denominator comes from) and we end up with the surprisingly simple result:

$$SNR \approx OSNR \cdot \frac{BW_O}{2BW}. \tag{3.17}$$

The SNR can be obtained simply by scaling the OSNR with the ratio of the optical and 2× the electrical bandwidth. For example for a 7.5 GHz receiver, an OSNR of 14.7 dB measured in a 0.1 nm bandwidth (12.5 GHz at $\lambda = 1.55\,\mu$m) translates into an SNR of 13.9 dB. Later in Section 4.3 we will use Eq. (3.17) to analyze amplified transmission systems.

[3] In the literature, spontaneous-spontaneous beat noise is sometimes given as $4\mathcal{R}^2 S'^2_{ASE} BW_O$ [Agr97] and sometimes as $2\mathcal{R}^2 S'^2_{ASE} BW_O$ [Ols89] ($S'_{ASE} = S_{ASE}/2$, the ASE spectral density in one polarization mode) which may be quite confusing. The first equation applies to EDFA/p-i-n systems *without* a polarizer in between the amplifier and the p-i-n detector; the second equation applies to EDFA/p-i-n systems *with* a polarizer. In practice polarizers are not usually employed because this would require polarization control of the signal.

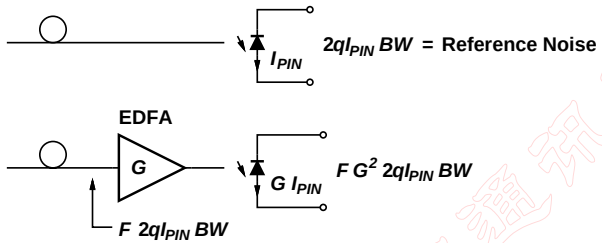

Figure 3.10: Definition of the noise figure of an optical amplifier.

Noise Figure of an Optical Amplifier. Just like electrical amplifiers, optical amplifiers are characterized by a noise figure F. A typical value for an EDFA noise figure is $F = 5\,\text{dB}$ and the theoretical limit is 3 dB. But what is the meaning of noise figure for an *optical* amplifier?

In an electrical system, noise figure is the input-referred mean-square noise normalized to the mean-square noise of the source. Usually, the noise of the source is the thermal noise of a $50\,\Omega$ resistor. (For more information on the electrical noise figure, see Section 6.2.3). Now, an optical amplifier doesn't get its signal from a $50\,\Omega$ source so the definition of its noise figure cannot be based on $50\,\Omega$ noise. What reference noise is it based on? The shot noise of an ideal p-i-n photodiode!

The noise figure of an optical amplifier is the input-referred mean-square noise normalized to the mean-square noise of a ideal p-i-n photodiode receiving the same optical signal as the amplifier. Input-referred mean-square noise is simply the mean-square noise current at the output of the amplifier-plus-ideal-photodiode system divided by the squared amplifier gain, G^2. The photodiode used in this definition is ideal in the sense that the quantum efficiency $\eta = 1$.[4]

Figure 3.10 illustrates the various noise quantities. An ideal photodiode in place of the optical amplifier, produces a current I_{PIN} and a mean-square shot-noise current $2qI_{PIN} \cdot BW$. This is our reference noise upon which the noise figure acts as a multiplier. The input-referred mean-square noise of an amplifier with noise figure F therefore is $F \cdot 2qI_{PIN} \cdot BW$. At the output, the mean-square noise current of an optically preamplified ideal p-i-n detector becomes:

$$\overline{i^2_{n,OA}} = F \cdot G^2 \cdot 2qI_{PIN} \cdot BW \tag{3.18}$$

where I_{PIN} is the current produced by an ideal p-i-n diode receiving the same amount of light as the optical preamplifier. Now, what is the noise current produced by a *real* p-i-n detector with $\eta < 1$? We have to substitute $\overline{i^2_{n,OA}} \to \eta^2 \cdot \overline{i^2_{n,OA}}$ and $I_{PIN} \to \eta \cdot I_{PIN}$ resulting in:

$$\overline{i^2_{n,OA}} = \eta F \cdot G^2 \cdot 2qI_{PIN} \cdot BW. \tag{3.19}$$

[4] An equivalent definition of the noise figure is: Input SNR divided by output SNR, where both SNRs are measured in the electrical domain with an ideal photodetector and the input SNR contains only shot noise.

As usual, the mean-square noise on zeros and ones is different and written as a function of the average received power $\overline{P}$ is:

$$\overline{i^2_{n,OA,0}} \approx 0, \tag{3.20}$$

$$\overline{i^2_{n,OA,1}} = \eta F \cdot G^2 \cdot 4q\mathcal{R}\overline{P} \cdot BW. \tag{3.21}$$

However, spontaneous-spontaneous beat noise may cause the $\overline{i^2_{n,OA,0}}$ noise term to become significantly non-zero.

It is instructive to compare the noise expression Eq. (3.10) for the APD detector to Eq. (3.19) for the optically preamplified p-i-n detector. We discover that the excess noise factor F of the APD plays the same role as the product ηF of the optical preamplifier!

Noise Figure and ASE Noise. In Eq. (3.15) we expressed the electrical noise in terms of the optical ASE noise, in Eq (3.19) we expressed the electrical noise in terms of the amplifier's noise figure. Now let's combine the two equations and write F as a function of ASE noise. With the assumption that all electrical noise at the output of the optically preamplified p-i-n detector is caused by ASE noise ($\overline{i^2_{n,OA}} = \overline{i^2_{n,ASE}}$), i.e., ignoring shot noise, $P_S = GP$ where P is the input power, and $\mathcal{R} = \lambda q/(hc)$ for the ideal photodiode, we arrive at:

$$F = \frac{\lambda}{hc} \cdot \left(\frac{S_{ASE}}{G} + \frac{S^2_{ASE}}{2 \cdot G^2 P} \cdot BW_O \right). \tag{3.22}$$

The first term is caused by signal-spontaneous beat noise while the second term is caused by spontaneous-spontaneous beat noise. The second term can be neglected for large input power levels P and small optical bandwidths BW_O. Sometimes a restrictive type of noise figure $\tilde{F}$ is defined based on just the first term in Eq. (3.22).

$$\tilde{F} = \frac{\lambda}{hc} \cdot \frac{S_{ASE}}{G}. \tag{3.23}$$

This noise figure is known as *Signal-Spontaneous Beat Noise Limited Noise Figure* or *Optical Noise Figure* and is only approximately equal to the noise figure F defined on the electrical SNR ratio. (The fact that there are two similar but not identical noise figure definitions can be quite confusing.)

Let's go one step further. Experts have calculated that the ASE power spectral density can be written as follows [Agr97]:

$$S_{ASE} = 2(G-1) \cdot \frac{N_2}{N_2 - N_1} \cdot \frac{hc}{\lambda} \tag{3.24}$$

where N_1 is the number of Erbium atoms in the ground state and N_2 is the number of Erbium atoms in the excited state. Thus the stronger the amplifier is pumped, the more atoms will be in the excited state resulting in $N_2 \gg N_1$. Combining the first term of Eq. (3.22), i.e., the optical noise-figure expression, with Eq. (3.24) and taking $G \gg 1$, we find the following simple approximation for the EDFA noise figure:

$$F \approx 2 \cdot \frac{N_2}{N_2 - N_1}. \tag{3.25}$$

This equation means that increasing the pump power will decrease the noise figure until it reaches the theoretical limit of 3 dB (cf. Fig. 3.9).

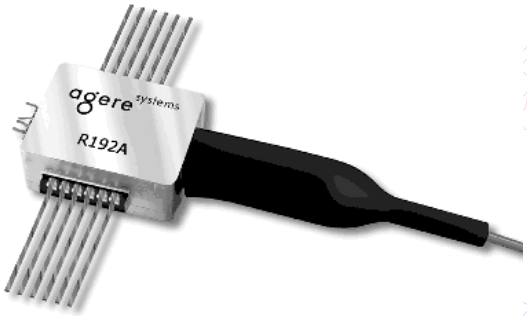

Figure 3.11: 10 Gb/s photodiode and TIA in a 16-pin surface-mount package with a single-mode fiber pigtail (1.6 cm × 1.3 cm × 0.7 cm).

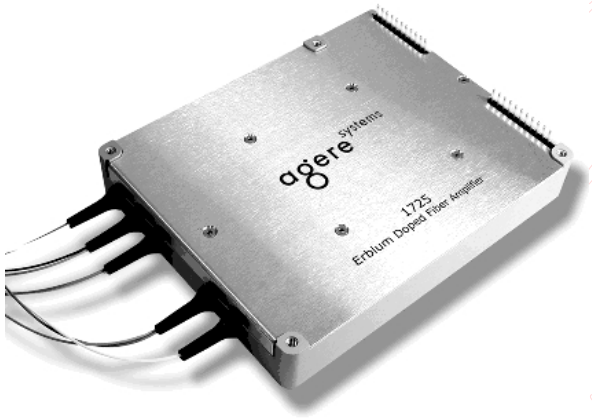

Figure 3.12: A packaged two-stage erbium-doped fiber amplifier with single-mode fiber pigtails for the input, output, interstage access, and tap monitor ports (12 cm × 10 cm × 2 cm).

Negative Noise Figure? What would an optical amplifier with a *negative* noise figure do? It would *improve* the SNR beyond the limit given by shot noise, i.e., beyond the limit given by the fact that electro-magnetic energy is quantized into photons. How could that be possible? Now you may be surprised to hear that you can actually *buy* optical amplifiers with negative noise figures. You can buy a Raman amplifier with $F = -2\,\text{dB}$ or even less, if you are willing to pay more!

Let's look at this situation. A fiber span with loss $1/G$ has a noise figure of G. The same fiber span followed by an EDFA with noise figure F has a combined noise figure of $G \cdot F$. You can easily prove both facts with the noise figure definition given earlier. For example a 100 km fiber span with 25 dB loss followed by an EDFA with a noise figure of 5 dB has a total noise figure of 30 dB.

There is a type of optical amplifier, called *Raman Amplifier*, which provides distributed gain *in* the fiber span itself. The fiber span is "pumped" from the receive end with a strong laser (1 W or so) and a mechanism called *Stimulated Raman Scattering* (SRS) provides the gain. For example, by pumping a 100 km fiber span the loss may reduce from 25 dB to 15 dB and the noise figure may improve from 25 dB to 23 dB. How do you sell such an amplifier? Right, you compare it to an EDFA and say it has a gain of 10 dB and a noise figure of −2 dB. O.k., I'll order one but please ship it without the fiber span ...

3.4 Summary

Three types of photodetectors are commonly used for optical receivers today:

- The p-i-n detector with a typical responsivity in the range $0.6 - 0.9\,\text{A/W}$ (for an InGaAs detector) is used in short-haul applications.
- The APD detector with a typical responsivity in the range $5 - 20\,\text{A/W}$ (for an InGaAs detector) is used in long-haul applications up to 10 Gb/s.
- The optically preamplified p-i-n detector with a responsivity in the range $6 - 900\,\text{A/W}$ is used in ultra-long-haul applications and for speeds at or above 10 Gb/s.

All detectors produce a current *amplitude* which is proportional to the received optical *power*, i.e., a 3-dB change in optical power produces a 6-dB change in current.

All three detectors generate a *signal dependent* noise current, specifically, the noise power $\overline{i^2_{n,PD}}$ grows proportional to the signal current I_{PD}. As a result, received one bits contain more noise than zero bits. The p-i-n detector generates shot noise which is usually negligibly small. The APD detector generates avalanche noise, quantified by the excess noise factor F. The optical preamplifier generates ASE noise which is converted into several electrical noise components by the p-i-n detector. The noise characteristics of the optical amplifier is specified by a noise figure F.

3.5 Problems

3.1 Photodiode vs. Antenna. An ideal photodiode ($\eta = 1$) and antenna are both exposed to the same continuous-wave electromagnetic radiation at power level P.

(a) Calculate the power level P at which the signal from the photodiode becomes equal to the noise. (b) Calculate the corresponding power level (sensitivity) for the antenna. (c) How do the sensitivities of photodiode and antenna compare?

3.2 Power Conservation in the Photodiode. The photodiode produces a current which is proportional to the received optical power P_{opt}. When this current flows into a resistor it creates a voltage drop which is also proportional to the received power P_{opt}. Thus the electrical power dissipated in the resistor is proportional to P_{opt}^2! (a) Is this a violation of energy conservation? (b) What can you say about the maximum forward voltage V_F of a photodiode?

3.3 Shot Noise. The current produced by a photodiode contains shot noise because this current consists of moving chunks of charge (electrons). (a) Does a battery loaded by a resistor also produce shot noise? (b) Explain the answer!

3.4 Optical Signal-to-Noise. Equations (3.16) and (3.17) state the relationship between SNR and OSNR for a continuous-wave signal with power P_S. How does this expression change for a DC-balanced, NRZ-modulated signal with average power $\overline{P}_S$?

Chapter 4

Receiver Fundamentals

An understanding of basic receiver theory is required before we can discuss the components of the receiver such as the TIA and the MA. In the following we will analyze important receiver properties such as noise, bit-error rates, bandwidth, sensitivity, and power penalties.

4.1 Receiver Model

The basic receiver model used in this chapter is shown in Fig. 4.1. It consists of (i) a detector model, (ii) a linear channel model which comprises the TIA, the MA, and optionally a filter, and (iii) a binary decision circuit with a fixed threshold. Later in Sections 4.7-4.11 we will extend this basic model to include an adaptive equalizer, an adaptive decision threshold, and a multi-level decision circuit.

The *Detector Model* consist of a signal current source i_{PD} and a noise current source $i_{n,PD}$. The signal current is linearly related to the received optical power. The noise current is typically white and signal dependent. The characteristics of these two currents have been discussed in Chapter 3 for the p-i-n detector, the APD detector and the optically preamplified p-i-n detector.

The *Linear Channel* can be modeled as a complex transfer function $H(s)$ which relates the output voltage v_O to the input current i_{PD}. This transfer function can be decomposed

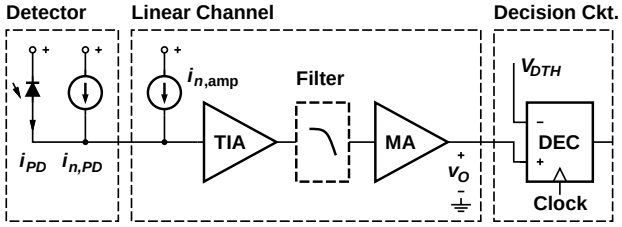

Figure 4.1: Basic receiver model.

into a product of three transfer functions: one for the TIA, one for the filter, and one for the MA. But for now we are concerned with the receiver as a whole. The noise characteristics of the linear channel are modeled by a single noise current source $i_{n,\text{amp}}$ at the input of the channel. The noise spectrum of this source is chosen such that after passing through the noiseless channel $H(s)$ the output noise of the actual noisy channel is reproduced. The linear-channel noise $i_{n,\text{amp}}$ is almost completely given by the input-referred noise of the TIA, because the TIA is the first element of the linear channel. It is important to distinguish the different characteristics of the two noise current sources:

- The detector noise, $i_{n,PD}$, is *nonstationary* (the rms value is varying with the bit pattern), however, it is *white* to a good approximation. The power spectrum of the detector noise must be written as a function of frequency and *time*:

$$I^2_{n,PD}(f,t) \sim \text{Bit Pattern}. \tag{4.1}$$

- The linear-channel noise, $i_{n,\text{amp}}$, is *stationary* (the rms value is independent of time) but usually *not white*. In Section 5.2.3 we will calculate the spectrum of this noise source (Eqs. (5.33), (5.34), and (5.37)) and we will see that its two main components are a constant part (white noise) and a part increasing with f^2. This is true no matter if the receiver is built with a FET or BJT front end. The power spectrum of the linear-channel noise can therefore be written in the general form:

$$I^2_{n,\text{amp}}(f) = a + b \cdot f^2 + \ldots \tag{4.2}$$

The last element in our receiver model, the *Decision Circuit*, compares the output voltage from the linear channel, v_O, to a fixed threshold voltage, V_{DTH}. If the output voltage is larger than the threshold, a one bit is detected; if it is smaller, a zero bit is detected. The comparison is triggered by a clock signal supplied to the decision circuit.

We mentioned in Chapter 1 that the MA can be implemented as an AGC amplifier or a limiting amplifier. Now a LA is *not linear* and you may wonder if the model of a *linear channel* is still applicable in this case. It is, as long as the input signal is small enough such that limiting does not occur in the LA. Typically, a LA reaches limiting only for signals much larger than the receiver's own noise voltage and therefore the linear model is appropriate for the sensitivity calculations we are going to carry out in this chapter.

4.2 Bit-Error Rate

The output voltage v_O consists of a superposition of the desired *signal* voltage v_S and the undesired *noise* voltage v_n ($v_O = v_S + v_n$). The noise voltage v_n, of course, is caused by the detector noise and the amplifier noise. Occasionally the instantaneous noise voltage $v_n(t)$ is so large that it corrupts the received signal $v_S(t)$ leading to a decision error or *Bit Error*. In this section we first want to calculate the rms value of the output noise voltage v_n^{rms} and then derive the bit-error rate *BER* caused this noise.

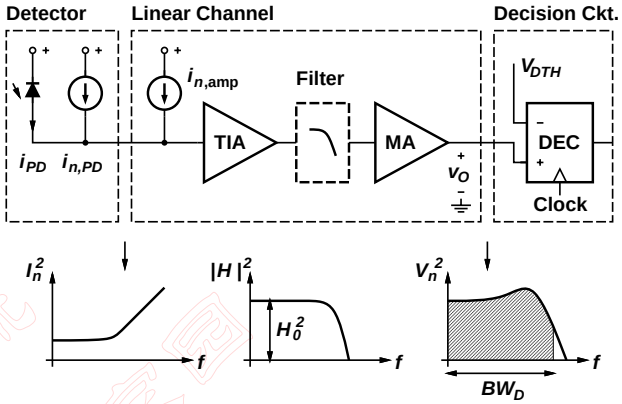

Figure 4.2: Calculation of the total output-referred noise.

Output Noise. The output noise power can be written as the sum of two components one caused by the detector noise and one caused by the amplifier noise. Let's start with the amplifier noise which is stationary and therefore easier to deal with. Given the input-referred power spectrum $I^2_{n,\text{amp}}$ of the amplifier noise and the transfer function of the linear channel $H(f)$, we can easily calculate the power spectrum at the output:

$$V^2_{n,\text{amp}}(f) = |H(f)|^2 \cdot I^2_{n,\text{amp}}(f). \tag{4.3}$$

Integrating this noise spectrum over the bandwidth of the decision circuit, BW_D, gives us the total noise power that the decision circuit sees:

$$\overline{v^2_{n,\text{amp}}} = \int_0^{BW_D} |H(f)|^2 \cdot I^2_{n,\text{amp}}(f)\, df. \tag{4.4}$$

This equation is illustrated by Fig. 4.2. The input noise spectrum, which increases with frequency due to the f^2 component, is shaped by the $|H(f)|^2$ frequency response producing an output spectrum rolling off rapidly at high frequencies. Because of the rapid roll-off the precise value of the upper integration bound (BW_D) is uncritical and is sometimes chosen to be infinity.

Next we have to deal with the nonstationary detector noise. Visualize the input spectrum as a two-dimensional surface with a time and frequency dimension. This 2d-spectrum is mapped to the output of the linear channel according to [Liu96]:

$$V^2_{n,PD}(f,t) = H(f) \cdot \int I^2_{n,PD}(f, t-\tau) \cdot h(\tau) \cdot e^{j2\pi f\tau}\, d\tau \tag{4.5}$$

where $h(t)$ is the step response of $H(f)$. Not only does the spectrum get "shaped" along the frequency axis, but it also gets "smeared out" along the time axis! Potentially, this is

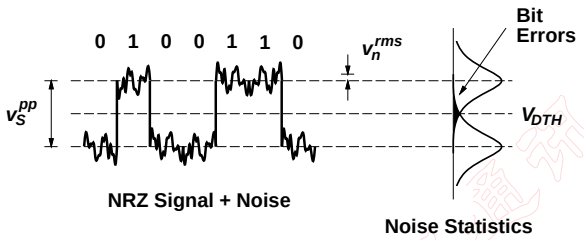

Figure 4.3: Relationship between signal, noise, and bit-error rate.

a complex situation, because the noise during the n-th bit slot depends on the values of all the preceding bits. In some books this noise analysis is carried out to the full extend [Per73, SP82], however here we will take the easy way out and assume that the noise is approximately stationary by the time it gets sampled. This lets us use the same equation for the detector noise as we used for the amplifier noise, except that it is time dependent:

$$\overline{v_{n,PD}^2}(t) = \int_0^{BW_D} |H(f)|^2 \cdot I_{n,PD}^2(f,t)\, df. \tag{4.6}$$

The total noise power at the output of the linear channel is obtained by adding the noise powers given in Eq. (4.4) and Eq. (4.6):

$$v_n^{rms} = \sqrt{\overline{v_{n,PD}^2} + \overline{v_{n,\text{amp}}^2}} = \sqrt{\int_0^{BW_D} |H(f)|^2 \cdot (I_{n,PD}^2 + I_{n,\text{amp}}^2)\, df}. \tag{4.7}$$

Signal, Noise, and Bit-Error Rate. Now that we have derived the value of the output rms noise, how is it related to the bit-error rate? The situation for a clean NRZ signal v_S with noise v_n is illustrated in Fig. 4.3. In the following we assume that the instantaneous noise voltage $v_n(t)$ follows a Gaussian distribution, which turns out to be a pretty good approximation in practice. With this assumption we can identify the standard deviation of the Gaussian distribution with the rms value of the noise voltage v_n^{rms} which we calculated in Eq. (4.7). For now, we assume further that (i) the noise is signal independent, i.e., that we have the same amount of noise on zeros and ones and (ii) that the output signal v_S is a clean NRZ signal without distortions. Later we will drop the last two assumptions.

The decision circuit determines if a bit is a zero or a one by comparing the output voltage v_O to the threshold voltage V_{DTH} which is located at the midpoint between zero and one (i.e., at the crossover point of the two distributions). This threshold value results in the lowest bit-error rate. The *Bit-Error Rate* (BER) is the probability that a zero is misinterpreted as a one or the other way round.[1]

[1]In fact, the term "bit-error rate" is misleading as it suggests a measurement of bit errors per time interval. A more accurate term would be "bit-error ratio" or "bit-error probability", however, because of the widespread use of the term "bit-error rate" we will stick with it.

$\mathcal{Q}$	*BER*	$\mathcal{Q}$	*BER*
0.0	1/2	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}

Table 4.1: Numerical relationship between $\mathcal{Q}$ and bit-error rate.

Given the above assumptions, we can derive a mathematical expression for the BER. We calculate the shaded areas under the Gaussian curves (which are equal) and sum them with a weight of 0.5 each (because zeros and ones are equally probable).

$$BER = \int_{\mathcal{Q}}^{\infty} \text{Gauss}(x)\, dx \quad \text{with} \quad \mathcal{Q} = \frac{v_S^{pp}}{2 \cdot v_n^{rms}} \tag{4.8}$$

where v_S^{pp} is the peak-to-peak value of the NRZ signal and v_n^{rms} is the rms value of the noise. The lower bound of the integral is the difference between the one (or zero) signal level and the decision threshold, $v_S^{pp}/2$, normalized to the standard deviation v_n^{rms} of the Gaussian. The $\mathcal{Q}$ parameter in the above equation, a.k.a. the *Personick Q*, is a measure of the ratio between signal and noise. (But it is not *identical* to SNR, as we will see later.) The integral in the above equation can be written as:

$$\int_{\mathcal{Q}}^{\infty} \text{Gauss}(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{\mathcal{Q}}^{\infty} e^{-\frac{x^2}{2}}\, dx = \frac{1}{2} \operatorname{erfc}\left(\frac{\mathcal{Q}}{\sqrt{2}}\right) \approx \frac{1}{\sqrt{2\pi}} \cdot \frac{\exp(-\mathcal{Q}^2/2)}{\mathcal{Q}}. \tag{4.9}$$

The approximation on the right is correct within 10 % for $\mathcal{Q} > 3$. The precise numerical relationship is tabulated in Table 4.1 for some commonly used BER values.

A Generalization: Unequal Noise Distributions. We know that the rms value of the output noise is time dependent because the detector noise depends on the value of the transmitted bit. To include this effect into our BER calculation we define two rms noise values, one for when a zero is transmitted, $v_{n,0}^{rms}$, and one for when a one is transmitted, $v_{n,1}^{rms}$. For this case of unequal noise distributions we can generalize Eq. (4.8) to:

$$BER = \int_{\mathcal{Q}}^{\infty} \text{Gauss}(x)\, dx \quad \text{with} \quad \mathcal{Q} = \frac{v_S^{pp}}{v_{n,0}^{rms} + v_{n,1}^{rms}}. \tag{4.10}$$

(See [Agr97] for a derivation.) This generalization is important for cases where the detector noise is significant compared to the amplifier noise, i.e., for receivers with an optically preamplified p-i-n detector or an APD detector.

Signal-to-Noise Ratio. We said earlier that $\mathcal{Q}$ is similar to the *Signal-to-Noise Ratio* (SNR), but not quite the same. So, what exactly is the relationship? SNR is defined as

the average signal power divided by the average noise power.[2] The average signal power is calculated as $\overline{v_S^2} - \overline{v}_S^2$ which turns out to be $(v_S^{pp}/2)^2$ for a DC-balanced NRZ signal.[3] The noise power is calculated as $\overline{v_n^2}$ which is $1/2 \cdot (\overline{v_{n,0}^2} + \overline{v_{n,1}^2})$ for an NRZ signal with an equal number of zeros and ones. The SNR follows as:

$$SNR = \frac{(v_S^{pp})^2}{2 \cdot (\overline{v_{n,0}^2} + \overline{v_{n,1}^2})}. \tag{4.11}$$

Comparing Eq. (4.10) and (4.11), we realize that there is no simple relationship between $\mathcal{Q}$ and *SNR*. However, there are two important special cases: (i) If the noise on zeros and ones is equal (noise dominated by amplifier) and (ii) if the noise on ones is much larger than on zeros (noise dominated by detector):

$$\begin{aligned} SNR &= \mathcal{Q}^2, && \text{if } \overline{v_{n,1}^2} = \overline{v_{n,0}^2} \\ SNR &= 1/2 \cdot \mathcal{Q}^2, && \text{if } \overline{v_{n,1}^2} \gg \overline{v_{n,0}^2}. \end{aligned} \tag{4.12}$$

For example, to achieve a BER of 10^{-12} ($\mathcal{Q} = 7.0$) we need an SNR of 16.9 dB in the first case and 13.9 dB in the second case.

At this point, you may wonder if you should use $10 \log \mathcal{Q}$ or $20 \log \mathcal{Q}$ to expressed $\mathcal{Q}$ in dBs. The above SNR discussion suggests $20 \log \mathcal{Q}$. But an equally strong case can be made for $10 \log \mathcal{Q}$ (for example look at Eq. (4.17) below). So, my advice is use $\mathcal{Q}$ on a linear scale whenever possible and if you must express $\mathcal{Q}$ in dBs, *always* clarify if you used $10 \log \mathcal{Q}$ or $20 \log \mathcal{Q}$ as a conversion rule.

SNR for TV Signals. Although this text is focused on the transmission of NRZ signals over optical fiber it is useful to be aware of other signal types as well. For example, in HFC/CATV systems multiple analog and/or digital TV signals are transported over an optical fiber using sub-carrier multiplexing (cf. Chapter 1). For these signals much higher SNRs than the $14 - 17$ dB for an NRZ signal are required. Actually, we should refer to *Carrier-to-Noise Ratio* (CNR) instead of SNR for those types of signals: cable-television engineers use the term CNR for RF modulated signals and reserve the term SNR for baseband signals such as NRZ. For analog TV, based on AM-VSB modulation, the National Association of Broadcasters recommends $CNR > 46$ dB. For digital TV, based on QAM-256 with FEC, typically $CNR > 30$ dB is required.

4.3 Sensitivity

Rather than asking: "What is the bit-error rate given a certain signal level?" we could ask the other way round: "What is the minimum signal level to achieve a given bit-error rate?". This minimum signal, when referred back to the input of the receiver, is known as the *Sensitivity*. The sensitivity is one of the most important characteristics of the optical receiver. It tells us to what level the transmitted signal can get attenuated during transmission over a long fiber and still be detected reliably by the receiver. We define an electrical and optical receiver sensitivity.

Definitions. *Electrical Receiver Sensitivity*, i_S^{pp} is defined as the minimum peak-to-peak signal current at the input of the receiver, necessary to achieve a specified BER. The current i_S^{pp} at the input of the linear channel causes as output signal voltage $v_S^{pp} = H_0 \cdot i_S^{pp}$ where H_0 is the passband value of $H(f)$ (see Fig. 4.2). We can now obtain the electrical sensitivity by solving Eq. (4.8) for v_S^{pp}:

$$i_S^{pp} = \frac{2\mathcal{Q} \cdot v_n^{rms}}{H_0}. \tag{4.13}$$

Before continuing it is useful to define the *input-referred* rms noise value:

$$i_n^{rms} = \frac{v_n^{rms}}{H_0}. \tag{4.14}$$

Similarly we can define the input-referred rms noise values: $i_{n,0}^{rms} = v_{n,0}^{rms}/H_0$, $i_{n,1}^{rms} = v_{n,1}^{rms}/H_0$, $i_{n,PD}^{rms} = v_{n,PD}^{rms}/H_0$, and $i_{n,\text{amp}}^{rms} = v_{n,\text{amp}}^{rms}/H_0$. In Section 4.4 we will discuss the properties of this noise, which is also known at total input-referred noise, in more detail. Now with the definition Eq. (4.14) we can rewrite the electrical sensitivity Eq. (4.13) in the simpler form:

$$i_S^{pp} = 2\mathcal{Q} \cdot i_n^{rms}. \tag{4.15}$$

In a situation with different amounts of noise on zeros and ones we can use Eq. (4.10) to obtain the more general sensitivity expression:

$$i_S^{pp} = \mathcal{Q} \cdot (i_{n,0}^{rms} + i_{n,1}^{rms}). \tag{4.16}$$

Optical Receiver Sensitivity, $\overline{P}_S$ is defined as the minimum optical power, averaged over time, necessary to achieve a specified BER.[4] Assuming a DC-balanced signal ($\bar{i}_S = i_S^{pp}/2$) and a photodetector with responsivity $\mathcal{R}$, this sensitivity is

$$\overline{P}_S = \frac{\mathcal{Q} \cdot i_n^{rms}}{\mathcal{R}}, \tag{4.17}$$

or more generally

$$\overline{P}_S = \frac{\mathcal{Q} \cdot (i_{n,0}^{rms} + i_{n,1}^{rms})}{2\mathcal{R}}. \tag{4.18}$$

Note that the optical sensitivity is defined as the *average* received power to meet a certain BER. This alone gives an RZ receiver an apparent sensitivity advantage of 3 dB over an NRZ receiver, given the same peak-to-peak signal and noise currents. The factor two

[2] In some books on optical communication [Sen85, Agr97], the SNR is defined as *peak* signal power divided by average noise power. Of course, the results derived with this SNR definition differ from ours. Here we are following the standard definition of SNR because it yields results consistent with digital communication theory.

[3] If the signal is time varying (modulated) like here, the mean power $\overline{v}_S^2$ is removed from the total power $\overline{v_S^2}$ to compute the signal power. However, if the signal is constant (unmodulated), the mean power is *not* removed, or else the signal would vanish. Cf. the SNR calculations in Sections 3.1 and 3.3 where the signal is constant.

[4] The optical receiver sensitivity defined in regulatory standards also takes into account power penalties caused by the use of a transmitter with a worst-case output (e.g., low extinction ratio). More on this in Section 7.1.

Parameter		2.5 Gb/s	10 Gb/s
Typical rms noise from amplifier	$i_{n,\text{amp}}^{rms}$	400 nA	1.2 μA
Input p-p signal for $BER = 10^{-12}$	i_S^{pp}	5.6 μA	16.8 μA
Sensitivity for p-i-n Detector	$\overline{P}_{S,PIN}$	−24.3 dBm	−19.5 dBm
Sensitivity for APD Detector	$\overline{P}_{S,APD}$	−34.3 dBm	−29.5 dBm
Sensitivity for OA + p-i-n Detector	$\overline{P}_{S,OA}$	−44.3 dBm	−39.5 dBm

Table 4.2: Approximate receiver sensitivity at $BER = 10^{-12}$ for various detectors. Only the amplifier noise is considered.

derives from the difference between $\bar{i}_S = i_S^{pp}/4$ for RZ modulation with pulses filling 50 % of the bit interval and $\bar{i}_S = i_S^{pp}/2$ for NRZ modulation.

Sometimes the optical sensitivity is specified for a receiver with an *ideal photodetector*. This optical sensitivity is designated by $\eta\overline{P}_S$ and, similar to the electrical sensitivity, is useful to compare the electrical performance of different receivers. With Eq. (3.2) we can express this sensitivity as:

$$\eta\overline{P}_S = \frac{hc}{\lambda q} \cdot \mathcal{Q} \cdot i_n^{rms}, \tag{4.19}$$

or more generally

$$\eta\overline{P}_S = \frac{hc}{\lambda q} \cdot \mathcal{Q} \cdot \frac{i_{n,0}^{rms} + i_{n,1}^{rms}}{2}. \tag{4.20}$$

Bit-Error Rates. When specifying a sensitivity we have to do that with reference to a BER. The following BERs are commonly used: The SONET OC-48 standard (2.5 Gb/s) requires a system bit-error rate of $\leq 10^{-10}$ which corresponds to $\mathcal{Q} \approx 6.4$. The faster SONET OC-192 standard (10 Gb/s) is even stricter and requires a system bit-error rate of $\leq 10^{-12}$ corresponding to $\mathcal{Q} \approx 7.0$. Component manufacturers usually aim at even lower BERs such as 10^{-15} to meet the system BERs quoted above.

The problem with BERs as low as 10^{-15} is that they are very time consuming to measure. For example to collect 10 errors at a BER of 10^{-15} and a bit rate of 10 Gb/s we have to wait for 10^6 seconds or 12 days! By the way, how many errors should we collect to obtain a good BER estimate? Is 10 errors enough? If we assume that the error statistics follows a Poisson distribution, its standard deviation is $\sqrt{n}$ for n collected errors. So, for 10 collected errors the standard deviation is 3.2 or we have a 32 % uncertainty, for 100 errors the uncertainty reduces to 10 %, etc.

Sensitivity Analysis Based on Amplifier Noise Only. To get a feeling for sensitivity numbers we want to carry out some numerical calculations. For this first calculation we will ignore the detector noise and use Eq. (4.17) with $i_n^{rms} = i_{n,\text{amp}}^{rms}$ to estimate the sensitivity based on the amplifier noise only. For the p-i-n detector we get:

$$\overline{P}_{S,PIN} = \frac{\mathcal{Q} \cdot i_{n,\text{amp}}^{rms}}{\mathcal{R}}. \tag{4.21}$$

The APD and the optically preamplified p-i-n detector have responsivities which are enhanced by their respective gains:

$$\overline{P}_{S,APD} = \frac{1}{M} \cdot \frac{\mathcal{Q} \cdot i_{n,\text{amp}}^{rms}}{\mathcal{R}} \tag{4.22}$$

and

$$\overline{P}_{S,OA} = \frac{1}{G} \cdot \frac{\mathcal{Q} \cdot i_{n,\text{amp}}^{rms}}{\mathcal{R}}. \tag{4.23}$$

For our numerical calculations we use the typical amplifier noise values $i_{n,\text{amp}}^{rms} = 400$ nA at 2.5 Gb/s and 1.2 μA at 10 Gb/s which are based on the TIA noise data from Section 5.1.4. We continue to use the typical detector values ($\mathcal{R} = 0.75$ A/W, $M = 10$, $G = 100$) introduced earlier. Our reference BER is 10^{-12}. The resulting approximate sensitivities for all three detector types are listed in Table 4.2. We see how the sensitivity improves along with the responsivity as we go from the p-i-n detector to the APD detector and finally to the optically preamplified p-i-n detector.

How important is a difference of 1 dB in sensitivity? Is it worth a lot of trouble to improve the sensitivity by a single dB? In a system without optical amplifiers for signal regeneration, a 1 dB improvement in sensitivity means an extended reach of 4 km. This is simply because the attenuation of a fiber at 1.55 μm is about 0.25 dB/km. For this reason system designers usually care about small sensitivity degradations such as 0.1 or 0.2 dB.

Sensitivity Analysis Including Detector Noise. Now we want to repeat the sensitivity calculations, but this time taking into account the detector noise and its unequal noise distributions for zeros and ones. This exercise will show us the relative significance of the detector noise. Using Eq. (4.18) with $\overline{i_{n,0}^2} = \overline{i_{n,PIN,0}^2} + \overline{i_{n,\text{amp}}^2}$ and $\overline{i_{n,1}^2} = \overline{i_{n,PIN,1}^2} + \overline{i_{n,\text{amp}}^2}$ and Eqs. (3.6), (3.7), we can derive the sensitivity for the p-i-n diode receiver:

$$\overline{P}_{S,PIN} = \frac{\mathcal{Q} \cdot i_{n,\text{amp}}^{rms}}{\mathcal{R}} + \frac{\mathcal{Q}^2 \cdot q \cdot BW_n}{\mathcal{R}} \tag{4.24}$$

where BW_n is the noise bandwidth of the receiver front-end. We will discuss this bandwidth in more detail in Section 4.4. For now we can just take it as 0.75 times the bit rate. The first part of this equation is the same expression as in Eq. (4.21), the second term is due to the shot noise of the photodiode. Using the APD noise equation we can derive the sensitivity of a receiver with an APD detector:

$$\overline{P}_{S,APD} = \frac{1}{M} \cdot \frac{\mathcal{Q} \cdot i_{n,\text{amp}}^{rms}}{\mathcal{R}} + F \cdot \frac{\mathcal{Q}^2 \cdot q \cdot BW_n}{\mathcal{R}}. \tag{4.25}$$

Similarly, we obtain the sensitivity of an optically preamplified p-i-n detector:

$$\overline{P}_{S,OA} = \frac{1}{G} \cdot \frac{\mathcal{Q} \cdot i_{n,\text{amp}}^{rms}}{\mathcal{R}} + \eta F \cdot \frac{\mathcal{Q}^2 \cdot q \cdot BW_n}{\mathcal{R}}. \tag{4.26}$$

Comparing these three equations, we observe that the first term, which contains the amplifier noise, is suppressed with increasing detector gain (p-i-n → APD → OA). The second term, caused by the detector noise, however grows proportional to the noise figure (or excess noise factor) of the detector.

Parameter		2.5 Gb/s	10 Gb/s
Typical rms noise from amplifier	$i_{n,\mathrm{amp}}^{rms}$	400 nA	1.2 μA
Sensitivity for p-i-n Detector	$\overline{P}_{S,PIN}$	−24.3 dBm	−19.5 dBm
Sensitivity for APD Detector	$\overline{P}_{S,APD}$	−33.1 dBm	−28.0 dBm
Sensitivity for OA + p-i-n Detector	$\overline{P}_{S,OA}$	−41.3 dBm	−35.8 dBm

Table 4.3: Receiver sensitivity at $BER = 10^{-12}$ for various detectors. Amplifier and photodetector noise is considered.

Now we want to evaluate these equations for the typical detector values introduced earlier ($F = 6$ [7.8 dB] for APD and $F = 3.16$ [5 dB] and $\eta = 0.6$ for OA). We obtain the numbers shown in Table 4.3. When comparing these numbers with the earlier approximations, we notice the following: The sensitivity for the p-i-n detector did not change at all, which means that the shot noise contributed by the p-i-n photodiode is negligible compared to the amplifier noise. The sensitivity of the APD detector degraded by a little more than 1 dB, which means that neglecting the noise of the APD detector gets us only a rough sensitivity estimate. The sensitivity estimate of the optically preamplified p-i-n detector was about $3 - 4$ dB too optimistic, i.e., we definitely need to include the noise of the optical preamplifier.

BER Plots. To characterize the performance of an optical receiver it is very useful to measure the BER as a function of received power. From Eq. (4.24) and its discussion we know that for a p-i-n receiver the $\mathcal{Q}$-factor corresponding to the measured BER should be linearly related to the received optical power $\overline{P}_S$. It is thus convenient to use graph paper which presents this relationship as a straight line. One possibility is to plot $\mathcal{Q}$ on the y-axis and $\overline{P}_S$ on the x-axis. For easy reading of the graph, the y-axis may still be labeled in BER units, but note that these labels are neither linearly nor logarithmically spaced. Alternatively, if we prefer to represent the power in dBm rather than in mW, we can plot $10 \log \mathcal{Q}$ on the y-axis and $10 \log \overline{P}_S$, i.e., the power in dB, on the x-axis. Figures 4.4 and 4.5 illustrate both types of plots. For receivers with an APD or optically preamplified p-i-n detector, the plots on this graph paper are not exact straight lines (cf. Eqs. (4.25) and (4.26)).

Data points plotted on these graph papers can easily be *extrapolated* down to very low BERs which would be hard (slow) to measure. For example Figs. 4.4 and 4.5 show an extrapolation down to $BER = 10^{-15}$ (same data is used in both plots). However, we have to be very cautious when carrying out such extrapolations! The result is only correct if the system noise closely follows a Gaussian distribution over a range of many sigmas. Some receivers exhibit a BER floor, which means that even for a very high received power the BER never goes below a certain value, the BER floor. The extrapolation goes right past this floor and predicts the wrong result!

BER plots, like the ones in Figs. 4.4 and 4.5 are also a very useful analytical tool. For example, from a few measured data points we can see if the data follow a straight line or if there is a "bend" which may indicate a BER floor. From the slope of the line in the linear BER plot we can infer the receiver noise (assuming $\mathcal{R}$ is known), and from a horizontal shift we can infer an offset problem, etc. Of course, BER plots don't necessarily have to

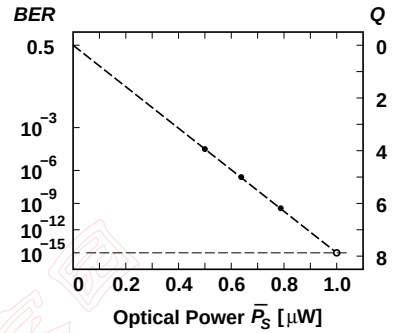

Figure 4.4: BER plot in the linear $(\overline{P}_S, \mathcal{Q})$ coordinate system with an extrapolation down to $BER = 10^{-15}$.

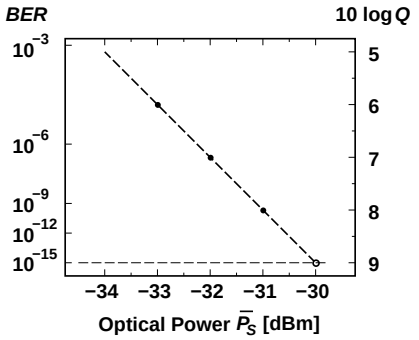

Figure 4.5: BER plot in the $(\log \overline{P}_S, \log \mathcal{Q})$ coordinate system with an extrapolation down to $BER = 10^{-15}$.

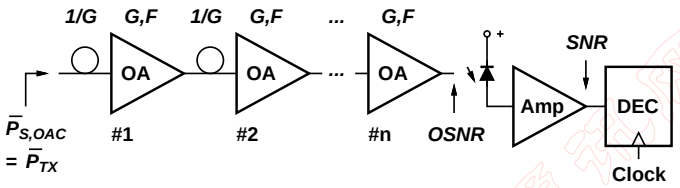

Figure 4.6: A cascade of optical amplifiers followed by a p-i-n receiver.

be a function of received optical power, as in the figures, but can also be a function of the input current (for TIAs) or the input voltage (for MAs).

Optimum APD Gain. From Eq. (4.25) we can clearly see that there is an optimum APD gain, M_{opt}, which yields the best sensitivity. If M is chosen too low, the first term containing $1/M$ limits the sensitivity; if M is chosen too high the second term starting with F limits the sensitivity. Remember that the excess noise factor F increases with gain M according to Eq. (3.11). Note that the situation is different for the optically preamplified p-i-n detector where the noise figure F decreases with increasing gain G.

Combining Eqs. (4.25) and (3.11) and solving for the M that minimizes $\overline{P}_{S,APD}$ yields:

$$M_{\text{opt}} = \sqrt{\frac{i_{n,\text{amp}}^{rms}}{\mathcal{Q} \cdot k_A \cdot q \cdot BW_n} - \frac{1-k_A}{k_A}}. \tag{4.27}$$

M_{opt} increases with amplifier noise, which makes sense because the APD gain helps to suppress this noise. M_{opt} decreases with k_A which means that the optimum gain for an InGaAs APD ($k_A \approx 0.6$) is smaller than that for a Si APD ($k_A \approx 0.03$).

Cascade of Optical Amplifiers. In ultra long-haul fiber links, for example connecting two continents, many optical amplifiers are inserted periodically to regenerate the signal (so-called inline amplifiers). To be specific, let's assume an 8,000 km long fiber link. To compensate the fiber loss of 0.25 dB/km we insert an optical amplifier with gain $G = 20$ dB every 80 km. This makes a total of 100 amplifiers; and they all contribute noise! In such a scenario the optical amplifier noise becomes absolutely dominant.

To analyze this situation more quantitatively, let's consider the whole 8,000 km link including all the optical amplifiers to be part of the detector (see Fig. 4.6). In other words, the input of our receiver is right at the transmitter end. Furthermore, we assume that the fiber loss is exactly balanced by the optical amplifier gain. Then we can write the "sensitivity" as:

$$\overline{P}_{S,OAC} = \frac{\mathcal{Q} \cdot i_{n,\text{amp}}^{rms}}{\mathcal{R}} + \eta \cdot nGF \cdot \frac{\mathcal{Q}^2 \cdot q \cdot BW_n}{\mathcal{R}}. \tag{4.28}$$

The first term is similar to that in Eq. (4.26), but lacking the $1/G$ factor because the amplifier gain is offset by the span loss. The second term, again comparing it to Eq. (4.26), corresponds to an optical preamplifier with the terribly bad noise figure $F' = nGF$. For

100 amplifiers ($n = 100$) and the typical values introduced earlier we get $nGF = 45$ dB and the "sensitivity" $\overline{P}_{S,OAC}$ becomes +1.7 dBm, essentially regardless of the electrical amplifier noise $i_{n,\text{amp}}^{rms}$. This "sensitivity" is very low, but recall that it refers to the channel input which is the transmitter end! So all this means is that the transmitter must launch a power of at least 1.7 dBm into the fiber and we are fine.

We draw two conclusions from this last example: (i) It is in fact possible to send an optical signal over 8,000 km of fiber trough a chain of 100 EDFAs and receive it with a low BER of 10^{-12}! For example, the commercial transpacific cable TPC-5 segment J is 8,620 km long and contains about 260 EDFA-type repeaters spaced 33 km apart. (ii) The concept of receiver sensitivity looses its meaning in a situation where many in-line EDFAs contribute most of the system noise. Remember, receiver sensitivity is the minimum power required to achieve a certain BER based on the *receiver* noise.

In optically amplified transmission systems, the system designer is interested in the minimum OSNR required at the receiver end rather than the sensitivity of the receiver. To illustrate this, let's repeat the above calculation but now thinking in terms of OSNR. First we want to know how much OSNR we need at the receiver to achieve a certain BER? After a chain of many EDFAs the noise on ones will be much larger than the noise on zeros, therefore Eq. (4.12) gives the required SNR as $1/2 \cdot \mathcal{Q}^2$. The required OSNR can be derived from this SNR with the help of Eq. (3.17) and we obtain [dSBB+99]:

$$OSNR \approx \frac{BW_n}{BW_O} \cdot \mathcal{Q}^2. \tag{4.29}$$

For example given a 7.5 GHz receiver noise bandwidth, we need an OSNR of 14.7 dB measured in a 0.1 nm optical bandwidth (12.5 GHz at $\lambda = 1.55\,\mu$m) to achieve a BER of 10^{-12}. Next we want to know how much OSNR we have got at the end of a chain of n amplifiers. This OSNR can be expressed in terms of amplifier noise figure by using the first term of Eq. (3.22):

$$OSNR = \frac{\overline{P}_{TX}}{S_{ASE} \cdot BW_O} \approx \frac{\overline{P}_{TX}}{nGF \cdot hc/\lambda \cdot BW_O} \tag{4.30}$$

where $\overline{P}_{TX}$ is the mean output power launched by the transmitter, F is the amplifier noise figure, and G is the amplifier gain which is equal to the span loss. If Eq. (4.30) is transformed to the log domain and specialized for $\lambda = 1.55\,\mu$m and a BW_O corresponding to 0.1 nm we obtain the useful engineering rule [ZNK97]:

$$OSNR[\text{dB}] \approx 58\,\text{dB} + \overline{P}_{TX}[\text{dBm}] - G[\text{dB}] - F[\text{dB}] - 10\log n. \tag{4.31}$$

For example with the familiar values $n = 100$, $F = 5$ dB, $G = 20$ dB we find that we need $\overline{P}_{TX} = 1.7$ dBm to achieve the required OSNR of 14.7 dB in a 0.1 nm optical bandwidth. Well that's the same transmit power we have obtained before with Eq. (4.28)!

If we want to design a practical long-haul transmission system the above idealized OSNR calculations must be refined as follows: (i) Transmission penalties due to fiber dispersion, polarization effects, nonlinear pulse-shape distortions, nonlinear signal/noise mixing, and crosstalk in WDM systems must be considered. (ii) Margins for system

Parameter		2.5 Gb/s	10 Gb/s
p-i-n Detector Limit	$\overline{P}_{S,PIN}$	−47.1 dBm	−41.1 dBm
APD Detector Limit	$\overline{P}_{S,APD}$	−39.3 dBm	−33.3 dBm
OA + p-i-n Detector Limit	$\overline{P}_{S,OA}$	−44.3 dBm	−38.3 dBm

Table 4.4: Maximum receiver sensitivity at $BER = 10^{-12}$ for various detectors. A noiseless amplifier is assumed.

Parameter		2.5 Gb/s	10 Gb/s
Quantum Limit	$\overline{P}_{S,\text{quant}}$	−53.6 dBm	−47.6 dBm

Table 4.5: Quantum limits for $BER = 10^{-12}$.

aging, repairs, etc. must be allocated. (iii) If forward error correction (FEC) is used, the required $\mathcal{Q}$ value is lower than given in Table 4.1 (cf. Section 4.11).

We are now finished with the main part of this section. We understand the significance of the detector noise relative to the amplifier noise as well as its impact on receiver sensitivity and BER. If you are tired of this subject you can skip over the rest if this section. In the remainder we want to explore the theoretical sensitivity limits of an optical receiver.

Sensitivity Analysis for a Noiseless Amplifier. To study how sensitive we can make a receiver in theory, we will repeat our sensitivity calculations once more but this time neglecting the amplifier noise. We can calculated these sensitivities, which are based on the noise produced by the detector only, from Eqs. (4.24), (4.25), and (4.26) while setting $i_{n,\text{amp}}^{rms} = 0$. Numerical values for all three detectors are listed in Table 4.4.

The first observation is that the p-i-n detector gives us the best sensitivity in theory. This is because the noise of the p-i-n detector is so low. But as we have seen, in reality, the sensitivity of the p-i-n detector is degraded by about 22 dB because of the amplifier noise. The second observation regarding the optically preamplified p-i-n detector is that although the noise from the optical preamplifier dominates the amplifier noise, we could still squeeze out another $2-3$ dB of sensitivity with an ultra low-noise amplifier.

The above calculations also give us an interesting interpretation of the detector excess noise factor and noise figure. Once we have calculated the p-i-n detector limit, the APD limit can be obtained by simply adding the excess noise factor F in dB: At 2.5 Gb/s we get $-47.1\,\text{dBm} + 7.8\,\text{dB} = -39.3\,\text{dBm}$ for an APD with $F = 7.8$ dB. Similarly, the OA limit can be obtained by adding ηF in dB: At 2.5 Gb/s we get $-47.1\,\text{dBm} - 2.2\,\text{dB} + 5\,\text{dB} = -44.3\,\text{dBm}$ for an OA with $F = 5$ dB and $\eta = 0.6$. This fact can be explained with Eqs. (4.24), (4.25), and (4.26). In other words, F (for APD) and ηF (for OA) tells us how much less sensitive the detector is compared to a p-i-n detector in the limit of zero amplifier noise.

Quantum Limit. Can we build a receiver with arbitrarily high sensitivity, at least in theory? Maybe we can use some fancy detector with a noiseless amplifier and such? No! There is a quantum limit that cannot be surpassed. Next, we will derive this limit.

The quantum limit is obtained from the observation that at least one photon must be received for each transmitted one bit to have error-free reception. The number of photons, n, contained in a one bit follows the Poisson distribution:

$$\text{Poisson}(n) = e^{-M} \cdot \frac{M^n}{n!} \tag{4.32}$$

where M is the mean of the distribution. The total error probability equals half the error probability for zeros plus half the error probability for ones. Since nothing is transmitted for zeros, the probability of error for zeros is zero. The probability of error for ones is Poisson(0), corresponding to the situation when zero photons are received for a one bit. Thus $BER = 1/2 \cdot \text{Poisson}(0)$. Given a bit-error rate of 10^{-12} we find that an average number of $M = -\ln(2 \cdot BER) = 27$ photons are required per one bit. Assuming 50 % mark density, an average of $M/2$ photons are required per bit. The quantum limit sensitivity follows as:

$$\overline{P}_{S,\text{quant}} = \frac{-\ln(2 \cdot BER)}{2} \cdot \frac{hc}{\lambda} \cdot B \tag{4.33}$$

where B is the bit rate.[5] Numerical values of this expression with $\lambda = 1.55\,\mu$m and $BER = 10^{-12}$ are is listed in Table 4.5.

We see that with an optically preamplified p-i-n detector we are coming within about 12 dB of the quantum limit! This means that such a detector has the ability to correctly recognize 16 photons (or more) as a one bit.

4.4 Personick Integrals

Total Input-Referred Noise. We have seen in the previous section that the input-referred rms noise values (i_n^{rms}, $i_{n,0}^{rms}$, $i_{n,1}^{rms}$) play a key role in determining the receiver sensitivity. In Eq. (4.14) this noise value has been defined as $i_n^{rms} = v_n^{rms}/H_0$. Therefore, with Eq. (4.7), the input-referred noise power $(i_n^{rms})^2$ can be calculated from the power spectrum as follows:

$$\overline{i_n^2} = \frac{1}{H_0^2} \int_0^{BW_D} |H(f)|^2 \cdot I_n^2(f)\, df \tag{4.34}$$

where $I_n^2(f) = I_{n,PD}^2(f) + I_{n,\text{amp}}^2(f)$, the noise power spectrum of the combined detector and amplifier noise.

Equation (4.34) is easy to apply when running simulations, but it looks quite cumbersome for analytical hand calculations. Is there an easier way to calculate the total input-referred noise from the input-referred spectrum? It is tempting to just integrate the input-referred noise spectrum over all frequencies:

$$\overline{i_n^2} \stackrel{?}{=} \int_0^\infty I_n^2(f)\, df. \tag{4.35}$$

The problem with this integral is that it does not converge giving us an infinite noise current. This can't be right! For this integral to converge the power spectrum needs to roll off with frequency, not to increase.

[5] The quantum limit derived here is for on-off keying (OOK). Other modulation formats such as *Phase-Shift Keying* (PSK) or *Frequency-Shift Keying* (FSK) can be detected with a slightly better sensitivity [HLG88].

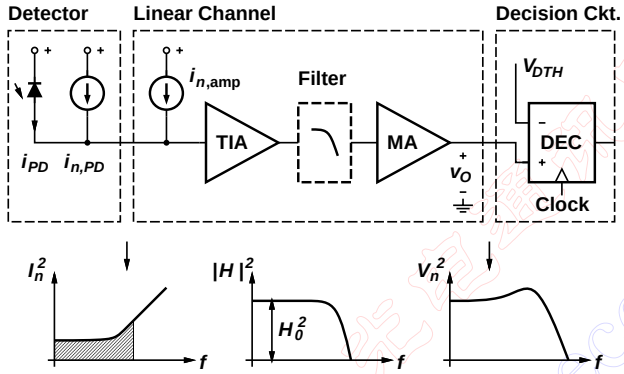

Figure 4.7: How *not* to calculate total input-referred noise.

Maybe, we may think, we should integrate only up to the 3-dB bandwidth of the receiver:

$$\overline{i_n^2} \stackrel{?}{=} \int_0^{BW_{3dB}} I_n^2(f)\, df. \tag{4.36}$$

This value is shown as the cross hatched area in Fig. 4.7. At least, now we get a finite result. But the result is very sensitive to the upper bound of the integration (the 3-dB bandwidth) because it lies in the rising part of the spectrum. So if we were to use the 1-dB bandwidth instead of the 3-dB bandwidth our noise result would come out quite a bit different. This doesn't sound right either.

Noise Bandwidths. Actually, there is a way to easily calculate the total input-referred noise from the associated noise spectrum. We start out by writing the input-noise spectrum in the general form introduced in Section 4.1:

$$I_n^2(f) = a + b \cdot f^2. \tag{4.37}$$

Constant a describes the white part of the spectrum, and b describes the f^2-noise part (we are neglecting possible $1/f$ and f-noise terms here). Now we plug this spectrum into Eq. (4.34)

$$\overline{i_n^2} = \frac{1}{H_0^2} \int_0^{BW_D} |H(f)|^2 \cdot (a + bf^2)\, df \tag{4.38}$$

expand it

$$\overline{i_n^2} = a \cdot \frac{1}{H_0^2} \int_0^{BW_D} |H(f)|^2\, df + b \cdot \frac{1}{H_0^2} \int_0^{BW_D} |H(f)|^2 \cdot f^2\, df \tag{4.39}$$

and rewrite it as

$$\overline{i_n^2} = a \cdot BW_n + b/3 \cdot BW_{n2}^3 \tag{4.40}$$

$H(f)$	BW_n	BW_{n2}
1st-order low pass	$1.57 \cdot BW_{3dB}$	∞
2nd-order low pass, crit. damped ($Q = 0.500$)	$1.22 \cdot BW_{3dB}$	$2.07 \cdot BW_{3dB}$
2nd-order low pass, Bessel ($Q = 0.577$)	$1.15 \cdot BW_{3dB}$	$1.78 \cdot BW_{3dB}$
2nd-order low pass, Butterworth ($Q = 0.707$)	$1.11 \cdot BW_{3dB}$	$1.49 \cdot BW_{3dB}$
Brick wall low pass	$1.00 \cdot BW_{3dB}$	$1.00 \cdot BW_{3dB}$
Rectangular filter	$0.500 \cdot B$	∞
Raised-cosine filter (FRC-NRZ)	$0.564 \cdot B$	$0.639 \cdot B$

Table 4.6: Numerical values for BW_n and BW_{n2}.

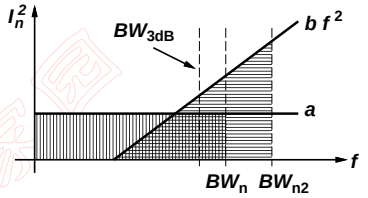

Figure 4.8: Interpretation of BW_n and BW_{n2} as integration bounds.

where

$$BW_n = \frac{1}{H_0^2} \int_0^{BW_D} |H(f)|^2\, df, \tag{4.41}$$

$$BW_{n2}^3 = \frac{3}{H_0^2} \int_0^{BW_D} |H(f)|^2 \cdot f^2\, df. \tag{4.42}$$

The bandwidths BW_n and BW_{n2} depend *only* on the frequency response $H(f)$ of the receiver and the bandwidth of the decision circuit BW_D. The latter is uncritical as long as it is larger than the receiver bandwidth and the receiver has a steep roll-off. In the following we will assume the decision-circuit bandwidth to be infinite. The bandwidths BW_n and BW_{n2} for some simple receiver responses are listed in Table 4.6. Once these two bandwidths are known we can easily calculate the total input-referred noise with Eq. (4.40).

Why did we choose the peculiar $b/3$ term in Eq. (4.40)? Because it results in a neat interpretation of the bandwidths BW_n and BW_{n2}. If we were to integrate the input spectrum Eq. (4.37) up to the 3-dB point we would get:

$$\overline{i_n^2} = a \cdot BW_{3dB} + b/3 \cdot BW_{3dB}^3. \tag{4.43}$$

Equation (4.40) can thus be interpreted as the result of integrating the white-noise component of the input-referred spectrum up to BW_n and the f^2-noise component up to BW_{n2}. This interpretation is illustrated graphically in Fig. 4.8. You may already have noticed it: BW_n is identical to the *Noise Bandwidth* of the receiver's frequency response. In a way,

BW_{n2} could be called the 2nd-order noise bandwidth, because it plays the same role as the noise bandwidth but with the white noise replaced by f^2 noise.

Now let's go back and see how different this is from integrating the input-referred spectrum up to the 3-dB point. Integrating up to the 3-dB point means that we set $BW_n = BW_{n2} = BW_{3\text{dB}}$. By consulting Table 4.6 we see that in the case of the brick-wall frequency response we obtain the correct result. But in the case of a second-order Butterworth response, we underestimate the white-noise power by 1.11× and the f^2-noise power by 3.33×, quite a significant difference!

Personick Integrals. For an Electrical Engineer the BW_n and BW_{n2} bandwidths have an intuitive meaning and this is why I introduced them here. However, in the literature, such as [Per73, SP82, Kas88, BM95], you will find the so-called *Personick Integrals* instead, usually designated with I_1, I_2, and I_3. The second and third Personick integrals are directly related to our BW_n and BW_{n2} bandwidths:

$$I_2 = BW_n/B, \tag{4.44}$$
$$I_3 = BW_{n2}^3/(3B^3) \tag{4.45}$$

where B is the bit rate. (The first Personick Integral relates to the nonstationary detector noise, which we wiped under the rug in Section 4.2.) With these integrals we can write the input-referred noise power as a function of the bit rate:

$$\overline{i_n^2} = a \cdot I_2 B + b \cdot I_3 B^3. \tag{4.46}$$

4.5 Power Penalty

In this section we want to introduce the important concept of *Power Penalty* which is a useful tool to quantify impairments in the receiver, transmitter, and fiber.

The basic idea is the following: We have seen in the previous section that the receiver noise determines the *basic* receiver sensitivity. The *actual* sensitivity is lower than this because of a variety of impairments such as distortion (ISI) introduced by the linear channel of the receiver, offset errors in the decision threshold, secondary noise sources, etc. The power penalty PP is defined as the loss in optical sensitivity due to a particular impairment or, equivalently, the amount of additional transmit power needed to achieve the same BER as in the absence of the impairment.[6] Power penalties are usually expressed in dBs using the conversion rule $10 \cdot \log PP$. Table 4.7 lists examples of impairments in various parts of an optical communication system. For many of these impairments we will later calculate the associated power penalty.

Example 1: Decision-Threshold Offset. To illustrate this concept let's make an example and calculate the power penalty for the case that V_{DTH} is not exactly centered

[6] We will assume in the following that the act of increasing the transmit power does not introduce impairments of its own, such as an increase in system noise or nonlinear distortions. In systems with optical amplifiers or APDs this may not be true and the power penalties must be modified. [RS98]

Transmitter:	Extinction ratio
	Relative intensity noise
	Output power variations
Fiber:	Dispersion
	Nonlinear effects
Detector:	Dark current
TIA:	Distortions (ISI)
	Offset
MA:	Distortions (ISI)
	Offset
	Noise figure
	Low-frequency cutoff
CDR:	Decision-threshold offset
	Decision-threshold ambiguity
	Sampling-time offset
	Sampling-time jitter

Table 4.7: Examples of impairments leading to power penalties.

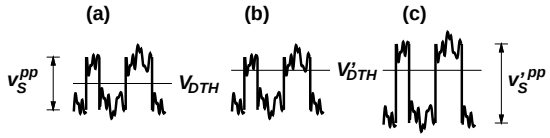

Figure 4.9: Power penalty due to decision-threshold offset: (a) without offset, (b) with offset, (c) with offset and increased signal amplitude to restore the original bit-error rate.

between the zero and one level. Figure 4.9(a) shows the situation with an ideally centered threshold. (We'll assume a linear channel and equal noise on zeros and ones for this example.) In Fig. 4.9(b) the threshold voltage V'_{DTH} is off center:

$$V'_{DTH} = V_{DTH} + \epsilon \cdot v_S^{pp} \tag{4.47}$$

where ϵ is the threshold offset relative to the signal value in percent. The offset error in Fig 4.9(b) will cause many ones to be misinterpreted as zeros thus significantly increasing the bit-error rate.

To restore the original BER we need to increase the signal voltage from v_S^{pp} to $v_S'^{pp}$ as shown in Fig. 4.9(c). The signal voltage is increased such that the voltage difference between the one level and the decision threshold is the same as in Fig. 4.9(a):

$$v_S'^{pp} = v_S^{pp} + 2\epsilon \cdot v_S^{pp} = v_S^{pp}(1 + 2\epsilon). \tag{4.48}$$

As a result the probability of misinterpreting a one as a zero is now the same as in Fig. 4.9(a). The probability of misinterpreting a zero as a one is now even better, so the overall BER is somewhat lower than in Fig. 4.9(a), but we'll ignore this effect here as it has little impact on the power penalty. Knowing that the signal voltage is proportional to the received optical power, the power penalty for a decision-threshold error ϵ must be:

$$PP = 1 + 2\epsilon. \tag{4.49}$$

For example, a 10 % decision-threshold error causes a power penalty of 0.79 dB.

The power-penalty concept is especially useful to derive receiver specifications. If we solve Eq. (4.49) for ϵ we find:

$$\epsilon = \frac{PP - 1}{2}. \tag{4.50}$$

This means that if the largest acceptable power penalty is PP, we must control the decision threshold to a precision better than ϵ given in Eq. (4.50). For example, given a worst-case power penalty of 0.2 dB ($PP = 1.047$) the decision-threshold error must be less than 2.4 %.

Example 2: Dark Current. In Chapter 3 we mentioned the detector dark current and how it interferes with the received signal. Now we have the necessary tools to quantify this effect! The dark current by itself does not negatively impact the received signal, it just adds an offset but leaves the peak-to-peak value unchanged. There is no power penalty for this. However, the *noise* associated with the dark current will enhance the receiver noise and cause a power penalty. Let's calculate it!

According to Eq. (3.5) the dark current I_{DK} in a p-i-n detector causes the shot noise current:

$$\overline{i^2}_{n,DK} = 2qI_{DK} \cdot BW_n. \tag{4.51}$$

This noise power adds to the receiver noise, which we assume is dominated by the amplifier noise $\overline{i^2}_{n,\text{amp}}$. (Neglecting the detector noise overestimates the power penalty somewhat.) So the dark current noise increases the noise power by:

$$\frac{\overline{i^2}_{n,\text{amp}} + \overline{i^2}_{n,DK}}{\overline{i^2}_{n,\text{amp}}} = 1 + \frac{2qI_{DK} \cdot BW_n}{\overline{i^2}_{n,\text{amp}}}. \tag{4.52}$$

We know from Eq. (4.17) that the receiver sensitivity is proportional to the rms noise current and thus we have found the power penalty:

$$PP = \sqrt{1 + \frac{2qI_{DK} \cdot BW_n}{\overline{i^2}_{n,\text{amp}}}}. \tag{4.53}$$

With the typical numbers for our 2.5 Gb/s receiver ($i^{rms}_{n,\text{amp}} = 400$ nA, $BW_n = 1.9$ GHz) and a worst-case dark current of 20 nA we find the power penalty to be 0.00017 dB ($PP = 1.000038$). As expected this is very, very small. For an APD detector we had to replace I_{DK} with $F \cdot M^2 \cdot I_{DK}$ where I_{DK} is now the *primary* dark current. In this case the power penalty would be larger.

Now we can turn this game around and ask: "What is the maximum allowable dark current for a given power penalty?" A little algebra reveals:

$$I_{DK} < (PP^2 - 1) \cdot \frac{\overline{i^2}_{n,\text{amp}}}{2q \cdot BW_n}. \tag{4.54}$$

With the same typical numbers as before we find that the dark current must be less than 25 μA to keep the power penalty below 0.2 dB ($PP = 1.047$). No problem!

With these two examples we have illustrated how to compute power penalties and how to derive specifications from them.

4.6 Bandwidth

Should we make the bandwidth of the receiver wide or narrow? If we make it wide, the receiver adds very little distortion to the signal. But at the same time a wideband receiver produces a lot of noise and we know from Section 4.2 that this noise reduces the receiver sensitivity. Alternatively, if we choose a narrow bandwidth the noise is reduced which is good. But now we have distortions called *Inter-Symbol Interference* (ISI) in the output signal. ISI also reduces the sensitivity because the worst-case output signal is reduced. We can conclude from this thought experiment that there must be an *Optimum Receiver Bandwidth* for which we get the best sensitivity. As a rule of thumb this 3-dB bandwidth is:

$$BW_{3\text{dB}} \approx \frac{2}{3} \cdot B \tag{4.55}$$

where B is the bit rate and NRZ modulation is assumed.

Example: Butterworth Receiver. Figure 4.10 illustrates the trade-off between ISI and noise with an example for a 10 Gb/s receiver. The received input signal is a clean NRZ waveform, the noise is white, and the receiver has a 2nd-order Butterworth response (i.e., maximally flat amplitude response). The output signals for 3 different bandwidths are shown from top to bottom in the form of eye diagrams.

For the moment, let's ignore the gray stripes in the eye diagrams. The top eye diagram shows the output voltage signal from the wideband receiver with a 3-dB bandwidth of 4/3 the bit rate, i.e., twice the optimum bandwidth. As expected, we get a clean eye

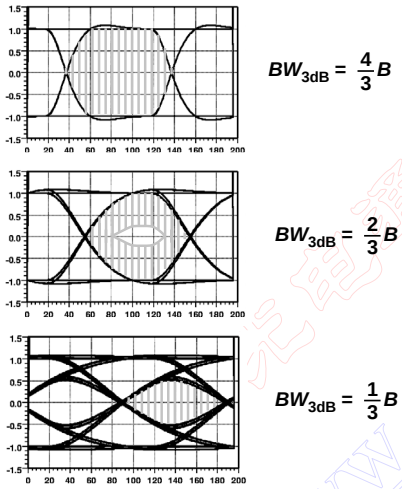

Figure 4.10: Trade-off between ISI and noise in a receiver.

pattern with almost no ISI. The next eye diagram is the output of the receiver with the optimum bandwidth, 2/3 the bit rate. Finally, the bottom eye diagram is the output of the narrowband receiver with half the optimum bandwidth. In this case we observe severe ISI. In particular, we can see a trace with the full amplitude corresponding to the bit sequence "00110011..." and another trace with only about half the amplitude corresponding to the bit sequence "01010101...". The result of ISI is a partially closed eye pattern.

Now let's add some white noise. The received noise power is approximately proportional to the receiver bandwidth. Therefore the noise voltage is proportional to the square root of the bandwidth. The gray stripes in Fig. 4.10 represent the noise voltage which gets smaller as we go from the wideband to the narrowband receiver. For clarity only the noise inside the eye is drawn, of course in reality noise is present on both sides of the signal trace.

But wait a minute, what exactly does the height of these gray stripes represent? Doesn't Gaussian noise have a potentially unlimited amplitude? Yes, the trick is to define the amplitude based on a BER. For example, if we are ready to accept a BER of 10^{-12} then we don't care about noise voltages exceeding $7.0 \cdot v_n^{rms}$ (this happens only with a probability of 10^{-12}) and so we can take the amplitude of the noise signal as $7.0 \cdot v_n^{rms}$ or more generally $\mathcal{Q} \cdot v_n^{rms}$ (cf. Appendix **??**)

Going back to Fig. 4.10, we recognize that while the eyes (including the noise) for the wide and narrowband receivers are completely closed, the optimum receiver has an eye which is open at the center. Thus only with the optimum receiver it is possible to recover

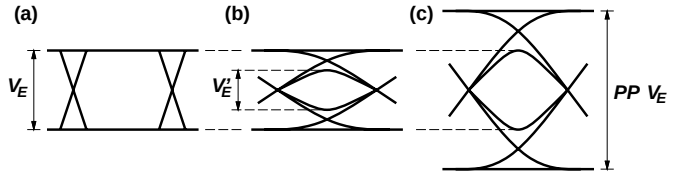

Figure 4.11: Eye diagram (a) without ISI, (b) with ISI, (c) with ISI and increased signal to restore the original bit-error rate.

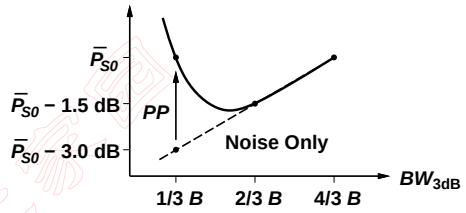

Figure 4.12: Sensitivity as a function of receiver bandwidth.

the data at the desired BER. To do this the decision circuit must operate in the open part of the eye.

Power Penalty due to ISI. The ISI due to the finite receiver bandwidth can be quantified as a power penalty. The idea is simply that ISI reduces the worst-case output signal (corresponding to the bit sequence "01010101...") which can be determined from the vertical opening in the noise-less eye pattern. In Fig. 4.11(a) the signal without ISI has a vertical eye opening of V_E. In Fig. 4.11(b) the eye opening is reduced to V_E' due to ISI. As a result we need to transmit a stronger signal to achieve the same BER as before (see Fig. 4.11(c)). Here is the power penalty!

$$PP = \frac{V_E}{V_E'} \tag{4.56}$$

Inspecting the noise-less eye diagrams in Fig. 4.10 we see that, when sampling at the center of the eye, there is no significant vertical eye closure for the top and center cases, however in the bottom case, the eye is about halfway closed (50 % vertical eye closure). Therefore the power penalties due to ISI from top to bottom are $PP = 0\,\text{dB}$, $PP = 0\,\text{dB}$, and $PP = 3\,\text{dB}$, respectively. Actually, the power-penalty as defined in Eq. (4.56) is somewhat pessimistic because it was derived for the worst-case bit pattern. For a typical bit pattern some bits will have a higher signal amplitude and therefore are detected with a lower BER.

Now let's combine the basic, noise-based sensitivity with the power penalty to derive the actual receiver sensitivity. A graphical representation of the result is shown in

Fig. 4.12. The basic sensitivity decreases as we increase the bandwidth. Specifically, if we double the bandwidth the rms-noise increases by a factor $\sqrt{2}$ (for white noise) which reduces the sensitivity by 1.5 dB according to Eq. (4.17). The basic sensitivity is represented by the dashed line moving up from $\overline{P}_{S0} - 3.0\,\text{dB}$ to $\overline{P}_{S0} - 1.5\,\text{dB}$ and to $\overline{P}_{S0}$ ($\overline{P}_{S0}$ is an arbitrary sensitivity reference.) Next, we correct the basic sensitivity with the power penalty due to ISI and we end up with the solid line. Not surprisingly, the best sensitivity is reached near the bandwidth $2/3 \cdot B$.

Here is an interesting observation: It is possible to build a practical receiver with a bandwidth of only 1/3 of the bit rate if the receiver's phase linearity is good enough. As can be seen from the bottom case of Fig. 4.10 the *horizontal* eye opening is nearly 100 % and from Fig. 4.12 we see that the loss in sensitivity is just about 1.5 dB compared to the optimum. In a 40 Gb/s system with optical preamplifier this loss may be acceptable, if in return the receiver can be built from 13 GHz electronic components [Rei01].

Bandwidth Allocation. So far we have been talking about the bandwidth of the complete receiver. The receiver consists of several building blocks: photodetector, TIA, Filter (optional), MA, and decision circuit. It is the *combination* of all these blocks that should have a bandwidth of about $2/3 \cdot B$. The combined bandwidth can be calculated approximately by adding the inverse-square bandwidths of the individual blocks: $1/BW^2 = 1/BW_1^2 + 1/BW_2^2 + \ldots$. There are several strategies of assigning bandwidths to the individual blocks to achieve this goal. Here are three practical bandwidth allocation strategies:

- The whole receiver (p-i-n/APD, TIA, MA, CDR) is built with a bandwidth much higher than the desired receiver bandwidth. A *precise filter* is inserted, typically after the TIA, to control the receiver bandwidth and its frequency response. Often a 4th-order Bessel-Thomson filter, which exhibits good phase linearity, is used. This method is typically used for lower-speed receivers (2.5 Gb/s and below).

- The TIA is designed to have the desired receiver bandwidth and all other components (p-i-n/APD, MA, CDR) are built for a much higher bandwidth. No filter is used. This approach has the advantage that the TIA bandwidth specification is relaxed permitting a higher transimpedance and better noise performance. (We will see why in Section 5.2.2.) But the receiver's frequency response is less well controlled compared to when a filter is used.

- All components together (p-i-n/APD, TIA, MA, CDR) provide the desired receiver bandwidth. Again, no filter is used. This approach is typically used for high-speed receivers (10 Gb/s and above). At these speeds it is very hard to design electronic circuits and we cannot afford the luxury of overdesigning them. If an APD detector is used, the bandwidth of this device may also significantly affect the overall receiver bandwidth.

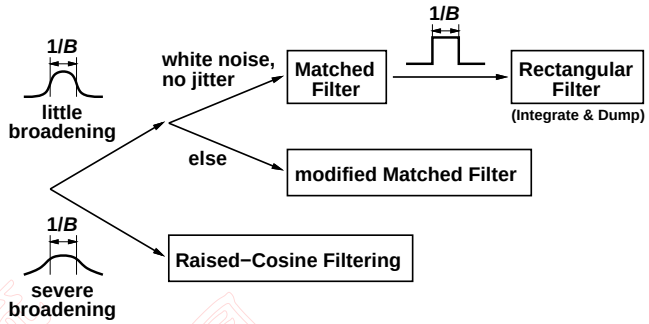

Figure 4.13: Decision tree to determine the optimum receiver response.

Optimum Receiver Response. Is there an optimum frequency response for the receiver of an NRZ signal? Yes, but the answer depends on many factors, such as the shape of the pulses at the receiver input, the spectrum of the input-referred noise, the timing jitter of the sampler, the bit estimation technique, etc. Figure 4.13 shows a decision tree distinguishing the most important cases. For this discussion it is assumed that we estimate each bit independently by comparing the sampled output voltage to a threshold as indicated in Fig. 4.3.[7]

If the NRZ pulses at the input of the receiver are well shaped, in particular if the pulses are broadened by less than 14 % of the bit interval ($1/B$), a matched-filter response or a modification thereof is the best choice [Liu96]. If we further assume white noise and no sampling jitter, the *Matched-Filter Response* optimizes the sampled signal-to-noise ratio and results in the lowest BER. The frequency response of a matched filter matches the spectral shape of the input signal, hence the name. If we idealize further and assume that the received input signal is an undistorted NRZ signal, the matched filter becomes the *Rectangular Filter*, named so because its impulse response is rectangular. We will discuss this case as well as a possible implementation (integrate and dump) in a moment.

If the noise spectrum is not white or if the sampler (decision circuit) exhibits timing jitter, the matched-filter response is not the optimum and needs to be modified [BM95].

In long-haul transmission the signal at the input of the receiver is usually not a well-shaped NRZ signal but consists of severely broadened pulses, e.g., due to fiber dispersion. When the input pulses are broader than the bit interval, the matched filter response discussed earlier produces ISI resulting in a power penalty. For significantly broadened pulses (more than 20 % of the bit interval), *Raised-Cosine Filtering* gives the best results [Liu96]. In raised-cosine filtering, the receiver response is chosen such that the (broadened) input pulses are transformed into pulses with a raised-cosine spectrum. Note that this

[7] It is also possible, and in fact better, to make a *joint* decision on a sequence of bits, e.g., by using a Viterbi decoder. In this case the optimum receiver response is always the matched-filter response (c.f. Section 4.7).

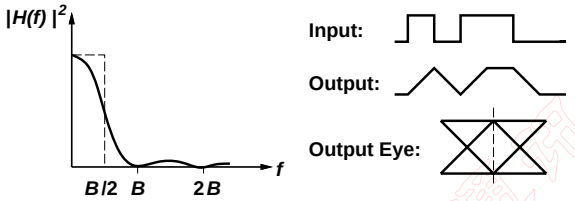

Figure 4.14: Rectangular-filter receiver response and corresponding waveforms.

does *not* mean that the receiver itself has a raised-cosine response! The fact that the output pulses have a raised-cosine spectrum makes them free of ISI as desired, but also makes the receiver response non-causal which means that it can be realized only as an approximation.

The advantage of raised-cosine filtering over the matched-filter approach is that the output pulses are free of ISI, however, the receiver's noise bandwidth is wider compared to that of the matched-filter response. Note that both the raised cosine-filtering and the matched-filter approach depend on the expected pulse shape at the input of the receiver.

Rectangular Filter. Now let's consider the special case of the rectangular-filter response, which is optimal when receiving undistorted NRZ pulses embedded in white noise. For example a clean optical signal received by a wideband p-i-n photodiode (without amplifier noise) fulfills these conditions. The transfer function of the rectangular filter is [Ein96]:

$$H(f) = \frac{\sin(\pi f/B)}{\pi f/B} \cdot e^{-j\pi f/B}. \tag{4.57}$$

The frequency response of Eq. (4.57) is plotted in Fig. 4.14 on a lin-lin scale. The NRZ input signal, the filtered output signal, and the corresponding eye diagram are shown in the same figure. An interesting observation is that this frequency response introduces no ISI at all, when sampling exactly at the center of the eye (dashed line in the eye diagram). The noise bandwidth of the rectangular-filter response is only $BW_n = B/2$. (The 3 dB bandwidth is slightly less than this $BW_{3\text{dB}} = 0.443B$.) The combination of a small noise bandwidth and zero ISI are the characteristics of an ideal receiver response. However, the triangular eye shape implies that we have to sample *exactly* at the center of the eye to have zero ISI. In other words, any sampling offset or sampling jitter will translate into a power penalty.

Integrate and Dump. In the time domain a filter convolves the input signal with its impulse response. The impulse response $h(t)$ of a rectangular filter is one from $t = 0$ to $t = T$, where $T = 1/B$, and zero everywhere else. Written formally we have:

$$y(t) = \int_0^t x(\tau) \cdot h(t-\tau)\, d\tau = \int_{t-T}^t x(\tau)\, d\tau \tag{4.58}$$

where $x(t)$ and $y(t)$ are the filter input and output signals, respectively. Note that this filter computes a moving average. Sampling the output signal at the center of the eye means sampling the output signal at the end of each bit period with $t = nT$. So, the sampling result for the n-th bit is:

$$y(nT) = \int_{(n-1)T}^{nT} x(\tau)\, d\tau. \tag{4.59}$$

From this analysis we see that the filter and the sampler can be replaced by a circuit that integrates the received signal over the bit period T. The result is identical to that of a rectangular filter followed by a sampler. At the end of the bit period the integrator needs to be reset quickly before processing the next bit. For this reason this method is called *Integrate and Dump* [Liu96].

The integrate-and-dump arrangement has the advantage that it lends itself well to monolithic integration. Its frequency response is well controlled and a decision circuit with "instantaneous" sampling can be avoided. For CMOS implementations see [SH97, SR99]. However, just like the rectangular filter, integrate-and-dump is optimal only when receiving clean rectangular pulses with white noise which is rarely the case in practice.

A related issue is the implementation of clock recovery in an integrated-and-dump receiver. If the integrate-and-dump mechanism is part of the decision circuit, standard CDR techniques can be used. However, if the integrate-and-dump mechanism is part of the TIA, as proposed in [Jin90], it is not obvious how to obtain the phase information for the clock recovery PLL. One solution is to sample the analog output of the integrator at the middle of each bit and compute $[y(nT+1) - y(nT+0.5)] - [y(nT+0.5) - y(nT)]$ which is zero if the phase is correctly adjusted.

Raised-Cosine Filtering Example. To illustrate the concept of raised-cosine filtering let's make a simple example. We want to calculate the transfer function that transforms undistorted NRZ pulses into pulses with a full raised-cosine spectrum (this transfer function is called FRC-NRZ in Table 4.6). The full raised-cosine spectrum for the output pulses is defined as

$$H_{FRC}(f) = \frac{1+\cos(\pi f/B)}{2} \cdot e^{-j2\pi f/B} \quad \text{for} \quad f < 1/B \tag{4.60}$$

and $H_{FRC}(f) = 0$ for $f \geq 1/B$. This spectrum guarantees that the output pulses are free of ISI. The spectrum of the incoming clean NRZ pulses is:

$$H_{NRZ}(f) = \frac{\sin(\pi f/B)}{\pi f/B} \cdot e^{-j\pi f/B}. \tag{4.61}$$

The transfer function of the desired filter is just the quotient of these two spectra:

$$H(f) = \frac{H_{FRC}(f)}{H_{NRZ}(f)} = \frac{1+\cos(\pi f/B)}{2} \cdot \frac{\pi f/B}{\sin(\pi f/B)} \cdot e^{-j\pi f/B} \quad \text{for} \quad f < 1/B. \tag{4.62}$$

The 3-dB bandwidth of this filter is $0.580B$ and the noise bandwidth is $0.564B$. This is about 13 % more than that of the rectangular filter which was also free of ISI! This filter

is clearly not optimal for clean NRZ pulses with white noise. Raised-cosine filtering is most attractive when the received pulses are significantly broadened, as we have pointed out earlier. Nevertheless, this filter and its associated Personick integrals ($I_2 = 0.5638$, $I_3 = 0.0868$) are frequently used for theoretical analysis.

Bandwidth of a Receiver for RZ Signals. What is the optimum bandwidth of a receiver for a 50 %-RZ signal? One way to approach this question is to observe that an RZ signal at bit rate B is like an NRZ signal at bit rate $2B$ where every second bit is a zero. Thus we would expect that the optimum bandwidth is about twice that for an NRZ signal, i.e., $BW_{3\text{dB}} \approx 4/3 \cdot B$. Another way to approach this question is the matched filter view: since the spectral width of the RZ signal is twice that of the NRZ signal we would expect again that we have to double the receiver bandwidth (from $0.443B$ to $0.886B$). Finally, what does the raised-cosine approach recommend? Going through the math we find that we have to *reduce* the bandwidth from $0.58B$ for NRZ to $0.39B$ for RZ [Kas88]! How can we explain this? Remember that the raised-cosine approach is producing the same output pulses no matter if the input signal is NRZ or RZ. Therefore, the RZ receiver has to broaden the pulses more than the NRZ receiver which explains the narrower bandwidth of the RZ receiver.

So, we have the following options: (i) Use a wide-bandwidth receiver ($\approx 4/3 \cdot B$) which has high sensitivity but requires a CDR that can process an RZ signal. In particular, the sampling instant must be controlled well to sample the narrow RZ pulse at its maximum value. (ii) Use a narrow-bandwidth receiver which converts the received RZ signal to an NRZ signal permitting a standard CDR, however, the narrow bandwidth lowers the signal amplitude significantly leading to a suboptimal receiver sensitivity.

4.7 Adaptive Equalizer

The signal at the output of the linear receiver channel invariably contains some ISI. This ISI is caused, among other things, by optical dispersion in the fiber (modal, chromatic, or polarization dispersion) as well as the frequency response of the linear channel. In principle, it would be possible to remove ISI with raised-cosine filtering (cf. Section 4.6), but in practice it is often impossible to predict the precise pulse shape upon which the filter depends. The pulse shape varies with the length of the fiber link, the quality of the fiber, laser chirp, etc. and it may even change over time. For example *Polarization-Mode Dispersion*, which is significant in long-haul transmission at high speed (10 Gb/s or more) over older (already installed) fiber, may change slowly with time. For these reasons it is preferred to use a linear channel that approximates a matched filter followed by an adaptive *ISI Canceller*.

The optimum implementation for the ISI canceller is a *Viterbi Decoder* which performs a maximum-likelihood sequence detection on the sampled received signal. However, the implementation of such a decoder is usually too complex and an *Adaptive Equalizer* is used instead. A popular equalizer type is the adaptive *Decision Feedback Equalizer* (DFE) which consists of two adaptive *Finite Impulse Response* (FIR) filters as shown in Fig. 4.15

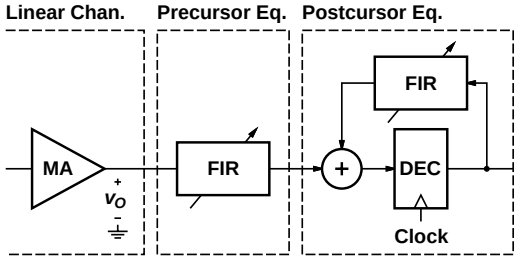

Figure 4.15: The linear channel of Fig. 4.1 followed by a decision-feedback equalizer.

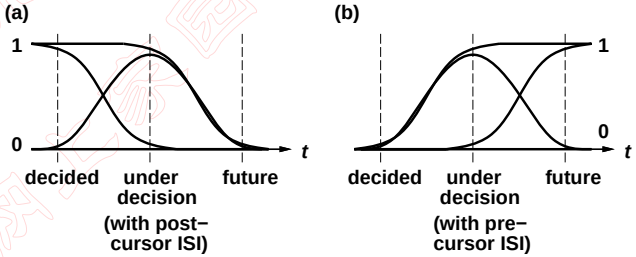

Figure 4.16: (a) Postcursor and (b) precursor ISI.

[GHW92, LM94].[8] In contrast to the simpler *Feed Forward Equalizer* (FFE) which consists of only one FIR filter, the DFE provides minimal noise enhancement. Of course, the MA in Fig. 4.15 must be linear (AGC amplifier) to preserve the shape of the received waveform.

How does a DFE remove ISI? In Fig. 4.16(a) we see how the bit *before* the bit currently under decision influences the signal value. This disturbance is called *Postcursor ISI*. If the preceding bit, a.k.a. the decided bit, is a one, the signal levels of the current bit are slightly shifted *upwards* compared to when the bit is a zero. So if we know the value of the decided bit, we could just subtract a small fraction of its value from the current signal level and remove the postcursor ISI. This is exactly what the 1-tap *Postcursor Equalizer* in Fig. 4.17 does. The decided bit is available at the output of the decision circuit and the (negative) fraction c_2 of this bit is added to the current signal.

Now there is also some influence of the bit *after* the bit currently under decision. This disturbance is called *Precursor ISI*. This may at first sound like a violation of causality, but since the transmission system has a latency of many bits, precursor ISI is possible. The influence of the future bit on the current signal levels is shown in Fig. 4.16(b). Again, if the future bit is a one, the signal levels of the current bit are slightly shifted upwards.

[8] Here we use the term DFE to mean the combination of a pre- and postcursor equalizer. However, some authors use DFE for the postcursor equalizer *only* and use FFE for the precursor equalizer.

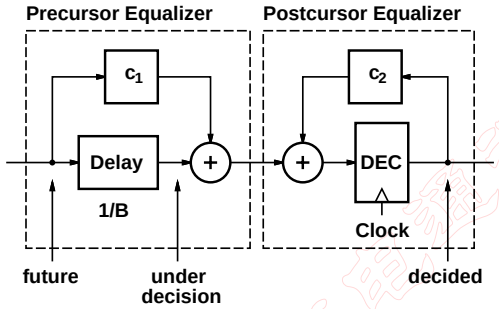

Figure 4.17: Simple DFE equalizer to illustrate the operating principle.

So if we know the value of the future bit, we could just subtract a small fraction of its value from the current signal level and remove the precursor ISI. The 2-tap *Precursor Equalizer* shown in Fig. 4.17 does this by delaying the signal by one bit period so it can look into the "future" of the decision circuit and then adds the (negative) fraction c_1 of the future bit to the current signal.

In practice more taps than shown in Fig. 4.17 are used to take the effects of additional bits before and after the current bit into account. Furthermore several algorithms exist to automatically find and adapt the coefficients $c_1, c_2, \ldots$ to changing ISI conditions. For equalizer implementations for optical applications see [WG90, WK92, AHK+02].

4.8 Nonlinearity

In Section 4.1 we introduced the concept of *linear channel* comprising the TIA followed by an optional filter and the main amplifier. How linear does this channel have to be? If the linear channel is followed directly by a decision circuit, as shown in Fig. 4.1, the linearity requirements are relaxed. In this case amplitude distortions don't matter, but we have to make sure that the nonlinearities don't introduce pulse-width distortions and jitter which reduce the horizontal eye opening. If the linear channel is followed by some type of signal processor, such as the equalizer shown in Fig. 4.15, linearity does matter. In this case we want to design the linear channel such that amplitude compression and harmonic distortions remain small. If the linear channel is part of a receiver for analog AM-VSB or QAM signals, for example in a HFC/CATV application, then the linearity requirements are very strict. In this case we must design the linear channel such that all intermodulation products and their composite effects are kept very low.

In the following we want to discuss how to characterize and quantify nonlinearity. The most straightforward way to describe a nonlinear DC transfer curve $y = f(x)$ is by expanding it into a power series:

$$y = A \cdot (x + a_2 \cdot x^2 + a_3 \cdot x^3 + a_4 \cdot x^4 + a_5 \cdot x^5 + \ldots) \tag{4.63}$$

where A is the small-signal gain and a_i are the normalized power-series coefficients characterizing the nonlinearity. The nonlinear AC characteristics can be described by writing Eq. (4.63) as a Volterra series and making the coefficients frequency dependent: $A(f)$, $a_2(f, f')$, $a_3(f, f', f'')$, etc.

Gain Compression. A simple measure of nonlinearity is the loss of gain caused by large signals relative to the small-signal gain, which is known as *Gain Compression*. For an input signal with amplitude X, swinging from $-X$ to X, we find the large-signal gain with Eq. (4.63) to be $[y(X) - y(-X)]/[X - (-X)] = A \cdot (1 + a_3 \cdot X^2 + a_5 \cdot X^4 + \ldots)$. When normalized to the small-signal gain A, we obtain the gain compression GC:

$$GC = 1 + a_3 \cdot X^2 + a_5 \cdot X^4 + \ldots \tag{4.64}$$

For practical amplifiers a_3 is usually negative meaning that the gain is reduced for large signals. Frequently the input amplitude X for which $GC = -1\,\text{dB}$ ($0.89\times$) is specified, this amplitude is known as the 1-dB gain compression point.

Harmonic Distortions. A more sophisticated method for describing nonlinearity in broadband amplifiers is the specification of *Harmonic Distortion*. In this case the input signal is a sine wave $x(t) = X \cdot \sin(2\pi f \cdot t)$ with amplitude X and frequency f. The n-th order harmonic distortion HDn is defined as the ratio of the output-signal component (distortion product) at frequency $n \cdot f$ to the fundamental at f. For small signals X we can derive the following expressions from Eq. (4.63) [GM77, Lee98]:

$$HD2 \approx 1/2 \cdot |a_2| \cdot X, \tag{4.65}$$

$$HD3 \approx 1/4 \cdot |a_3| \cdot X^2. \tag{4.66}$$

From these equations we see that a 1-dB increase in input signal X causes the $HD2$ to increase by 1 dB while the $HD3$ increases by 2 dB. In general, higher-order harmonics depend more strongly on the input signal amplitude: the nth harmonic distortion product is proportional to X^n or, equivalently, the nth harmonic distortion HDn is proportional to X^{n-1}. Usually only $HD2$ and $HD3$ are considered because the higher-order harmonics drop off fast for small signals. Also note that nth-order harmonic distortions originate from nth-order coefficients in the power series. This means that for a differential circuit, which has small even-order coefficients, $HD2$ is usually small compared to $HD3$ while the reverse is true for a single-ended circuit. Often *Total Harmonic Distortion* (THD) is used to specify the nonlinearity with a single number:

$$THD = \sqrt{HD2^2 + HD3^2 + \ldots} \tag{4.67}$$

The THD can be expressed as a percentage value (amplitude ratio) or in dB using the conversion rule $20 \log THD$. The input dynamic range of an amplifier can be specified, for example, as the maximum value of X for which $THD < 1\,\%$

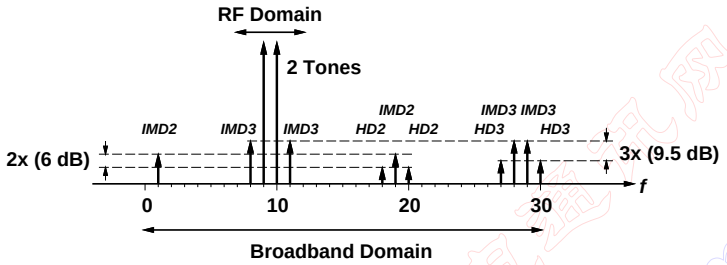

Figure 4.18: 2nd- and 3rd-order distortion products caused by two tones with frequencies $f_1 = 9$ and $f_2 = 10$ in the presence of a nonlinearity.

Intermodulation Distortions. In most applications the input signal to the amplifier is not a pure sine wave and therefore we have to deal with *Intermodulation Distortions* besides the harmonic distortions. Let's start with the two-tone case, i.e., a superposition of two equally strong sine waves at frequencies f_1 and f_2 at the input of the amplifier: $x(t) = X[\sin(2\pi f_1 \cdot t) + \sin(2\pi f_2 \cdot t)]$. With Eq. (4.63) we find that there are two 2nd-order intermodulation products at $f_1 + f_2$ and $|f_1 - f_2|$ and four 3rd-order intermodulation products at $2f_1 + f_2$, $2f_1 - f_2$, $2f_2 + f_1$, and $2f_2 - f_1$. Interestingly, both 2nd-order products have the same amplitude and all four 3rd-order products have equal amplitudes as well. If normalized to the amplitude of the two tones we obtain the following intermodulation distortions:

$$IMD2 \approx |a_2| \cdot X, \tag{4.68}$$

$$IMD3 \approx 3/4 \cdot |a_3| \cdot X^2 \tag{4.69}$$

where X is the amplitude of each individual tone which is assumed to be small. See [Lee98] for a derivation. Compared to the harmonic distortions in Eqs. (4.65) and (4.66) we find the same dependence on input-signal amplitude X and power-series coefficients a_i. However, the $IMD2$ is twice as strong as $HD2$ and the $IMD3$ is three times as strong as $HD3$. In addition to the intermodulation products, of course, we still have the harmonic distortion products corresponding to each tone. Figure 4.18 summarizes all the 2nd- and 3rd-order distortion products for the two-tone case.

RF engineers, who design narrow-band systems, typically worry only about the 3rd-order intermodulation products $2f_1 - f_2$ and $2f_2 - f_1$ which fall back into the band of interest (see Fig. 4.18). The other intermodulation and harmonic distortion products are "out-of-band" and can be ignored. Therefore the value of X for which $IMD3 = 1$ (extrapolated from $IMD3(X)$ where $IMD3 \ll 1$) is commonly used as a measure for the input dynamic range and is known as the *Input-Referred 3rd-Order Intercept Point* (IIP3). Unfortunately, life is not so easy for the broadband engineer and we have to worry about all those distortion products. Actually things get worse as we add more tones!

Let's add a third tone with frequency f_3. We again get n harmonic distortion products for each of the three tones at the frequencies $n \cdot f_1$, $n \cdot f_2$, and $n \cdot f_3$. Then, we get 2nd-order intermodulation products at all permutations of $|f_i \pm f_j|$ (6 products in total). Then, we get 3rd-order intermodulation products at all permutations of $2f_i \pm f_j$ (12 products in total). But then there is also something new: we get additional 3rd-order intermodulation products at all combinations of $|f_1 \pm f_2 \pm f_3|$ (4 products in total). These are the so called *Triple Beat* products and they have twice the amplitude of the two-tone 3rd-order intermodulation products $IMD3$. For small signals the triple-beat distortions can be written as:

$$TBD3 \approx 3/2 \cdot |a_3| \cdot X^2. \tag{4.70}$$

Note that $TBD3$ is six times (15.6 dB) stronger than $HD3$.

Composite Distortions. In HFC/CATV applications we have as many carriers (or tones) as TV channels (e.g., 80). All these carriers produce a huge number of harmonic and intermodulation products in the presence of a nonlinearity. To measure the total effect, one channel is selected and turned off while all the other channels are operating. Then the composite distortion products in the turned-off channel are measured. Finally, this procedure is repeated for all channels until the worst-case channel is found. In the North American Standard channel plan the carriers are spaced 6 MHz apart and are offset 1.25 MHz upwards from harmonics of 6 MHz. As a result all even-order products fall 1.25 MHz above or below the carrier, while all odd-order products fall on the carrier or 2.5 MHz above or below the carrier. Thus the composite even- and odd-order products have different effects on the picture quality and can be measured separately with a bandpass filter. [CFL99]

The composite even-order products are usually dominated by 2nd-order products. When normalized to the carrier amplitude they are called *Composite Second Order* distortion (CSO). The composite odd-order products are usually dominated by triple-beat products. When normalized to the carrier amplitude they are called *Composite Triple Beat* distortion (CTB). These composite distortions can be calculated by summing the power of the individual distortions (assuming phase-incoherent carriers). In the case of equal-power carriers we can write:

$$CSO = \sqrt{N_{CSO}} \cdot IMD2 \approx \sqrt{N_{CSO}} \cdot |a_2| \cdot X, \tag{4.71}$$

$$CTB = \sqrt{N_{CTB}} \cdot TBD3 \approx \sqrt{N_{CTB}} \cdot 3/2 \cdot |a_3| \cdot X^2 \tag{4.72}$$

where N_{CSO} and N_{CTB} are the number of 2nd-order intermodulation products and triple-beat products, respectively, falling onto the turned-off channel. These beat counts can be fairly high,[9] for example in a 80-channel system the maximum N_{CSO} is 69 and occurs for channel 2 while the maximum N_{CTB} is 2,170 and occurs for channel 40 [PD97]. CSO and CTB are usually expressed in dBc using the $20 \log CSO$ and $20 \log CTB$ conversion rules, respectively. The National Association of Broadcasters recommends that both CSO and CTB should be less than -53 dBc for analog TV [PD97]. CATV amplifiers are usually designed for less than -70 dBc.

[9] A rough estimate is $N_{CSO}(\max) \approx N/2$ and $N_{CTB}(\max) \approx 3/8 \cdot N^2$ where N is the number of channels.

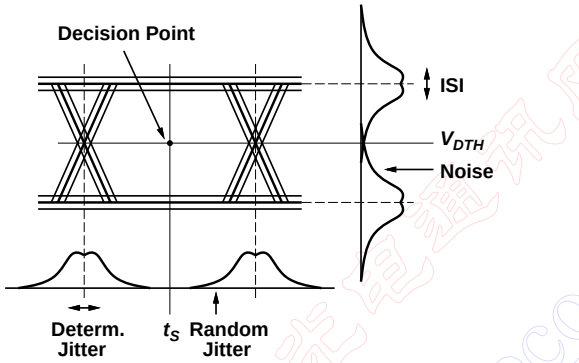

Figure 4.19: Eye diagram at the input of the decision circuit with ISI, noise, deterministic, and random jitter.

4.9 Jitter

So far we have talked about how ISI and noise affects the signal *amplitude* and how we have to set the slice level to minimize bit errors. However, the decision process does not only involve the signal amplitude but also the signal *timing*. In Fig. 4.19 we see how the decision process is controlled by a decision level V_{DTH} as well as a sampling instant t_S. The decision level slices the eye diagram horizontally, while the sampling time slices the eye diagram vertically. The two slicing line intersect at the *Decision Point*. ISI and noise does not only occur in the amplitude domain but also in the time domain. ISI in the time domain is known as *Data Dependent Jitter* and noise in the time domain is known as *Random Jitter*.

Data Dependent Jitter. Data dependent jitter means that the signal edge is moving slightly depending on the value of the surrounding bits. For example, the sequence "...110" may have a falling edge which is a little bit retarded relative to the sequence "...010". Data dependent jitter is caused for example by a receiver with insufficient bandwidth or phase linearity. A TIA operated beyond its overload limit or an MA operated outside its dynamic range is likely to produce data dependent jitter. The histogram of data dependent jitter is bounded and usually discrete (non Gaussian) similar to the histogram of ISI in the amplitude domain. It is therefore natural to specify the peak-to-peak value t_J^{pp} for this type of jitter. Data dependent jitter belongs to a larger class of jitter known as *Deterministic Jitter* which further includes *Duty-Cycle Distortion Jitter*, *Sinusoidal Jitter*, and *Bounded Uncorrelated Jitter* [NCI00].

Random Jitter. Random jitter, as the name implies, is random, i.e., not related to the data pattern transmitted or any other specific cause. Random jitter is produced by noise on signal edges with a finite slope. The finite slope transforms the amplitude uncertainty into timing (zero-crossing) uncertainty. The histogram of random noise can be well approximated by a Gaussian distribution similar to the histogram of noise in the amplitude domain. It is therefore natural to specify the rms value t_J^{rms} for this type of jitter. In principle, we had to consider amplitude noise as well as random jitter to accurately calculate the BER. In analogy to the discussion in Section 4.2, the BER due to random jitter is determined by

$$Q = \frac{1}{2 \cdot B \cdot t_J^{rms}}. \tag{4.73}$$

For example, with an rms jitter of 7.1 % of the bit interval, the BER due to this jitter is 10^{-12}. However, in practice the BER is mostly determined by amplitude noise and we can usually neglect the influence of random jitter on BER.

Composite Jitter. A typical wideband jitter histogram, as shown in in Fig. 4.19, contains both types of jitter. The inner part of the histogram is due to deterministic jitter, the Gaussian tails are due to random jitter. In this case it is harder to accurately specify the amount of jitter. Two commonly used methods are: (i) Decompose the composite jitter into a random and a deterministic part. This can be done by first transmitting a clock-like pattern to determine the random jitter only which is specified as an rms value. Then, by transmitting a PRBS pattern the deterministic jitter is determined and specified as a peak-to-peak value (the random jitter must be subtracted from the latter value). (ii) Perform a so-called *BERT Scan* using a *Bit-Error Rate Tester* (BERT). In this measurement the BERT is used to observe the BER while scanning the sampling instant t_S horizontally across the eye. The BER will be low when sampling in the middle of the eye and go up at both ends when approaching the eye crossing, hence this curve is known as the "bathtub curve" (cf. Appendix **??**). The composite jitter amount is defined as the separation of the two points to the left and the right of the eye crossing where the bathtub curve crosses a certain reference BER such as 10^{-12}.

Jitter Bandwidth. The jitter that we discussed so far is so-called *Wideband Jitter*. It is also possible, and required by some standards such as SONET, to measure jitter in a specified *Jitter Bandwidth*. How do we do that? Should we pass the data signal through a filter with the specified bandwidth and then measure the jitter? No, we are not supposed to filter the signal itself, but the *jitter* (phase error) on the signal! Conceptually, we could use a high-quality *Phase-Locked Loop* (PLL), a.k.a. golden PLL, with a jitter transfer characteristics of 0 dB and a bandwidth equal to the upper corner of the desired jitter bandwidth. Now the recovered clock from the PLL contains only the low-frequency jitter and we can determine the narrowband jitter by displaying the histogram of the clock signal on a scope. A second golden PLL with a bandwidth equal to the lower corner of the desired jitter bandwidth can be used to suppress low-frequency jitter. This is done by connecting the output of this PLL to the trigger input of the scope. Since the scope input and the trigger input both get the same amount of low-frequency jitter, it is suppressed (a common-mode signal in the time domain) and only the desired high-frequency jitter appears in the histogram.

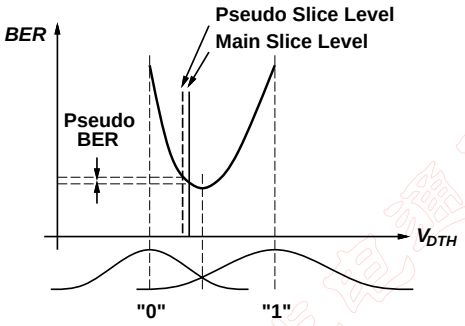

Figure 4.20: Optimum decision threshold for unequal noise distributions.

In practice, a so-called *Jitter Analyzer* can measure the jitter in the desired bandwidth. Alternatively, a *Time Interval Analyzer* or *Spectrum Analyzer* can be used. The time interval analyzer accurately measures the delay between threshold crossings of the output signal. From the collected statistical data it is possible to calculate the power-density spectrum of the jitter [NCI00]. Finally, a spectrum analyzer can be used to measure the phase noise for a clock-like data pattern. From the spectral noise data it is possible to calculate the time-domain rms jitter in the desired bandwidth [NCI00].

Deterministic and random jitter observed in the receiver may also originate in the transmitter or regenerators along the way. Limiting the jitter generation in the transmitter is important and we will talk more about this in Section 8.1.5.

Another type of jitter, relevant to the CDR portion of the receiver, is the *Sampling Jitter*. This jitter can be visualized as an uncertainty on the sampling instant t_S. Sampling jitter causes a power penalty because the eye is not always sampled at the instant of maximum vertical opening. Assuming raised-cosine pulses, this power penalty is about 0.42 dB given an rms sampling jitter of 10 % of the bit interval [Agr97].

4.10 Decision Threshold Control

The optimum slice level (decision threshold) is at the point where the probability distributions of the zero and one bits intersect (cf. Fig. 4.3). In the case where the noise distributions have equal widths the slice level is halfway between the zero and one levels. This slice level is automatically attained in an AC-coupled receiver, a DC-balanced signal and no offset presumed. If there is more noise on the ones than the zeros, due to optical amplifiers or an APD detector, the optimum slice level is below the center (see Fig. 4.20) and a simple AC-coupled receiver is not optimal.

In that case we can make the slice level adjustable and manually adjust it until optimum performance is reached. For example a variable offset voltage in the decision circuit or the preceding MA can be used for this purpose. This method is called *Slice-Level Adjust* and we will discuss its implementation in Section 6.3.3.

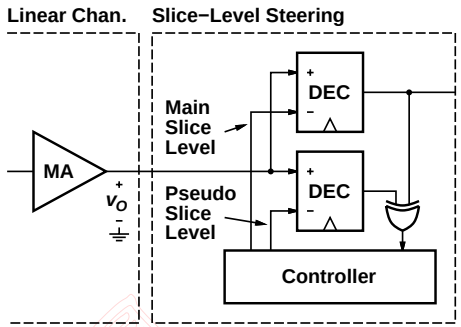

Figure 4.21: The linear channel of Fig. 4.1 followed by a circuit for slice-level steering.

Slice-level adjust is fine, but it would be more convenient to have a system that *automatically* adjusts the slice level for the lowest possible bit-error rate. An automatic mechanism has the additional advantage that it can track variations in the signal strength and noise statistics over time making the system more robust. This automatic method is called *Slice-Level Steering*. It can be extended to find not only the optimum *slice level* but also the optimum *sampling instant* and is then known as *Decision-Point Steering* [SD89, KWOY89].

The difficulty with finding the optimum slice level automatically is that we normally have no means to measure the BER which we seek to minimize (i.e., we do not know the transmitted bit sequence). However, there is a trick: we can measure a *Pseudo Bit-Error Rate* and minimize the latter. This scheme can be implemented by simultaneously slicing the received signal at two slightly different levels: a main slice level and a pseudo slice level (see Figs. 4.21 and 4.20). The output from the main slicer (decision circuit) is the regular data output, the output from the pseudo slicer is compared to the data output with an XOR gate. Any disagreement is counted as a pseudo error. Now the controller has to do the following: First put the pseudo slice level a small amount above the main level and measure the pseudo BER. Then put the pseudo slice level below the main level by the same small amount and measure the pseudo BER again. Then adjust the main slice level into whichever direction that gave the smaller pseudo BER. Iterate this procedure until both pseudo BERs are the same. Congratulations, you have found a close approximation to the optimum slice level! (The smaller the difference between main- and pseudo-slice level, the better the approximation.)

4.11 Forward Error Correction

In ultra-long haul optical transmission systems (e.g. undersea lightwave systems) the transmit data is often *encoded* before it is sent out over the fiber. Coding improves the

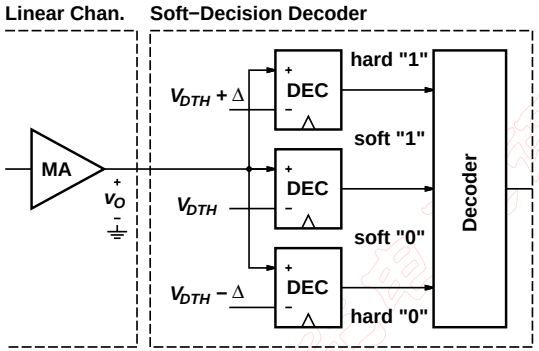

Figure 4.22: The linear channel of Fig. 4.1 followed by a 2-bit soft-decision decoder.

bit-error rate at the receiver for a given signal-to-noise ratio because many transmission errors can be detected and corrected. This method is known as *Forward Error Correction* (FEC). Popular codes used for this purpose are *Reed-Solomon Codes* (RS) and *Bose-Chaudhuri-Hocquenghem Codes* (BCH).

For example, the Reed-Solomon Code RS(255,239) is frequently used in SONET systems [IT00]. The data stream into the encoder, the so-called payload, is cut up into blocks of 238 data bytes. Each data block is concatenated with a framing byte forming a 239 byte block. The latter block is then encoded with the RS(255,239) code which adds 16 bytes of redundancy producing a 255 byte block. Before transmitting the encoded blocks they are run through a 16× *Interleaver*, i.e., rather than transmitting a complete 255 byte block at a time, the interleaver transmits one byte from a first block, then one byte from a second block, etc., until block 16 is reached, then the process continues with the next byte from the first block, etc. Interleaving spreads out burst errors into multiple blocks thus increasing the error-correcting capacity for bursts.

The RS(255,239) encoder increases the transmitted bit rate by 7 % (255/238 = 15/14) which means that slighly faster hardware is required. However, the benefit of the code is that up to 8 byte errors can be corrected per block thus significantly lowering the BER after error correction. Furthermore, burst errors up to 1024 bits in length ($16 \times 8 \times 8$) can be corrected. The precise improvement in BER depends on the BER and the distribution of the bit errors in the received signal. In a typical transmission system with RS(255,239) coding a BER at the output of the CDR of 10^{-4} can be boosted to $BER = 10^{-12}$ after error correction. This is an improvement of eight orders of magnitude!

Commonly the effectiveness of FEC is measured in *Coding Gain*. Continuing our example from above, the incoming BER of 10^{-4} corresponds to $\mathcal{Q} = 3.7$ while the outgoing BER of 10^{-12} corresponds to $\mathcal{Q} = 7.0$ thus the coding gain in $\mathcal{Q}$ is 1.9 ×. This gain can be translated into a coding gain in SNR using Eq. (4.12) giving about 5.5 dB ($20 \cdot \log 1.9$). Alternatively, it can be translated into a coding gain in sensitivity using Eq. (4.17) giving about 2.8 dB ($10 \cdot \log 1.9$).

The use of FEC in a receiver has little impact on the front-end circuits which are the focus of this text. However two considerations must be kept in mind: (i) the CDR must be able to operate at unusually high BERs ($\approx 10^{-4}$) and (ii) a multi-level slicer is required in the case that *Soft-Decision Decoding* is used. Although many transmission errors can be corrected after the received signal has been sliced to a binary value (hard decision), *more* errors can be corrected if the analog value of each received sample is known (soft decision) [GHW92, LM94]. The soft-decision decoder can then take the confidence with which the bit was received into account when performing the error correction. For example a system using the RS(255,239) code together with a soft-decision decoder can correct more than 8 erroneous bytes per block. Typically, a soft-decision decoder achieves about 2 dB better coding gain in SNR than a hard-decision decoder.

Figure 4.22 shows a receiver with a soft-decision decoder in which the linear channel of Fig. 4.1 is followed by a slicer with four different output states (similar to a 2-bit flash A/D converter). These states correspond to "hard zero", "soft zero", "soft one", and "hard one". They can be encoded into 2 bits such that one bit represents the likely bit value and the other bit the confidence level. (Of course, more than 2 bits can be used with a soft-decision decoder, but the coding gains are usually small.) The slicer outputs are fed into the decoder logic which detects and corrects the errors. The MA in Fig. 4.22 must be linear to preserve the analog values and thus must be realized as an AGC amplifier.

4.12 Summary

Bit errors are caused by noise corrupting the data signal at the decision circuit. The two main contributions to the receiver noise are (i) amplifier noise (mostly from the TIA) and (ii) detector noise. The receiver noise can often be modelled as Gaussian, leading to a simple mathematical expression which relates the bit-error rate *BER* to the peak-to-peak signal strength v_S^{pp} and the rms-value of the noise v_n^{rms}.

Electrical sensitivity is the minimum input signal (peak-to-peak) needed to achieve a certain BER. Optical receiver sensitivity is the minimum optical power (average) needed to achieve a certain BER. Electrical sensitivity depends on the total input-referred receiver noise. Optical sensitivity depends on the total input-referred receiver noise, detector noise, and detector responsivity. While the noise of a p-i-n detector has little impact on the optical sensitivity, the noise of an APD or optically preamplified p-i-n detector is significant. Power penalties describe the reduction in optical sensitivity due to system impairments such as a decision-threshold offset in the receiver, dispersion in the fiber, or a finite extinction ratio in the transmitter.

Total input-referred noise is calculated by integrating the output-referred noise spectrum at the decision circuit and then referring it back to the input. Alternatively, Personick integrals can be used to compute the total input-referred noise from the input-referred noise spectrum.

The optimum receiver bandwidth for NRZ modulation is about 2/3 of the bit rate. The optimum frequency response of the receiver depends on the shape of the received pulses. Adaptive equalizers can be used to cancel ISI and automatically respond to varying pulse shapes.

For digital modulation formats such as NRZ or RZ nonlinearities in the receiver and transmitter are of secondary importance, however, for analog transmission as used in CATV applications they are critical. ISI and noise appear not only in the amplitude domain but also in the time domain where they are known as data-dependent jitter and random jitter, respecively.

In systems with APDs or optical amplifiers, the optimum decision threshold is not centered in between the zero and one levels. In this situation either a manual slice-level adjustment or an automatic decision-point steering circuit must be used for optimum performance.

Some transmission systems employ forward error correction (FEC) to boost the BER performance. One type of decoder (soft-decision decoder) requires a receiver with a multi-level slicer instead of the regular binary decision circuit.

4.13 Problems

4.1 SNR and Q. Let ξ be the ratio between the rms noise on the zeros and the ones of an NRZ signal. Now, derive a more general relationship between *SNR* and $\mathcal{Q}$ than the one given in Eq. (4.12).

4.2 SNR Requirement for RZ. (a) Derive the SNR requirement for an RZ signal with duty cycle ξ ($\xi = 1$ corresponds to NRZ) such that it can be detected with a given BER. Assume additive Gaussian noise. (b) What is the SNR value for $\xi = 0.5$ and $BER = 10^{-12}$?

4.3 SNR Requirement for PAM-4. (a) Derive the SNR requirement for a PAM-4 signal such that it can be detected with a given BER. Assume additive Gaussian noise. (b) What is the SNR value for $BER = 10^{-12}$?

4.4 Receiver Sensitivity. Engineers use the following rule to estimate the sensitivity of a p-i-n receiver:

$$\overline{P}_S\,[\mathrm{dBm}] \approx -21.53\,\mathrm{dBm} + 10\log(i_n^{rms}\,[\mu\mathrm{A}]) - 10\log(\mathcal{R}\,[\mathrm{A/W}]) \tag{4.74}$$

Explain the origin of this equation! What is the meaning of $-21.53\,\mathrm{dBm}$?

4.5 OSNR and Q. The approximation in Eq. (4.29) takes only signal-spontaneous noise into account. Derive a more precise equation which includes the effect of spontaneous-spontaneous noise.

4.6 Quantum Limit. Consider the following "alternative derivation" of the quantum limit in Eq. (4.33). We start with the p-i-n detector sensitivity in Eq. (4.24), set the amplifier noise to zero ($i_{n,\mathrm{amp}}^{rms} = 0$), insert the responsivity of an ideal detector (Eq. (3.2) with $\eta = 1$), and use the optimum noise bandwidth $BW_n = B/2$ (Section 4.6) which lead us to:

$$\overline{P}'_{S,\mathrm{quant}} = \frac{\mathcal{Q}^2}{2} \cdot \frac{hc}{\lambda} \cdot B. \tag{4.75}$$

Compared to Eq. (4.33) we have got a $\mathcal{Q}^2$ term instead of the $-\ln(2 \cdot BER)$ term. Given a bit-error rate of 10^{-12}, $\mathcal{Q}^2 = 49$ while $-\ln(2 \cdot BER) = 27$. What is going on here?

Chapter 5

Transimpedance Amplifiers

5.1 TIA Specifications

Before looking into the implementation of TIAs, it is useful to discuss their main specifications: the transimpedance, the input overload current, the maximum input current for linear operation, the input-referred noise current, the bandwidth, and the group-delay variation.

5.1.1 Transimpedance

Definition. The *Transresistance* is defined as the output voltage change, Δv_O, per input current change, Δi_I (see Figs. 5.1 and 5.2):

$$R_T = \frac{\Delta v_O}{\Delta i_I}. \tag{5.1}$$

Thus, the higher the transresistance, the more output signal we get for a given input signal. The transresistance is either specified in units of Ω or dBΩ. In the latter case, the value in dBΩ is calculated as $20 \cdot \log(R_T/\Omega)$, for example, 1 kΩ corresponds to 60 dBΩ. When operated with an AC signal, the TIA is characterized by a transresistance R_T and a phase shift Φ between the input and the output signal. The complex quantity $Z_T = R_T \cdot \exp(i\Phi)$ is known as the *Transimpedance*.

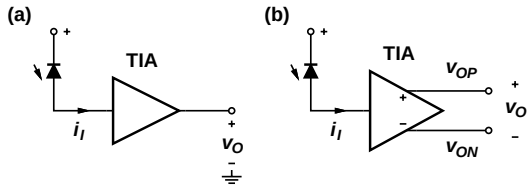

Figure 5.1: Input and output signals of a TIA: (a) single-ended and (b) differential.

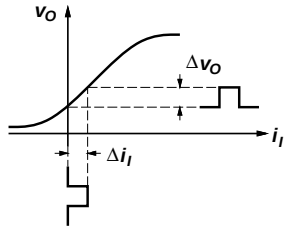

Figure 5.2: Transfer function of a TIA.

The transimpedance is usually measured with a small input signal for which the transfer function is approximately linear. For large input signals the transimpedance drops and eventually pulse-width distortions and jitter may set in (cf. Section 5.1.2).

Some TIAs have differential outputs as shown in Fig. 5.1(b) and therefore the output voltage can be measured single-endedly, $v_O = v_{OP}$ or $v_O = v_{ON}$, or differentially, $v_O = v_{OP} - v_{ON}$. It is important to specify which way the transimpedance was measured because the differential transimpedance is twice as large as the single-ended one.

Most high-speed TIAs have $50\,\Omega$ outputs which must be properly terminated with $50\,\Omega$ resistors when measuring R_T or else the value will come out too high.

Typical Values. In practice it is desirable to make the transimpedance of the TIA as high as possible because this relaxes the gain and noise requirements for the main amplifier. However, we will see later in Section 5.2.2 that there is an upper limit to the transimpedance that can be achieved with the basic shunt-feedback topology. This limit depends on many factors including the bit rate and the technology used. For higher bit rates it is harder to get a high transimpedance with the required bandwidth. Therefore we typically see a lower R_T for 10 Gb/s parts than for 2.5 Gb/s parts. Typical values for the single-ended transimpedance are:

$$2.5\,\text{Gb/s TIA:} \qquad R_T = 1.0\,\text{k}\Omega \ldots 2.0\,\text{k}\Omega \tag{5.2}$$

$$10\,\text{Gb/s TIA:} \qquad R_T = 500\,\Omega \ldots 1.0\,\text{k}\Omega. \tag{5.3}$$

Higher transimpedance values than these can be obtained with a post amplifier following the basic TIA. We will discuss this topology and its characteristics in Section 5.2.5.

5.1.2 Input Overload Current

The input current signal i_I into by the TIA (supplied by either a p-i-n or APD photodiode) is shown schematically in Fig. 5.3. It is important to distinguish between the signal's peak-to-peak value i_I^{pp} and its average value $\overline{i}_I$. In the case of a DC-balanced NRZ signal the relationship is $i_I^{pp} = 2 \cdot \overline{i}_I$.

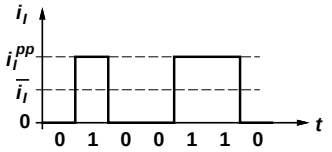

Figure 5.3: TIA input current signal: peak-to-peak value and average value.

Definition. Some TIA implementations produce severe pulse-width distortions and jitter, if the input current signal exceeds a critical value: $i_I^{pp} > I_{OVL}$. This current is called the *Input Overload Current*.

Typical Value. The input overload requirement for a TIA is given by the maximum optical signal at the receiver end and the responsivity of the photodiode. For a maximum optical signal of 0 dBm (e.g., given by the SONET OC-192 short reach specification) and a responsivity of 0.75 A/W the value is $I_{OVL} = 1.5\,\text{mA}$ peak-to-peak.

5.1.3 Maximum Input Current for Linear Operation

A related specification is the *Maximum Input Current for Linear Operation*, I_{LIN}. This current is always less than the overload current I_{OVL}.

Definition. In practice, several definitions are in use for I_{LIN}. One commonly used definition is the input current $I_{\text{LIN}} = i_I^{pp}$ for which the transimpedance drops 1 dB (about 11 %) below the small-signal value. Another one is the input current $I_{\text{LIN}} = i_I^{pp}$ for which the output is 80 % of its fully limited value. In analog applications (HFC/CATV) harmonic distortions and intermodulation products are significant hence I_{LIN} is defined such that these distortions remain small (cf. Section 4.8).

A TIA with a fixed R_T must have a large output voltage swing in order to accommodate a high I_{LIN}, in particular the output swing must be larger than $R_T \cdot I_{\text{LIN}}$ to avoid signal compression. Alternatively, R_T can be made adaptive such that its value reduces for large input signals (cf. Section 5.2.4).

Typical Values. The I_{LIN} current is relevant if linear signal processing, such as equalization, is performed on the output signal. In this case the requirement for this current is also around 1 mA peak-to-peak. In other applications, where the output signal is processed by a (nonlinear) limiting amplifier, I_{LIN} may be as smaller as $10\,\mu\text{A}$.

5.1.4 Input-Referred Noise Current

The input-referred noise current is one of the most critical TIA parameters. Often the noise of the TIA dominates all other noise sources (e.g. photodetector, main amplifier, etc.) and therefore determines the performance of the receiver. However, in long-haul

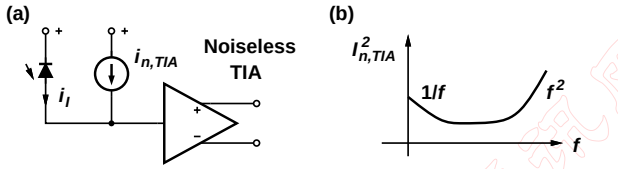

Figure 5.4: (a) Input-referred noise current and (b) typical power spectrum.

transmission systems with many optical in-line amplifiers, the noise accumulated during transmission is so large that the TIA noise becomes less critical.

Definition. Figure 5.4(a) shows a noiseless TIA with an *Equivalent Noise Current Source*, $i_{n,TIA}$, at the input. This current source is chosen such that, together with the noiseless TIA, it produces the same noise at the output as the real noisy TIA.

There is a small complication with this definition for TIAs with differential outputs. Should we choose the equivalent noise current such that it reproduces the single-ended output noise or the differential output noise? These two equivalent noise currents are not necessarily the same because the noise at the two outputs may be only partially correlated. In a particular differential TIA design, the input noise current was 30 % higher when referring it to the single-ended rather than the differential output. For a differential-output TIA it is reasonable to refer the equivalent noise current to the differential output since the TIA will presumably drive the next block differentially.

The *Input-Referred Noise Current Spectrum*, $I^2_{n,TIA}(f)$, has the typical frequency dependence shown in Fig. 5.4(b) (see Section 5.2.3 for mathematical expressions). Since this spectrum is *not white* it is not sufficient to specify a value at a single frequency. For a meaningful comparison of the TIA noise performance it is necessary to look at the whole spectrum up to about $2\times$ the TIA bandwidth. Furthermore, the sensitivity of the TIA depends on a *combination* of the noise spectrum and the frequency response $|Z_T(f)|$.

The *Input-Referred RMS Noise Current*, or *Total Input-Referred Noise Current*, $i^{rms}_{n,TIA}$, is determined by integrating the *output-referred* noise spectrum up at least $2\times$ the bandwidth of the TIA (or measuring the output rms noise voltage in the same bandwidth) and dividing it by the passband transimpedance value (cf. Section 4.4). Written in the squared (power) form, we have:

$$\overline{i^2}_{n,TIA} = \frac{1}{|Z_T|^2}\int |Z_T(f)|^2 \cdot I^2_{n,TIA}(f)\, df \tag{5.4}$$

where $|Z_T(f)|$ is the frequency response of the transimpedance. The total input-referred noise current directly determines the sensitivity of the TIA:

$$i_S^{pp} = 2\mathcal{Q}\cdot i^{rms}_{n,TIA}. \tag{5.5}$$

The value of $i^{rms}_{n,TIA}$ is therefore a good metric to compare TIAs designed for the same bit rate. The TIA noise is usually the main contribution to the linear-channel noise, $i^{rms}_{n,\mathrm{amp}}$, and thus $i^{rms}_{n,TIA}$ also determines the electrical receiver sensitivity.

The *Averaged Input-Referred Noise Current Density* is defined as the input-referred rms noise current, $i^{rms}_{n,TIA}$, divided by the square-root of the 3-dB bandwidth. Again, it is important that the averaging is carried out over the *output-referred* noise spectrum rather than the input-referred noise spectrum, $I^2_{n,TIA}(f)$. In a particular 10 Gb/s TIA design, the input-referred noise current spectrum, $I_{n,TIA}$, was 8.8 pA/$\sqrt{\text{Hz}}$ at 100 MHz, the same input-referred spectrum averaged up to the 3-dB bandwidth of 7.8 GHz (not recommended) was 15.4 pA/$\sqrt{\text{Hz}}$, and the input-referred rms noise divided by the square-root of the bandwidth (correct way to determine the averaged input-referred noise) was 18.0 pA/$\sqrt{\text{Hz}}$.

Typical Values. Typical values for the input-referred rms noise current seen in commercial parts are:

$$2.5\,\text{Gb/s TIA:} \qquad i^{rms}_{n,TIA} = 400\,\text{nA} \tag{5.6}$$

$$10\,\text{Gb/s TIA:} \qquad i^{rms}_{n,TIA} = 1200\,\text{nA}. \tag{5.7}$$

We see that 10 Gb/s parts are quite a bit noisier than 2.5 Gb/s parts. For white noise we would expect only a doubling of the rms-noise value for a quadrupling of the bandwidth. But because of the f^2 noise and a smaller feedback resistor at 10 Gb/s the noise is higher than that. See Section 5.2.3 for a more detailed analysis.

5.1.5 Bandwidth and Group-Delay Variation

Definition. The TIA bandwidth, $BW_{3\text{dB}}$, is defined as the frequency at which the amplitude response of $Z_T(f)$ dropped by 3 dB below its passband value. This bandwidth is therefore also called the 3-dB bandwidth to distinguish it from the noise bandwidth.

The bandwidth specification does not say anything about the phase response. Even if the amplitude response is flat up to a given frequency, distortions may occur below this frequency if the phase linearity of $Z_T(f)$ is insufficient. A common measure for phase linearity is the variation of the group delay with frequency. Group delay, τ, is related to the phase, Φ, as $\tau(\omega) = -d\Phi/d\omega$.

Typical Values. The 3-dB bandwidth of the TIA depends on which bandwidth allocation strategy is chosen for the receiver (cf. Section 4.6). If the TIA sets the receiver bandwidth, its bandwidth is chosen around $0.6 - 0.7\cdot B$. If the receiver bandwidth is controlled in another way (e.g., with a filter), the TIA bandwidth is chosen wider around $0.9 - 1.2\cdot B$:

$$2.5\,\text{Gb/s TIA:} \qquad BW_{3\text{dB}} = 1.7\,\text{GHz}\ldots 3\,\text{GHz} \tag{5.8}$$

$$10\,\text{Gb/s TIA:} \qquad BW_{3\text{dB}} = 6.7\,\text{GHz}\ldots 12\,\text{GHz}. \tag{5.9}$$

In optically amplified transmission systems it is important that the TIA bandwidth remains constant, in particular it should not depend on signal strength. An increase in electrical receiver bandwidth decreases the SNR at the decision circuit causing the BER to go up, and that for a constant received OSNR (cf. Eq. (3.17)).

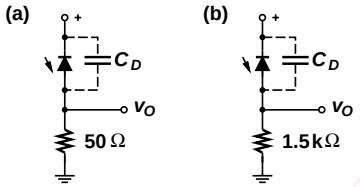

Figure 5.5: (a) Low-impedance and (b) high-impedance front-end.

Typically, a group delay variation, $\Delta\tau$, of less than $\pm 10\,\%$ of the bit period (± 0.1 UI) over the specified bandwidth is required. This corresponds to:

$$2.5\,\text{Gb/s TIA:} \quad |\Delta\tau| < 40\,\text{ps} \tag{5.10}$$

$$10\,\text{Gb/s TIA:} \quad |\Delta\tau| < 10\,\text{ps}. \tag{5.11}$$

5.2 TIA Circuit Principles

In the following section, we will have a look at the design principles for TIAs: How can we get a high transimpedance, high bandwidth, low noise current, high overload current, and so on? Then in the subsequent section we will examine concrete circuit implementations to illustrate these principles.

5.2.1 Low- and High-Impedance Front-Ends

The shunt-feedback TIA is by far the most popular circuit to convert the photodiode current i_I into a voltage signal v_O. But there are two other possibilities which we want to mention at this point for completeness and as a motivation for the shunt-feedback TIA: the *Low-Impedance Front-End* and the *High-Impedance Front-End*. Both front-ends basically consist of just a resistor from the photodiode to ground (see Fig. 5.5). This resistor is either a $50\,\Omega$ resistor (low-impedance front-end) or a resistor in the kΩ-range (high-impedance front-end). The circuits in Fig. 5.5 may be followed by a post amplifier and, in the case of the high-impedance front-end, by an equalizer.

Let's first consider the low-impedance front-end. Obviously, the voltage output signal is $v_O = 50\,\Omega \cdot i_I$. Thus this front end has a very low transimpedance $R_T = 50\,\Omega$ which is not giving us much of an output signal. (If the front end is loaded directly by a $50\,\Omega$ system the transimpedance reduces to $25\,\Omega$, but in the following we assume that the front end is followed by an ideal buffer.) Furthermore, it is a noisy front end because the input-referred current is that of a $50\,\Omega$ resistor and small resistors have a large noise current. (Remember, the thermal noise current of a resistor is given by $\overline{i^2_{n,\text{res}}} = 4kT/R \cdot BW$). On the plus side the bandwidth is very good: with a photodiode capacitance $C_D = 0.15\,\text{pF}$ the bandwidth of this front-end is a respectable 21 GHz.[1]

[1]The low-impedance front end has an interesting application in high-speed receivers. Let's assume that the optical signal is preamplified to +9 dBm and then converted to a voltage using a $50\,\Omega$ low-impedance front-end directly loaded by a $50\,\Omega$ system. The resulting voltage signal is about $300\,\text{mV}_{\text{pp}}$, enough to drive a CDR directly! This is a practical solution at very high speeds for which TIAs and MAs are not yet available.

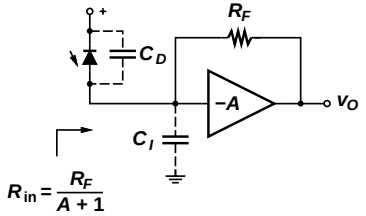

Figure 5.6: Basic shunt-feedback transimpedance amplifier.

Alternatively, if we use a much larger resistor, let's say $1.5\,\text{k}\Omega$, the transimpedance goes up to a reasonable value $R_T = 1.5\,\text{k}\Omega$ and the noise current is reduced too. But now we have other problems: For one, the bandwidth is reduced to 710 MHz. Furthermore, the dynamic range is reduced because the photodiode reverse bias drops below its limit for a smaller photodiode current. Is there a way to get high bandwidth, high transimpedance, low noise, and high dynamic range all at the same time? Yes, this is exactly what the shunt-feedback transimpedance amplifier does for us!

5.2.2 Shunt Feedback TIA

Simple Analysis. The circuit of the basic shunt-feedback TIA is shown in Fig 5.6. The photodiode is connected to an inverting amplifier with a shunt-feedback resistor R_F. The current from the photodiode flows into R_F and the amplifier output responds in such a way that the input remains at a virtual ground. Therefore $v_O \approx -R_F \cdot i_I$ and the photodiode reverse voltage remains approximately constant.

Now let's analyze the frequency response. For now we assume that the feedback amplifier has gain $-A$ and and an infinite bandwidth; later, we will drop the latter assumption. The input resistance of the feedback amplifier is taken to be infinite, a good assumption for a FET amplifier. Besides the parasitic photodiode capacitance, C_D, we also have to consider the input capacitance of the amplifier, C_I. Since both capacitances appear in parallel we can combine them into a single capacitance $C_T = C_D + C_I$. Given these assumptions we can calculate the transimpedance as:

$$Z_T(s) = -R_T \cdot \frac{1}{1 + s/\omega_p} \tag{5.12}$$

where

$$R_T = \frac{A}{A+1} \cdot R_F, \tag{5.13}$$

$$\omega_p = \frac{A+1}{R_F C_T}. \tag{5.14}$$

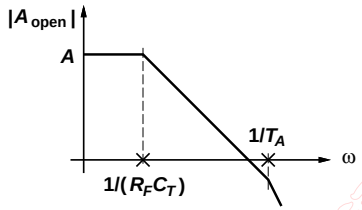

Figure 5.7: Open-loop frequency response of a TIA with a single-pole feedback amplifier.

From Eq. (5.13) we conclude that for a gain A much larger than one, the transresistance R_T is close to the value of the feedback resistor R_F. The 3-dB bandwidth of the TIA follows from Eq. (5.14) as:

$$BW_{\text{3dB}} = \frac{\omega_p}{2\pi} = \frac{1}{2\pi} \cdot \frac{A+1}{R_F C_T}. \tag{5.15}$$

In other words, the bandwidth is $A+1$ times *larger* than that of a high-impedance front-end with resistor R_F and parasitic capacitance C_T. This bandwidth boost can be understood by calculating the input resistance of the TIA. Due to the feedback action, this input resistance is $R_{\text{in}} = R_F/(A+1)$ as indicated in Fig. 5.6. So, the input pole is sped up by a factor $A+1$ over the situation without a feedback amplifier.

This increase in bandwidth is the reason for the popularity of the TIA. We can get a high transimpedance ($\approx R_F$) and a low noise similar to those of the high-impedance front-end, and at the same time we get a high bandwidth and dynamic range similar to that of the low-impedance front-end. Well, at least this is true if we can build the feedback amplifier with the necessary bandwidth ...

Effects of Finite Amplifier Bandwidth. Next we want to consider a more realistic feedback amplifier. Let's replace our infinite-bandwidth amplifier with one that has a single dominant pole at the frequency $f_A = 1/(2\pi \cdot T_A)$, a good approximation for a single-stage amplifier. Now the open-loop response has two poles, one from the feedback amplifier at f_A and one from the low pass formed by R_F and C_T as shown in Fig. 5.7. In a second-order system like this, there is the possibility of undesired peaking in the closed-loop frequency response. Such peaking leads to ringing in the time-domain which reduces the eye opening. The closed-loop transimpedance can be expressed as:

$$Z_T(s) = -R_T \cdot \frac{1}{1 + s/(\omega_0 Q) + s^2/\omega_0^2} \tag{5.16}$$

where

$$R_T = \frac{A}{A+1} \cdot R_F, \tag{5.17}$$

$$\omega_0 = \sqrt{\frac{A+1}{R_F C_T \cdot T_A}}, \tag{5.18}$$

$$Q = \frac{\sqrt{(A+1) \cdot R_F C_T \cdot T_A}}{R_F C_T + T_A}. \tag{5.19}$$

In these equations R_T has the same value as before, ω_0 is the pole (angular) frequency, and Q is the pole quality factor[2] which controls the peaking and ringing (not to be confused with the Personick $\mathcal{Q}$ of Eq. (4.8)). In particular two values for Q are of interest: For $Q = 1/\sqrt{3} = 0.577$ we get a maximally flat *delay* response (a.k.a. *Bessel Response*), i.e., the response with the smallest group-delay variation. For $Q = 1/\sqrt{2} = 0.707$ we get a maximally flat *amplitude* response (a.k.a. *Butterworth Response*). But now there is a little bit of peaking (≈ 0.06 UI) in the delay response. If we go above $1/\sqrt{2}$, there will be peaking in the amplitude response and more and more ringing in the step response.

In the following we will require that $Q \le 1/\sqrt{2}$ to avoid any amplitude peaking. With a little bit of algebra we find that to satisfy this condition we need to make the feedback amplifier faster than:

$$f_A \ge \frac{1}{2\pi} \cdot \frac{2A}{R_F C_T}. \tag{5.20}$$

The interpretation of this equation is that the two open-loop poles in Fig. 5.7 must be spaced apart by at least $2A$. (If we want a Bessel response they must be spaced apart by $3A+1$.) Equivalently, the second pole ($1/T_A$) must be at least a factor two above the zero-crossing frequency of the open-loop response. In the case of equality in Eq. (5.20) (Butterworth response), we find the TIA bandwidth BW_{3dB} ($= \omega_0/(2\pi)$) with Eqs. (5.18), (5.20), and $T_A = 1/(2\pi \cdot f_A)$ to be:

$$BW_{\text{3dB}} = \frac{1}{2\pi} \cdot \frac{\sqrt{2A(A+1)}}{R_F C_T}. \tag{5.21}$$

There are two interesting things to note about this equation. First, the bandwidth of the TIA *increased* when we made the bandwidth of the feedback amplifier finite. By comparing Eqs. (5.15) and (5.21) we find a bandwidth increase of about $\sqrt{2}$. It is remarkable that a system gets faster by making one of its components slower!

Second, by combining Eqs. (5.17), (5.20) and (5.21) we can derive the following inequality which is known as the *Transimpedance Limit* [MHBL00]:

$$R_T \le \frac{A \cdot f_A}{2\pi \cdot C_T \cdot BW_{\text{3dB}}^2}. \tag{5.22}$$

This inequality tells us the following: If we want to double the bit rate, which means that we must double BW_{3dB}, then, all other things left unchanged, the transimpedance will degrade by a factor of four! This is the reason why high bit-rate TIAs generally have a lower transimpedance. Alternatively, if we want to achieve the same transimpedance while doubling the bit rate, we need to half C_T and double $A \cdot f_A$, or leave C_T the same and quadruple $A \cdot f_A$, or etc. The expression $A \cdot f_A$ is the gain-bandwidth product of the feedback amplifier which is roughly proportional to the technology parameter f_T. In other words, doubling $A \cdot f_A$ means that we need a technology which is twice as fast.

[2] $Q = 1/(2\zeta)$ where ζ is the damping factor.

We can further conclude from Eq. (5.22) that given a certain technology (amplifier and photodiode), the product $R_T \cdot BW_{3\text{dB}}^2$ is approximately constant. Note that unlike in the gain-bandwidth product, the bandwidth appears *squared* in this product. For this reason the simple product $R_T \cdot BW_{3\text{dB}}$ measured in ΩHz, which has been proposed as a figure of merit for TIAs, is inversely proportional to the bandwidth and can assume very large values for low-speed TIAs.

It should be pointed out that the transimpedance limit has been derived based on the basic shunt-feedback topology of Fig. 5.6 and is not a fundamental limit. Other topologies, in particular the TIA with post amplifier (see Section 5.2.5), the TIA with common-base input stage (see Section 5.2.6), and the TIA with inductive input coupling (see Section 5.2.7) may achieve higher transimpedances.

Multiple Amplifier Poles. The assumption of a single dominant amplifier pole in the feedback amplifier is often too simplistic in practice. Additional amplifier poles are caused by cascode transistors, buffers, level shifters, or, in the case of a multistage amplifier, by every single stage. Under these circumstances detailed transistor-level simulations are necessary to determine the TIA bandwidth and to make sure amplitude peaking and group-delay variations are within specifications.

What can we do to promote stability in the presence of multiple amplifier poles? One approach is to place the feedback amplifier poles at a higher frequency than given by Eq. (5.20) such that the phase margin of the open-loop response remains sufficient. Another technique is to add a small capacitor C_F in parallel to the feedback resistor R_F. This capacitor introduces a zero in the open-loop response but not in the closed-loop response. This zero is therefore known as *Phantom Zero*. The zero in the open-loop response can be used to stabilize the amplifier [Nor83, Ran01].

Is there a reason to implement the feedback amplifier with multiple stages? Yes, in low-voltage FET technologies the maximum gain achievable with a single stage is limited to relatively small values. With a resistive load and the quadratic FET model, the gain is given by $A \approx 2V_R/(V_{GS} - V_{TH})$ where V_R is the DC-voltage drop across the load resistor and $V_{GS} - V_{TH}$ is the FET's overdrive voltage. For example with a power supply voltage of 1 V, headroom considerations limit V_R to about 0.5 V while $V_{GS} - V_{TH}$ should be at least 0.3 V for speed reasons (cf. Section 6.3.2). Thus this stage is limited to a gain of less than 3.3×. By cascading multiple stages it is possible to overcome this limit which in turn permits the use of a larger feedback resistor resulting in better noise performance. (Note that this is much less of a problem for a bipolar stage where the gain is given by $A \approx V_R/V_T$. In the previous example this gain would be about 19×.)

A Numerical Example. To get a better feeling for numerical values we want to illustrate the foregoing theory with a 10 Gb/s TIA design example. Figure 5.8 shows our familiar TIA circuit annotated with some typical values. The photodiode and amplifier input capacitance are 0.15 pF each, the feedback amplifier has a gain of 14 dB (5×), and the feedback resistor is 600 Ω.

We can easily calculate the transimpedance of this TIA:

$$R_T = \frac{5}{5+1} \cdot 600\,\Omega = 500\,\Omega \tag{5.23}$$

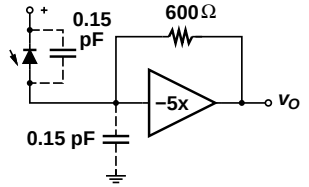

Figure 5.8: Numerical example values for a 10 Gb/s TIA.

which is equal to 54 dBΩ. This is somewhat lower than R_F, as expected. The TIA bandwidth turns out to be

$$BW_{\text{dB}} = \frac{\sqrt{2 \cdot 5 \cdot (5+1)}}{6.28 \cdot 600\,\Omega \cdot 0.3\,\text{pF}} = 6.85\,\text{GHz}, \tag{5.24}$$

which is suitable for a 10 Gb/s system where the TIA sets the receiver bandwidth. For comparison, a high-impedance front-end with the same transimpedance (500 Ω) and the same total capacitance (0.3 pF) would have a bandwidth of only 1.06 GHz. To achieve a flat passband (without peaking) we need the feedback amplifier pole to be faster than:

$$f_A \geq \frac{2 \cdot 5}{6.28 \cdot 600\,\Omega \cdot 0.3\,\text{pF}} = 8.85\,\text{GHz}. \tag{5.25}$$

Thus we need a technology in which we can realize an amplifier with the gain-bandwidth product:

$$A \cdot f_A = 5 \cdot 8.85\,\text{GHz} = 44\,\text{GHz}. \tag{5.26}$$

Project Leapfrog. Just as you are getting ready for the tape out of the 10 Gb/s TIA design your boss appears at your cubicle. "We have to implement initiative Leapfrog", she announces. "Can you beef up your 10-gig TIA to 40 gig before tape out on Friday?"

What do you do now? You can't switch to a faster technology, but you could reduce the transimpedance. O.k., you need to boost the bandwidth by a factor 4, from 6.85 GHz to 27.4 GHz, so you start by reducing R_F by a factor 4, from 600 Ω to 150 Ω. Now you've got to check if the feedback amplifier is fast enough to avoid peaking:

$$f_A \geq \frac{2 \cdot 5}{6.28 \cdot 150\,\Omega \cdot 0.3\,\text{pF}} = 35.4\,\text{GHz}. \tag{5.27}$$

Wow, this means a gain-bandwidth product of $5 \cdot 35.4 = 177$ GHz, that's not possible in your 44 GHz technology. But if you lower the feedback amplifier gain by a factor 2, from 5 × to 2.5 × then you can do it. Now $f_A \geq 17.7$ GHz and $A \cdot f_A = 2.5 \cdot 17.7\,\text{GHz} = 44\,\text{GHz}$. Flat response! Let's check the TIA bandwidth:

$$BW_{\text{dB}} = \frac{\sqrt{2 \cdot 2.5 \cdot (2.5+1)}}{6.28 \cdot 150\,\Omega \cdot 0.3\,\text{pF}} = 14.8\,\text{GHz}. \tag{5.28}$$

Cold sweat is dripping from your forehead. This is not enough! So far you gained just a factor 2.2 in speed. You hear your boss' foot steps in the aisle. She is leaving! Good, there will be no more interruptions tonight. In desperation you lower R_F from $150\,\Omega$ to $50\,\Omega$. Again peaking is an issue and you have to drop the feedback amplifier gain even lower, this time from $2.5\times$ to $1.44\times$. Soon this amplifier will be a buffer! Now you have:

$$f_A \geq \frac{2 \cdot 1.44}{6.28 \cdot 50\,\Omega \cdot 0.3\,\text{pF}} = 30.6\,\text{GHz} \tag{5.29}$$

and $A \cdot f_A = 1.44 \cdot 30.6\,\text{GHz} = 44\,\text{GHz}$. Fine. What's the bandwidth now?

$$BW_{\text{dB}} = \frac{\sqrt{2 \cdot 1.44 \cdot (1.44+1)}}{6.28 \cdot 50\,\Omega \cdot 0.3\,\text{pF}} = 28.1\,\text{GHz}. \tag{5.30}$$

Hurray, you've got it! Tomorrow you'll be the "40-Gig Hero." But before you leave, you want to check the transimpedance:

$$R_T = \frac{1.44}{1.44+1} \cdot 50\,\Omega = 29.5\,\Omega. \tag{5.31}$$

Lousy, $17\times$ less than what you had before! Or should that be 16? Maybe it goes down with bandwidth squared. Anyway, time to go home. You know your boss only cares about 40 gig and that's that.

5.2.3 Noise Optimization

In Section 4.3 we emphasized the importance of the input-referred noise current and its impact on receiver sensitivity. Now we want to analyze and optimize this important noise quantity for the shunt-feedback TIA.

Figures 5.9 and 5.11 show the most important noise sources in TIAs with a FET and bipolar front-end, respectively. For complete transistor-level circuits see Section 5.3. All noise sources associated to individual devices ($i_{n,\text{res}}$, $i_{n,G}$, $i_{n,D}$, $i_{n,B}$, $i_{n,Rb}$, $i_{n,C}$) can be condensed into a single equivalent noise source $i_{n,TIA}$ at the input of the TIA. In the following we want to study the power spectrum $I^2_{n,TIA}(f)$ of this equivalent noise source, then we want focus on the *total* input-referred noise, $i^{rms}_{n,TIA}$, which is the relevant quantity for sensitivity calculations.

Input-Referred Noise Power Spectrum. The input-referred noise spectrum of the TIA, can be written as the sum of two components: The noise from the feedback resistor(s) and the noise from the amplifier front-end:

$$I^2_{n,TIA}(f) = I^2_{n,\text{res}}(f) + I^2_{n,\text{front}}(f). \tag{5.32}$$

In high-speed receivers the front-end noise contribution is typically larger than the contribution from the feedback resistor. However, in low-speed receivers the resistor noise may become dominant. The noise spectrum of the feedback resistor is white (frequency independent) and given by the well-known thermal-noise equation:

$$I^2_{n,\text{res}}(f) = \frac{4kT}{R_F}. \tag{5.33}$$

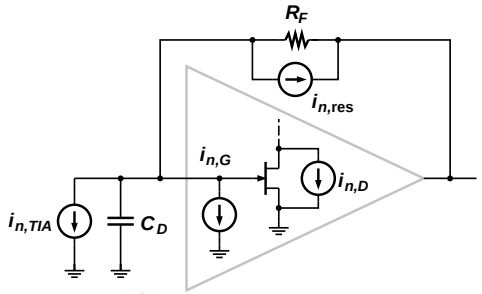

Figure 5.9: Significant noise sources in a TIA with FET front end.

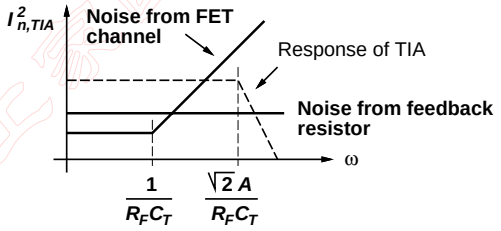

Figure 5.10: Noise spectrum components of a TIA with FET front end.

This noise current appears directly in the equivalent TIA noise equation because $i_{n,\text{res}}$ is connected to the same node as $i_{n,TIA}$. Clearly, increasing R_F helps to lower the TIA noise.

Next we want to analyze the amplifier front-end noise contribution. For a FET common-source input stage, as shown in Fig. 5.9, the main terms are [SP82, Kas88]:

$$I^2_{n,\text{front}}(f) = 2qI_G + 4kT\Gamma \cdot \frac{(2\pi C_T)^2}{g_m} \cdot f^2 + \ldots \tag{5.34}$$

The first term describes the shot noise ($\overline{i^2_{n,G}}$) generated by the gate current I_G which contributes directly to the input-referred noise. This component is negligible for MOSFETs but may be significant for MESFET or JFET transistors which have a larger gate leakage current. The second term is due to the FET channel noise. The channel-noise factor Γ is 2/3 for long-channel transistors but is much larger for submicron devices. For GaAs MESFETs $\Gamma = 1.1 - 1.75$, for Si MOSEFETs $\Gamma = 1.5 - 3.0$, and for Si JFETs $\Gamma = 0.7$ [Kas88]. Note that the second noise term is *not white* but increases proportional to f^2.

Where does this peculiar frequency dependence come from? The transfer function from the input current source $i_{n,TIA}$ to the drain current source $i_{n,D}$ under the condition of zero output signal has a low-pass characteristics. Therefore the inverse function, which

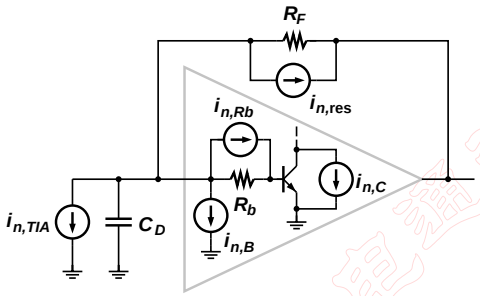

Figure 5.11: Significant noise sources in a TIA with bipolar front end.

refers the drain current back to the input, has a *high-pass* characteristics [JM97]:

$$H(s) = \frac{1 + sR_FC_T}{g_mR_F}. \tag{5.35}$$

The white channel noise at the drain, $\overline{i^2_{n,D}} = 4kT\Gamma g_m$, is referred back to the input through this high-pass which gives:

$$\begin{aligned} I^2_{n,\text{front},D}(f) &= \frac{1 + (2\pi f \cdot R_FC_T)^2}{(g_mR_F)^2} \cdot 4kT\Gamma g_m \\ &= 4kT\Gamma \cdot \frac{1}{g_mR_F^2} + 4kT\Gamma \cdot \frac{(2\pi C_T)^2}{g_m} \cdot f^2. \end{aligned} \tag{5.36}$$

The first term in Eq. (5.36) has been neglected in Eq. (5.34) but the second term is exactly the f^2-noise term that we are trying to explain! Figure 5.10 illustrates the channel-noise component of Eq. (5.36) and the feedback-resistor noise component of Eq. (5.33) graphically. It is interesting to observe that the input-referred channel noise starts to rise at the frequency $1/(2\pi \cdot R_FC_T)$ given by the zero in Eq. (5.35). This frequency is *lower* than the 3 dB bandwidth of the TIA which is $\sqrt{2A(A+1)}/(2\pi \cdot R_FC_T)$ (cf. Eq. (5.21)). As a result the output-referred noise spectrum has a "hump" as shown in Fig. 4.2.

Besides the noise terms given in Eq. (5.34) there are several other noise terms which have been neglected. For example the white noise term in Eq. (5.36) caused by the input transistor has been neglected because for $g_mR_F \gg \Gamma$ it is small compared to the feedback resistor noise. However, for small values of R_F this noise can be significant. Another reason to try and make R_F as large as possible! The FET also produces $1/f$ noise, which when referred back to the input turns into f noise at high frequencies and $1/f$ noise at low frequencies. Furthermore, there are additional device noise sources which also contribute to the equivalent noise such as the FET load resistor and subsequent gain stages. However, if the gain of the first stage is sufficiently large these sources can be neglected.

In the case of a BJT common-emitter front-end, as shown in Fig. 5.11, the main noise

terms are [SP82, Kas88]:

$$I^2_{n,\text{front}}(f) = \frac{2qI_C}{\beta} + 2qI_C \cdot \frac{(2\pi C_T)^2}{g_m^2} \cdot f^2 + 4kTR_b \cdot (2\pi C_D)^2 \cdot f^2 + \ldots \tag{5.37}$$

Here again, we have a sum of white-noise and f^2-noise terms. The first term describes the shot noise $\overline{i^2_{n,B}}$ caused by the base current I_C/β. The second term is due to the shot noise $\overline{i^2_{n,C}}$ caused by the collector current I_C. This noise is white at the collector but gets emphasized proportional to f^2 when referred back to the input. The last noise term is due to the thermal noise $\overline{i^2_{n,Rb}}$ of the intrinsic base resistance R_b. This noise too gets emphasized proportional to f^2 by the familiar process.

In summary, from Eqs. (5.33), (5.34), and (5.37) we conclude that the input-referred noise power spectrum $I^2_{n,TIA}(f)$ of a FET or bipolar TIA is constant (white) up to a certain frequency and then rises rapidly proportional to f^2. For a more detailed discussion of the noise spectrum (e.g., including $1/f$ noise) see [SP82, Kas88, BM95].

Total Input-Referred Noise Current. As we have seen in Section 4.4 there are two equivalent methods to obtain the *total* input-referred noise current from the input-referred noise spectrum: (i) We can calculate the following expression:

$$\overline{i^2_{n,TIA}} = \frac{1}{|Z_T|^2} \int |Z_T(f)|^2 \cdot I^2_{n,TIA}(f)\, df \tag{5.38}$$

where $|Z_T(f)|$ is the frequency response of the TIA. The integration must be carried out over at least $2\times$ the TIA bandwidth. This method is most useful for simulations. (ii) We can use the noise bandwidths BW_n and BW_{n2} to integrate the input-referred white noise up to BW_n and the f^2 noise up to BW_{n2}. This is the method of choice for analytical calculations and we will use it in the following example and discussion.

A Numerical Example. To illustrate the foregoing theory with an example, let's calculate the total input-referred noise current for a differential 10 Gb/s TIA realized with bipolar transistors. The input-referred noise power spectrum follows from Eqs. (5.33) and (5.37):

$$I^2_{n,TIA} \approx 2\left(\frac{4kT}{R_F} + \frac{2qI_C}{\beta} + 2qI_C \cdot \frac{(2\pi C_T)^2}{g_m^2} \cdot f^2 + 4kTR_b \cdot (2\pi C_D)^2 \cdot f^2\right). \tag{5.39}$$

The factor two in front of the expression is because the differential TIA structure features two input transistors and two feedback resistors (we assume a balanced circuit, cf. Section 5.2.8). Applying the noise bandwidths BW_n and BW_{n2} according to Eq. (4.40) we can write the total input-referred noise:

$$\begin{aligned} \overline{i^2_{n,TIA}} &\approx 2\left(\frac{4kT}{R_F} \cdot BW_n + \frac{2qI_C}{\beta} \cdot BW_n + \right. \\ &+ \left. 2qI_C \cdot \frac{(2\pi C_T)^2}{g_m^2} \cdot \frac{BW_{n2}^3}{3} + 4kTR_b \cdot (2\pi C_D)^2 \cdot \frac{BW_{n2}^3}{3}\right). \end{aligned} \tag{5.40}$$

To evaluate this equation numerically, we choose the same values as in our example from Section 5.2.2: $C_D = C_I = 0.15\,\text{pF}$, $C_T = 0.3\,\text{pF}$, $R_F = 600\,\Omega$ and $BW_{3\text{dB}} = 6.85\,\text{GHz}$. Assuming the TIA has a 2nd-order Butterworth frequency response, we find with the help of Table 4.6 $BW_n = 1.11 \cdot 6.85\,\text{GHz} = 7.60\,\text{GHz}$ and $BW_{n2} = 1.49 \cdot 6.85\,\text{GHz} = 10.21\,\text{GHz}$. Furthermore, with the typical BJT parameters $\beta = 100$, $I_C = 1\,\text{mA}$, $g_m = 40\,\text{mS}$, $R_b = 80\,\Omega$, and $T = 300\,\text{K}$ we arrive at the following noise value:

$$\overline{i_{n,TIA}^2} \approx (648\,\text{nA})^2 + (221\,\text{nA})^2 + (710\,\text{nA})^2 + (913\,\text{nA})^2 = (1344\,\text{nA})^2. \tag{5.41}$$

The input-referred noise current $i_{n,TIA}^{rms} = 1344\,\text{nA}$ agrees with the typical numbers seen for commercial 10 Gb/s TIAs. Note that in this example a large amount of noise originates in the BJT's intrinsic base resistance R_b.

Noise Optimization. Now that we have analytical expressions for the input-referred noise current we can go ahead and minimize it through an appropriate choice of component values and operation point. The noise current of a (single-ended) TIA with *FET front-end* is obtained from Eqs. (5.33), (5.34), and (4.40):

$$\overline{i_{n,TIA}^2} = \frac{4kT}{R_F} \cdot BW_n + 2qI_G \cdot BW_n + 4kT\Gamma \frac{(2\pi C_T)^2}{g_m} \cdot \frac{BW_{n2}^3}{3} + \dots \tag{5.42}$$

The first term can be minimized by choosing R_F as large as possible. The second term suggests the use of a FET with a low gate-leakage current. The third term contains the expression C_T^2/g_m which has a minimum for a particular input capacitance C_I. To understand this we rewrite $C_T = C_D + C_I$ and $g_m \approx 2\pi f_T \cdot C_I$. Now the third noise term is proportional to $(C_D + C_I)^2/C_I$ which has a minimum at

$$C_I = C_D. \tag{5.43}$$

Therefore the design rule for FET front-ends is to choose the FET such that its capacitance C_I ($C_{gs} + C_{gd}$, to be precise) matches the parasitic photodiode capacitance C_D plus other stray capacitances. Given the photodiode and stray capacitances, the transistor technology, and the gate length (usually minimum length), the gate width of the front-end FET is determined by this rule.

The noise current of a (single-ended) TIA with *BJT front-end* is obtained from Eqs. (5.33), (5.37), and (4.40):

$$\begin{aligned} \overline{i_{n,TIA}^2} &= \frac{4kT}{R_F} \cdot BW_n + \frac{2qI_C}{\beta} \cdot BW_n + \\ &+ 2qI_C \cdot \frac{(2\pi C_T)^2}{g_m^2} \cdot \frac{BW_{n2}^3}{3} + 4kTR_b \cdot (2\pi C_D)^2 \cdot \frac{BW_{n2}^3}{3} + \dots \end{aligned} \tag{5.44}$$

As before, the first term can be minimized by choosing R_F as large as possible. The second term (base shot noise) *increases* with the collector current I_C, while the third term (collector shot noise) *decreases* with I_C. (Remember that for bipolar transistors $g_m = I_C/V_T$.) Therefore there is an optimum bias current for which the total noise expression is minimized. In practice, the bias current optimization is complicated by the fact that C_I, which is part of C_T, also depends on I_C. The fourth term can be minimized by choosing a technology with low intrinsic base resistance, such as SiGe HBT. For more details on TIA noise optimization see [Kas88].

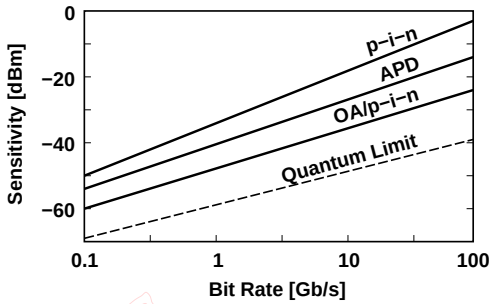

Figure 5.12: Scaling of receiver sensitivity (at $BER = 10^{-9}$) with bit rate [ZNK97].

Given a choice, should we prefer a FET or bipolar front-end? The analysis in [Kas88] concludes that at low speeds ($< 100\,\text{Mb/s}$) the FET front-end outperforms the bipolar front-end by a significant margin. While at high speeds both front-ends perform about the same, with the GaAs MESFET front-end being slightly better.

Scaling of Noise and Sensitivity with Bit Rate. How does the total input-referred noise current of a TIA scale with bit rate? This is an interesting question because the answer to it will also tell us how the sensitivity of a p-i-n receiver scales with bit rate. How sensitive can we make a 10 Gb/s, 40 Gb/s, 160 Gb/s, etc. receiver?

If we assume simplistically that the TIA noise is white and that all TIAs regardless of speed have the same noise density, then the total rms noise is proportional to $\sqrt{B}$ (the noise power is proportional to the bandwidth). Correspondingly, the sensitivity of a p-i-n receiver should drop by 5 dB for every decade of speed increase (assuming a fixed detector responsivity). This was the assumption that went into the plot of Fig. 4.12. However, from Table 5.1 we see that the total rms noise of commercially available TIAs scales roughly like $B^{0.85}$ corresponding to a sensitivity drop of a p-i-n receiver of about 8.5 dB per decade. The fit to the experimental receiver-sensitivity data presented in [ZNK97] (see Fig. 5.12) shows a slope of about 15.8 dB per decade. Both numbers are significantly larger than 5 dB per decade. How can we explain this?

From Eqs. (5.42) and (5.44) we see that for a given technology (constant C_D, C_I, g_m, Γ, and R_b) many noise terms scale with BW^3. Exceptions are the input shot-noise and the feedback-resistor noise terms which both scale with BW. However, remember that the feedback resistor, R_F, is *not* bandwidth independent. As we go to higher bandwidths we are forced to reduce R_F. With the transimpedance limit Eq. (5.22) and Eq. (5.13) we can derive that R_F scales following a $1/BW^2$ law.[3] Thus the resistor-noise term scales

[3] This scaling law leads to extremely large resistor values for low bit-rate receivers (e.g., 1 Mb/s). In

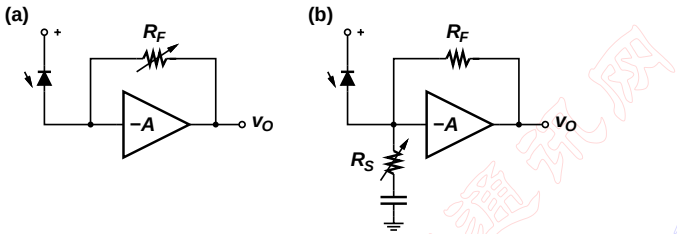

Figure 5.13: TIA with adaptive transimpedance: (a) variable feedback resistor and (b) variable input shunt resistor.

with BW^3 as well! Overall we would expect the total rms noise to be about proportional to $B^{3/2}$. Correspondingly, the sensitivity of a p-i-n receiver should drop by about 15 dB for every decade of speed increase. This agrees well with the data in Fig. 5.12. If the technology is not kept constant across bit-rates, but higher f_T technologies are used for higher bit rates, the slope is reduced.

If we use an APD detector or an optically preamplified p-i-n detector, the sensitivity is determined by TIA noise *and* detector noise (cf. Eqs. (4.25) and (4.26)). In the extreme case where detector noise dominates, the sensitivity scales proportional to B or 10 dB per decade. In practice there is some noise from the TIA and the scaling law is somewhere between $B^{1.0} - B^{1.5}$ corresponding to $10 - 15$ dB per decade. The experimental data in Fig. 5.12 confirms this expectation: we find a slope of about 13.5 dB per decade for APD receivers and 12 dB per decade for optically preamplified p-i-n receivers. The quantum limit has a slope of 10 dB per decade.

For a detector-noise limited receiver with a slope of 10 dB per decade, the number of photons per bit or the energy per bit is independent of the bit rate. However, for a receiver with TIA noise, in particular a p-i-n receiver, we need more and more photons or energy per bit as we go to higher bit rates.

So far we have discussed the basic shunt-feedback TIA. Next we want to look at some variations of this theme.

5.2.4 Adaptive Transimpedance

To design TIAs with a high input overload current, I_{OVL}, or a high maximum input current for *linear* operation, I_{LIN}, the transimpedance can be made adaptive. For large input signals the transimpedance is automatically reduced to prevent the TIA from overloading or limiting. Such TIAs are useful in applications where linear signal processing, such as equalization, is performed on the output signal.

practice, dynamic-range considerations may force the choice of a smaller than optimum resistor [KMJT88]. Under these circumstances the resistor noise may dominate the front-end noise.

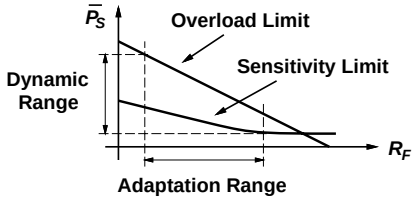

Figure 5.14: Extension of the dynamic range with an adaptive feedback resistor.

Variable Feedback Resistor. A variable feedback resistor R_F, as shown in Fig. 5.13(a), can be used to vary the transimpedance R_T as a function of the signal strength [KTT95, MM96]. Remember that the transimpedance is related to R_F like:

$$R_T = \frac{A}{A+1} \cdot R_F. \tag{5.45}$$

With this approach the input dynamic range is extended without reducing the sensitivity. For small input currents the feedback resistor assumes its maximum value keeping the noise current contributed by R_F small and the sensitivity high. For large input currents the feedback resistor is reduced preventing the amplifier from limiting or overloading. Figure 5.14 shows the overload limit ($\sim 1/R_F$) and the sensitivity limit ($\sim 1/\sqrt{R_F}$ for a small R_F) of a TIA as a function of R_F. It can be seen clearly how an adaptive feedback extends the dynamic range over what can be achieved with any fixed value of R_F. The variable feedback resistor can be implemented with a FET operating in the linear mode. An important consideration for this topology is to make sure that the circuit is stable and peaking remains within specifications for all values of the feedback resistor.

Variable Input Shunt Resistor. Alternatively, a variable input shunt resistor R_S, as shown in Fig. 5.13(b) can be used to achieve the same goal [Yod98]. Here, R_T is controlled by R_S in the following way:

$$R_T = \frac{A}{A+1+R_F/R_S} \cdot R_F. \tag{5.46}$$

For small input currents the shunt resistor R_S assumes a high value and virtually all current from the photodiode flows into the TIA. For large input currents the shunt resistor is reduced, diverting some of the photodiode signal current to AC-ground, thus preventing the amplifier from limiting or overloading. Like in the variable-feedback method, the variable shunt resistor can be implemented with a FET operating in the linear mode. The shunt method has the advantage of causing fewer stability and peaking problems than the variable-feedback method.

For both methods, the bandwidth of the TIA is not constant and usually increases for large input signals. For input signals with little or constant noise this is no problem, because the SNR at the output of the TIA improves for large input signals. But for

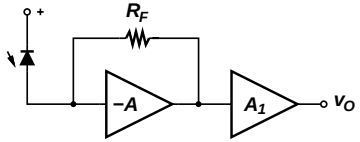

Figure 5.15: TIA with post amplifier.

optical signals with a constant OSNR, as it occurs in optically amplified transmission systems, the bandwidth extension is detrimental because the SNR at the output of the TIA degrades for large input signals. In the latter case a filter after the TIA may be used to stabilize the receiver's frequency response.

5.2.5 Post Amplifier

Post amplifier or main amplifiers are commonly used to further amplify the signal from the TIA. Sometimes, TIA chips contain a "hidden" post amplifier A_1 as shown in Fig. 5.15, often combined with the 50 Ω buffer to drive off-chip loads. The combination of a TIA and a voltage amplifier still behaves like a transimpedance amplifier but with the boosted transimpedance

$$R_T = A_1 \cdot \frac{A}{A+1} \cdot R_F \tag{5.47}$$

where A_1 is the gain of the post amplifier. Thus, with this topology the transimpedance limit, Eq. (5.22), can be broken. However, the noise performance of the boosted TIA is worse than that of the basic shunt-feedback TIA with the same transimpedance. This is so because the boosted TIA has a lower, i.e., noisier, feedback resistor compared to the equivalent TIA without post amplifier (the resistor is lower by a factor A_1). Therefore the transimpedance of the basic shunt-feedback TIA should be made as high as possible. The addition of a post amplifier also reduces the dynamic range for linear operation (I_{LIN}), unless R_F or A_1 are made adaptive.

The post amplifier can be implemented with any of the circuits and techniques described in Chapter 6 on main amplifiers. The paper [OMI+99] describes the combination of a TIA and a limiting amplifier to obtain a high transimpedance at 10 Gb/s.

5.2.6 Common-Base/Gate Input Stage

An interesting TIA variation is the use of a common-base (or common-gate) stage in front of the shunt-feedback topology as shown in Fig. 5.16. Ideally, the transimpedance is not affected by this addition since the current gain of the common-base stage (Q_1, R_C, and R_E) is close to one.

The common-base stage serves to isolate the parasitic photodiode capacitance C_D from the critical node x. As a result, the photodiode capacitance has less of an impact on the bandwidth of the TIA. Therefore this topology is useful if a fixed bandwidth must be achieved with a variety of photodiodes.

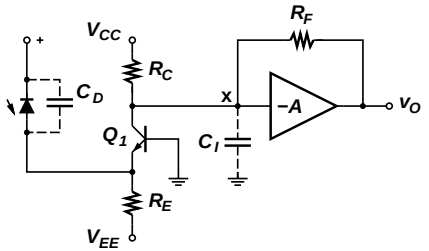

Figure 5.16: TIA with common-base input stage.

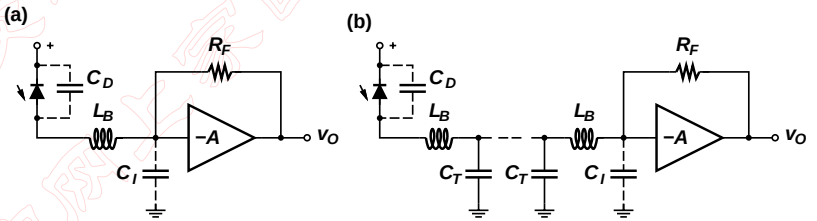

Figure 5.17: TIA with (a) an inductor and (b) an artificial transmission line to couple the photodetector to the input.

Furthermore, we may benefit from the reduced capacitance at node x (ideally reduced from $C_D + C_I$ to C_I). For example we can now increase R_F to get the original bandwidth resulting in a higher transimpedance and less noise from the resistor. Alternatively we can increase the size of the transistor(s) in the amplifier front-end lowering its noise. Potentially, this topology lets us design TIAs with low noise and a transimpedance higher than given by the transimpedance limit Eq. (5.22). However, in practice there are several difficulties: (i) the common-base input stage, in particular Q_1, adds noise and parasitic capacitances of its own, (ii) the input stage may have a current gain of less than one, enhancing the input-referred noise and reducing the transimpedance, and (iii) the input stage is adding to the power dissipation. A comparison of TIAs with common-emitter and common-base input stages can be found in [VT95].

A related approach is the use of a current mirror in front of the shunt-feedback TIA. Such a current-mirror input stage can also be used to decouple the photodiode capacitance from the TIA bandwidth. But just like the common-base stage, the mirror is more likely to increase than to reduce the noise.

5.2.7 Inductive Input Coupling

The photodiode and the TIA are usually located side by side in the same package and connected by a short bond wire. This bondwire introduces a parasitic inductance L_B as shown in Fig. 5.17(a). If the bond wire has the right length (optimum value for L_B), the TIA bandwidth as well as its noise characteristics are improved! This can be understood as follows: Near the resonance frequency of the tank circuit formed by C_D, L_B, and C_I, the current through L_B into the TIA will be *larger* than that generated by the intrinsic photodiode. Thus, if the resonance is placed close to the 3-dB point of the TIA, this peaking mechanism can extend the bandwidth. Furthermore, the current gain in the input network reduces the input-referred noise at high frequencies. The recommended bond-wire inductance is [NRW96]:

$$L_B \approx \frac{1}{(2\pi \cdot BW_{3\text{dB}})^2 \cdot C_D}. \tag{5.48}$$

The inductor in Fig. 5.17(a) can be extended to an artificial (discrete) transmission line between the photodiode and the TIA as shown in Fig. 5.17(b). The idea is to absorb both the detector capacitance and the TIA input capacitance into the transmission line (cf. Section **??**) which is matched to the TIA input impedance $R_{\text{in}} = R_F/(A+1)$. The characteristic impedance of an infinite, artificial transmission line below the cutoff frequency is $Z_{TL} = \sqrt{L_B/C_T}$ and with the matching condition we find:

$$L_B = R_{\text{in}}^2 C_T. \tag{5.49}$$

With this inductor value the cutoff frequency of the artificial transmission line becomes:

$$f_{\text{cutoff}} = \frac{1}{2\pi} \cdot \frac{2}{\sqrt{L_B C_T}} = \frac{1}{2\pi} \cdot \frac{2}{R_{\text{in}} C_T}. \tag{5.50}$$

In conclusion, the original input pole at $1/(2\pi \cdot R_{\text{in}} C_T)$ disappeared and the speed is now limited by the cutoff frequency of the transmission line which is $2\times$ higher. See Section 5.3.3 for a TIA implementation based on this idea [KCBB01].

5.2.8 Differential Output and Offset Control

Most modern TIAs have differential outputs which makes the output signal less susceptible to power-supply noise. In a perfectly balanced circuit power-supply noise only affects the output common-mode and leaves the signal-carrying differential mode undisturbed. Immunity to power-supply noise is particularly important if the TIA is integrated with other digital circuits on the same chip. Differential outputs also permit a larger (differential) output voltage swing, up to twice the power-supply voltage. This is especially helpful in low-voltage designs.

Figure 5.18 shows how the basic shunt-feedback TIA can be expanded into a fully differential topology.[4] For the moment, we ignore the current source I_{OS} shown with

[4] In this and the following circuits we always assume that the differential-output feedback amplifier includes some means to keep the output common-mode voltage at a constant level. For an implementation example see Section 5.3.3.

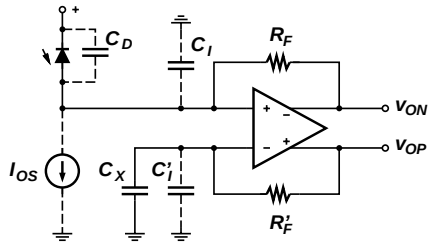

Figure 5.18: TIA with differential outputs.

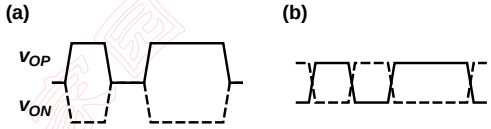

Figure 5.19: TIA output signals: (a) without and (b) with offset control.

dashed lines. Since the input signal from the photodiode is single-ended we have to decide what to do with the unused amplifier input. There are two possibilities each with its own set of advantages and disadvantages.

Capacitive Input Termination. One possibility is to connect a small capacitor C_X that matches the photodiode capacitance, $C_X = C_D$, to the unused input. In this case our circuit is fully balanced and we get excellent power-supply noise rejection. We may want to realized C_X with a dummy photodiode kept in the dark to obtain good matching. The differential transimpedance of this topology is about R_F and therefore the single-ended one is $R_F/2$, only half of what we had for the single-ended TIA. The input-referred noise current contains the noise of *both* feedback resistors, R_F and R'_F, making this differential TIA less sensitive than the single-ended one.

Alternatively, we can use a large capacitor $C_X \to \infty$ to short the unused input to AC ground. The properties of the resulting circuit are very similar to the single-ended TIA: The differential transimpedance is now about $2 \cdot R_F$ and thus the single-ended one is R_F, just like what we had for the single-ended TIA. Furthermore, we only have the noise contribution of resistor R_F, the noise of R'_F is shorted out by the large capacitor. However, because of the asymmetric input capacitances, power-supply noise couples differently to the two inputs causing noise to leak into the differential mode.

Offset Control. The two output signals of the TIA in Fig. 5.18 with $I_{OS} = 0$ are vertically offset against each other as shown in Fig 5.19(a). Recall the characteristics of the current signal from the photodiode shown in Fig. 5.3. When the photodiode is dark the input current is zero and the two output voltages are equal, when the diode is

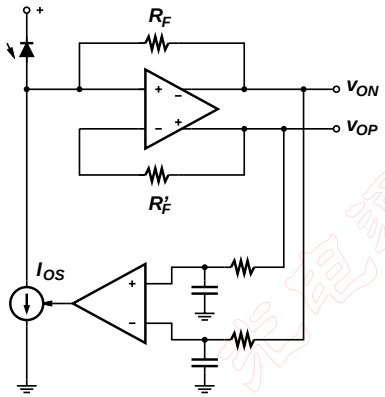

Figure 5.20: TIA with offset control.

illuminated, v_{ON} (dashed) moves down and v_{OP} (solid) moves up.

This undesirable output offset voltage can be eliminated by AC coupling the TIA outputs to the inputs of the next block (the MA). However, there are two disadvantages to this method: (i) the offset voltage unnecessarily reduces the output swing of the TIA and (ii) in an integrated solution (TIA and MA on the same chip) AC coupling may not be a viable option because it requires large (external) capacitors. The offset control circuit shown in Fig. 5.20 eliminates the output offset voltage producing output signals as shown in Fig. 5.19(b). This circuit measures the output offset voltage by subtracting the average (low-pass filtered) values of the two output signals, and in response to this difference controls the current source I_{OS} until the output offset is eliminated. Alternatively, the output offset can be determined from the difference of the output peak values or the average voltage drop across R_F.

In order to reduce the TIA's input capacitance one may consider to connect an offset-control current source with opposite polarity, $-I_{OS}$, to the unused input. Although this alternative does remove the output offset voltage, it is not a very practical solution because now the amplifier's average input common-mode voltage becomes dependent on the received power level and as a result the amplifier's common-mode range may be violated at high input power levels.

5.2.9 Burst-Mode TIA

How does a burst-mode TIA differ from what we have discussed so far? A burst-mode TIA has to cope with an input signal without DC balance and an amplitude that varies significantly (up to 30 dB in PON systems) from burst to burst. To remove the output offset voltage we cannot use AC coupling because of the lack of DC balance. The circuit in Fig. 5.20 doesn't work either, for the same reason. So, we could leave the output

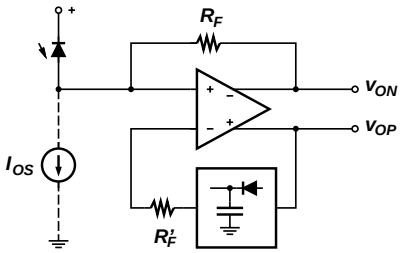

Figure 5.21: TIA for burst-mode operation.

offset voltage the way it is and deal with it in the main amplifier. Or we can use the offset-control circuit shown in Fig. 5.21 [OS90, OSA+94, MS96].

Offset Control. The offset-control circuit in Fig. 5.21 removes the output offset on a burst-by-burst basis, i.e., the offset is adjusted at the beginning of every burst. This circuit is also known as an *Adaptive Threshold Control* (ATC) circuit, because after offset removal, the differential output signal can be sliced at the crossover point without the need of any further threshold.

How does this circuit eliminate the output offset voltage? For now let's ignore the current source I_{OS}. Before the burst arrives, the peak detector is reset to a voltage equal to the output common-mode voltage of the amplifier. The differential output voltage is now 0 V. Then, when the first one bit of the burst arrives, v_{OP} moves up and v_{ON} moves down. The peak value of v_{OP} is stored in the peak detector and fed back to the inverting input. During the next zero bit, the value of the peak detector appears at the v_{ON} output. Why? Because there is no voltage drop across R'_F (no input current), no voltage across the inputs of the amplifier (for a large gain), and no voltage drop across R_F (no photo current). Thus, the peak values of both output signals are equal and the output offset is eliminated!

The differential transimpedance of this burst-mode TIA is about $2 \cdot R_F$, the same value as for the continuous-mode TIA in Fig. 5.18 with a large $C_X \to \infty$.

Chatter Control. Besides offset control, there is another problem with burst-mode TIAs: Extended periods of time may elapse in between bursts during which no optical signal is received at all. If the peak detector in the TIA is in its reset state during these periods (reset is usually done at the end of each burst), the amplifier noise causes output signal crossovers leading to random bit sequences called *Chatter* at the output of the receiver.

One way to fix this problem is to introduce a small artificial offset voltage by means of the current source I_{OS} shown in Fig. 5.21 [OS90]. This offset voltage must be larger than the peak noise voltage to suppress the chatter, but it should not be too high either in order not to degrade the sensitivity.

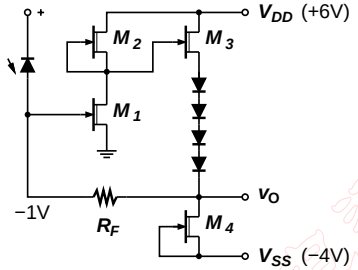

Figure 5.22: MESFET/HFET implementation of a TIA.

5.3 TIA Circuit Implementations

In the following section we will examine some representative transistor-level TIA circuits taken from the literature and designed for a variety of technologies (cf. Appendix ??). These circuits will illustrate how the design principles discussed in the previous section are implemented in practice.

5.3.1 MESFET & HFET Technology

Figure 5.22 shows a classical single-ended TIA implemented in a depletion-mode MESFET or HFET technology. For instance the TIAs reported in [SBL91, Est95] are based on this topology and realized in GaAs technology.

The gain of the feedback amplifier is provided by a common-source stage M_1 with a current-source load M_2. Note that for depletion-mode devices, shorting the gate and the source yields a constant current source. A source follower M_3 buffers the output signal and four Schottky diodes shift the output DC-level to a level that is suitable to bias the input transistor M_1. M_4 is another current source to bias the source follower. The feedback resistor R_F closes the loop back to the inverting input.

5.3.2 BJT & HBT Technology

Figure 5.23 shows a single-ended TIA implemented in a bipolar (BJT or HBT) technology. Circuits based on this topology can be found for example in [KTT95, STS+94].

The operation is very similar to that of the MESFET implementation. The common-emitter stage Q_1 provides the gain of the feedback amplifier which is determined by the ratio of the collector resistor R_C to the emitter resistor R_E. Following this gain stage is a cascade of three emitter followers ($Q_2 - Q_4$) which buffer and level-shift the output signal. The feedback resistor R_F closes the loop. The output signal is taken from the collector of the last transistor Q_4 which has a higher amplitude than the emitter signal.

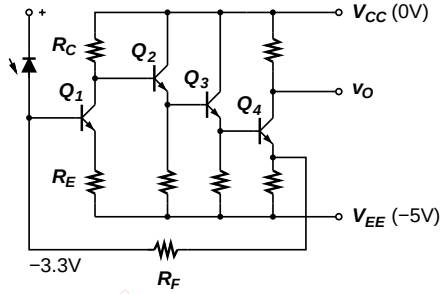

Figure 5.23: BJT/HBT implementation of a TIA.

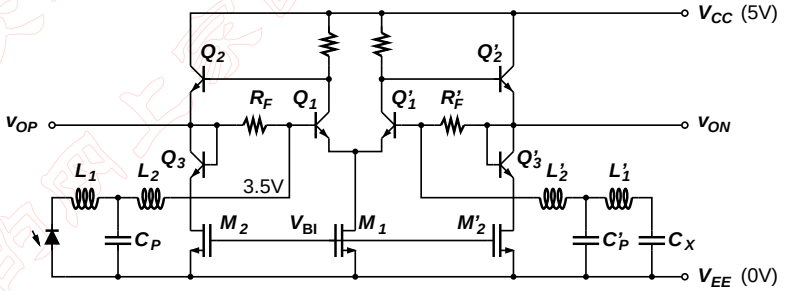

Figure 5.24: BiCMOS implementation of a TIA.

5.3.3 BiCMOS Technology

Figure 5.24 shows a differential TIA implemented in BiCMOS technology taken from [KCBB01]. The signal path is implemented with BJTs while the bias network is implemented with MOS transistors. This partitioning is typical for BiCMOS implementations: BJTs exhibit good high-frequency and matching characteristics. MOS current sources are easy to bias, provide a high output impedance, and have low noise. For a noise comparison of MOS and BJT current sources see [Raz96].

The feedback amplifier consists of the differential pair Q_1, Q_1', the tail current source M_1, and poly-resistor loads. The constant tail current together with the linear load resistors guarantee a constant common-mode output voltage. Each output is buffered with an emitter follower (Q_2, Q_2'). The emitter followers are biased by the MOS current sources M_2, M_2' which have diode-connected bipolar transistors Q_3, Q_3' in series to keep the drain-source voltage below the breakdown voltage. Two feedback resistors R_F and R_F' are closing the loop.

As discussed in Section 5.2.7 the photodiode is coupled to the TIA input with an arti-

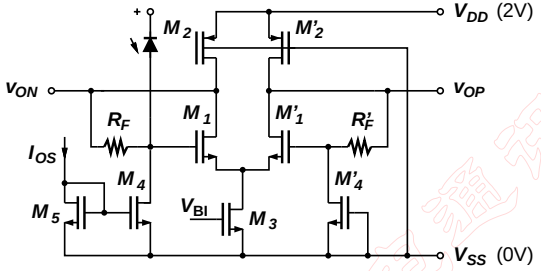

Figure 5.25: CMOS implementation of a TIA.

ficial transmission line to improve noise and bandwidth characteristics. The transmission line (or filter) consists of the bond-wire inductor L_1, the bond-pad capacitance C_P, and a spiral inductor L_2. As explained in Section 5.2.8, the capacitor C_X on the unused input can be either very large or matched to the photodiode capacitance. In this case a balanced approach is used replicating the filter and the photodiode capacitance on the unused side. No offset control mechanism is present in this TIA implementation.

5.3.4 CMOS Technology

Low-Voltage TIA Circuit. A differential CMOS TIA circuit for low-voltage operation, taken from [TSN+98b], is shown in Fig. 5.25. The feedback amplifier consists of the CMOS differential pair M_1 and M'_1, the tail current source M_3, and the p-MOS load transistors M_2 and M'_2. The latter have grounded gates and operate in the linear mode. The output signals of the differential stage are fed back to the inputs with the feedback resistors R_F and R'_F. No source-follower buffers are used, because of the limited headroom in this 2 V design. M_5 and M_4 form a current mirror to supply the offset control current I_{OS} to the input. Dummy transistor M'_4 at the other input balances out the capacitance of M_4. Removal of the output offset voltage, to increase the output dynamic range, is important in this low-voltage design.

Common-Gate TIA Circuit. A differential CMOS TIA circuit with common-gate input stage, taken from [MHBL00], is shown in Fig. 5.26. Again, a differential pair M_1 and M'_1 is at the center of the feedback amplifier but now, additionally, cascode transistors M_2 and M'_2 as well as inductive loads L and L' are used to extend the bandwidth. Broadband techniques such as these will be discussed in Section 6.3.2. Source-follower buffers M_3 and M'_3 drive the TIA outputs as well as the feedback resistors R_F and R'_F. So far this circuit is quite similar to what we have seen before. However, note that the photodiode is connected to the source of M_4 instead of the gate of M_1. Transistor M_4 is a common-gate input stage used to decouple the photodiode capacitance from the critical node at the gate of M_1. Transistor M_4 also increases the reverse-bias of the photodiode by V_{DS4} which may

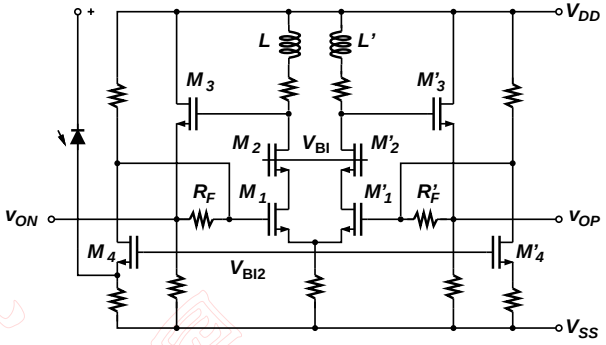

Figure 5.26: CMOS implementation of a TIA with common-gate stages.

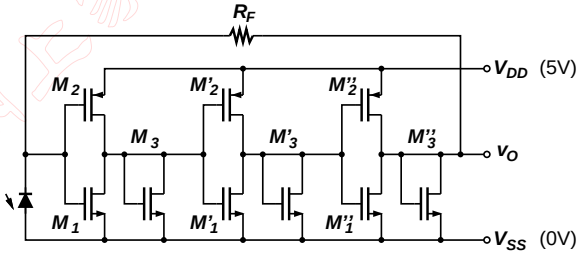

Figure 5.27: CMOS implementation of a multistage TIA.

help to improve the photodiode speed in low-voltage designs. To make the TIA topology as balanced as possible the common-gate input stage is replicated at the unused input.

A CMOS TIA with a common-gate input stage using a regulated-cascode is described in [PT00].

Multistage TIA Circuit. A CMOS TIA circuit with three stages in the feedback amplifier, taken from [IVCS94], is shown in Fig. 5.27. Each stage of the amplifier consists of three transistors (M_1, M_2, and M_3). Transistors M_1 and M_2 act as a push-pull transconductor and M_3 represents the load resistor. The stage gain is approximately $(g_{m1} + g_{m2})/g_{m3}$ and thus well defined by the device geometry. Stability of this TIA is an issue because we now have three rather than one non-dominant pole. These poles must be placed at a higher frequency than the single pole in Fig. 5.7 to ensure a flat closed-loop frequency response.

5.4 Product Examples

Table 5.1 summarizes the main parameters of some commercially available TIA chips. The numbers have been taken from data sheets of the manufacturer which were available to me at the time of writing. For up-to-date product information please contact the manufacturer directly. For consistency, the tabulated R_T is always the *single-ended* transimpedance, for TIAs with single-ended as well as differential outputs. Both the overload current (I_{OVL}) and the maximum current for linear operation (I_{LIN}) are measured peak-to-peak.

Comparing the 2.5 and 10 Gb/s parts we see that in general higher speed parts have a lower transimpedance in accordance with the transimpedance limit Eq. (5.22). We also find a few parts that have an unusually high transimpedance, such as the FOA1251B1. These parts have an on-chip post amplifier as discussed in Section 5.2.5. The MAX3866 chip includes a TIA and a LA which explains its higher power dissipation compared to the MAX3267. The LG1628AXA can sustain a particularly high overload current due to its adaptive transimpedance as discussed in Section 5.2.4 [Yod98].

Again comparing the 2.5 and 10 Gb/s parts we observe that the total input-referred noise current is higher for the faster parts. This is not surprising since their bandwidth is larger and their feedback resistor is lower.

5.5 Research Directions

The research effort going on around the world can be divided roughly into four categories: higher speed, higher integration, lower cost, and lower noise.

Higher Speed. Traditionally, every new generation of optical telecommunication equipment runs at 4× the speed of the previous generation: 155 Mb/s (OC-3), 622 Mb/s (OC-12), 2.5 Gb/s (OC-48), 10 Gb/s (OC-192), etc. In the world of data communication the steps taken are even larger. Every new generation is 10× faster than the previous one. Ethernet comes at the speed grades 10 Mb/s, 100 Mb/s, 1 Gb/s, 10 Gb/s, etc.

The next speed of great interest is 40 Gb/s (OC-768). Many research groups are aiming at that speed and beyond. To achieve the necessary TIA bandwidth in the $30-40$ GHz range aggressive high-speed technologies are needed. Such technologies are SiGe, GaAs, and InP combined with heterostructure devices such as HBTs and HFETs.

Here are some examples from the literature:

- In SiGe-HBT technology a 40 Gb/s, a 35 GHz, and a 45 GHz TIA have been reported in [MMR+98], [MOO+98], and [MOA+00], respectively.
- In GaAs-HFET technology a 22 GHz TIA has been reported in [LBH+97].
- In GaAs-HBT technology a 40 Gb/s and a 25 GHz TIA have been reported in [SSA+99] and [RZP+99], respectively.
- In InP-HFET technology a 49 GHz TIA has been reported in [SSS+01].

Company & Product	Speed	R_T	I_{OVL}	I_{LIN}	$BW_{3\mathrm{dB}}$	$i_{n,TIA}^{rms}$	Power	Technology
Agere LG1628AXA	2.5 Gb/s	5.8 kΩ	8.0 mA		1.6 GHz	300 nA	728 mW	GaAs HFET
Anadigics ATA30013D1C	2.5 Gb/s	1.8 kΩ		0.6 mA	2.0 GHz	433 nA	850 mW	GaAs MESFET
Infineon FOA1251B1	2.5 Gb/s	13.5 kΩ	2.2 mA	10 µA	1.8 GHz		112 mW	Si BJT
Maxim MAX3267	2.5 Gb/s	1.0 kΩ	1.0 mA	40 µA	1.9 GHz	485 nA	86 mW	SiGe HBT
Maxim MAX3866	2.5 Gb/s				1.8 GHz	433 nA	165 mW	
Nortel AB89	2.5 Gb/s	1.3 kΩ	2.6 mA		2.5 GHz	330 nA	182 mW	Si BJT
Nortel AC89	2.5 Gb/s	2.0 kΩ	3.0 mA		2.2 GHz	270 nA	325 mW	Si BJT
Agere TTIA110G	10 Gb/s	1.0 kΩ	2.5 mA	0.72 mA	10 GHz	1500 nA	850 mW	GaAs HFET
AMCC S3090	10 Gb/s	0.7 kΩ	2.2 mA		10 GHz	1580 nA	416 mW	SiGe HBT
Anadigics ATA7600D1	10 Gb/s	0.63 kΩ	0.8 mA		9 GHz		300 mW	GaAs HBT
Giga GD19906	10 Gb/s	0.5 kΩ	2.5 mA		10 GHz	1200 nA	750 mW	GaAs HBT
Maxim MAX3970	10 Gb/s	0.3 kΩ	1.6 mA	0.13 mA	9 GHz	1100 nA	150 mW	SiGe HBT
Nortel D68	10 Gb/s	0.5 kΩ	2.5 mA		10 GHz		750 mW	GaAs HBT
Philips CGY2110	10 Gb/s	2.0 kΩ	1.8 mA		9 GHz		290 mW	GaAs HFET
AMCC S76800	40 Gb/s	0.22 kΩ	3.0 mA		45 GHz	4000 nA	600 mW	SiGe HBT

Table 5.1: Examples for 2.5 Gb/s+ TIA products.

- In InP-HBT technology 40 Gb/s TIAs have been reported in [SSO+97] and [MTM+00].

Higher Integration. Another promising research direction, aiming at higher integration, is to combine the photodetector and electronic circuits on the same chip, so-called *Optoelectronic Integrated Circuits* (OEIC) [Sze98, Wal99]. The simplest form of an OEIC is the so-called *pin-FET* which consists of a p-i-n photodetector and a FET integrated on the same substrate. The OEIC approach reduces size and saves packaging cost but the resulting receivers are often less sensitive than their hybrid counterparts. This is because of compromises that must be made to integrate a photodiode and transistors in the *same* technology. However, with the increasing use of EDFAs this loss in sensitivity may to be a major drawback.

In some InP circuit technologies it is possible to build high quality photodiodes which are sensitive in the desired $1.3 - 1.6\,\mu$m range by reusing the base-collector junction of an HBT. Bringing the photodiode and the TIA on the same chip reduces the critical capacitance C_T and therefore is particularly advantageous for high-speed receivers. If the circuit technology does not permit the integration of good photodiodes it is sometimes possible to use *Metal-Semiconductor-Metal Photodetectors* (MSM). This is an attractive possibility for FET and HFET technologies. Another OEIC approach is to use a flip-chip photodiode on top of a circuit chip. In the latter case the circuit technology may be based on silicon or GaAs while the flip-chip is fabricated on an InP substrate in order to be sensitive to long wavelengths. An advantage of this approach is that detectors and transistors can be optimized individually.

In silicon technologies it is possible to monolithically integrate a photodiode which is sensitive in the $0.85\,\mu$m wavelength range together with a TIA which is useful for data-communication applications [WK98], [Ste01].

Lower Cost. Another area of research is focusing on the design of high-performance TIAs in low-cost, mainstream technologies, in particular digital CMOS. In order to achieve this goal "tricky" analog circuits are invented. Besides reducing cost, a CMOS TIA has the advantage that it can be combined with dense digital logic on the same chip. This system-on-a-chip approach reduces the chip count and power dissipation but poses a challenge with regard to noise isolation between the analog and digital sections.

A SONET compliant 10 Gb/s TIA has been implemented in a low-cost "modular BiCMOS" technology [KCBB01]. A *modular* BiCMOS technology is basically a CMOS technology with a few masks (a module) added to provide bipolar transistors. A 2.4 Gb/s, $0.15\,\mu$m CMOS TIA integrated with an AGC amplifier and a CDR has been reported in [TSN+98b]. A 1.2 GHz, $0.5\,\mu$m CMOS TIA has been reported in [MHBL00].

Lower Noise. Recognizing that the feedback resistor(s) contribute a large portion of the overall noise in low-speed receivers, researchers have looked into building TIAs with a noise-free feedback mechanism. Approaches which have been studied are (i) capacitive feedback [Raz00], (ii) optical feedback [KMJT88], and (iii) no feedback at all. In the last case a switch is used to precharge the amplifier input at the beginning of every bit (integrate and dump) [Jin90].

5.6 Summary

The main specifications for the TIA are:

- The transimpedance which we want to be as large as possible to simplify the MA design.
- The input overload current which must be large enough to avoid pulse-width distortion and jitter when receiving a large optical signal.
- The maximum current for linear operation which is important in applications with linear signal processing (e.g., equalization).
- The input-referred noise current which must be as small as possible to obtain good sensitivity, in particular if a p-i-n photodiode is used. The total input-referred noise current, not the spectral density determines the sensitivity.
- The TIA bandwidth which is chosen between $0.6 - 1.2 \cdot B$ depending on how bandwidth is allocated in the system.
- The group-delay variation which must be kept small to minimize signal distortions.

The shunt-feedback principle is used in virtually all TIA designs because it simultaneously provides high bandwidth, high transimpedance, high dynamic range, and low noise. Noise in a FET front-end can be optimized by matching the TIA input capacitance to the photodiode capacitance. Noise in a bipolar front-end can be optimized with the appropriate choice of collector bias current.

Circuits with adaptive transimpedance are used to improve the overload current and the maximum current for linear operation. Post amplifiers can be used to boost the transimpedance. Common-base/gate input stages can be used to decouple the photodiode capacitance from the TIA bandwidth. A bondwire inductance or LC filter between the photodiode and TIA can improve the bandwidth and noise characteristics. Differential outputs are becoming increasingly popular because of the improved power-supply rejection ratio and increased output voltage swing. An offset control circuit is often used in differential-output TIAs to improve the output dynamic range. Burst-mode TIAs require a special offset control mechanism that can deal with a bursty input signal with varying amplitude and no DC balance.

TIAs have been implemented in a wide variety of technologies including MESFET, HFET, BJT, HBT, BiCMOS, and CMOS.

Currently, researchers are working on 40 Gb/s TIAs, TIAs integrated with the photodetector on the same chip, TIAs in low-cost technologies such as CMOS, and ultra low-noise TIAs

5.7 Problems

5.1 Low-Impedance Front-End. A $50\,\Omega$ low-impedance front-end is followed by an amplifier with $50\,\Omega$ inputs and outputs, gain $A = 40$ dB, noise figure $F = 2$ dB,

and noise bandwidth $BW_n = 10\,\text{GHz}$. (a) How large is the transimpedance of this arrangement? (b) How large is the input-referred rms noise current? (b) How does the optical sensitivity of this front end compare to a TIA front-end with $i_n^{rms} = 1.2\,\mu\text{A}$?

5.2 TIA Stability. For a 2-pole TIA the open-loop pole spacing, $R_F C_T / T_A$, must be $2A$ to obtain a Butterworth response after closing the loop (Eq. (5.20)). (a) Given, an arbitrary closed-loop Q value, what is the required open-loop pole spacing? (b) What is the required pole spacing for a Bessel response? (c) What is the required pole spacing for a critically damped response?

5.3 TIA Noise. The "noise corner" occurs at the frequency where the white noise and f^2 noise are equally strong. (a) Derive an expression for the noise-corner frequency of a TIA with a MOSFET front-end. (b) How is this corner frequency related to the bit rate?

Chapter 6

Main Amplifiers

6.1 Limiting vs. Automatic Gain Control (AGC)

The job of the *Main Amplifier* (MA) is to amplify the small input signal v_I from the TIA to a level which is sufficient for the reliable operation of the CDR. The required level for the output signal v_O is typically a few 100 mV peak-to-peak. Almost all MAs feature differential inputs as well as differential outputs, as shown in Fig. 6.1. The MA is also known as *Post Amplifier* since it follows the TIA.

Depending on the application, nonlinear distortions caused by the MA may or may not be acceptable (cf. Chapter 4). If good linearity is required, an *Automatic Gain Control Amplifier* (AGC) must be used. If nonlinear distortions can be tolerated, the simpler *Limiting Amplifier* (LA) design is preferred.

Limiting Amplifier. For small input signals, most amplifiers have a fairly linear transfer function. For large signals, however, nonlinear effect become apparent. Specifically, if the output amplitude approaches the power-supply voltage, severe distortions in the form of clipping set in. The DC transfer function of a LA is shown schematically in Fig. 6.2(a). The linear and the limiting regimes can be distinguished clearly. No special amplifier design is required to obtain the limiting characteristics of a LA, it happens naturally because of the finite power-supply voltage and other signal swing constraints. However, the designer needs to make sure that the limiting happens in a *controlled* way, i.e., pulse-width distortions, jitter, and delay variations when the LA transitions from the linear to the limiting regime must be minimized.

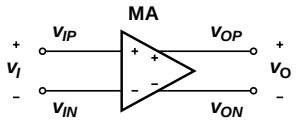

Figure 6.1: Input and output signals of a fully differential MA.

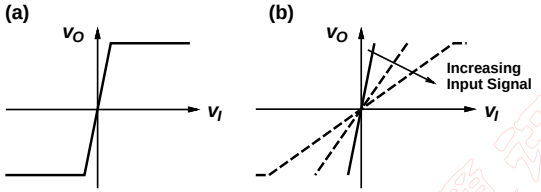

Figure 6.2: DC transfer characteristics of (a) a LA and (b) an AGC amplifier.

AGC Amplifier. An AGC amplifier has a variable gain which is controlled by a feedback mechanism such that the output amplitude remains constant for a wide range of input amplitudes. Whereas a LA starts to distort (limit) for large input signals, the AGC amplifier reduces its gain and thus manages to stay in the linear regime. Figure 6.2(b) shows the DC transfer function of an AGC amplifier and its dependence on the input signal amplitude. For very large signals the AGC amplifier cannot reduce the gain any further and limiting will occur eventually. The system designer must make sure that the input dynamic range of the AGC amplifier is not exceeded.

The LA can be regarded as an AGC amplifier for which the gain-control mechanism is stuck at the maximum gain. This suggests that a LA is easier to design than an AGC amplifier because the gain-control mechanism can be omitted. Furthermore, a LA can be designed with better electrical characteristics such as power dissipation, bandwidth, noise, etc. compared to an AGC amplifier realized in the same technology. The advantage of the AGC amplifier is its linear transfer function which preserves the signal waveform and permits analog signal processing on the output signal. Examples for such signal processing tasks are equalization, slice-level steering, and soft-decision decoding which have been discussed in Sections 4.7, 4.10, and 4.11, respectively.

6.2 MA Specifications

Before examining LA and AGC circuits we will discuss their main specifications: gain, bandwidth, group-delay variation, noise figure, input offset voltage, low-frequency cutoff, input dynamic range, and AM-to-PM conversion.

How do we find appropriate values for these specification? We could just look at the data sheet of an existing chip that is known to work well and use the same values. But there is a more systematic way to do this. First, we establish a mathematical relationship between each specification and the associated power penalty. Then, given an acceptable value for the power penalty, we can derive a limit for each specification. In the following we will illustrate this method, which was introduced in Section 4.5, with several examples.

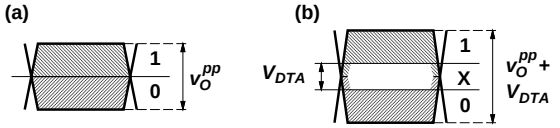

Figure 6.3: MA output signal for (a) an ideal DEC and (b) a DEC with finite sensitivity.

6.2.1 Gain

Definition. The *Voltage Gain* of the MA is defined as:

$$A = \frac{\Delta v_O}{\Delta v_I} \tag{6.1}$$

where Δv_I is the input-voltage change produced by a small signal such that the amplifier operates in the linear regime and Δv_O is the output-signal change (see Fig. 6.1). In the case of an AGC amplifier the input signal must be small enough such that the amplifier assumes its maximum gain, $A_{\max}$. When operated with an AC signal, the MA is characterized by a gain magnitude $|A|$ and a phase shift Φ between the input and the output signal. The two can be combined into a complex gain value $A = |A| \cdot \exp(i\Phi)$.

The voltage gain A is closely related to the S parameter S_{21}. If the input matching of the amplifier is perfect ($S_{11} = 0$), the two quantities are identical (cf. Appendix **??**).

For differential-output amplifiers it is important to specify if the gain is measured single-endedly, with $v_O = v_{OP}$ or $v_O = v_{ON}$, or differentially, with $v_O = v_{OP} - v_{ON}$. The differential gain is 6 dB higher than the singe-ended gain. Fortunately, the gain depends little on how the input is driven: single-endedly or differentially. The input voltage v_I of a differential amplifier is *always* defined as the voltage between the non-inverting v_{IP} and inverting input v_{IN}, no matter if they are both driven or if one of them is grounded. The only difference between single-ended and differential drive is in the excitation of the common mode.

Most high-speed MAs have 50 Ω outputs which must be properly terminated with 50 Ω resistors when measuring A or else the value will come out too high. Next we want to derive the gain requirement for the MA.

Power Penalty. The MA gain specification can be derived from the minimum signal $v_I^{pp}(\min)$ provided by the TIA and the minimum signal $v_O^{pp}(\min)$ required by the CDR for reliable operation.

The minimum input signal to the MA is given by the smallest TIA output signal for which the required BER can be met, i.e., the TIA output signal at the sensitivity limit. This voltage can be expressed in terms of the TIA's noise and transresistance values using Eqs. (5.1) and (5.5):

$$v_I^{pp}(\min) = 2\mathcal{Q} \cdot R_T \cdot i_{n,TIA}^{rms}. \tag{6.2}$$

The minimum input signal required by the CDR depends on the decision circuit sensitivity and the power penalty that we are willing to accept. Let's define the *Decision Circuit*

Sensitivity a.k.a. *Decision Threshold Ambiguity Width* (DTAW), V_{DTA}, as the peak-to-peak input voltage below which the decision circuit makes random decisions ($BER = 0.5$). We further assume that the DEC operates without error if we go just slightly above this voltage (i.e., we assume the DEC is noise free). From Fig. 6.3 we see that in order to obtain the same BER for $V_{DTA} > 0$ as in the case for $V_{DTA} = 0$, we need to increase the DEC input signal, which is the MA output signal v_O^{pp}, by V_{DTA}. This means we incur the power penalty:

$$PP = \frac{v_O^{pp} + V_{DTA}}{v_O^{pp}} = 1 + \frac{V_{DTA}}{v_O^{pp}}. \tag{6.3}$$

If we solve this equation for the minimum MA output voltage, we find:

$$v_O^{pp} > \frac{1}{PP - 1} \cdot V_{DTA}. \tag{6.4}$$

For example, if the DEC sensitivity is 10 mV and we accept a power penalty of 0.2 dB ($PP = 1.047$) the MA must deliver 213 mV peak-to-peak or more to the CDR. In addition to the above DEC-sensitivity considerations, we also have to make sure that this voltage is large enough to drive the CDR's phase detector. But this is usually not a problem and we will assume here that the phase detector presents no limitation.

Now we can divide the output voltage, Eq. (6.4), by the input voltage, Eq. (6.2), to get an expression for the minimum MA gain:

$$A > \frac{V_{DTA}}{(PP - 1) \cdot 2\mathcal{Q} \cdot R_T \cdot i_{n,TIA}^{rms}}. \tag{6.5}$$

As expected, if the DEC has perfect sensitivity ($V_{DTA} = 0$), no gain is required. A TIA with high transimpedance reduces the gain requirement for the MA.

Typical Values. Next we want to calculate some typical numbers based on a 10 mV DEC sensitivity, 0.2 dB power penalty ($PP = 1.047$), and $BER = 10^{-12}$ ($\mathcal{Q} = 7$). For a typical 2.5 Gb/s system ($R_T = 1.5\,\text{k}\Omega$, $i_{n,TIA}^{rms} = 400\,\text{nA}$) we obtain the numerical value:

$$A > \frac{10\,\text{mV}}{0.047 \cdot 14 \cdot 1.5\,\text{k}\Omega \cdot 400\,\text{nA}} = 28.1\,\text{dB}. \tag{6.6}$$

And for a typical 10 Gb/s system ($R_T = 0.5\,\text{k}\Omega$, $i_{n,TIA}^{rms} = 1200\,\text{nA}$) we get:

$$A > \frac{10\,\text{mV}}{0.047 \cdot 14 \cdot 0.5\,\text{k}\Omega \cdot 1200\,\text{nA}} = 28.1\,\text{dB}. \tag{6.7}$$

In conclusion, a typical MA has a gain of around 30 dB.

6.2.2 Bandwidth and Group-Delay Variation

Definition. The MA bandwidth $BW_{3\text{dB}}$ is defined as the (upper) frequency at which the small-signal gain $A(f)$ has dropped by 3 dB below its passband value. This definition does not say anything about the phase response. Even if the amplitude response is flat up to a given frequency, distortions may occur below this frequency if the phase linearity of $A(f)$ is insufficient. A common measure for phase linearity is the variation of the group delay with frequency. Group delay, τ, is related to the phase, Φ, as $\tau(\omega) = -d\Phi/d\omega$.

You may wonder: "What is the meaning of bandwidth for a nonlinear circuit such as a LA?" You are right, if the LA is operated in the limiting regime the concept of bandwidth does no longer apply must be replaced by other concepts such as slew rate. But it is always possible to reduce the input signal to the point where the LA becomes linear and then bandwidth is defined as usual.

Typical Values. As discussed in Section 4.6, the MA bandwidth is often made much wider than the receiver bandwidth, which is determined by either the TIA or a filter. This measure guarantees that the MA does not reduce the desired receiver bandwidth. As a rule of thumb, the MA bandwidth is chosen:

$$BW_{3\text{dB}} = 1.0 \cdot B \ldots 1.2 \cdot B \tag{6.8}$$

where B is the bit rate. This is nearly twice the recommended receiver bandwidth ($2/3 \cdot B$). Numerical values for the MA bandwidth based on Eq. (6.8) are:

$$2.5\,\text{Gb/s MA:} \quad BW_{3\text{dB}} = 2.5\,\text{GHz} \ldots 3\,\text{GHz} \tag{6.9}$$

$$10\,\text{Gb/s MA:} \quad BW_{3\text{dB}} = 10\,\text{GHz} \ldots 12\,\text{GHz}. \tag{6.10}$$

Typically, a group delay variation, $\Delta\tau$, of less than $\pm 10\,\%$ of the bit period (± 0.1 UI) over the specified bandwidth is required. This corresponds to:

$$2.5\,\text{Gb/s MA:} \quad |\Delta\tau| < 40\,\text{ps} \tag{6.11}$$

$$10\,\text{Gb/s MA:} \quad |\Delta\tau| < 10\,\text{ps}. \tag{6.12}$$

6.2.3 Noise Figure

Noise generated in the MA adds to the total receiver noise and therefore degrades the receiver sensitivity. Usually the dominant noise source in the receiver is the TIA and for this reason the MA noise should be made small compared to the TIA noise. In the following we will first review the definition of noise figure and then go on to calculate the power penalty incurred due to the MA noise. The latter calculation leads up to an expression for the highest acceptable noise figure.

Definition. The noise characteristics of the MA can be described with a so-called *Equivalent Noise Voltage Source* which is attached to one input of the amplifier (see Fig. 6.4). The equivalent noise source, $v_{n,\text{eq}}$, is chosen such that together with a noiseless amplifier and noiseless signal source it produces the same output noise as the real, noisy amplifier. The rms value of this noise source, $v_{n,\text{eq}}^{rms}$, is determined, just like for the TIA, by integrating the output-referred noise spectrum and referring the result back to the input.

There is a potential pitfall which is worth pointing out. The noise voltage that is superimposed to the signal (v_S) at the input of the amplifier is *only half* of the equivalent noise voltage $v_{n,\text{eq}}$. This is because the equivalent noise voltage is split by two 50 Ω resistors: the source resistance and the input termination resistor inside the amplifier (see

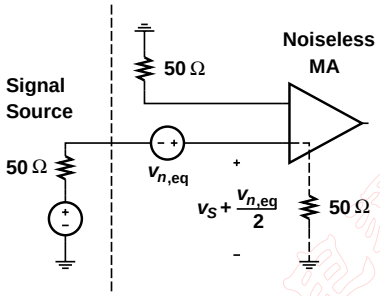

Figure 6.4: Single-ended noise figure: The 50 Ω input resistor to the right of the dashed line is not part of the source.

Fig. 6.4). The noise available at the input of the amplifier is known as the *Available Noise Voltage*:

$$v_{n,\mathrm{av}}^{rms} = \frac{1}{2} \cdot v_{n,\mathrm{eq}}^{rms}. \tag{6.13}$$

This may be obvious to the microwave engineer, but it comes as a surprise to the op-amp designer. There is no such distinction for op amps because they have a high-impedance input. If a data sheet quotes "Input Referred Noise" it is most likely available noise. Sometimes there is a note saying "output noise divided by low-frequency gain", then you know for sure that it is available noise! Equivalent noise is rarely quoted in data sheets.

Now with this out of the way, we can define the *Single-Ended Noise Figure*:

$$F = \left(\frac{v_{n,\mathrm{eq}}^{rms}}{v_{n,\mathrm{src}}^{rms}}\right)^2 = \frac{\overline{v_{n,\mathrm{eq}}^2}}{\overline{v_{n,\mathrm{src}}^2}} \tag{6.14}$$

where $v_{n,\mathrm{src}}^{rms}$ is the output noise due to the *source resistance* at a temperature of 290 K referred back to the input. For a 50 Ω source this is:

$$v_{n,\mathrm{src}}^{rms} = 0.8949\,\mathrm{nV}/\sqrt{\mathrm{Hz}} \cdot \sqrt{BW_n} \tag{6.15}$$

where BW_n is the noise bandwidth of the MA. In other words, noise figure is "equivalent input-referred noise" normalized to a convenient reference value, namely the thermal noise of the signal source resistance which is usually 50 Ω.[1] The use of noise figure eliminates the guesswork regarding equivalent vs. available noise: the noise figure is *always* based on equivalent noise. Noise figures are usually expressed in dBs using the conversion rule $10 \log F$.

Why do we say *single-ended* noise figure? Is there also a differential noise figure? Yes! The difference lies in the value of the source resistance. If we use a single-ended 50 Ω source as shown in Fig. 6.4, the source noise is that of a 50 Ω resistor, however, if we use a differential 50 Ω source then the source noise is that of *two* 50 Ω resistors thus the differential source outputs twice the noise power. For this reason the differential noise figure is 3 dB *lower* than the single-ended noise figure. If a data sheet quotes a noise figure without saying single-ended or differential, it is most likely the single-ended one, simply because most noise-figure test sets measure only the single-ended noise figure.

[1] An equivalent definition of the noise figure is: Input SNR divided by output SNR, where the input SNR contains only the thermal noise of the source resistance.

What about the output of the amplifier, do we also have to distinguish between a single-ended and a differential measurement? Fortunately there is almost no difference, because of the multistage topology typically used for MAs the noise signals at the two MA outputs are usually highly correlated, i.e., $v_{n,OP} = -v_{n,ON}$. If we measure differentially, we get twice the output noise and twice the gain, as a result the input referred noise remains the same. (Remember that this is not usually the case for TIAs, cf. Section 5.1.4.)

Power Penalty. Finally we are ready to calculate the power penalty due to the MA noise contribution. The total mean-square noise current referred back to the input of the TIA can be written as:

$$\overline{i_{n,\mathrm{amp}}^2} = \overline{i_{n,TIA}^2} + \overline{i_{n,MA}^2} \tag{6.16}$$

where $\overline{i_{n,TIA}^2}$ is the familiar input-referred noise of the TIA and $\overline{i_{n,MA}^2}$ is the noise of the MA referred all the way back to the *input of the TIA*. Thus by including the MA noise, the input-referred rms noise goes up from $\sqrt{\overline{i_{n,TIA}^2}}$ to $\sqrt{\overline{i_{n,\mathrm{amp}}^2}}$ and, according to Eq. (4.17), we need to increase the optical input power by the same amount to maintain the BER. Therefore the power penalty is:

$$PP = \sqrt{\frac{\overline{i_{n,\mathrm{amp}}^2}}{\overline{i_{n,TIA}^2}}} = \sqrt{1 + \frac{\overline{i_{n,MA}^2}}{\overline{i_{n,TIA}^2}}}. \tag{6.17}$$

The noise of the MA referred back to the TIA input can be expressed in terms of the MA noise figure F and the TIA transresistance R_T:

$$\overline{i_{n,MA}^2} = \frac{\overline{v_{n,\mathrm{av}}^2} - 1/4 \cdot \overline{v_{n,\mathrm{src}}^2}}{R_T^2} = \frac{\overline{v_{n,\mathrm{eq}}^2} - \overline{v_{n,\mathrm{src}}^2}}{4R_T^2} = \frac{(F-1) \cdot \overline{v_{n,\mathrm{src}}^2}}{4R_T^2}. \tag{6.18}$$

Here are the steps: First, we refer the available noise power of the MA back to the TIA input by dividing by R_T^2, but we have to be careful because the available noise contains some of the source noise which is already accounted for by the TIA noise. We have to subtract that out. Second, we use Eq. (6.13) to convert available noise to equivalent noise. Third, we apply Eq. (6.14) to convert equivalent noise to noise figure, and we are done! Now, for the grand finale, we can combine Eqs. (6.17) and (6.18) and solve for the maximum allowable single-ended noise figure:

$$F < 1 + (PP^2 - 1) \cdot \frac{4R_T^2 \cdot \overline{i_{n,TIA}^2}}{\overline{v_{n,\mathrm{src}}^2}}. \tag{6.19}$$

A TIA with high transimpedance helps relaxing the noise-figure specification for the MA. No big surprise.

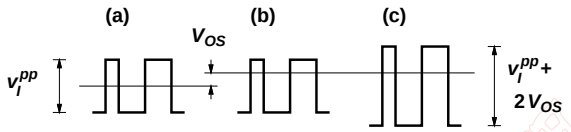

Figure 6.5: Effect of the LA offset voltage: (a) without offset, (b) with offset, (c) with offset and increased signal amplitude to restore the original bit-error rate.

Typical Values. Next we want to calculate some typical numbers based on a 0.2 dB power penalty ($PP = 1.047$) and a noise bandwidth $BW_n = 1.2 \cdot B$. For a typical 2.5 Gb/s system ($R_T = 1.5\,\mathrm{k\Omega}$, $i_{n,TIA}^{rms} = 400\,\mathrm{nA}$) we obtain the numerical value:

$$F < 1 + 0.096 \cdot \frac{4 \cdot (1.5\,\mathrm{k\Omega})^2 \cdot (400\,\mathrm{nA})^2}{(49.0\,\mu\mathrm{V})^2} = 17.7\,\mathrm{dB}. \tag{6.20}$$

And for a typical 10 Gb/s system ($R_T = 0.5\,\mathrm{k\Omega}$, $i_{n,TIA}^{rms} = 1200\,\mathrm{nA}$) we get:

$$F < 1 + 0.096 \cdot \frac{4 \cdot (0.5\,\mathrm{k\Omega})^2 \cdot (1200\,\mathrm{nA})^2}{(98.0\,\mu\mathrm{V})^2} = 11.9\,\mathrm{dB}. \tag{6.21}$$

From a wireless LNA designer's point of view these noise figures are very large, nevertheless a careful design is required to meet these noise figures in a *broadband* amplifier design.

6.2.4 Input Offset Voltage

An undesired offset voltage at the input of the LA causes the slicing level to be in a non-optimal position and as a result the receiver sensitivity is degraded. A small offset voltage in an AGC amplifier is less severe because it can be corrected at the output of the amplifier. In particular, if decision-point steering is used after the AGC amplifier, the offset voltage is eliminated automatically. Next we want to calculate the power penalty caused by an offset voltage of a LA. As we already know, this result will give us an idea of how much offset voltage can be tolerated.

Definition. The input-referred offset voltage V_{OS} is the input voltage for which the output voltage of the MA becomes zero. If the input-offset voltage is small, i.e., if the amplifier is not driven into limiting by the offset itself, we can also take the output offset voltage and divide it by the voltage gain A.

Power Penalty. In the presence of an offset voltage V_{OS} we need to increase the peak-to-peak value of the input signal v_I^{pp} by nearly $2V_{OS}$ to maintain the same bit error rate as without offset. The thought experiment leading up to this conclusion is illustrated in Fig. 6.5. It is assumed that the signal is sliced at the horizontal line. The power penalty comes out as:

$$PP = \frac{v_I^{pp} + 2V_{OS}}{v_I^{pp}} = 1 + \frac{2V_{OS}}{v_I^{pp}}. \tag{6.22}$$

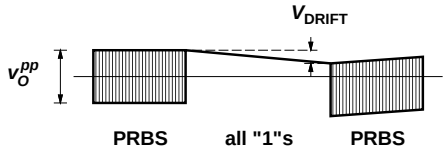

Figure 6.6: Effect of a low-frequency cutoff: The output signal drifts during a long string of ones and subsequently causes a slicing error.

This is, of course, the same result that we have already found in Eq. (4.49) when we first introduced the concept of power penalty. We see from this equation that the penalty is worst for small input signals. The smallest meaningful input signal to the MA, $v_I^{pp}(\mathrm{min})$, is the TIA output signal at the sensitivity limit. An expression for $v_I^{pp}(\mathrm{min})$ has already been derived in Eq. (6.2):

$$v_I^{pp}(\mathrm{min}) = 2\mathcal{Q} \cdot R_T \cdot i_{n,TIA}^{rms}. \tag{6.23}$$

Combining Eqs. (6.22) and (6.23) and solving for the maximum allowable offset voltage reveals:

$$V_{OS} < (PP - 1) \cdot \mathcal{Q} \cdot R_T \cdot i_{n,TIA}^{rms}. \tag{6.24}$$

As so many times before, a TIA with high transimpedance helps relaxing the MA specifications.

Typical Values. Next we want to calculate some typical numbers based on a 0.2 dB power penalty ($PP = 1.047$) and $BER = 10^{-12}$ ($\mathcal{Q} = 7$). For a typical 2.5 Gb/s system ($R_T = 1.5\,\mathrm{k\Omega}$, $i_{n,TIA}^{rms} = 400\,\mathrm{nA}$) we obtain the numerical value:

$$V_{OS} < 0.047 \cdot 7 \cdot 1.5\,\mathrm{k\Omega} \cdot 400\,\mathrm{nA} = 0.198\,\mathrm{mV}. \tag{6.25}$$

And for a typical 10 Gb/s system ($R_T = 0.5\,\mathrm{k\Omega}$, $i_{n,TIA}^{rms} = 1200\,\mathrm{nA}$) we get:

$$V_{OS} < 0.047 \cdot 7 \cdot 0.5\,\mathrm{k\Omega} \cdot 1.2\,\mu\mathrm{A} = 0.198\,\mathrm{mV}. \tag{6.26}$$

In conclusion, a good LA design has an offset voltage of around 0.1 mV or less. This is a fairly low offset voltage, even for bipolar implementations, which means that an offset compensation technique must be employed. More on that subject in Section 6.3.3.

6.2.5 Low-Frequency Cutoff

Many MAs have a *Low-Frequency Cutoff*, i.e., a frequency below which the gain drops off. This effect can be caused by AC coupling between the TIA and the MA or by some types of offset-compensation circuits (cf. Section 6.3.3).

Definition. The low-frequency cutoff f_{LF} is defined as the lower frequency at which the gain $A(f)$ has dropped by 3 dB below its passband value. See Fig. 6.23 for an illustration.

When amplifying a long string of zeros or ones, the output signal of the amplifier *drifts* and the first bit after the string is sliced with an offset error. Thus a low-frequency cutoff results in an effect similar to that of an offset voltage. Figure 6.6 illustrates this situation. The drift of the output signal can also produce data dependent jitter. Next we want to calculate the power penalty and then derive a limit for f_{LF}.

Power Penalty. What is the longest string of zeros or ones that we will encounter in a data stream? In SONET/SDH systems, which use scrambling as a line code, the run length is potentially unlimited. However, SONET/SDH equipment is tested with a particular bit sequence which puts the system under stress: the so-called "consecutive identical digit immunity measurement" [IT94]. This sequence consists of a long pseudo-random bit sequence (PRBS) with more than 2000 bits and 50 % mark density followed by 72 consecutive bits of zero; then again the PRBS followed by 72 bits of one. So it is reasonable to design a SONET/SDH system for a maximum run length $r = 72$. In Gigabit Ethernet systems, which use 8B10B encoding, the longest runs of zeros or ones are strictly limited to $r = 5$.

Assuming a linear system with a single-pole, high-pass transfer function we can calculate the drift caused by r consecutive zeros or ones:

$$V_{\text{DRIFT}} = \frac{v_O^{pp}}{2}\left[1 - \exp\left(-2\pi f_{LF} \cdot \frac{r}{B}\right)\right] \approx \frac{v_O^{pp}}{2} \cdot r \cdot \frac{2\pi f_{LF}}{B} \tag{6.27}$$

where B is the bit rate. The drift is approximately linear in time if the time constant $1/(2\pi f_{LF})$ is much larger than the drift time r/B. Similar to the offset voltage discussed in Section 6.2.4, the drift voltage causes the power penalty:

$$PP = 1 + \frac{2V_{\text{DRIFT}}}{v_O^{pp}}. \tag{6.28}$$

Combining Eqs. (6.27) and (6.28) yields:

$$PP = 1 + r \cdot \frac{2\pi f_{LF}}{B}. \tag{6.29}$$

Solving for the highest allowable LF-cutoff frequency reveals:

$$f_{LF} < (PP - 1) \cdot \frac{B}{2\pi \cdot r}. \tag{6.30}$$

Typical Values. Next we want to calculate some typical numbers based on a 0.2 dB power penalty ($PP = 1.047$). For a 2.5 Gb/s SONET system ($r = 72$) we obtain the numerical value:

$$f_{LF} < 0.047 \cdot \frac{2.5\,\text{Gb/s}}{6.28 \cdot 72} = 260\,\text{kHz}. \tag{6.31}$$

And for a 10 Gb/s SONET system we get:

$$f_{LF} < 0.047 \cdot \frac{10\,\text{Gb/s}}{6.28 \cdot 72} = 1.04\,\text{MHz}. \tag{6.32}$$

In conclusion, the low-frequency cutoff for a SONET system should be $f_{LF} < B/10,000$. It the case of a Gigabit Ethernet, Fiber Channel, etc. system with $r = 5$ the low-frequency cutoff specification is relaxed to $f_{LF} < B/700$, in accordance with [NCI00].

In practice the low-frequency cutoff is often set much lower than the numbers derived above (e.g. 2.5 kHz for 2.5 Gb/s and 25 kHz for 10 Gb/s). The cutoff frequency is usually set by an external capacitor, e.g., a coupling capacitor (DC block) or a capacitor part of an offset compensation circuit such as C in Fig. 6.21. In this case there is no cost involved in making this capacitor larger and reducing the power penalty below 0.2 dB.

6.2.6 Input Dynamic Range and Sensitivity

Definition. The *Input Dynamic Range* of the MA describes the minimum and maximum input signal for which the MA performs a useful function. The definition of the minimum signal (lower end of the dynamic range) is identical to that of the *Sensitivity*. In other words, the minimum input signal is that for which the required BER can just be met. Similar to the TIA sensitivity, the MA sensitivity v_S^{pp} is given by:

$$v_S^{pp} = 2\mathcal{Q} \cdot v_{n,\text{av}}^{rms} \tag{6.33}$$

where $v_{n,\text{av}}^{rms}$ is the available rms noise voltage of the MA (cf. Section 6.2.3).

The definition of the maximum signal (upper end of the dynamic range) depends on the type of amplifier. An AGC amplifier is supposed to operate linearly and therefore the maximum input-signal amplitude is reached when the amplifier starts to limit or distort. A commonly used measure is the 1-dB compression point, the amplitude at which the gain has dropped by 1 dB below its small-signal value. In analog applications (HFC/CATV) harmonic distortions and intermodulation products are significant hence the maximum signal is defined such that these distortions remain small (cf. Section 4.8). Obviously, those definitions cannot be applied to LAs since they are operated in the nonlinear limiting regime on purpose. The maximum signal for a LA is somewhat vaguely defined as the signal at which pulse-width distortions and jitter become unacceptable. For example in BJT implementations, a large input signal can cause the collector-base diodes to become forward biased leading up to such distortions [GS99].

Next we want to derive the power penalty caused by a finite MA sensitivity. This result will provide us with a rule for the required sensitivity.

Power Penalty. This one is easy! We already know the power penalty due to MA noise from Eqs. (6.17) and (6.18) and we know that the sensitivity is a function of this noise from Eq. (6.33). So all we have to do is put these equations together and solve for the MA sensitivity:

$$v_S^{pp} < 2\mathcal{Q} \cdot \sqrt{(PP^2 - 1) \cdot R_T^2 \cdot \overline{i_{n,TIA}^2} + 1/4 \cdot \overline{v_{n,\text{src}}^2}} \approx \sqrt{PP^2 - 1} \cdot 2\mathcal{Q} \cdot R_T \cdot i_{n,TIA}^{rms}. \tag{6.34}$$

In the approximation we neglected the 50-Ω noise component $1/4 \cdot \overline{v_{n,\text{src}}^2}$. The approximate result has an interesting interpretation: the term after the square root is just the TIA output signal at the sensitivity limit (Eq. 6.2). Thus the square-root term tells us how

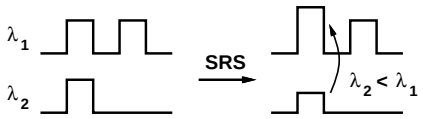

Figure 6.7: Amplitude modulation in a WDM system caused by SRS.

much smaller we have to make the MA sensitivity. For example for a 0.2 dB power penalty ($PP = 1.047$) the MA sensitivity must be made 3.2× lower than the output signal from the TIA at the sensitivity limit.

Typical Values. Next we want to calculate some typical MA sensitivity numbers based on a 0.2 dB power penalty ($PP = 1.047$) and a noise bandwidth $BW_n = 1.2 \cdot B$. For a typical 2.5 Gb/s system ($R_T = 1.5\,\mathrm{k\Omega}$, $i_{n,TIA}^{rms} = 400\,\mathrm{nA}$) we obtain the numerical value:

$$v_S^{pp} < 14 \cdot \sqrt{0.096 \cdot (1.5\,\mathrm{k\Omega})^2 \cdot (400\,\mathrm{nA})^2 + 1/4 \cdot (49.0\,\mu\mathrm{V})^2} = 2.6\,\mathrm{mV}. \tag{6.35}$$

And for a typical 10 Gb/s system ($R_T = 0.5\,\mathrm{k\Omega}$, $i_{n,TIA}^{rms} = 1200\,\mathrm{nA}$) we get:

$$v_S^{pp} < 14 \cdot \sqrt{0.096 \cdot (0.5\,\mathrm{k\Omega})^2 \cdot (1200\,\mathrm{nA})^2 + 1/4 \cdot (98.0\,\mu\mathrm{V})^2} = 2.7\,\mathrm{mV}. \tag{6.36}$$

The requirements for the upper end of the dynamic range depends on the maximum output signal from the TIA and can be several volts. A typical input dynamic range for a MA is:

$$v_I = 2\,\mathrm{mV_{pp}} \ldots 2\,\mathrm{V_{pp}}. \tag{6.37}$$

6.2.7 AM-to-PM Conversion

The received optical signal sometimes contains rapid amplitude variations which may cause phase or delay variations in the MA. Rapid delay variations in the MA result in jitter in the output signal which interferes with the clock and data recovery process. Therefore it is necessary to limit the conversion from *Amplitude Modulation* (AM) to *Phase Modulation* (PM) in the MA.

The amplitude modulation is caused by a variety of effects in the optical fiber. For example, the combination of *Self-Phase Modulation* (SPM) and chromatic dispersion causes intensity overshoots at the beginning and end of the optical pulses. Furthermore, *Stimulated Raman Scattering* (SRS) in a system that carries multiple wavelength in a single fiber (WDM system) causes amplitude modulation as well. The SRS effect transfers optical energy from channels with shorter wavelengths to channels with longer wavelengths. Thus a transmitted one bit on one channel that is accompanied by a one on a shorter-wavelength channel grows in amplitude as opposed to a one that is accompanied by a zero (see Fig. 6.7). For more information on SPM and SRS see [Agr97, RS98].

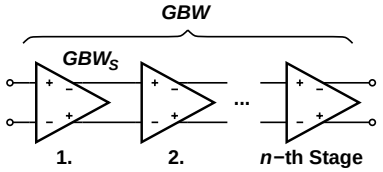

Figure 6.8: An n-stage amplifier with overall gain-bandwidth product GBW and stage gain-bandwidth product GBW_S.

Definition. AM-to-PM conversion of a MA is usually specified by the maximum delay variation, $\Delta\tau_{AM}$, observed when varying the input-signal amplitude over the entire dynamic range. Although the signal amplitude is not expected to vary that much in practice, this delay variation serves as an upper bound for AM-to-PM induced jitter. A LA may exhibit a significant delay variation when it transitions from the linear regime to the limiting regime.

Typical Values. In practice the delay variation due to AM-to-PM conversion is often limited to about ±10 % of the bit period or ±0.1 UI. This corresponds to:

$$\text{2.5 Gb/s MA:} \quad |\Delta\tau_{AM}| < 40\,\mathrm{ps} \tag{6.38}$$

$$\text{10 Gb/s MA:} \quad |\Delta\tau_{AM}| < 10\,\mathrm{ps}. \tag{6.39}$$

6.3 MA Circuit Principles

In the following section, we will have a look at the design principles for MAs: How can we get a high gain-bandwidth product, low offset voltage, low cutoff frequency, automatic gain control, and so on? Then in the next section we will examine concrete circuit implementations to illustrate these principles.

6.3.1 Multistage Amplifier

From the discussion of MA specifications we know that the *Gain-Bandwidth Product* (GBW) requirements for MAs are astounding. For example, a 2.5 Gb/s MA requires 30 dB gain and 3 GHz bandwidth which results in a GBW of nearly 100 GHz, a corresponding 10 Gb/s MA requires a GBW of nearly 400 GHz. The GBW required is much higher than the f_T of the technology we would like to realize the MA in. Is it possible to build an amplifier with $GBW \gg f_T$? Yes, if we use a multistage design.

Contrary to an op-amp design, which needs to be stable under unity feedback conditions, we do not require a single dominant pole for a MA. This permits us to cascade multiple stages, as shown in Fig. 6.8, and boost the gain-bandwidth product (GBW) way beyond that of a single stage (GBW_S). How does this work? To get an intuition for this mechanism, let's start with a simple example. Let's assume that all stages are

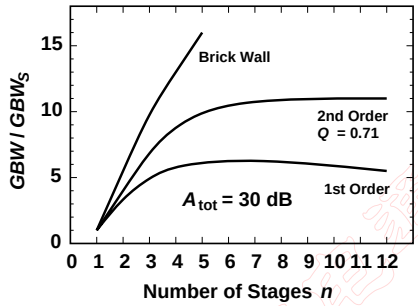

Figure 6.9: Gain-bandwidth extension as a function of the number of stages.

identical and have a brick-wall, low-pass frequency response with a bandwidth of 3 GHz. As a result, the total amplifier bandwidth is also 3 GHz. Our goal is to build an amplifier with a total gain $A_{\text{tot}} = 30\,\text{dB}$ (31.6×). Thus our total gain-bandwidth product is $GBW = 31.6 \times 3\,\text{GHz} = 95\,\text{GHz}$. Now let's consider two possible designs:

- Single-stage architecture ($n = 1$). In this case the GBW of the stage is equal to the GBW of the total amplifier $GBW_S = GBW = 95\,\text{GHz}$.
- Three-stage architecture ($n = 3$). In this case each stage needs only a gain of 10 dB (3.16×) and thus the gain-bandwidth per stage is $GBW_S = 3.16 \times 3\,\text{GHz} = 9.5\,\text{GHz}$.

In conclusion, the 3-stage design requires 10× less GBW per stage! We could also say that cascading 3 stages gave us a *Gain-Bandwidth Extension* GBW/GBW_S of 10×. How far can we push this? Could we build our amplifier from stages with $GBW_S = 0.95\,\text{GHz}$? No! As we use more and more stages the GBW requirement per stage is reduced, but even with infinitely many stages we still need a stage gain of slightly more than 1.0 and thus a minimum GBW per stage of slightly more than 3 GHz is required to build the amplifier in our example.

We can easily generalize from the above example and write the GBW extension as a function of the number of stages, n:

$$\frac{GBW}{GBW_S} = A_{\text{tot}}^{1-1/n} < A_{\text{tot}}. \tag{6.40}$$

This function is plotted in Fig 6.9 and labeled "Brick Wall". We can see from Eq. (6.40) that the maximum GBW extension that can be achieved with a multistage architecture is limited by A_{tot}.

In real amplifiers we don't have stages with a brick-wall frequency response, the stages are more likely to have a 1st or 2nd order response: A simple transistor stage has a first-order roll-off, while a stage with local feedback or inductive load has a second-order response. Curves for these two cases are also plotted in Fig. 6.9. All curves are computed for a total amplifier gain $A_{\text{tot}} = 30\,\text{dB}$. The GBW extension is less dramatic than in

the brick-wall case, but we can still achieve a respectable 6× or 10× improvement. The GBW extension is lower because the total amplifier bandwidth *shrinks* as we cascade more and more stages with a slow roll-off. We can see from the plots that there is an optimal number of stages beyond which GBW/GBW_S decreases.

The mathematical expression for the GBW extension in the 1st-order case is well known [Jin87, Feu90]:

$$\frac{GBW}{GBW_S} = A_{\text{tot}}^{1-1/n} \cdot \sqrt{2^{1/n} - 1}. \tag{6.41}$$

The first term is the same as in the brick-wall case Eq. (6.40) and the second term is due to the bandwidth shrinkage. The optimum number of stages turns out to be $n_{\text{opt}} \approx 2 \ln A_{\text{tot}}$. The optimum stage gain follows as $A_S = \sqrt{e} = 4.34\,\text{dB}$. A similar expression can be found for 2nd-order Butterworth stages ($Q = 1/\sqrt{2}$, no zeros) often used in practice:

$$\frac{GBW}{GBW_S} = A_{\text{tot}}^{1-1/n} \cdot \sqrt[4]{2^{1/n} - 1}, \tag{6.42}$$

and the optimum number of stages is $n_{\text{opt}} \approx 4 \ln A_{\text{tot}}$. The optimum stage gain follows as $A_S = \sqrt[4]{e} = 2.17\,\text{dB}$.

Now we understand that a 2.5 Gb/s MA with a GBW of nearly 100 GHz can be built from stages with a GBW of only around 12 GHz. This latter GBW is well below the f_T of, for example, a 0.25 μm CMOS technology ($f_T \approx 25\,\text{GHz}$) and therefore it is possible to implement a 2.5 Gb/s MA in this technology [SF00].

6.3.2 Techniques for Broadband Stages

Although the use of a multistage topology greatly relaxes the gain-bandwidth requirement per stage, we still need to make each stage as fast as possible, with a GBW approaching the technologies' f_T, to meet the specifications of multi-gigabit parts. In this section we review the most important techniques to build broadband amplifier stages. These techniques are also applicable to the feedback and post amplifiers of a TIA.

Use Fast Transistors. The first step towards a fast amplifier is to choose fast transistors. The speed of a transistor is usually measured by f_T, the frequency where the current gain becomes unity, and f_{max}, the frequency where the power gain becomes unity, a.k.a. the maximum frequency of oscillation. The speed optimizations techniques for bipolar and FET transistors are quite different and we will discuss them one after the other.

The f_T of a bipolar transistor (BJT or HBT) is given by the following expression:

$$f_T = \frac{1}{2\pi} \cdot \frac{g_m}{C_{be} + C_{bc}} = \frac{1}{2\pi} \cdot \frac{1}{\tau_F + (C_{je} + C_{jc}) \cdot V_T / I_C} \tag{6.43}$$

where τ_F is the transit time, C_{je} and C_{jc} are parasitic junction capacitances, V_T is the thermal voltage ($kT/q \approx 25\,\text{mV}$), and I_C is the collector current. To obtain a fast bipolar transistor the emitter area A_E must be chosen carefully because. The smaller the emitter area, the smaller is the transistor and with it the parasitic junction capacitances C_{je} and

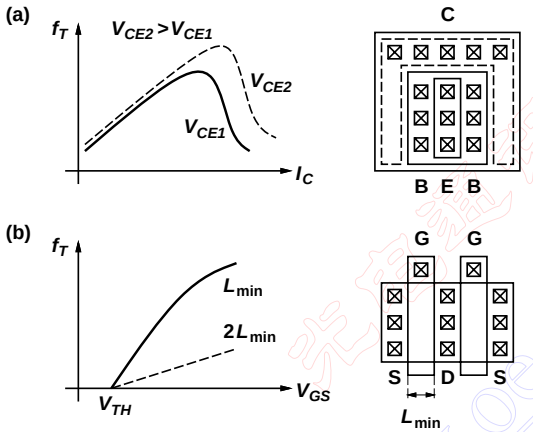

Figure 6.10: f_T and layout of (a) BJT (b) MOSFET transistor.

C_{jc}. However, a small emitter size also leads to a high collector current density[2] I_C/A_E. If this density exceeds a critical value ($1-2\,\mathrm{mA}/\mu\mathrm{m}^2$), the transit time τ_F increases rapidly due to an extension of the base region into the collector region known as *Base Pushout* or *Kirk Effect*. The critical current density at which the Kirk effect kicks in, increases with the collector-emitter voltage V_{CE}. For this reason a large V_{CE} is desirable for high-speed transistors. See Fig. 6.10(a) for a graphical illustration of the function $f_T(I_C, V_{CE})$.

The $f_{\max}$ of a bipolar transistor (BJT or HBT) is given by:

$$f_{\max} = \frac{1}{2} \cdot \sqrt{\frac{f_T}{2\pi \cdot R_b C_{bc}}}. \tag{6.44}$$

To obtain a high $f_{\max}$ the base spreading resistance R_b must be minimized. For high-frequency transistors, a narrow stripe geometry with two base contact stripes on each side of the emitter stripe is recommended (see Fig. 6.10(a)). For more information on high-speed bipolar transistors see [RM96, Gre84].

The f_T of a FET (MESFET, HFET, or MOSFET) in saturation, assuming low electric fields and the quadratic model, is approximated by the following expression:

$$f_T = \frac{1}{2\pi} \cdot \frac{g_m}{C_{gs}+C_{gd}} \approx \frac{3}{4\pi} \cdot \frac{\mu_n}{L^2} \cdot (V_{GS} - V_{TH}) \tag{6.45}$$

where V_{GS} is the gate-source voltage, V_{TH} is the threshold voltage, L is the channel length, and μ_n is the carrier mobility. See Fig. 6.10(b) for a graphical illustration of the function

[2] More precisely, the collector current density is I_C/A_C with the collector area $A_C = (W_E + 2\gamma) \cdot L_E \approx A_E$ where W_E is the emitter width, L_E the emitter length, and γ the current spreading from the emitter into the collector. The current spreading effect can be significant in submicron technologies, e.g., for $W_E = 0.3\,\mu\mathrm{m}$ and $\gamma = 0.2\,\mu\mathrm{m}$ A_C is 2.3× larger than A_E.

$f_T(V_{GS}, L)$. For the quadratic FET model, f_T increases proportional to the gate overdrive voltage $V_{GS}-V_{TH}$. However, in short-channel MOSFETs carrier velocity saturation causes the curve to flatten out as shown in Fig. 6.10(b) for $L_{\min}$. In MESFETs and HFETs the curve typically reaches a *maximum value* after which it declines due to the turn on of the Schottky diode at the gate and other effects. To obtain a fast FET the smallest channel length available in the target technology must be chosen. Furthermore, the gate overdrive voltage should be chosen either as large as possible, considering headroom limitations (for MOSFETs) or set to the value where f_T is maximum (for MESFETs and HFETs). A typical overdrive voltage is around 400 mV.

The $f_{\max}$ of a FET (MESFET, HFET, or MOSFET) is given by:

$$f_{\max} = \frac{1}{2} \cdot \sqrt{\frac{f_T}{2\pi \cdot R_g C_{gd}}}. \tag{6.46}$$

To obtain a high $f_{\max}$ is is necessary to keep the gate resistance R_g low. This is achieved by breaking wide transistors, e.g. if $W > 6\,\mu\mathrm{m}$, into several smaller, parallel transistors. The result is a finger-structure layout as shown in Fig. 6.10(b). Contacting the gate fingers on both sides further reduces the resistance. With these techniques the gate resistance can be made very small such that $f_{\max}$ becomes larger than f_T.

The f_T of deep sub-micron CMOS transistors is comparable to that of fast bipolar transistors. Does that mean that we can build circuits operating at about the same speed in either technology? No, bipolar circuits have an advantage over FET circuits even for identical f_T and $f_{\max}$ parameters. To understand this we have to analyze how f_T degrades if we take parasitic capacitances such as wiring and junction capacitances into account. If we add a parasitic capacitance C_P at the gate or base and calculate the f'_T of the joint system (transistor plus parasitic capacitance) we find for either FET or bipolar transistor:

$$f'_T = \frac{1}{1/f_T + 2\pi \cdot C_P/g_m}. \tag{6.47}$$

Now we see that the transistor with the higher g_m is more resilient to f_T degradation. For bipolar transistors $g_m = I_C/V_T$ which is quite large even for modest collector currents. For FETs $g_m = 2I_D/(V_{GS} - V_{TH})$ which is typically around 8× smaller than that of a bipolar transistor at the same bias current: $V_T \approx 25\,\mathrm{mV}$ while $(V_{GS} - V_{TH})/2 \approx 200\,\mathrm{mV}$.

Boost f_T. The fact that f_T is related to the carrier transit time may suggest that f_T is a fundamental property of the underlying technology. However, this is not the case and it is indeed possible to increase f_T by means of circuit techniques. To understand this we have to realize that f_T is determined by the ratio of the transconductance g_m to the input capacitance under shorted-output conditions (cf. left side of Eqs. (6.43) and (6.45)). Thus if we can come up with a circuit that lowers the input capacitance while maintaining the same transconductance we can increase f_T. The so-called *f_T-Doubler* circuit shown in Fig. 6.11(a) does exactly this [Feu90]. The input signal (B-E) is split into two signals each with half the voltage (B-x and x-E), then each half is amplified with a separate transistor (Q_1 and Q_2), and the output currents are combined at the collector. The overall transconductance is the same as that of a single transistor, but

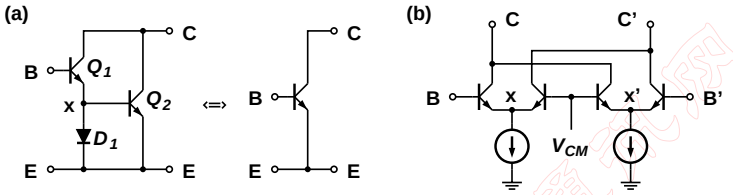

Figure 6.11: The f_T-doubler: (a) principle of operation, (b) practical differential implementation.

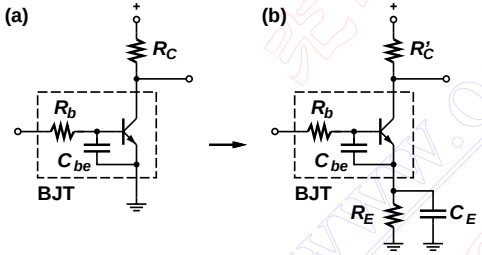

Figure 6.12: BJT gain-stage (a) without and (b) with series feedback.

the base-emitter capacitance C_{be} is cut in half due to the Miller effect caused by the signal at the emitter (node x) which is half the input voltage. Thus f_T of the compound structure is approximately doubled. Figure 6.11(b) shows a more practical differential implementation of an f_T doubler [WHKY98]. The differential input signal between B and B' is split in two halves and amplified by separate differential pairs. The bias voltage V_{CM} is set to the input common-mode voltage. The differential output currents from both pairs are combined and therefore the overall transconductance is the same as that of a simple differential pair. But unlike a simple differential pair, the nodes x and x' are not virtual grounds but carry about half the AC signals of nodes B and B', respectively. Thus the input capacitance is approximately halved and f_T is approximately doubled.

In practice f_T is not exactly doubled because only C_{be} and not C_{bc} is reduced by the feedback mechanism. A drawback of the f_T doubler is that the collector-substrate capacitance is doubled because of the two transistors. Since the latter capacitance is much smaller than the base-emitter capacitance, which is halved, the doubler still works fine when incorporated into a larger circuit. Although it is possible to build a corresponding f_T-doubler circuit with MOSFETs it usually does not provide an advantage once put into a larger circuit. The reason is that the gate-source capacitance, which is halved, is very similar in magnitude to the drain-bulk capacitance which is doubled.

Speed up the BJT Input Pole. A major speed bottleneck in BJTs is the input pole formed by the base resistance R_b and the base-emitter capacitance C_{be}. In BJTs the base is lightly doped which leads to a relatively high R_b. In HBTs (e.g., SiGe technology) the base is more heavily doped and the input pole is less of a problem (cf. Appendix **??**). The base-emitter capacitance is bias dependent and increases with collector current: $C_{be} = I_C/V_T \cdot \tau_F + C_{je}$. For example, a silicon BJT transistor in a 0.25 μm modular BiCMOS technology at $I_C = 1\,\text{mA}$ has $R_b = 170\,\Omega$ and $C_{be} = 170\,\text{fF}$. Thus, the low-pass filter formed by R_b and C_{be} has a bandwidth of 5.5 GHz, much lower than the transistor's $f_T \approx 30\,\text{GHz}$.

A well-known technique to speed up the input pole is to apply *Series Feedback* [CH63]. The implementation of this technique with an emitter degeneration resistor R_E is illustrated in Fig. 6.12. Since the emitter voltage is now approximately following the base voltage the current into C_{be} is reduced. Ideally, for $R_E \to \infty$ the emitter voltage follows the base voltage exactly and the effect of C_{be} is completely suppressed. It can be shown that series feedback reduces the effective input capacitance by $1 + g_m R_E$ and thus speeds up the input pole by the same amount:

$$\frac{1}{2\pi} \cdot \frac{1}{R_b C_{be}} \to \frac{1}{2\pi} \cdot \frac{1 + g_m R_E}{R_b C_{be}}. \tag{6.48}$$

As a side effect of adding the emitter degeneration, the DC gain drops but it can be restored to its original value by increasing the collector resistor from R_C to R'_C. The resistor R_E also introduces a new high-frequency pole (emitter pole) in the transfer function. This undesired pole can be cancelled with a zero by adding an emitter capacitor with the value $C_E = 1/(2\pi f_T \cdot R_E)$. Continuing the BiCMOS example, we now add an emitter resistor $R_E = 100\,\Omega$. As a result, the input pole is sped up by a factor 5× to 27.5 GHz and is now comparable to the transistor's f_T. To compensate the emitter pole we need to add an emitter capacitor $C_E = 53\,\text{fF}$.

A variation of this technique, called *Emitter Peaking*, uses the zero introduced by C_E to cancel the output pole of the stage rather than the emitter pole. In this case the emitter capacitor is made much larger than before. Emitter peaking can be used to boost the amplifier bandwidth, but undesirable peaking occurs if C_E is made too large.

Besides speeding up the input pole, the emitter degeneration resistor R_E also improves the input dynamic range. For a bipolar differential stage without emitter degeneration an input voltage of just a few temperature voltages, e.g., $3 \cdot V_T \approx 75\,\text{mV}$, causes the output signal to saturate. By inserting emitter resistors this range can be greatly enhanced.

In FET circuits the input pole usually presents no speed limitation and source degeneration to speed up this pole is not necessary. However, there are other uses for a source degeneration such as controlling the stage gain, linearizing the stage response, and providing peaking in the frequency response.

Reduce the Capacitive Load: Buffering and Scaling. After boosting the input pole of the stage, the output pole is now limiting the bandwidth. The output pole is determined by the output resistance ($\approx R_C$ in Fig. 6.12) and the total load capacitance of the stage. The latter consists of three components: the stage self loading (C_O), the

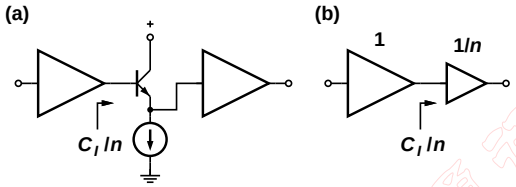

Figure 6.13: (a) Stages with interstage buffer and (b) inversely scaled stages.

interconnect capacitance, and the next-stage loading (C_I). For simplicity, we will ignore the interconnect capacitance in the following analysis.

The next-stage loading can be reduced by inserting a buffer in between the stages. Figure 6.13(a) shows two stages with an emitter-follower buffer in between them. If we assume that the buffer has an input capacitance that is n times smaller than the load it is driving, the next-stage loading is reduced from C_I to

$$C_I' = C_I/n \tag{6.49}$$

thus speeding up the output pole of the first stage as follows:

$$\frac{1}{2\pi}\cdot\frac{1}{R_C(C_I+C_O)} \rightarrow \frac{1}{2\pi}\cdot\frac{1}{R_C(C_I/n+C_O)}. \tag{6.50}$$

Note that the buffer acts as a *Capacitance Transformer*: the buffer's input capacitance is a fraction of the load capacitance, while the input and output voltages are the same.

In practice, the bandwidth gained by this technique is partially offset by bandwidth lost due to the finite buffer bandwidth, parasitic buffer capacitances, and signal attenuation in the buffer. In particular, there is a trade-off between the capacitance transformation ratio n and the buffer bandwidth: a buffer with a high n has a low bandwidth while a wideband buffer has an n close to unity. For example, the capacitance transformation ratio of a MOSFET buffer stage with bandwidth BW can be approximated by: $n \approx \alpha\cdot(f_T/BW-\beta)$ where α and β are constants which depend on the buffer topology (source follower or common source) and the technology [SF00]. Typically n reduces to unity for a bandwidth around $0.4-0.7\cdot f_T$. The use of interstage buffers is very popular in BJT designs but it is less effective in MOSFET designs because of the considerable attenuation and level shifting inherent to a MOSEFET source follower.

Alternatively, the next-stage loading C_I can be reduced by scaling down the transistor sizes and currents in the next stage [SF00]. Figure 6.13(b) shows two stages where the driven stage is made n times smaller than the driving stage thus reducing the next-stage loading to C_I/n. For example, when scaling all transistor widths in a MOSFET stage by $1/n$ and all resistors by n, the capacitances and currents are reduced by n while the node voltages, gain, and bandwidth remain mostly unaffected. The limitations of this technique lie in the large input capacitance of the first stage and/or the limited drive capability of the last stage in a multistage amplifier.

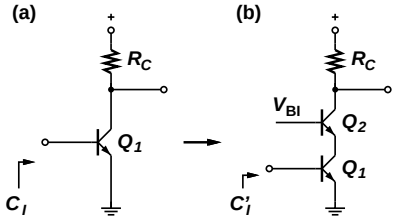

Figure 6.14: Stage (a) without and (b) with cascode transistor.

Suppress Capacitances with Cascode. The input capacitance C_I is comprised of two components. For a BJT stage as shown in Fig. 6.14(a), these components are (i) the base-emitter capacitance C_{be1} and (ii) the *Miller Capacitance* which is the base-collector capacitance C_{bc1} multiplied by the voltage gain plus one:

$$C_I = C_{be1} + (|A|+1)\cdot C_{bc1}. \tag{6.51}$$

When discussing the input pole in Fig. 6.12 we considered only the first component, but depending on the stage gain A, the Miller component can become quite large and may even dominate the first one. Of course, a corresponding relationship applies to FETs as well.

The addition of a *Cascode Transistor* as shown in Fig. 6.14(b) reduces the voltage gain of the input transistor to about unity ($\approx g_{m1}/g_{m2}$) and thus the input capacitance is reduced to:

$$C_I' = C_{be1} + 2\cdot C_{bc1}. \tag{6.52}$$

The stage gain A, given by the transconductance of the input transistor Q_1 (g_{m1}) and the collector resistor R_C, remains mostly unaffected. Obviously, this technique is most effective for stages with a high gain. On the down side, the cascode transistor introduces an additional pole, which may offset some of the bandwidth gained from reducing the input capacitance. Also the cascode transistor reduces the voltage headroom which may be a concern in low-voltage designs.

Suppress Capacitances with TIA load. Another method, which has been popularized by Cherry and Hooper [CH63], speeds up the output pole and reduces the input capacitance as well. Figure 6.15 shows the principle: the load resistor R_C is replaced by a transimpedance amplifier (Q_2, Q_3) with a transresistance equal to R_C, i.e., $R_F = (|A|+1)/|A|\cdot R_C$ where A is the gain of the TIA feedback amplifier. Note that under this condition both stages in Fig 6.15(a) and (b) have the same DC gain. We easily recognize the TIA circuit in Fig. 6.15(b) which has been discussed in the previous chapter. The role of the photodiode is now played by the input transistor Q_1. We know that the TIA circuit presents the low input resistance $R_F/(|A|+1) = R_C/|A|$ to node x. Thus compared to the original stage with a load resistor, the output pole (at node x) is sped up approximately by A.

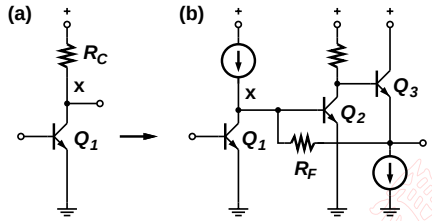

Figure 6.15: Stage (a) with resistor load and (b) TIA load.

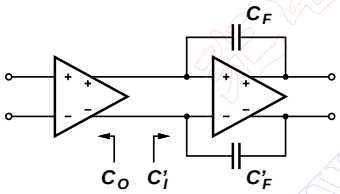

Figure 6.16: Cancelling parasitic capacitances with a negative capacitance.

A more detailed analysis of this stage can be obtained by realizing that its voltage gain is given by $A(s) = g_{m1} \cdot Z_T(s)$ where g_{m1} is the transconductance of Q_1 and $Z_T(s)$ is the transimpedance of the TIA load. Fortunately, the latter has already been calculated in Eqs. (5.16) – (5.19) for the case of a single-pole feedback amplifier (Q_2, Q_3). From these equations we conclude that the stage gain must have a conjugate-complex pole pair with the possibility of undesired peaking but also with the advantage of a fast 2nd-order roll-off which helps to reduce bandwidth shrinkage in multistage amplifiers. Using Eq. (5.21) we can estimate the bandwidth improvement over a stage with resistive load:

$$\frac{1}{2\pi} \cdot \frac{1}{R_C(C_I + C_O)} \rightarrow \frac{1}{2\pi} \cdot \frac{\sqrt{2A(A+1)}}{R_C(C_I + C_O)}. \tag{6.53}$$

Like the cascode transistor Q_2 in Fig. 6.14(b), the TIA load significantly reduces the voltage gain of Q_1. As a result, the Miller component of the input capacitance is reduced too which helps the bandwidth of the previous stage.

A disadvantage of this topology is its higher power dissipation due to the extra TIA in each stage. And of course, just like pointed out in Section 5.2.2, the feedback amplifier must have enough bandwidth to avoid peaking.

Cancel Capacitances. If we had a *negative* capacitor with the value $-(C_I + C_O)$, we could connect it to the output of each stage and compensate both the self-loading C_O and the next-stage loading C_I. This would completely eliminate the output pole! But what is a negative capacitor anyway? It is a capacitor whose voltage *drops* when we try

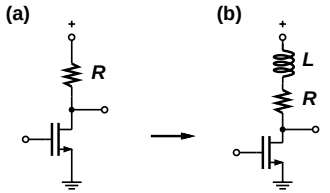

Figure 6.17: Stage with (a) a resistive load and (b) an inductive load.

to charge it *up*! In fact, the Miller mechanisms, which usually hurts us with additional undesired capacitance, can be used to generate a negative capacitance. We have to use *positive* feedback as shown in Fig. 6.16. The input capacitance of the second stage is

$$C_I' = C_I + (-|A| + 1) \cdot C_F \tag{6.54}$$

where C_I is the input capacitance of the stage without the feedback capacitors C_F and C_F'. We can see from Eq. (6.54) that for $|A| = 1$, C_F has no effect. This makes sense because $|A| = 1$ means that the output voltage exactly follows the input voltage (i.e., we have a buffer) and thus there is no voltage drop across C_F and no current flowing through it. If we increase the gain to $|A| > 1$, then the input capacitance is reduced. If we chose C_F such that

$$C_I + C_O + (-|A| + 1) \cdot C_F = 0, \tag{6.55}$$

the output pole of the first stage is eliminated.

In the case where C_F is made equal to C_{bc}, this technique is known as *Neutralization*. The negative capacitance $(-|A| + 1) \cdot C_F$ then cancels most of the Miller capacitance $(|A| + 1) \cdot C_{bc}$, part of C_I, but not the other capacitances.

The limitations of this technique lie in the fact that the input and output signals of the second stage are not exactly in phase at high frequencies and thus the input impedance caused by C_F is not a pure negative capacitance. Furthermore, if C_F is chosen too large the overall load capacitance becomes negative making the amplifier unstable. One way to address the latter issue is to use matched dummy transistors to implement C_F [Gre84].

Tune out Capacitances with Inductors: Shunt Peaking. Figure 6.17(b) shows a gain stage with an inductive load composed of R and L. The inductor L serves to "tune out" part of the load capacitance $C_I + C_O$. If the inductor value is chosen according to

$$L = 0.4 \cdot R^2 (C_I + C_O), \tag{6.56}$$

the bandwidth of the stage is extended by about 70 % without causing undesired peaking in the amplitude response as illustrated in Fig. 6.18(a). In addition, shunt-peaking improves the roll-off characteristics which helps to reduce bandwidth shrinkage in multistage amplifiers.

An intuitive way to visualize how this is happening is as follows: Just at the frequency when the gain would ordinarily roll off due to the output pole, the load impedance starts

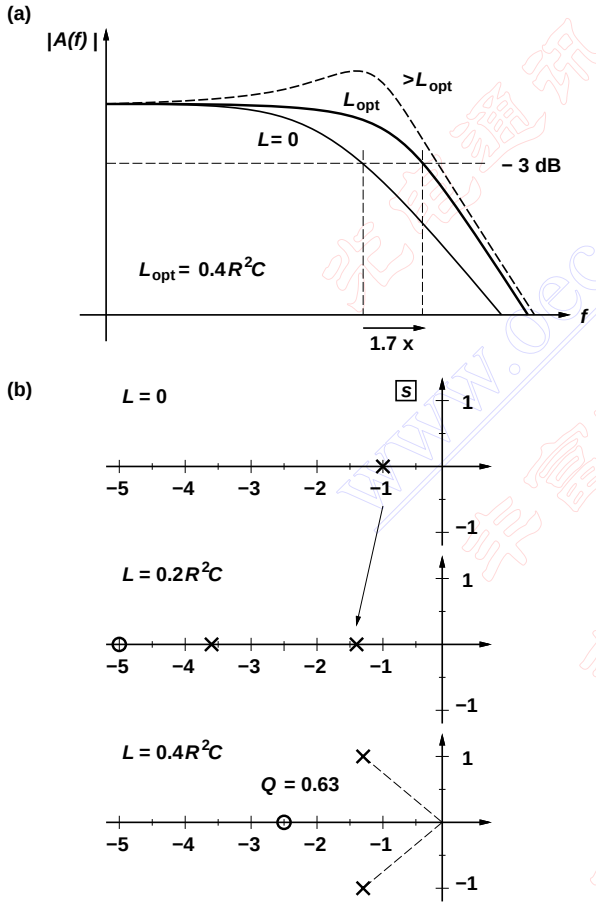

Figure 6.18: (a) Bode plot and (b) root-locus plot showing the effect of shunt-peaking inductor L.

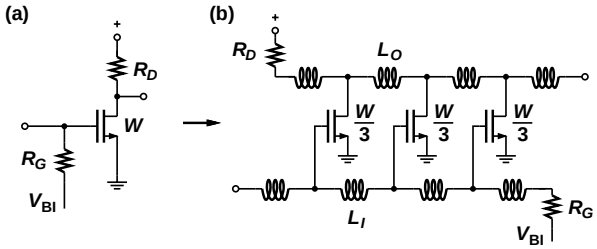

Figure 6.19: (a) Lumped amplifier and (b) distributed amplifier with inductors.

to go up due to the inductive component ($Z(s) = sL$). The increased load impedance boosts the gain and thus compensates for the gain roll-off. A more scientific approach is to study the pole/zero movement as a function of the inductor value as shown in Fig. 6.18(b). If the inductor is zero there is only one real pole (the output pole). For a small inductor, a pole/zero pair appears at high frequencies. When increasing the inductor further, the amplifier pole moves up (bandwidth extension!) and the new pole moves down until they eventually "collide" and move into conjugate-complex positions. The maximally flat amplitude response is reached for the pole quality factor $Q = 0.63$. In addition, the zero helps to extend the bandwidth further. More information on this method, which is known as *Shunt Peaking*, can be found in [Lee98].

The inductor can be realized with an on-chip spiral inductor, a bond wire inductor, or an active inductor (see Section 6.4.3 for an example). The bandwidth extension achieved in practice is usually less than the theoretical 70 %, either limited by the (spiral) inductor's self resonance or the maximum operating frequency of the active inductor ($\approx f_T/2$).

Tune out Capacitances with Inductors: Distributed Amplifier. The most aggressive broadband technique is the so-called *Distributed Amplifier*. The principle is illustrated in Figure 6.19. On the left side we have an ordinary lumped MOSFET amplifier stage with a bias resistor R_G at the gate and a load resistor R_D at the drain. On the right, the same transistor has been split into three smaller ones, each one 1/3 of the original size, and connected together with inductors L_I and L_O. Obviously, these two stages behave the same way at low frequencies where L_I and L_O behave like shorts. But there is an important difference at high frequencies: All parasitic gate and drain capacitances are absorbed into two artificial (discrete) transmission lines, one at the input and one at the output of the stage. Now the amplifier bandwidth is determined by the cutoff frequency of these transmission lines.

Clearly this technique can be generalized from three sections in the example above to n sections. Let's calculate the bandwidth for the case of a stage with n sections. The characteristic impedance of the output transmission line, Z_{TLO}, below the cutoff frequency is determined by the inductors L_O and the parasitic drain capacitances C_O/n (remember, each transistor is $1/n$-th the size of the original transistor). It can be approximated by

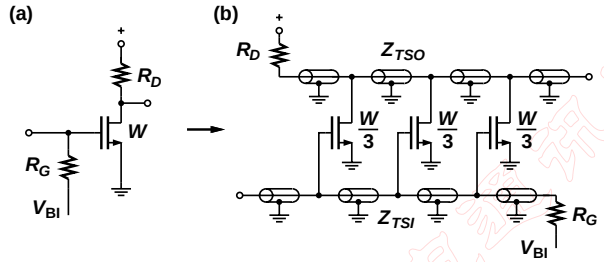

Figure 6.20: (a) Lumped amplifier and (b) distributed amplifier with transmission-line segments.

the expression for an infinite transmission line:

$$Z_{TLO} \approx \sqrt{\frac{L_O}{C_O/n}}. \tag{6.57}$$

Requiring that the transmission line impedance Z_{TLO} matches the termination resistor R_D to avoid reflections, we find that the inductors have to be $L_O = R_D^2 \cdot C_O/n$. The cutoff frequency of the output transmission line can be written:

$$f_{\text{cutoff}} = \frac{1}{2\pi} \cdot \frac{2}{\sqrt{L_O \cdot C_O/n}} = \frac{1}{2\pi} \cdot \frac{2 \cdot n}{R_D C_O}. \tag{6.58}$$

Thus, the output cutoff-frequency is $2 \cdot n$ times higher than the output pole of the lumped amplifier stage in Fig. 6.19(a). If we assume that the input transmission line is identical to the output line ($C_I = C_O$, $R_G = R_D$, same impedance, same cutoff frequency), the same improvement is obtained for the input pole and the overall amplifier bandwidth is increased by about $2 \cdot n$. In theory, there is no limit to the bandwidth if we choose a large enough number of sections ($n \to \infty$). In practice, the number of sections is limited to $4 - 7$, mostly due to losses in the artificial transmission lines.

Alternatively, the input and output transmission lines can be built from short (distributed) transmission-line segments instead of inductors as shown in Fig. 6.20. In this case the characteristic impedance of the transmission lines is lower than that of the segments due to the periodic capacitive loading by the transistors. For example, the loaded impedance of the output transmission line is:

$$Z_{TLO} \approx Z_{TSO} \cdot \sqrt{\frac{C_{TSO}}{C_{TSO} + C_O/n}} \tag{6.59}$$

where Z_{TSO} is the characteristic impedance of the unloaded transmission-line segment and C_{TSO} is the capacitance of the transmission-line segment which depends on its length [KDV+01]. For more information on distributed amplifiers see [Won93].

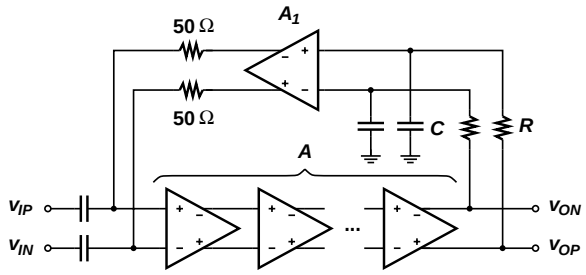

Figure 6.21: MA with offset compensation circuit.

6.3.3 Offset Compensation

We have seen in Section 6.2.4 that the offset voltage of a LA should be 0.1 mV or less. For larger offsets the data is sliced so much off center that the receiver sensitivity is substantially reduced. A typical BJT amplifier has a 3σ random offset voltage of a few mV while a RF MOSFET amplifier has an offset voltage of several 10 mV. In either case this is too much and we need an *Offset Compensation* scheme to reduce the offset voltage to the required value.

Figure 6.21 shows a MA with an offset-compensation circuit frequently used for bipolar implementations. The input signal is AC coupled to the MA and an error amplifier A_1 supplies a differential DC input voltage to compensate for the MA's offset voltage. The error amplifier senses the DC component of the MA's output signal by means of two low-pass filters (R and C) and adjusts its output voltage until the MA's differential output voltage becomes zero, i.e., the offset voltage is compensated. The 50 Ω resistors together with the output impedance of the error amplifier (assumed to be zero in Fig. 6.21) serve as the input termination for the MA.

Another approach to cancel the offset voltage is shown in Fig. 6.22. Now the first stage of the MA has two differential inputs, one pair is used for the input signal and the second pair is used for the offset-compensation voltage. Such a stage can be implemented for example with two differential pairs joined at the outputs and is known as a *Differential Difference Amplifier* (DDA)[3]. As before, the offset-compensation voltage is supplied by the error amplifier A_1 but now without the need for low-impedance outputs. The termination for the signal inputs is implemented with separate resistors to ground.

There are two reasons why the circuits in Figs. 6.21 and 6.22 do not completely eliminate the offset voltage: (i) the finite gain of the error amplifier A_1 and (ii) the offset voltage V_{OS1} of the error amplifier. A simple analysis shows that the MA offset voltage

[3]The classical DDA is characterized by two precisely matched input ports [SG87, Säc89], however, in offset-cancellation applications the port receiving the offset-compensation voltage often has a lower gain. For simplicity we assume in the following analysis that both ports have the same gain.

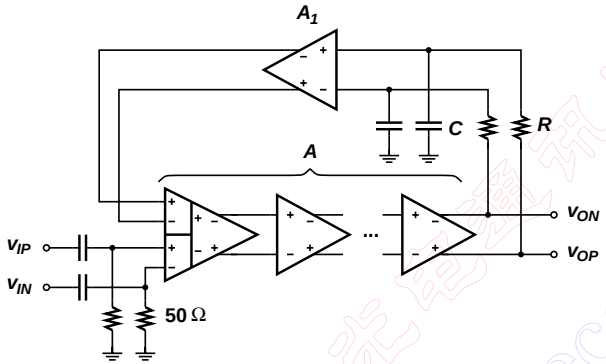

Figure 6.22: MA with DDA-type offset compensation circuit.

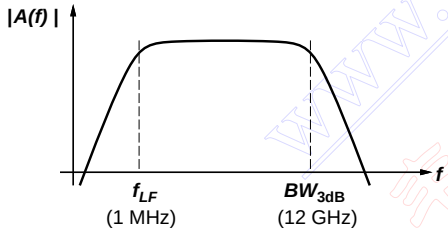

Figure 6.23: Frequency response of a MA with offset compensation.

V_{OS} is reduced to:

$$V'_{OS} = \frac{V_{OS} + A_1 \cdot V_{OS1}}{A \cdot A_1 + 1} \approx \frac{V_{OS}}{A \cdot A_1} + \frac{V_{OS1}}{A}. \tag{6.60}$$

Since the error amplifier doesn't have to be fast, large transistors with good matching properties can be used to make V_{OS1} very small. Depending on the amount of offset that must be compensated, A_1 can be just a buffer ($A_1 = 1$) or an amplifier ($A_1 > 1$). Typically, a buffer is sufficient for a BJT amplifier with low offset [MRW94], while MOSFET amplifiers require additional loop-gain to meet the offset specification.

Low-Frequency Cutoff. The offset-compensation circuits of Figs. 6.21 and 6.22 do not only suppress the offset voltage, but also the low-frequency components of the input signal. This undesired effect leads to a low-frequency cutoff in the MA frequency response $|A(f)|$ as illustrated in Fig. 6.23. See Section 6.2.5 for a discussion of the LF-cutoff specification. The 3 dB low-frequency cutoff for the circuit in Fig. 6.21 is:

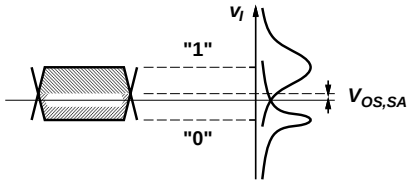

Figure 6.24: Slice-level adjustment is required in situations with unequal noise distributions for zeros and ones.

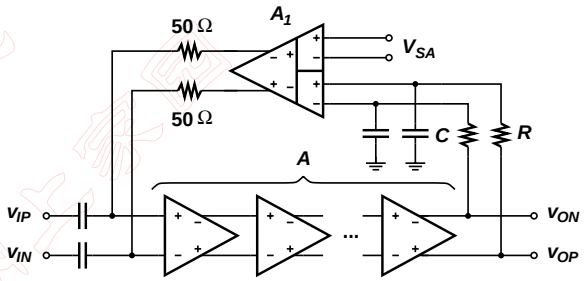

Figure 6.25: MA with offset compensation and slice-level adjustment.

$$f_{LF} = \frac{1}{2\pi} \cdot \frac{A \cdot A_1/2 + 1}{RC}. \tag{6.61}$$

The $A_1/2$ term is due to the fact that the error amplifier's AC output voltage is cut in half by the 50 Ω termination resistors and the 50 Ω source resistance. For the DDA-type offset compensation in Fig. 6.22 we would have the full gain A_1. From Eq. (6.61) we see that in order to get a desired cutoff frequency we need to make the loop bandwidth $1/(2\pi \cdot RC)$ much smaller than this frequency. For example, if $A \cdot A_1/2 + 1 = 100$, we need a loop bandwidth of 10 kHz to achieve a cutoff frequency of 1 MHz in the MA. The small loop bandwidth can be realized with feedback capacitors C_F around the error amplifier A_1 thus using the Miller effect to create large effective input capacitances $C = (|A_1| + 1) \cdot C_F$.

Slice-Level Adjustment. For a receiver with an APD photodetector or an optical preamplifier, the received ones contain more noise than the received zeros. The eye diagram and noise statistics looks like shown in Fig. 6.24. Under these circumstances the optimum slice level is a little bit below the mid-point between zeros and ones (the DC component of the received signal). To make the LA slice at this optimum level we would like to introduce a *controlled* amount of offset voltage $V_{OS,SA}$.

A simple modification to the offset-compensation circuit of Fig. 6.21 can introduce such an offset voltage. Figure 6.25 shows the offset compensation circuit from before, but now

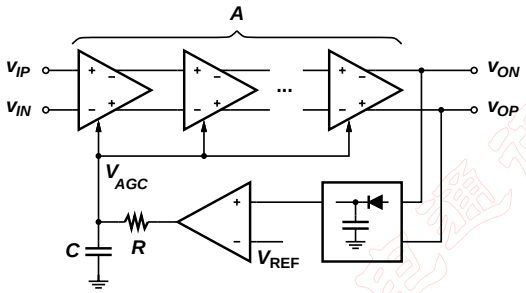

Figure 6.26: Block diagram of an AGC amplifier.

with a DDA in the feedback loop. While an op amp forces the input differential voltage to become zero, a DDA forces the two input-port voltages to become equal. Thus the feedback loop reaches steady state when the DC component of the MA's output voltage becomes equal to V_{SA}. Calculating the input offset voltage of the MA in Fig. 6.25 yields:

$$V_{OS,SA} = \frac{V_{OS} + A_1 \cdot (V_{SA} + V_{OS1})}{A \cdot A_1 + 1} \approx \frac{V_{SA}}{A}. \tag{6.62}$$

Thus if $A \cdot A_1 \gg 1$, $V_{OS1} \ll V_{SA}$, and $V_{OS} \ll A_1 \cdot V_{SA}$, the desired offset $V_{OS,SA}$ is solely controlled by the slice-level adjustment voltage V_{SA}.

Alternatively, slice-level adjustment can be implemented in the TIA or the decision circuit, if a linear amplifier is used.

6.3.4 Automatic Gain Control

An AGC amplifier, shown in Fig. 6.26, consist of a *Variable Gain Amplifier* (VGA) and an *Automatic Gain Control* (AGC) mechanism. The VGA consists of multiple stages just like any MA. But now a DC voltage V_{AGC} controls the gain of some or all stages. An *Amplitude Detector* determines the amplitude of the VGA's output signal. This value is compared to a reference voltage V_{REF} and in response to the difference, the gain-control voltage V_{AGC} is increased or decreased such that the output amplitude remains constant. The speed and stability of the gain-control loop is determined by the low-pass filter formed by R and C.

Let's have a closer look at the two main components of an AGC amplifier: the variable gain stages and the amplitude detector.

Variable Gain Stage. The gain of an amplifier stage can be controlled in a number of ways. Some popular gain-control methods are:

- Vary the transconductance g_m of the amplifying transistor. Since the gain of a simple gain stage (without feedback) is given by $g_m \cdot R_L$, varying g_m changes the gain. The value of g_m can be controlled for instance with the transistor's bias current.

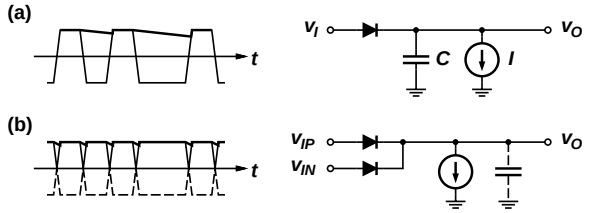

Figure 6.27: Amplitude detectors: (a) peak detector vs. (b) rectifier.

- Vary the load resistor R_L. This is an obvious alternative to what has been said before. The variable load can be implemented with a FET operating in the linear regime. But beware, changing R_L also changes the frequency response of the amplifier stage.
- Vary the amount of feedback. In particular, the gain of a stage with source degeneration (series feedback) can be controlled by varying the source resistor. Again, a FET can be used to do this job.
- Build the amplifier stage as a mixer (= multiplier). One input to the mixer is used for the signal and the other input is used for the gain-control voltage.
- Switch between two (or more) fixed gains. If we have two stages with different gains, we can digitally select one of the two gains by selectively enabling the corresponding stage.

We will discuss implementation examples of these gain-control methods in Section 6.4.

Amplitude Detector. There are two different approaches to designing an amplitude detector: the *Peak Detector* and the *Rectifier* approach. The principle of a peak detector is shown in Fig. 6.27(a). Whenever the input signal v_I exceeds the output voltage v_O the diode (assumed to be ideal) turns on and charges the capacitor C to the value of v_I. The current source I discharges the capacitor slowly causing the output voltage to droop. This is necessary for the amplitude detector to respond to decreasing amplitudes. Choosing the right values for C and I is critical in meeting response time and droop specifications.

The rectifier approach is shown in Fig. 6.27(b). The input signal is rectified, i.e., its absolute value is taken. The result is a DC voltage with periodic drops caused by the finite rise and fall times of the input signal. If the input signal is available in differential form (v_{IP}, v_{IN}), as is usually the case in MAs, the rectifier can be built with two (ideal) diodes. The output voltage v_O is equal to the larger one of the two input voltages: $v_O = \max(v_{IP}, v_{IN})$. A small capacitor can be added to the output to filter out the ripple. Note that this approach has no problem with response time or droop, but it requires differential signals with fast rise/fall times and low differential offset.

Figure 6.28 shows transistor-level implementations of both the peak detector and the rectifier. The peak detector circuit shown is similar to the one described in [OSA+94].

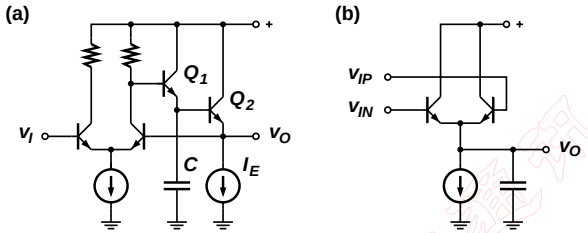

Figure 6.28: Realization of: (a) peak detector and (b) rectifier.

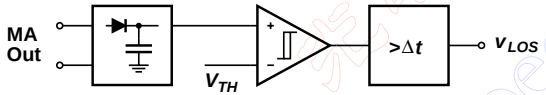

Figure 6.29: Block diagram of a loss-of-signal detector.

Transistor Q_1 is turned on by the differential stage if the input voltage v_I exceeds v_O thus charging up capacitor C. Emitter follower Q_2 buffers the output voltage across the capacitor and also provides a small discharge current $I = I_E/\beta$ to it. The feedback arrangement with the differential stage compensates for the base-emitter voltage drops of Q_1 and Q_2 and makes v_O track the peak of v_I without an offset. Peak detector circuits implemented in MOSFET technology can be found in [MM96, TSN+98b].

The rectifier circuit shown in Fig. 6.28(b) is similar to the one described in [RR89]. The emitter-coupled BJT pair operates as the "max" circuit and also provides a high-impedance input. Note that the output signal v_O is shifted down by a base-emitter voltage drop. For proper operation of this rectifier, the offset of the differential input signal must be very low. A MOSFET circuit combining a rectifier with an op amp is described in [PA94]. A current-mode rectifier is reported in [SDC+95].

6.3.5 Loss of Signal Detection

Loss of Signal (LOS) happens for instance when the fiber gets cut by accident or when the laser or its power supply fail, etc. The receiver must detect this situation in order to signal an alarm and/or to automatically restore the connectivity. For this reason some MAs include a loss-of-signal detector on the chip.

The block diagram of a LOS circuit is shown in Fig. 6.29. The amplitude of the MA output signal is monitored with an amplitude detector. This block has already been discussed in Section 6.3.4 on AGC amplifiers. The amplitude is compared to a threshold voltage V_{TH} with a comparator circuit. The threshold voltage is set equal to the amplitude that just meets the BER requirement. The comparator usually exhibits a small amount of hysteresis to avoid oscillations in the LOS output signal. Finally, a timer circuit suppresses short LOS events. For instance the SONET standard requires that only LOS events which

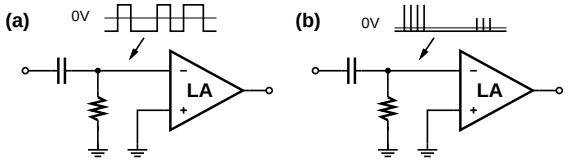

Figure 6.30: An AC-coupled LA operated with (a) continuous mode and (b) burst mode signals.

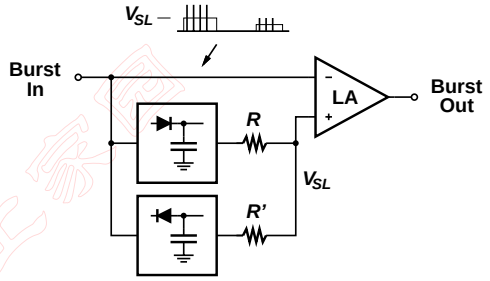

Figure 6.31: A DC-coupled burst-mode limiting amplifier.

persist for longer than 2.3 μs are signaled [Bel95].

6.3.6 Burst-Mode Amplifier

How does a burst-mode MA differ from a continuous-mode MA? Figure 6.30(a) shows a single-ended continuous-mode LA which receives the input signal through an AC-coupling network. As a result of the AC-coupling network the data signal is sliced at its average value. Since the continuous-mode signal is DC balanced, the average value corresponds to the vertical center of the eye and all is well. (We are disregarding the case of unequal noise distributions which has already been discussed in Section 6.3.3.) If we use the same AC-coupled arrangement to process burst-mode signals, as shown in Fig. 6.30(b), the slice level is still at the average value of the signal but now this level is nowhere near the vertical center of the eye! Many bit errors, pulse-width distortions, and jitter are the undesirable consequences.

In a burst-mode receiver the TIA and the MA (and the CDR) must be DC coupled and the various offset voltages produced in these components must be brought under control. One approach is to remove the offset voltage in the TIA, as explained in Section 5.2.9. Then, if the offset introduced by the MA is small, additional offset control may not be necessary. Alternatively, if the TIA has no offset control or is single ended, the MA must perform the offset control function.

A single-ended LA with slice-level control is shown in Fig. 6.31. The circuits described

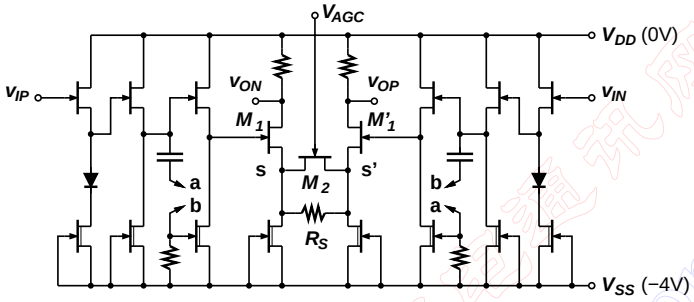

Figure 6.32: MESFET/HFET implementation of a MA stage.

in [INA+97, NIA98] are based on this topology. A peak and bottom detector determine the maximum and minimum value of the input signal, respectively. The average between these two values is obtained with two matched resistors R and R' which is then used as the slice level V_{SL}. This circuit slices the input data signal at the vertical center of the eye regardless of the signal's average value. The peak/bottom detectors must be reset in between bursts in order to acquire the correct amplitude for each individual burst. Remember that large amplitude variations from burst to burst are common in burst-mode systems such as ATM-PON and EPON.

6.4 MA Circuit Implementations

In the following section we will examine some representative transistor-level MA circuits taken from the literature and designed for a variety of technologies (cf. Appendix **??**). These circuits will illustrate how the design principles discussed in the previous section are implemented in practice.

6.4.1 MESFET & HFET Technology

Figure 6.32 shows a MA stage implemented in a MESFET or HFET technology with enhancement- and depletion-mode devices. The stage is shown with a provision for gain control (VGA stage) but can easily be modified into a LA stage. For example, the amplifiers reported in [LBH+97, LWL+97] are based on this topology and realized in GaAs technology.

The differential input signals v_{IP} and v_{IN} are buffered with a cascade of three source followers. The source followers are implemented with enhancement-mode FETs while the current-source loads are implemented with depletion-mode FETs which simplifies their biasing. The buffered signals drive the differential pair M_1 and M'_1 with a shared source degeneration resistor, R_S. The FET M_2 acts as a voltage-controlled resistor which controls the amount of series feedback and thus the gain of this stage. This stage can be

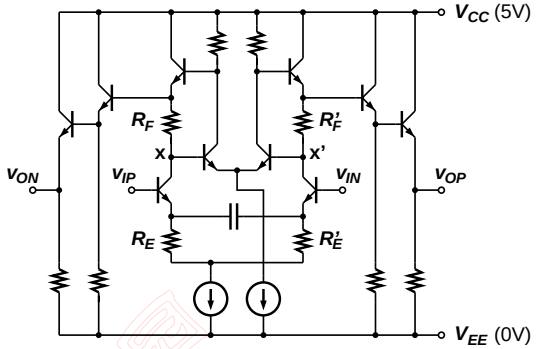

Figure 6.33: The "Cherry-Hooper" implementation of a LA stage.

turned into a LA stage by removing R_S and M_2 and shorting nodes s and s'. After this modification the stage operates at its maximum gain.

The stage further contains an RC network connected to the input buffers to speed up the transitions (see Figure 6.32). The operation of this network can be understood as follows: When the output of the left buffer (gate of M_1) is rising, the bias current in the buffer is momentarily reduced by means of the control voltage at node b. As a result, the transition is accelerated. The bias current reduction is caused by the falling edge at the output of the right buffer which is coupled to node b through an RC high-pass network. Of course, the implementation is symmetric and the right buffer is accelerated in the same way.

6.4.2 BJT & HBT Technology

Cherry-Hooper Stage. Figure 6.33 shows the so-called *Cherry-Hooper Stage* which has been used successfully since 1963 when the original Cherry-Hooper paper [CH63] was published.[4] For instance the amplifiers reported in [RR87, MSW+97, GS99, MOA+00], realized in BJT and HBT technologies, are based on this topology.

The differential input signal drives a differential pair with the emitter degeneration resistors R_E, R'_E. We have already discussed in Section 6.3.2 how emitter degeneration (series feedback) reduces the input capacitance and speeds up the BJT input pole. The load of the input differential pair is a transimpedance amplifier with the feedback resistors R_F, R'_F. The stage in Fig 6.33 is a differential version of that in Fig. 6.15 and we already know from Section 6.3.2 how a TIA load suppresses the parasitic capacitances at the nodes x, x' speeding up the output pole of the differential pair. Furthermore, the reduced swing at the nodes x, x' reduces the Miller contribution to the input capacitance. The

[4] In contrast to Fig. 6.33, the original paper [CH63] describes a single-ended version without emitter followers, but the basic idea of cascading a series feedback stage with a shunt feedback stage is the same.

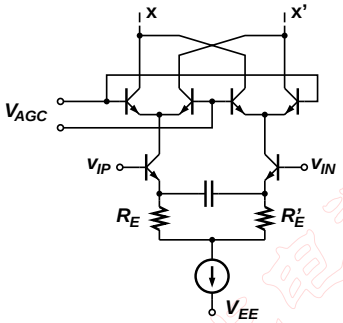

Figure 6.34: A variable-gain Cherry-Hooper stage based on a 4-quadrant mixer. The same transimpedance load as in Fig. 6.33 is connected to nodes x, x'.

differential output signal is buffered with a cascade of two emitter followers. The buffer reduces the loading and shifts the DC voltage down such that the next stage can operate at a high V_{CE} which improves the BJT's speed.

The input differential pair has a transconductance of $1/R_E$ (assuming $R_E \gg 1/g_m$), while the TIA has a transresistance of R_F (assuming $A \gg 1$). The product of these two quantities, R_F/R_E, is the voltage gain of the Cherry-Hooper stage. Since the gain depends on a resistor ratio it is insensitive to temperature, process, and supply voltage variations.

4-Quadrant Mixer VGA Stage. How can we control the gain of a Cherry-Hooper stage as required for the use in an AGC amplifier? Unfortunately, in a bipolar technology we don't have a voltage-controlled resistor to our disposal that we could use to control R_E or R_F (cf. Section 6.4.1). One solution is shown in Fig. 6.34 which has been used for example in the AGC amplifiers reported in [MRW94, OMOW99]. The TIA load is the same as shown in Fig. 6.33, but the input differential pair has been extended to a 4-quadrant mixer. If V_{AGC} is large, the output currents from the differential pair are directly routed to x and x'. But if V_{AGC} is small, nodes x and x' receive a combination of two currents one directly from the differential pair and the other one inverted. These two currents cancel each other partly and thus the gain is reduced. In particular for $V_{AGC} = 0$ the gain becomes zero (assuming all transistors are sized equally).

The gain control voltage for this stage must always be kept positive. If it drops below zero, the gain will start to increase again (and the data signal is inverted) which means that the negative-feedback AGC loop turns into a positive-feedback loop and the AGC becomes unstable.

The input dynamic range of this stage, just like that of the original Cherry-Hooper stage, is determined by the voltage drop across the emitter degeneration resistors R_E and R'_E and is independent of the gain-control voltage V_{AGC}. This may cause a problem at

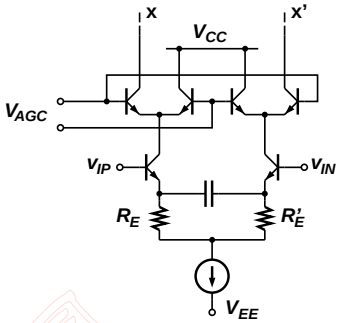

Figure 6.35: A variable-gain Cherry-Hooper stage based on a 2-quadrant mixer. The same transimpedance load as in Fig. 6.33 is connected to nodes x, x'.

low gains where the input signals become large.

2-Quadrant Mixer VGA Stage. So why not use a 2-quadrant multiplier instead? We don't need the quadrants corresponding to $V_{AGC} < 0$, in fact, they are harmful as we have just seen. The stage shown in Fig. 6.35 controls the gain by dumping a variable amount of output current to V_{CC} and has been used for example in the AGC amplifier reported in [STS+94]. As a result of the operating principle only positive gains in the range $0 - R_F/R_E$ are realized. A drawback of this topology is that the DC output current is varying with the gain-control voltage and thus a bias stabilization circuit is required.

4-Quadrant Mixer VGA Stage – The Sequel. Since the circuit in Fig. 6.34 is essentially a 4-quadrant multiplier cell, a.k.a. *Gilbert Cell*, we can swap its two input ports. In Fig. 6.36 we did exactly this and the lower inputs are now used for the gain-control voltage while the upper inputs are used for the input signal. The maximum gain of approximately R_F/R_{E1} is obtained when all of the tail current is directed into the left pair. As we divert some of the tail current into the other pair by means of lowering V_{AGC}, the gain of the left pair drops because g_m is reduced and at the same time a small signal is generated at the output of the right pair which is *subtracted* from the total output signal. As a result of both mechanisms the gain is reduced.

The stage in Fig. 6.36 is of interest because it can be made to have complementary characteristics to the original mixer stage in Fig. 6.34. While the gain of the stage in Fig. 6.34 drops with increasing input voltage, that of the stage in Fig. 6.36 increases. While the bandwidth of the stage in Fig. 6.34 shrinks with decreasing gain, that of the stage in Fig. 6.36 can be made to expand with the appropriate choice of R_{E1}, R_{E2}, C_{E1}, and C_{E2}. Thus in practice the two mixer stages are often combined in a single AGC amplifier. For example in [MRW94, OMOW99] the circuit of Fig. 6.34 is used for the first stage and the circuit of Fig. 6.36 is used for the second stage resulting in superior

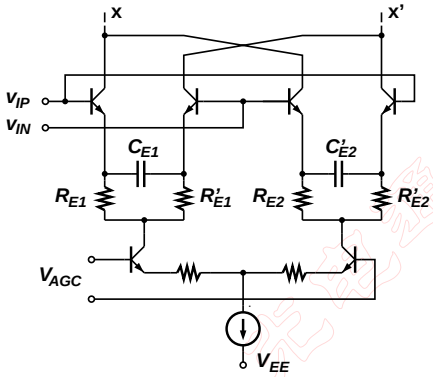

Figure 6.36: A variable-gain Cherry-Hooper stage based on a 4-quadrant mixer. The same transimpedance load as in Fig. 6.33 is connected to nodes x, x'.

linearity and a gain-independent bandwidth.

Selectable Gain Stage. A circuit to select one of two gain values with a Cherry-Hooper stage is shown in Fig. 6.37. Again, the TIA load is the same as in Fig 6.33. The input signal is applied simultaneously to two differential pairs, one with emitter degeneration resistors R_{E1}, R'_{E1} and the other with R_{E2}, R'_{E2}. The lower differential pair steers the tail current either into the left or right differential pair, depending on the select signal. The active pair determines the gain which is either R_F/R_{E1} or R_F/R_{E2}. This topology has the advantage that it has a wide input dynamic range when the lower gain is selected (large degeneration resistor), while it has a good noise figure when the higher gain is selected (small degeneration resistor).

It is also possible to combine this selectable-gain stage with the variable-gain stage in either Fig. 6.34 or Fig. 6.35 into a single stage by stacking them on top of each other. The select input is used to choose the gain range while the AGC input is used to continuously control the gain within each range. For example, such a VGA stage with two gain ranges, −7 dB to 7 dB and 7 dB to 20 dB, has been reported in [Gre01]. In the lower range the maximum input voltage is 1.7 V and the noise figure is 16 dB while in the upper range the maximum input voltage is 0.2 V and the noise figure improves to 12 dB.

6.4.3 CMOS Technology

Low-Voltage Gain Stage. Figure 6.38 shows a MOSFET stage with variable gain which has been used for example in the AGC amplifier reported in [TSN+98b]. The stage consists of an n-MOS differential pair, M_1, M'_1, and p-MOS load transistors, M_2, M'_2, which operate in the linear regime. The gain is controlled with a p-MOSFET, M_3, which

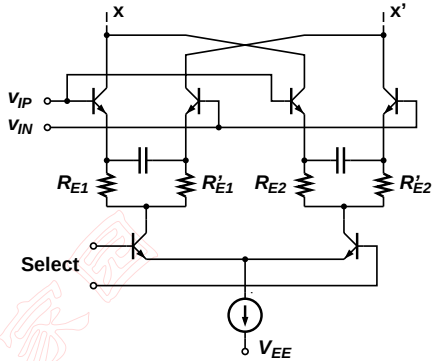

Figure 6.37: A Cherry-Hooper stage with selectable gain. The same transimpedance load as in Fig. 6.33 is connected to nodes x, x'.

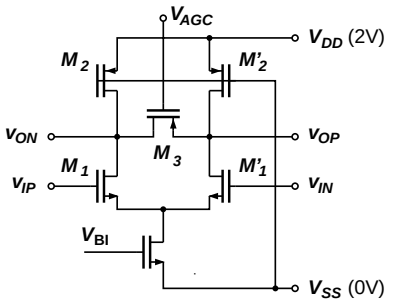

Figure 6.38: CMOS stage with variable gain.

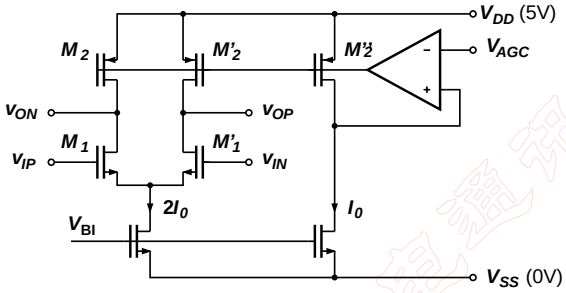

Figure 6.39: CMOS stage with variable gain and replica biasing circuit.

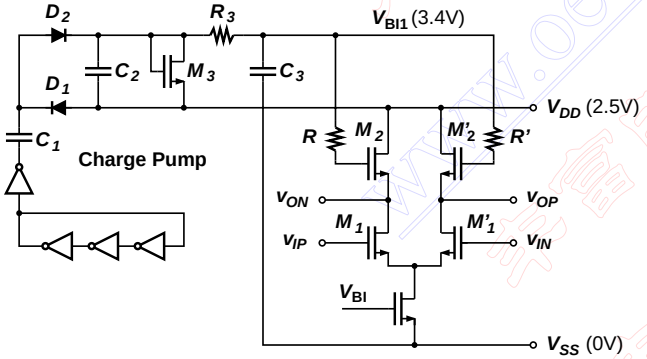

Figure 6.40: CMOS stage with active-inductor loads.

acts as a variable differential load resistance.

Gain Stage with Replica Biasing. Figure 6.39 shows another implementation of a variable-gain MOSFET stage which has been used for example in the AGC amplifier reported in [HG93]. The p-MOS load transistors, M_2, M'_2, operate in the linear regime and the gain is controlled by varying the resistance of these loads. The replica biasing circuit, on the right, generates a gate voltage such that the drain-source resistance of M_2, M'_2, and M''_2 all become equal to $(V_{DD} - V_{AGC})/I_0$. Thus, this amplifier features a well-controlled (temperature and process independent) variable load resistor. A potentially troublesome side effect of varying the load resistors (as opposed to the differential load as in Fig. 6.38) is the gain dependence of the common-mode output voltage, in fact, $v_{OCM} = V_{AGC}$.

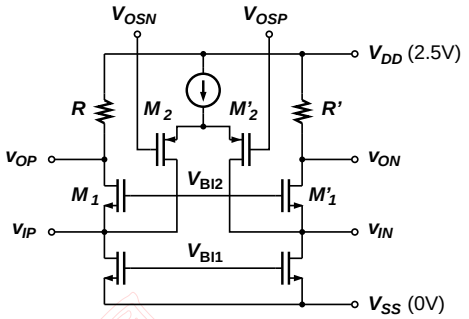

Figure 6.41: CMOS common-gate stage with offset control.

Gain Stage with Active Inductors. Figure 6.40 shows a MOSFET gain stage with active inductors to enhance the frequency response which has been used for example in the LA reported in [SF00]. The n-MOS load transistors, M_2, M'_2, with the gate resistors, R, R', constitute the active inductors. The equivalent inductance is

$$L \approx \frac{R\,C_{gs2}}{g_{m2}} \approx \frac{R}{2\pi \cdot f_T} \tag{6.63}$$

and thus can be controlled with the gate resistor R. The bandwidth extension that can be achieved with inductive loads has been discussed in Section 6.3.2. To alleviate headroom problems in this low-voltage design, the bias voltage $V_{\rm BI1}$ is set to one n-MOS threshold voltage above V_{DD}. As shown in Fig. 6.40 this bias voltage can easily be generated on chip with a charge pump: A ring oscillator drives a capacitor-diode charge pump (C_1, C_2, D_1, D_2) producing a voltage above V_{DD}. This voltage is clamped to the desired value by M_3. Oscillator ripples are filtered out with the low-pass network R_3, C_3. A welcome side effect of using active-inductor loads is that the DC gain of the stage is set by the transistor geometry $A = \sqrt{W_1/W_2}$ which is process, temperature, and supply voltage insensitive.

Common-Gate Stage with Offset Control. Figure 6.41 shows a common-gate MOSFET gain stage with an offset control circuit which has been used for example in the LA reported in [SF00]. This circuit is useful as the *input stage* of a LA. The common-gate n-MOS transistors, M_1, M'_1, provide a low-impedance input which can be made equal to $50\,\Omega$ with the appropriate choice of transistor dimensions and bias currents (constant-g_m biasing). Apart from providing the input termination, the common-gate input transistors also provide an effective ESD protection. Regular ESD networks are not suitable for high-speed inputs because they limit the signal bandwidth too much. It is therefore advisable to build the input stage such that the parasitic transistor diodes act as the ESD protection: in Fig. 6.41 there are three parasitic diodes, two to ground and one to V_{DD}. Finally, the input stage features a p-MOS differential pair, M_2, M'_2, which is used to control the offset voltage as shown in Fig. 6.22.

Company & Product	Speed	A	BW_{3dB}	f_{LF}	F	Power	Technology
Agere LG1605DXB	2.5 Gb/s	28 dB	3.0 GHz	2.5 kHz	15 dB	400 mW	GaAs HFET
AMCC S3051	2.5 Gb/s	26 dB	2.0 GHz	2.0 kHz	18 dB	375 mW	
Maxim MAX3265	2.5 Gb/s	43 dB		2.0 kHz		165 mW	
Philips OQ2538HP	2.5 Gb/s	40 dB	3.0 GHz		14 dB	270 mW	Si BJT
Agere TLMA0110G	10 Gb/s	33 dB	9 GHz	25 kHz	17 dB	700 mW	GaAs HFET
AMCC S3096	10 Gb/s	44 dB	10 GHz				SiGe HBT
Maxim MAX3971	10 Gb/s		10 GHz	40 kHz		155 mW	SiGe HBT
OKI GHAD4103	10 Gb/s	12 dB				1200 mW	GaAs HFET

Table 6.1: Examples for 2.5 and 10 Gb/s LA products.

Company & Product	Speed	A_{max}	BW_{3dB}	f_{LF}	F	Power	Technology
Nortel AC03	2.5 Gb/s	30 dB	2.0 GHz	25 kHz	9 dB	500 mW	Si BJT
Nortel AC10	2.5 Gb/s	24 dB	1.9 GHz		10 dB	198 mW	Si BJT
Giga GD19902	10 Gb/s	20 dB	10 GHz			2800 mW	GaAs HFET
OKI GHAD4102	10 Gb/s	10 dB	10 GHz			1300 mW	GaAs HFET

Table 6.2: Examples for 2.5 and 10 Gb/s AGC amplifier products.

6.5 Product Examples

Tables 6.1 and 6.2 summarize the main parameters of some commercially available LA and AGC amplifiers. The numbers have been taken from data sheets of the manufacturer which were available to me at the time of writing. For up-to-date product information please contact the manufacturer directly. For consistency, the values tabulated under A are the *single-ended* gains, even for MAs which have differential outputs. Similarly, the values tabulated under F are the *single-ended* noise figures (cf. Section 6.2.3).

Comparing the 2.5 and 10 Gb/s parts we see that in general the 10 Gb/s amplifiers consume significantly more power than the 2.5 Gb/s amplifiers. The reader may wonder about the difference between the two similar Nortel parts: The AC03 requires a 5.2 V power supply while the AC10 runs from a lower 3.3 V supply, which also explains the difference in power dissipation.

6.6 Research Directions

The research effort can be divided roughly into two categories: higher speed and lower cost.

Higher Speed. It has already been pointed out in Section 5.5 that many research groups are now aiming at the 40 Gb/s speed and beyond. To this end fast MAs, with a bandwidth of 40 GHz and more, must be designed. Usually SiGe, GaAs, and InP technologies combined with heterostructure devices such as HBTs and HFETs are used to reach this goal.

Here are some examples from the literature:

- In SiGe-HBT technology a 31 GHz and a 32.7 GHz AGC amplifier as well as a 40 Gb/s and a 49 GHz LA have been reported in [MOO+98], [OMOW99], [RDR+01], and [MOA+00], respectively.
- In GaAs-HFET technology a 18 GHz AGC amplifier as well as a 27.7 GHz and a 29.3 GHz LA have been reported in [LBH+97], [LBH+97], and [LWL+97], respectively.
- In GaAs-HBT technology a 26 GHz VGA has been reported in [RZP+99].
- In InP-HBT technology a 30 GHz LA has been reported in [MSW+97].

It can be seen, that in terms of speed, the LAs have an advantage over the more complex AGC amplifiers.

Lower Cost. Another area of research is focusing on the design of high-performance MAs in low-cost, mainstream technologies, in particular digital CMOS.

For example, a SONET compliant 10 Gb/s LA has been implemented in a low-cost "modular BiCMOS" technology [KB00]. A 2.4 Gb/s, 0.15 μm CMOS AGC amplifier has been reported in [TSN+98b] and a SONET compliant 2.5 Gb/s, 0.25 μm CMOS LA has

been demonstrated in [SF00]. The promise of a CMOS main amplifier is that it can be integrated with the CDR, DMUX, and the digital frame processing on single CMOS chips to provide a cost-effective and compact low-power receiver solution.

6.7 Summary

Two types of main amplifiers are used:

- The limiting amplifier which has superior speed and power-dissipation characteristics but is highly nonlinear which restricts its field of applications.
- The AGC amplifier which is linear over a wide range of input amplitudes making it suitable for receivers with an equalizer, decision-point steering, a soft-decision decoder, etc.

The main specifications for the MA are:

- The voltage gain which must be large enough to drive the subsequent CDR reliably.
- The bandwidth which must be large enough to avoid signal distortions due to ISI.
- The noise figure which must be low enough to avoid a degradation of the receiver sensitivity.
- The input offset voltage which must be very small compared to the input signal, especially if a LA is used.
- The low-frequency cutoff which must be very low, especially if signals with long runs of zeros and ones are to be amplified.
- The sensitivity which must be about 3.2× lower than the output signal of the TIA at the sensitivity limit.
- The input dynamic range which must be large enough to avoid distortions for large input signals.
- The AM-to-PM conversion factor which must be low enough to limit the generation of jitter in the presence of spurious amplitude modulation.

Virtually all MAs use a multistage topology because it allows the realization of very high gain-bandwidth products ($\gg f_T$). Furthermore, broadband techniques such as series feedback, emitter peaking, buffering, scaling, cascoding, transimpedance load, negative Miller capacitance, shunt peaking, and distributed amplifiers are applied to the MA stages to improve their bandwidth and shape their roll-off characteristics.

Most MAs include an offset compensation circuit to reduce the undesired offset voltage down to an acceptable value. Some MAs feature an adjustable amount of intentional offset (slice-level adjust) to optimize the BER performance in the presence of unequal noise distributions for zeros and ones. AGC amplifiers consist of a variable-gain amplifier,

an amplitude detector, and a feedback loop which controls the gain such that the output amplitude remains constant. Some MAs feature a loss-of-signal detector to detect faults such as a a cut fiber. Burst-mode LAs have special provisions to quickly adapt to an input signal with varying amplitude and no DC balance.

MAs have been implemented in a variety of technologies including MESFET, HFET, BJT, HBT, BiCMOS, and CMOS. The "Cherry-Hooper" architecture, which combines series feedback with a transimpedance load, is often chosen for BJT and HBT implementations.

Currently, researchers are working on 40 Gb/s MAs, as well as MAs in low-cost technologies such as CMOS.

Chapter 7

Optical Transmitters

Before we go on to discuss laser/modulator-driver circuits, it is useful to have some background information on optical transmitters and the main characteristics of commonly used lasers and modulators.

Types of Modulation. Figure 7.1 illustrates two fundamentally different ways to generate a modulated optical signal: (i) we can turn the laser on and off by modulating its current, this method is called *Direct Modulation* or (ii) we can leave the laser on at all times (continuous wave laser) and modulate the light beam with a sort of electrooptical shutter known as modulator, this method is called *External Modulation*. Direct modulation has the advantage of simplicity, compactness, and cost effectiveness while external modulation can produce higher-quality optical pulses permitting extended reach and higher bit rates.

7.1 Transmitter Specifications

In the following we want to look at two important specifications for optical transmitters: (i) the spectral linewidth and (ii) the extinction ratio. The values that can be achieved for these parameters depends on whether direct or external modulation is used.

Spectral Linewidth. If we assume a perfectly monochromatic light source followed by a perfect intensity modulator the optical spectrum at the output looks like that of an AM transmitter: a carrier and two sidebands corresponding to the spectrum of the NRZ baseband signal (see Fig. 7.1(b)). The 3-dB bandwidth of one NRZ sideband is about $B/2$ and thus the full linewidth is about equal to the bit rate.[1] If we convert this frequency linewidth to the commonly used wavelength linewidth we get:

$$\Delta\lambda = \frac{\lambda^2}{c}\Delta f \approx \frac{\lambda^2}{c}B \tag{7.1}$$

[1] The full bandwidth where the sidebands drop to virtually zero is about $2.5 \cdot B$, corresponding to a spectral efficiency of 0.4 b/s/Hz.

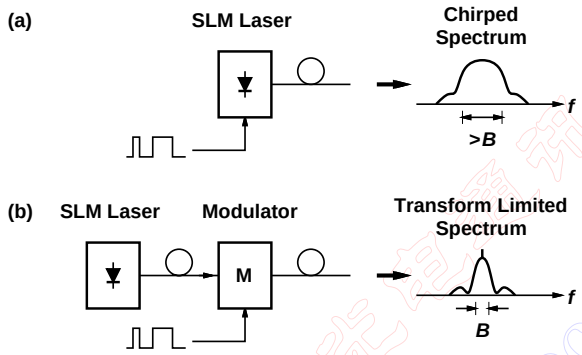

Figure 7.1: Optical transmitters: (a) direct modulation vs. (b) external modulation.

where c is the speed of light in vacuum ($c \approx 3 \cdot 10^8$ m/s). For example, at 10 Gb/s the frequency linewidth is about 10 GHz corresponding to a wavelength linewidth of 0.08 nm for $\lambda = 1.55\,\mu$m.

Optical pulses which have such a narrow linewidth are known as *Transform Limited Pulses*. In practice some external modulator can produce pulses which come very close to this ideal.

In most transmitters, especially when using directly modulated lasers, the modulation process not only changes the light's amplitude (AM) but also its phase (PM) or frequency (FM). This unwanted frequency modulation is called *Chirp* and causes the spectral linewidth to broaden (see Fig. 7.1(a)). Mathematically the effect of chirp on the linewidth can be approximated by:

$$\Delta\lambda \approx \frac{\lambda^2}{c}\sqrt{\alpha^2+1}\cdot B \tag{7.2}$$

where α is known as the *Chirp Parameter* or *Linewidth Enhancement Factor*. With the typical value $\alpha \approx 4$ for a directly modulated laser [HKV97] the linewidth of a 10 Gb/s transmitter broadens to about 41 GHz or 0.33 nm. External modulators also exhibit a small amount of chirp, but virtually all types of modulators can provide $|\alpha| < 1$ and some can readily obtain $|\alpha| < 0.1$ [HKV97].

So far we assumed that the *unmodulated* source is perfectly monochromatic (zero linewidth), or at least that the unmodulated linewidth is much smaller than those given in Eqs. (7.1) and (7.2). However, for some sources this is not the case. For example, a Fabry-Perot laser has a typical unmodulated linewidth of about 3 nm, an LED has an even wider linewidth in the range of $50 - 60$ nm. For such wide-linewidth sources, chirp and AM sidebands are mostly irrelevant and the transmitter linewidth is given by:

$$\Delta\lambda \approx \Delta\lambda_S \tag{7.3}$$

where $\Delta\lambda_S$ is the unmodulated linewidth of the source.

We know from Chapter 2 that optical pulses with a wide linewidth tend to spread out quickly in a dispersive medium such as a SMF operated at a wavelength of 1.55 μm. The power penalty due to this pulse spreading is known as *Dispersion Penalty*. Data sheets of lasers and modulators frequently specify this dispersion penalty for a given amount of fiber dispersion. For example, a 2.5 Gb/s, 1.55 μm laser may have a dispersion penalty of 1 dB for 2000 ps/nm fiber dispersion. We know from Section 2.2 that a fiber dispersion of 2000 ps/nm corresponds to about 120 km of SMF. With Eqs. (2.4) and (2.6) we can further estimate that the linewidth of this laser, when modulated at 2.5 Gb/s, must be around 0.1 nm.

The impact of fiber dispersion on the maximum bit rate and distance for the cases of direct and external modulation will be analyzed in more detail in Section 7.4. However, as a rough guide we can say that telecommunication systems at 10 Gb/s and above generally use external modulation, 2.5 Gb/s systems use direct or external modulation depending on the fiber length, and systems below 2.5 Gb/s generally use direct modulation. 10 Gb/s Ethernet systems use direct modulation even at 10 Gb/s to keep the cost low. In the latter case the 1.3 μm wavelength is used where dispersion in a SMF is small.

The narrow linewidth obtained with external modulation not only reduces the dispersion penalty but also permits a closer channel spacing in a *Dense Wavelength Division Multiplexing* (DWDM) system. To avoid interference, the channels must be spaced further apart than the linewidth of each channel. Current DWDM systems have a channel spacing of 200 or 100 GHz with a trend towards 50 GHz.

Extinction Ratio. Optical transmitters, no matter if direct or external modulation is used, do not shut off *completely* when a zero is transmitted. This undesired effect is quantified by the *Extinction Ratio* (ER), which is defined as follows[2]:

$$ER = \frac{P_1}{P_0} \tag{7.4}$$

where P_0 is the optical power emitted for a zero and P_1 the power for a one. Thus an ideal transmitter would have an infinite ER. The ER is usually expressed in dBs using the conversion rule $10 \cdot \log ER$. Typically, ERs for directly modulated lasers extend from 9 to 14 dB while ERs for externally modulated lasers can exceed 15 dB [FJ97]. SONET/SDH transmitters are typically required to have an ER in the range $8.2 - 10$ dB depending on the application.

It doesn't come as a surprise that a finite ER causes a power penalty. We observe in Fig. 7.2 that decreasing the ER reduces the peak-to-peak optical signal, $P_1 - P_0$, even if the average power $\overline{P} = (P_1 + P_0)/2$ is kept constant. Thus we have to increase the transmitted power by PP to restore the original peak-to-peak amplitude. The power penalty can easily be derived as [Agr97]:

$$PP = \frac{ER+1}{ER-1}. \tag{7.5}$$

[2]There is no consensus whether P_1/P_0 or P_0/P_1 should be used for ER. In this text we define ER such that it is larger than one.

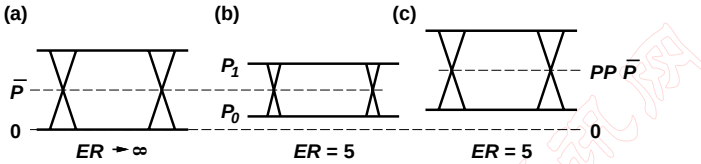

Figure 7.2: Eye diagram (a) with infinite ER, (b) with $ER = 5$, (c) with $ER = 5$ and increased average power (1.5×) to restore the peak-to-peak amplitude.

For example, an extinction ratio of 10 dB ($ER = 10$) causes a power penalty of 0.87 dB ($PP = 1.22$).

In deriving Eq. (7.5) we assumed that the amount of noise is not affected by the ER. This is the case for non-amplified p-i-n receivers, however, in systems with APD detectors or optical amplifiers the received noise increases for a reduced ER: (i) the non-zero value of P_0 causes noise on the zeros and (ii) the increase in power to compensate for the reduced amplitude adds noise on the zeros and ones necessitating an even larger power increase. As a result of these two mechanisms the power penalty will be larger than given in Eq. (7.5). If we assume the extreme case where the receiver noise is dominated by the detector (or amplifier) noise and that this noise power is proportional to the received signal level we find [FJ97]:

$$PP = \frac{ER+1}{ER-1} \cdot \frac{\sqrt{ER}+1}{\sqrt{ER}-1}. \tag{7.6}$$

For example, an extinction ratio of 10 dB ($ER = 10$) causes a power penalty up to 3.72 dB ($PP = 2.35$) in an amplified lightwave systems.

In regulatory standards the receiver sensitivity is specified for the worst-case extinction ratio. Therefore the power penalty due to finite extinction ratio must be deducted from the noise-based sensitivity as given by Eq. (4.18). Typically, 2.2 dB (for $ER = 6$ dB) must be deducted in short-haul applications and 0.87 dB (for $ER = 10$ dB) must be deducted in long-haul applications. [Gre01]

An alternative to the use of extinction ratio and average power is the *Optical Modulation Amplitude* (OMA) which is defined as $P_1 - P_0$ and measured in dBm. If the transmitter power and receiver sensitivity are specified in OMA rather than average power, there is no power penalty due to a finite extinction ratio. The OMA measure is used for example in 10GbE systems.

7.2 Lasers

In telecommunication systems the *Fabry-Perot Laser* (FP) and the *Distributed-Feedback Laser* (DFB) are the most commonly used lasers. In data communication systems, such as Gigabit Ethernet, the FP laser and the *Vertical-Cavity Surface-Emitting Laser* (VCSEL) are preferred because of their lower cost. In low-speed data communication (up to about 200 Mb/s) and consumer electronics *Light-Emitting Diodes* (LED) also find application as

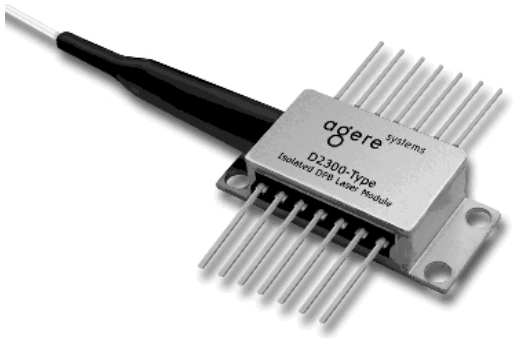

Figure 7.3: Cooled 2.5 Gb/s DFB laser in a 14-pin butterfly package with single-mode fiber pigtail (2.1 cm × 1.3 cm × 0.9 cm). The pins provide access to the laser diode, monitor photodiode, thermoelectric cooler, and temperature sensor.

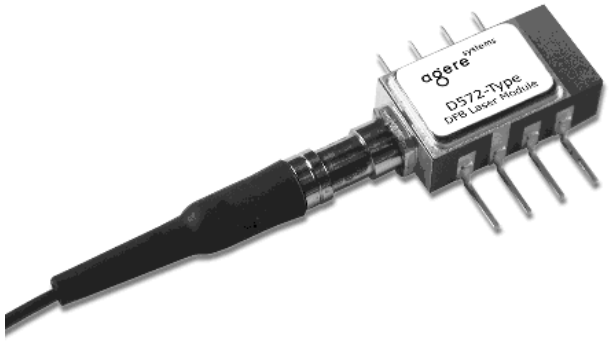

Figure 7.4: Uncooled 2.5 Gb/s DFB laser in a 8-pin package with single-mode fiber pigtail (1.3 cm × 0.7 cm × 0.5 cm).

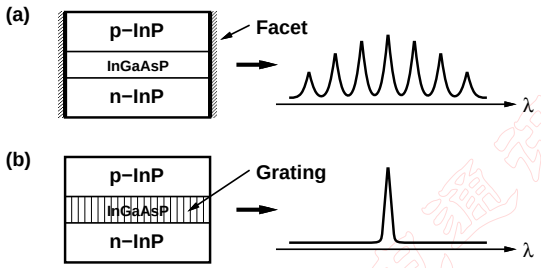

Figure 7.5: Edge-emitting lasers (schematically): (a) Fabry-Perot laser vs. (b) distributed-feedback laser. The light propagates in the direction of the arrow.

an optical source. Figures 7.3 and 7.4 show photos of a cooled and uncooled DFB lasers, respectively.

In the following we will first give a brief description of the FP, DFB, VCSEL, and LED, then we will summarize the laser's main characteristics. For more information on semiconductor lasers and their properties see [Agr97, Koc97, Sze81, Sze98].

Fabry-Perot Laser. The FP laser consists of an optical gain medium with two reflecting facets on each side (see Fig. 7.5(a)). The gain medium is formed by a forward biased p-n junction which injects carriers (electrons and holes) into a thin active region.[3] In Fig. 7.5 the active region consists of a layer of InGaAsP which can be lattice matched to the InP layers above and below. The bandgap of the $In_xGa_{1-x}As_yP_{1-y}$ compound can be controlled by the mixing ratios x and y to provide optical gain anywhere in the $1.0-1.6\,\mu$m range. The surrounding p and n regions are made from InP which has a wider bandgap than the active InGaAsP, helping to confine the carriers to the active region. A typical short-wavelength laser ($0.85\,\mu$m band) uses GaAs as the active layer material surrounded by the wider bandgap AlGaAs which is lattice matched to the GaAs substrate. In practical lasers the active region is usually structured as a *Multiple Quantum Well* (MQW) resulting in better performance than the simple structure shown in Fig. 7.5.

The distance between the two facets, the cavity length, determines the wavelengths at which the laser can operate. If the cavity contains a whole number of wavelengths and the net optical gain is larger than one, lasing occurs. Since the facets are many wavelengths apart (about $300\,\mu$m), there are multiple modes satisfying the cavity constraint. As a result the spectrum of the emitted laser light has multiple peaks as shown in Fig. 7.5(a) on the right. For this reason FP lasers are also known as *Multi-Longitudinal Mode* (MLM) lasers. The spectral linewidth of a FP laser is quite large, typically around $\Delta\lambda_S = 3\,$nm. FP lasers are therefore primarily used at the $1.3\,\mu$m wavelength where dispersion in a SMF is low.

[3]The gain medium by itself without the facets is also known as a *Semiconductor Optical Amplifier* (SOA).

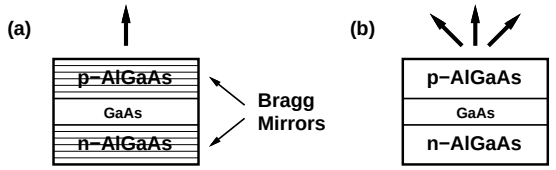

Figure 7.6: Surface-emitting sources (schematically): (a) vertical-cavity surface-emitting laser vs. (b) light-emitting diode. The light propagates in the direction of the arrow(s).

Most FP lasers are operated as *Uncooled Lasers*, which means that their temperature is not controlled and can go up to around 85 °C. This mode of operation simplifies the transmitter design and keeps its cost low but has the drawbacks of varying laser characteristics and reduced laser reliability.

Distributed-Feedback Laser. The DFB laser consist of a gain medium, similar to that in a FP laser, and a grating (or corrugation) as reflector. In contrast to the facets of the FP laser, the grating provides distributed feedback and selects exactly *one* wavelength for amplification (see Fig. 7.5(b)). For this reason DFB lasers are also known as *Single-Longitudinal Mode* (SLM) lasers. The emitted spectrum of an unmodulated DFB laser has a very narrow linewidth of less than $\Delta\lambda_S < 0.001\,$nm ($10 - 100\,$MHz). When the laser is directly modulated the linewidth broadens due to the AM sidebands and chirp as described by Eq. (7.2).

The *Distributed Bragg Reflector Laser* (DBR) is similar to the DFB laser in the sense that it also operates in a single longitudinal mode producing a narrow linewidth. In terms of structure it looks more like a FP laser, however, with the facets replaced by wavelength-selective Bragg mirrors (gratings).

DFB/DBR lasers can be modulated directly and are also very suitable as CW sources for external modulators. Because of their narrow linewidth they are ideal for WDM and DWDM systems. However, the wavelength emitted by a semiconductor lasers is slightly temperature dependent. A typical number is 0.4 nm/°C. Given that the wavelengths are spaced only 0.8 nm (100 GHz grid) apart, the laser temperature must be controlled precisely. Therefore many DFB/DBR lasers are operated as *Cooled Lasers* by mounting them on top of a *Thermoelectric Cooler* (TEC) and a thermistor for temperature stabilization.

Vertical-Cavity Surface-Emitting Laser. The *Vertical-Cavity Surface-Emitting Laser* (VCSEL) is also a SLM laser like the DFB/DBR laser, however, its linewidth is not as narrow (typical $\Delta\lambda_S \approx 1\,$nm). The distinguishing feature of a VCSEL is that it emits the light perpendicular to the wafer plane. The VCSEL consists of a very short vertical cavity (about $1\,\mu$m) with Bragg mirrors at the bottom and the top (see Fig. 7.6(a)).

The advantage of VCSELs is that they can be fabricated, tested, and packaged more easily and at a lower cost. However, the very short gain medium requires mirrors with a very high reflectivity to make the net gain larger than one. Currently, VCSELs are

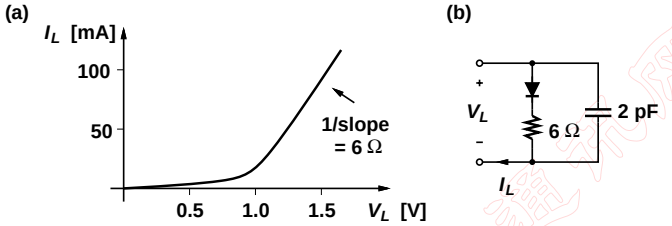

Figure 7.7: (a) Typical I/V curve of an edge-emitting InGaAsP laser, (b) corresponding large-signal AC model for a 10 Gb/s part.

commercially available only at short wavelengths (0.85 μm band) where fiber loss is appreciably high. Their application is mostly in data communication systems using MMF.

Light-Emitting Diode. The *Light-Emitting Diode* (LED) operates on the principle of *Spontaneous Emission* rather than *Stimulated Emission* and therefore is not a laser. The LED consists of a forward biased p-n junction without any mirrors or gratings (see Fig. 7.6(b)). As a result the light is emitted in all directions and it is difficult to couple much of it into a fiber. For example, a LED may couple 10 μW optical power into a fiber while a laser can easily produce 1 mW. Furthermore, because there is no mechanism to select a single wavelength the spectral linewidth is very wide ($\Delta\lambda_S = 50 - 60$ nm). The modulation speed of a LED is limited by the carrier lifetime to a few hundred Mb/s whereas a fast laser can be modulated in excess to 10 Gb/s.

On the plus side, LEDs are very low in cost, they are more reliable than lasers, and they are easier to drive because they lack the temperature-dependent threshold current typical for lasers. Their application is mostly in short-range data communication systems using MMF.

I/V Characteristics. From an electrical point of view, a semiconductor laser is a forward-biased diode. The relationship between the laser current I_L and the forward-voltage drop V_L, the so-called *I/V Curve*, is shown in Fig. 7.7(a) for the example of an edge-emitting InGaAsP laser. The typical small-signal resistance is in the range 3−8 Ω and the forward-voltage drop is in the range 0.7−1.6 V depending on the current, temperature, age, and semiconductor materials used.

A simple large-signal AC model for a laser is shown in Fig. 7.7(b), the values are typical for a 10 Gb/s, edge-emitting InGaAsP laser. The diode determines the forward-voltage drop at low currents. At high currents, the 6 Ω contact resistance dominates the small-signal diode resistance (V_T/I_L) and thus determines the slope of the I/V curve. The capacitor models the p-n junction capacitance (around 2 pF) which is much larger than that of a photodiode because the junction is forward biased. Additional elements (e.g., the bond-wire inductance) must be added to model a packaged laser.

Some packaged lasers contain a series resistor (e.g., 20 Ω or 45 Ω) to match the laser

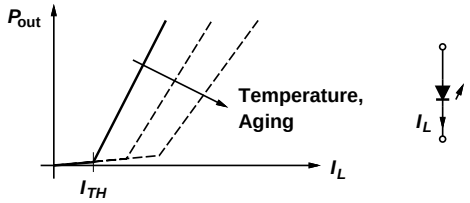

Figure 7.8: L/I curves for a semiconductor laser.

impedance to the transmission line connecting the driver and the laser. In that case an RF choke to apply the bias current directly to the laser (bypassing the matching resistor) is usually provided as well.

L/I Characteristics. The static relationship between the laser current I_L and the light output P_{out}, the so-called *L/I Curve*, is shown schematically in Fig. 7.8. Up to the so-called *Threshold Current*, I_{TH}, the laser outputs only a small amount of incoherent light (like a LED). In this regime the optical gain isn't large enough to sustain lasing. Above the threshold current the output power grows approximately linearly with the laser current as measured by the *Slope Efficiency*. Typical values for an edge-emitting InGaAsP MQW laser are $I_{TH} = 10$ mA, and a slope efficiency of 0.07 mW/mA. Thus for $I_L = 25$ mA (= 10 mA + 15 mA) we obtain about 1 mW optical output power. VCSELs are characterized by a lower threshold current and a better slope efficiency, while LEDs have zero threshold current and a much lower slope efficiency.

The slope efficiency is determined by the *Differential Quantum Efficiency* (DQE) which specifies how efficiently electrons are converted into photons. Just like in the case of the photodiode, we have the situation that optical *power* is linearly related to electrical *amplitude*, which means that we have to be careful to distinguish between electrical and optical dBs.

As indicated in Fig. 7.8 the laser characteristics, in particular I_{TH}, are strongly temperature and age dependent. For example a laser that nominally requires 25 mA to output 1 mW of optical power, may require in excess to 50 mA at 85 °C and near the end of life to output the same power. The temperature dependence of the threshold current is exponential and can be described by [Agr97]:

$$I_{TH}(T) = I_0 \cdot \exp\left(\frac{T}{T_0}\right) \tag{7.7}$$

where I_0 is a constant and T_0 is in the range of $50 - 70$ K for InGaAsP lasers. The temperature dependence of the laser's slope efficiency is less dramatic, a $30 - 40$ % reduction when heating the laser from 25 to 85 °C is typical.

Because of the strong temperature and age dependence of the laser's L/I curve all communication lasers have a built-in monitor photodiode which measures the optical output power at the back facet of the laser. The current from the photodiode closely

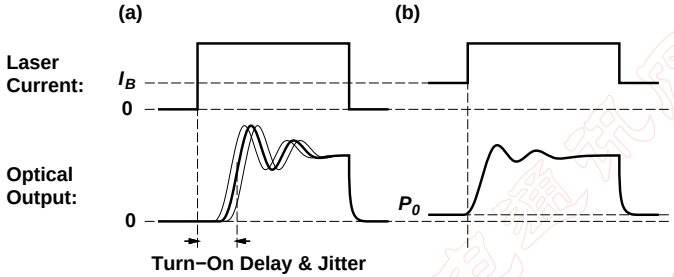

Figure 7.9: Turn-on delay, jitter, and relaxation oscillations in a directly modulated laser: (a) without bias and (b) with bias.

tracks the optical power coupled into the fiber with little dependence on temperature and age, a tracking error of $\pm 10\,\%$ is typical. The monitor photodiode can be used to implement an *Automatic Power Control* (APC).

For analog applications (HFC/CATV systems) the *linearity* of the L/I characteristics is of significance in addition to the slope efficiency and threshold current. A deviation from the linear characteristics causes harmonic distortions and intermodulation products which degrade the TV signal. Typically, the laser linearity is specified by the CSO and CTB parameters (cf. Section 4.8) measured at a certain laser current and modulation index.

Dynamic Behavior. The dynamic relationship between the laser current and the light output is quite complex. It is fully described by the so-called *Laser Rate Equations*, two coupled, nonlinear, differential equations relating the carrier density and photon density in the laser cavity [Agr97, Liu96]. It is beyond the scope of this text to discuss these equations, here we just want to give a verbal description of what happens in response to a laser-current pulse starting from zero current (see Fig. 7.9(a)).

At first the carrier density is zero and has to be built up by means of the injected current until the critical point is reached where the net optical gain becomes unity. Only after this so-called *Turn-On Delay* (TOD) does the laser begin to produce coherent light. The time it takes to reach this critical point is not entirely deterministic and therefore we experience a *Turn-On Delay Jitter* on the rising edge of the optical pulse. Once lasing sets in, carriers are rapidly converted to photons through stimulated emission which leads to a reduction of the carrier density and thus the optical gain. Due to this back-and-forth between elevated carrier and photon densities we observe a damped oscillation called *Relaxation Oscillation*. Finally, when the laser is turned off the photon density decays rapidly, following an exponential law, because the carriers are quickly removed through stimulated emission.

It is possible to transform the laser rate equations into an equivalent electrical circuit in which the carrier density and photon density are both represented by voltages. Such a circuit can be combined with the large-signal AC circuit in Fig. 7.7(b) permitting to simulate the optical output signal with a circuit simulator such as SPICE [Tuc85, MKD97].

Turn-On Delay and Extinction Ratio. From this brief description it should be intuitively clear that adverse effects such as turn-on delay, turn-on delay jitter, and overshoot can be reduced by starting the laser-current pulse at a current above zero (see Fig. 7.9(b)). If we bias the laser close to the threshold current, the carrier density is already built up and the net optical gain is close to unity, yet the photon density is still low. With just a little bit more current the laser will turn on quickly. A simple equation describes the TOD as follows [Sze81]:

$$t_{TOD} = \tau_c \cdot \ln \frac{I_{L,\text{on}} - I_B}{I_{L,\text{on}} - I_{TH}} \tag{7.8}$$

where τ_c is the carrier lifetime with a typical value of 3 ns, I_B is the off-current or bias current, and $I_{L,\text{on}}$ is the on-current. For zero bias current ($I_B = 0$) and the typical numbers $I_{L,\text{on}} = 25\,\text{mA}$ and $I_{TH} = 10\,\text{mA}$ we obtain a TOD of about 1.5 ns. Alternatively, if we bias the laser at $I_B = 10\,\text{mA}$ the TOD goes to zero. Note that lasers with a low threshold current have a small turn-on delay, even if they are not biased. In practical systems, laser are almost always biased, except in some low-speed transmitters (155 Mb/s and below) where it is possible to predistort the electrical signal to compensate for the TOD.

Biasing not only helps with the TOD but also reduces the turn-on delay jitter, the amplitude of the relaxation oscillations, and with it the frequency chirp. With zero bias current, the peak-to-peak turn-on delay jitter is typically around 0.3 ns.

However, there is one downside to biasing the laser near the threshold current: it lowers the extinction ratio. Even if the laser is not lasing at $I_B = I_{TH}$ it still produces a small amount of incoherent light with the power P_0 as shown in Fig. 7.9(b). To improve their speed lasers are often biased a little bit above I_{TH}, further reducing the ER. Typical ER values for high-speed semiconductor lasers are in the range $9 - 14\,\text{dB}$ [FJ97].

Modulation Bandwidth. The laser's modulation bandwidth determines the maximum bit rate for which the laser can be used. This bandwidth is closely related to the relaxation oscillations which depend on the interplay between the carrier density and the photon density described by the rate equations. The small-signal modulation bandwidth for $I_B > I_{TH}$ can be derived from the rate equations [Agr97] and is proportional to

$$BW \sim \sqrt{I_B - I_{TH}}. \tag{7.9}$$

Thus the higher the laser is biased above the threshold current, the faster it gets. For this reason, in high-speed transmitters the laser is biased as much above the threshold as allowed by the ER specification.

In contrast a LED's speed is mostly determined by the carrier lifetime τ_c which is a weak function of the bias current. The 3-dB modulation bandwidth of a LED is $BW = \sqrt{3}/(2\pi\tau_c)$ [Agr97].

Chirp. In Section 7.1 we introduced the chirp parameter α which describes the transmitter's spurious FM modulation. More precisely, the α parameter relates the frequency

excursion Δf to the change in optical output power P_{out} as follows (for direct or external modulation) [Koc97]:

$$\Delta f(t) \approx \frac{\alpha}{4\pi} \cdot \frac{d}{dt} \ln P_{\text{out}}(t). \tag{7.10}$$

With a positive value for α, as is the case for directly modulated lasers, the frequency shifts slightly towards the blue when the output power is rising (leading edge) and towards the red when the power is falling (trailing edge). When the laser current is modulated the optical gain varies as desired, but, as a side effect, the refractive index of the gain medium also varies slightly which leads to the unwanted chirp and $\alpha \neq 0$.

From Eq. (7.10) we can see that slowing the rise/fall times of the laser-current pulses, and thus lowering the rate of change of P_{out}, can help to reduce the amount of chirp [SA86]. We can further conclude that large relaxation oscillations exacerbate the chirping problem.

Noise. The stimulated emission of photons in the laser produces a coherent electromagnetic field. However, occasional spontaneous emissions add amplitude and phase noise to this coherent field. The results are a broadening of the (unmodulated) spectral linewidth and fluctuations in the intensity. The latter effect is known as *Relative Intensity Noise* (RIN). A photodetector receiving laser light with intensity noise produces a corresponding electrical RIN noise. To first order, the electrical RIN-noise power is proportional to the received signal power:

$$\overline{i^2_{n,RIN}} = RIN \cdot I^2_{PIN} \cdot BW. \tag{7.11}$$

This means that the received SNR due to RIN noise is fixed at $1/(RIN \cdot BW)$ for a CW laser, and cannot be improved by transmitting more power. Note that this is very different from the situation we had for the detector noise where $\overline{i^2_{n,PD}}$ was proportional to I_{PD}. For example, with the value $RIN = -135\,\text{dB/Hz}$, the SNR due to RIN noise is 35 dB in a 10 GHz bandwidth. For the reception of a digital NRZ signal this is more than enough and RIN noise can often be neglected in the analysis of digital optical transmission systems.

The effect of RIN noise on digital transmission can be quantified, as usual, with a power penalty. The RIN noise adds to the receiver noise reducing the receiver's sensitivity. This means that we need to transmit more power to achieve the same BER. The power penalty due to RIN noise is [Agr97]:

$$PP = \frac{1}{1 - \mathcal{Q}^2 \cdot RIN \cdot BW}. \tag{7.12}$$

With the example values $1/(RIN \cdot BW) = 35\,\text{dB}$ and $BER = 10^{-12}$ ($\mathcal{Q} = 7$) the power penalty is only 0.068 dB ($PP = 1.016$). However, if the SNR due to RIN noise approaches $\mathcal{Q}^2 = 16.9\,\text{dB}$, the power penalty becomes infinite. This can be explained as follows: We know from Eq. (4.12) that to receive an NRZ signal with much more noise on the ones than the zeros (as is the case for RIN noise) we need an SNR of at least $1/2 \cdot \mathcal{Q}^2$. The SNR due to RIN noise of the *unmodulated* optical signal is $1/(RIN \cdot BW)$, as given by Eq. (7.11). NRZ modulation reduces the signal power by a factor four and the noise power by a factor two, so the SNR due to RIN noise of the *modulated* signal becomes $1/(2 \cdot RIN \cdot BW)$. Since the transmitted SNR must be higher than the required SNR at the receiver we need $1/(RIN \cdot BW) > \mathcal{Q}^2$ in accordance with Eq. (7.12).

In multi-mode lasers, such as FP lasers, as well as single-mode lasers with an insufficient *Mode-Suppression Ratio* ($< 20\,\text{dB}$) another type of noise called *Mode-Partition Noise* (MPN) is of concern. This noise in caused by power fluctuations among the various modes (mode competition) and is harmless by itself as the *total* intensity remains constant. However, chromatic dispersion in the fiber can desynchronize the mode fluctuations and turn them into additional RIN noise. Furthermore, a weak optical reflection of the transmitted light back into the laser can create additional modes further enhancing the RIN noise. These RIN-noise enhancing effects may be strong enough to reduce the received SNR below the critical value of $1/2 \cdot \mathcal{Q}^2$. This situation manifests itself in a so-called *Bit-Error Rate Floor*, i.e., a minimum BER that cannot be reduced irrespective of the transmitted power [Agr97].

RIN noise is very important when transmitting *analog* signals. For example, HFC/CATV systems using AM-VSB modulation (cf. Chapter 1) are typically limited by RIN noise. The reason for this is that the receiver requires a very high *Carrier-to-Noise Ratio* (CNR) of typically more than 50 dB (cf. Section 4.2). This high CNR can be achieved by transmitting at a high optical power such that the receiver gets a signal of around 0 dBm(!). In this way, the CNR due to the detector and amplifier noise can be made arbitrarily small, but the CNR due to the RIN noise of the laser remains constant and ultimately becomes the dominant noise source.

Let's calculate the CNR for an optical source modulated with a sine-wave signal (RF carrier) with an amplitude m times the unmodulated optical signal where m is known as the *Modulation Index*. The RF signal power is $1/2 \cdot m^2 \cdot I^2_{PIN}$ while the noise power is $RIN \cdot I^2_{PIN} \cdot BW$, as in Eq. (7.11), as long as m is small. Thus the CNR is:

$$CNR = \frac{m^2}{2 \cdot RIN \cdot BW}. \tag{7.13}$$

With the typical numbers $RIN = -135\,\text{dB/Hz}$, $BW = 6\,\text{MHz}$ and $m = 10\,\%$ we obtain $CNR = 44.2\,\text{dB}$. This may not be sufficient for analog TV transmission and a laser with less than $-135\,\text{dB/Hz}$ RIN noise must be chosen.

Reliability. The *Mean-Time To Failure* (MTTF) of a semiconductor laser is between 1 and 10 years if continuously operated at 70 °C. When cooled down to room temperature this lifetime extends by about a factor 10× to between 10 and 100 years [Shu88]. For comparison, LEDs are much more reliable and, at room temperature, have a MTTF of more than 1000 years.

Laser reliability may not look so great but it is a huge improvement over the early lasers. The first "continuous-wave", semiconductor laser that could operate at room temperature was demonstrated by researchers at the *Ioffe Physical Institute* in Leningrad in 1970. However, the lifetime of these GaAs lasers was measured in seconds. Not very much of a continuous wave indeed! It took about 7 years of tedious work until *AT&T Bell Laboratories* could announce a diode laser with a 100 year (million hour) lifetime. [Hec99]

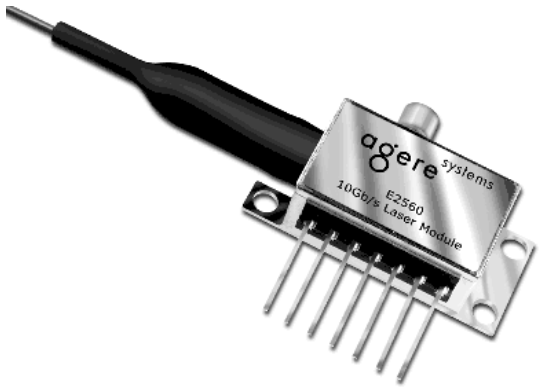

Figure 7.10: 10 Gb/s EML module, containing a DFB laser and an EA modulator, with a GPO connector for the RF signal, 7-pin connector for power-supply and control, and a single-mode pigtail (2.6 cm × 1.4 cm × 0.9 cm).

Figure 7.11: 40 Gb/s Lithium-Niobate Mach-Zehnder modulator with dual-drive inputs (V-type connectors), DC bias electrodes (pins), and two polarization-maintaining fiber (PMF) pigtails (12 cm × 1.5 cm × 1.0 cm).

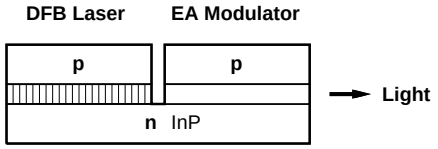

Figure 7.12: Integrated laser and electroabsorption modulator (schematically).

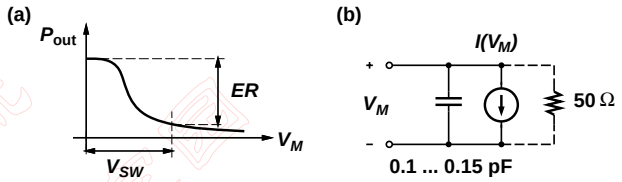

Figure 7.13: (a) Switching curve and (b) electrical equivalent circuit of an electroabsorption modulator.

7.3 Modulators

Two types of modulators are commonly used in communication systems: the *Electroabsorption Modulator* (EAM) and the *Mach-Zehnder Modulator* (MZM). The EAM is small and can be integrated with the laser on the same substrate, while the MZM is much larger but features superior chirp and ER characteristics. A laser combined with an EAM is known as an *Electroabsorption Modulated Laser* (EML). See Figs. 7.10 and 7.11 for photos of a packaged EML module and a MZ modulator, respectively.

In the following we will briefly describe each of the two modulators and summarize their main characteristics. For more information on modulators and their properties see [Koc97, HKV97, AT&95, Luc98, Luc00].

Electroabsorption Modulator. Figure 7.12 shows a schematic picture of an EML which consist of a DFB laser on the left and an EAM on the right. Both devices are monolithically integrated on the same InP substrate. The EAM consists of an active region sandwiched inside a reverse biased p-n junction. Without bias voltage, the bandgap of the active region is just wide enough to be transparent at the wavelength of the laser light. When a large enough reverse bias is applied across the p-n junction the effective bandgap is slightly reduced and the active region becomes opaque.

Figure 7.13(a) shows the optical output power of the EAM as a function of the applied reverse voltage V_M, the so-called switching curve. The voltage for switching the modulator from the on state to the off state, V_{SW} is typically in the range $2-4$ V.

The dynamic ER of the EAM is in the range $11-13$ dB [FJ97] and the magnitude of the chirp parameter α is typically smaller than one [HKV97]. A small on-state (bias) voltage ($0-1$ V) is often applied to minimize the modulator chirp [Koc97, Luc00].

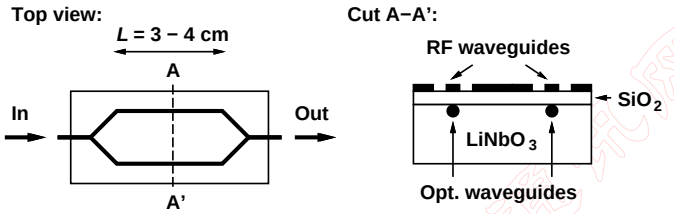

Figure 7.14: Dual-drive Mach-Zehnder modulator (schematically).

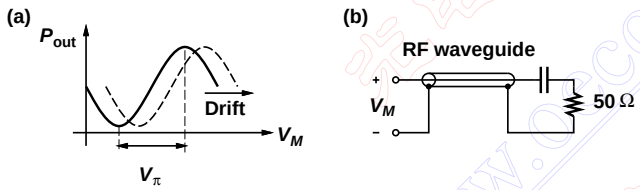

Figure 7.15: (a) Switching curve and (b) electrical equivalent circuit of a Mach-Zehnder modulator.

From an electrical point of view, the EAM is a reverse biased diode and, when the CW laser is off, presents a capacitive load of about $0.1-0.15\,\mathrm{pF}$ to the driver (see Fig. 7.13(b)). When the laser is on, the photons absorbed in the EAM generate a current pretty much like in a photodiode. This current is a function of how much light gets absorbed in the EAM and thus is also a function of V_M as indicated by $I(V_M)$ in Fig. 7.13(b). As a result the capacitive load appears shunted by a *nonlinear* resistance which has a high value when the EAM is completely turned on or off, but assumes a low value during the transition [KSC+98, Ran01].

Some EAMs are packaged together with a $50\,\Omega$ parallel resistor to match the impedance of the EAM to that of the transmission line between the driver and the modulator. However, it is difficult to achieve good matching (S_{11} parameter) over all voltage and frequency conditions. At low frequencies the nonlinear resistance spoils the match, while at high frequencies the shunt capacitance causes most of the S_{11} degradation.

Mach-Zehnder Modulator. A schematic picture of a MZ modulator is shown in Fig. 7.14. The incoming optical signal is split half and half and guided through two different paths. The delay (phase shift) in each path is controlled electrically. An RF waveguide (usually a coplanar transmission line) produces an electrical field which changes the refractive index of the optical waveguide. The latter is typically made from Titanium (Ti) diffused into Lithium Niobate ($LiNbO_3$) which is why these modulators are also known as *Lithium Niobate Modulators*. The two optical paths are then recombined where the electromagnetic fields interfere with each other. If the phase shift is 0°, the interference is constructive and the light intensity is high (on state); if the phase shift is 180° the interference is destructive and the light intensity is zero (off state).

The modulator shown in Fig. 7.14 is know as a *Dual-Drive MZM*, because each light path is controlled by a separate RF waveguide. Dual-drive MZMs can be driven in a push-pull fashion resulting in an excellent chirp performance. Alternatively, it is possible to control both light paths with a *single* RF waveguide, this arrangement is known as a *Single-Drive MZM*.

Figure 7.15(a) shows the optical output power of the MZM as a function of the voltage V_M applied to the RF waveguide. In the case of a dual-drive MZM, V_M is the differential voltage between the two arms. The theoretical switching curve is proportional to:

$$P_{\mathrm{out}} \sim \cos^2\left(\frac{\pi}{2}\cdot\frac{V_M}{V_\pi}\right) \tag{7.14}$$

where V_π is the switching voltage. In real MZMs the switching curve is shifted horizontally due to a delay mismatch in the two optical paths. Even worse, the horizontal shift is temperature and age dependent which is known as *Drift*. For this reason, MZMs require a bias controller which produces a bias voltage to compensate for this drift. The bias voltage can be added to the RF signal with a bias tee, alternatively, some MZMs provide separate electrodes for the RF signal and for the bias voltage [Luc98].

The switching voltage, V_π, is typically in the range $4-6\,\mathrm{V}$. A dual-drive MZM can be switched with half this voltage, $V_\pi/2$, on each arm, if operated in a push-pull fashion. The switching voltage of an MZM is inversely proportional to the length of the optical paths L. In other words, the product $V_\pi \cdot L$ is constant and has a typical value of $14\,\mathrm{Vcm}$ [HKV97]. For example, if we were to make the modulator 14 cm long, we could switch it with just 1 V, but such a long modulator would be very slow. The speed of a MZM depends inversely on its length L and the speed mismatch of the electrical and optical waves. For this reason high-speed modulators are short and require a high switching voltage. The large voltage swing needed for MZMs makes the design of high-speed drivers very challenging. The transmission line impedance is usually around $50\,\Omega$ and the end of each transmission line is terminated to avoid reflections as shown in Fig. 7.15(b).

The dynamic ER of a MZM is in the range $15-17\,\mathrm{dB}$ and the chirp parameter can be made very low $|\alpha| < 0.1$ [HKV97]. In fact, with a dual-drive MZM the chirp parameter can be controlled with the driving voltages [AT&95]:

$$\alpha = \frac{v_{M1}+v_{M2}}{v_{M1}-v_{M2}} \tag{7.15}$$

where v_{M1} and v_{M2} are the voltages on the first and second arm, respectively. We can see that if the two arms are driven with signals of the same amplitude but opposite phase (push-pull fashion), the chirp theoretically becomes zero. With a dual-drive MZM it is even possible to produce a *negative* chirp which leads to an initial *compression* rather than expansion of the optical pulse.[4] In practice, this effect can be used to achieve even longer transmission distances than with zero chirp.

[4] This effect is not described by our approximate Eqs. (2.4) and (7.2) and a more sophisticated theory is needed (e.g., see [Agr97]).

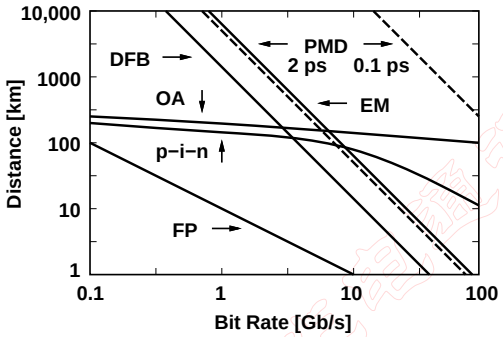

Figure 7.16: Limits due to chromatic dispersion (FP, DFB, EM), polarization-mode dispersion (PMD, dashed), and attenuation (p-i-n, OA) in a SMF link operated at the 1.55 μm wavelength

A dual-drive MZM can also be used to modulate the *phase* of the electromagnetic field in addition to the intensity. One interesting use of this feature is the transmission of a three-level signal known as *Optical Duobinary* modulation. The three levels are: (i) light on with no phase shift, (ii) light off, and (iii) light on with 180° phase shift. This signal has the interesting property that its spectral linewidth is half that of a two-level NRZ signal, lowering the dispersion penalty, yet it can be received with a simple two-level intensity detector [YKNS95].

The drawbacks of MZMs, besides the drift problem, the high switching voltage, and the rather large size (see Fig. 7.11), are their high optical insertion loss ($6 - 9$ dB) and their sensitivity to the polarization of the optical signal.

7.4 Limits in Optical Communication Systems

By combining results from Chapters 2, 5, and 7, we can now understand all major limits in an optical communication system: (i) the limits due to chromatic dispersion, (ii) the limits due to polarization-mode dispersion, and (iii) the limits due to attenuation.

Chromatic Dispersion Limits. For a transmitter that produces clean transform-limited pulses (e.g., with external modulation) we can approximate the maximum, dispersion-limited transmission distance by combining Eqs. (2.4), (2.6), and (7.1) [HLG88]:[5]

$$L < \frac{c}{2|D| \cdot \lambda^2 \cdot B^2}. \tag{7.16}$$

[5]This is only an approximation because Eq. (2.4) is not precisely valid for narrow-linewidth sources such as a DFB laser.

For example, given an externally modulated 1.55 μm laser without chirp transmitting over a SMF, we find from Eq. (7.16) that a 2.5 Gb/s system is limited to a span length of about 590 km and a 10 Gb/s system to about 37 km. A more precise analysis reveals that the attainable distances are even longer than that. A useful engineering rule states that we incur a 1-dB penalty after the distance [Koc97]:

$$L < \frac{17\,\text{ps/(nm}\cdot\text{km)}}{|D|} \cdot \frac{6000\,(\text{Gb/s})^2}{B^2}\ \text{km}. \tag{7.17}$$

According to this rule, a 2.5 Gb/s system is limited to about 960 km and a 10 Gb/s system to about 60 km distance. Note that in Eqs. (7.16) and (7.17) the transmission distance diminishes proportional to B^2. This is so because for a narrow-linewidth source, the linewidth increases with B (Eq. (7.1)) while simultaneously the allowed spreading amount decreases with B (Eq. (2.6)). So we are faced with two problems when going to higher bit rates and hence the B^2.

For a narrow-linewidth transmitter that produces a significant amount of chirp (e.g., a directly modulated DFB laser) the dispersion-limited transmission distance can be approximated by combining Eqs. (2.4), (2.6), and (7.2). Compared to Eq. (7.16), it is reduced by $\sqrt{\alpha^2+1}$:

$$L < \frac{c}{2|D| \cdot \lambda^2 \cdot \sqrt{\alpha^2+1} \cdot B^2}. \tag{7.18}$$

Again, a more precise analysis finds a somewhat longer distance and the following engineering rule applies [Koc97]:

$$L < \frac{1}{\sqrt{\alpha^2+1}} \cdot \frac{17\,\text{ps/(nm}\cdot\text{km)}}{|D|} \cdot \frac{6000\,(\text{Gb/s})^2}{B^2}\ \text{km}. \tag{7.19}$$

For example, given a directly modulated 1.55 μm DFB laser with $\alpha = 4$ transmitting over a SMF, we find that a 2.5 Gb/s system is limited to a span length of about 230 km and a 10 Gb/s system to about 15 km.

For a transmitter with a wide-linewidth source (e.g., a FP laser or LED) the dispersion-limited transmission distance can be found by combining Eqs. (2.4), (2.6) and (7.3) [HLG88]:

$$L < \frac{1}{2|D| \cdot \Delta\lambda_S \cdot B}. \tag{7.20}$$

For example, given a 1.55 μm FP laser with a 3-nm linewidth transmitting over a SMF, we find that a 2.5 Gb/s system is limited to a span length of about 4 km and a 10 Gb/s system to about 1 km. Note that in Eq. (7.20) the transmission distance diminishes proportional to B rather than B^2. This is so because in this case the linewidth is not affected by the bit rate.

The numerical results of the examples are summarized in Table 7.1. The bit-rate dependence for all three examples (FP, DFB, and EM) is plotted in Fig. 7.16. Note that all examples are based on a dispersion parameter equal to 17 ps/(nm · km). However, this value can be reduced by one of the following methods: (i) use a dispersion-compensation technique such as splicing in a fiber segment with negative dispersion (DCF), (ii) use a dispersion-shifted fiber (DSF or NZ-DSF), or (iii) operate the system at the 1.3 μm wavelength where dispersion is lower (but loss is higher).

Transmitter Type	2.5 Gb/s	10 Gb/s
Fabry-Perot Laser ($\Delta\lambda = 3$ nm)	4 km	1 km
Distributed Feedback Laser ($\alpha = 4$)	230 km	15 km
External Modulator ($\alpha = 0$)	960 km	60 km

Table 7.1: Maximum (unrepeatered) transmission distances over a SMF at 1.55 μm for various transmitter types based on Eqs. (7.20), (7.19), and (7.17) with $D = 17\,\mathrm{ps/(nm \cdot km)}$.

PMD Limit. By combining Eq. (2.2) which describes the pulse spreading due to PMD with the rule that we should make $\Delta T < 0.14/B$ to keep the power penalty due to PMD below 1 dB (cf. Section 2.2) we find the PMD-limited transmission distance:

$$L < \frac{(0.14)^2}{D_{PMD}^2 \cdot B^2}. \tag{7.21}$$

For example, given the parameter $D_{PMD} = 0.1\,\mathrm{ps}/\sqrt{\mathrm{km}}$ characteristic for new PMD optimized fiber, we find that a 2.5 Gb/s system is limited to a span length of about 310,000 km and a 10 Gb/s system to about 20,000 km. These are huge distances capable of connecting any two points on the earth! However, older already deployed fiber with $D_{PMD} = 2\,\mathrm{ps}/\sqrt{\mathrm{km}}$ poses more serious limits: a 2.5 Gb/s system is limited to a span length of about 780 km and a 10 Gb/s system to about 50 km. These latter numbers are smaller than the dispersion-limited transmission distance for transform limited pulses.

The bit-rate dependence of the PMD limit for both values of the PMD parameter are plotted in Fig. 7.16 with dashed lines.

Attenuation Limit. Given the launch power $\overline{P}_{TX}$ at the transmitter end and the receiver sensitivity $\overline{P}_S$, we can easily derive that fiber attenuation is limiting the transmission distance to [HLG88]:

$$L < \frac{10}{a} \cdot \log\left(\frac{\overline{P}_{TX}}{\overline{P}_S}\right) \tag{7.22}$$

where a is the fiber attenuation of fiber in dB/km. We can rewrite this equation in the more practical and intuitive form:

$$L < \frac{\overline{P}_{TX}[\mathrm{dBm}] - \overline{P}_S[\mathrm{dBm}]}{a}. \tag{7.23}$$

For example, given a fiber attenuation of 0.25 dB/km, typical for SMF at 1.55 μm, a launch power of 1 mW (0 dBm), and a 10 Gb/s p-i-n detector with a sensitivity of -18.5 dBm, we find the attenuation-limited transmission distance of 74 km. The bit-rate dependence of the attenuation limit for a p-i-n receiver (p-i-n) as well as an optically preamplified p-i-n receiver (OA) is plotted in Fig. 7.16 based the sensitivity data from Fig. 5.12. We can see that this limit is much less dependent on the bit rate than the other limits which can be explained by the log function in Eq. (7.22).

Note that this attenuation limit can be overcome with periodically spaced in-line optical amplifiers.

7.5 Summary

Two types of modulation are used in optical transmitters:

- Direct modulation where the laser current is directly modulated by the signal.
- External modulation where the laser is always on and a subsequent electrooptic modulator is used to modulate the laser light with the signal.

In general, external modulation produces higher quality optical signals with a narrower spectral linewidth and a higher extinction ratio, but is more costly and bulky than direct modulation.

The light sources which are commonly used in optical transmitters are:

- The FP laser which is low in cost, but has a wide linewidth which severely limits the transmission distance in a dispersive fiber.
- The DFB laser which has a very narrow linewidth is a good source for external modulator. When directly modulated its linewidth broadens due to chirp.
- The VCSEL which is low in cost is mostly used in data communication systems operating at short wavelengths.
- The LED which is very low in cost, but has a very wide linewidth, low output power, and small modulation bandwidth is mostly used for low-speed, short-range data communication.

Electrically, all four sources are forward biased p-n junctions.

The modulators which are commonly used are:

- The EA modulator which is small in size and can be driven with a reasonably small voltage swing. Electrically, it is a reverse-biased p-n junction.
- The MZ modulator which generates the highest-quality optical pulses with a controlled amount of chirp and high extinction ratio. Electrically, it is a terminated transmission line.

The maximum transmission distance which can be achieved in an optical communication system is determined by a combination of the chromatic dispersion limit, PMD limit, and attenuation limit.

Chapter 8

Laser and Modulator Drivers

8.1 Driver Specifications

Before examining laser- and modulator-driver circuits, we will discuss their main specifications: the modulation and bias current range (for laser drivers), the voltage swing and bias voltage range (for modulator drivers), rise and fall time, pulse-width distortion, and jitter generation. In this and the following sections we focus on laser drivers for digital applications (NRZ or RZ modulation) only.

8.1.1 Modulation and Bias Current Range

Definition. The *Bias Current* I_B is the DC current supplied by the driver to the laser when transmitting a zero. The *Modulation Current* I_M is the current added to the bias current when transmitting a one. Therefore, for a zero bit the laser current is I_B and for a one bit the current is $I_B + I_M$ as illustrated in Fig. 8.1. Note that for AC-coupled laser drivers the bias current is often redefined as the *average* current into the laser ($I_B + I_M/2$).

The bias and modulation currents of a laser driver are controlled either manually with trim pots, or automatically with the feedback signal from the monitor photodiode. Typically, the bias current is controlled automatically with an APC circuit, while the

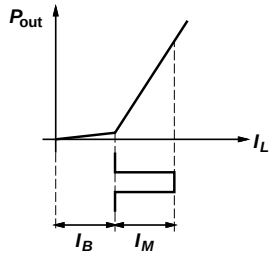

Figure 8.1: DC parameters of a laser driver.

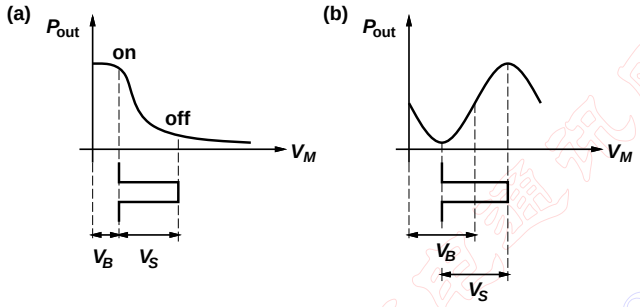

Figure 8.2: DC parameters of (a) EAM and (b) MZM driver.

modulation current is set with a trim pot.

Typical Values. The modulation-current range must be large enough to permit the required optical output power (e.g., 0 dBm) with the desired laser types under high-temperature and end-of-life conditions. A typical range seen in commercial 2.5 and 10 Gb/s laser drivers for uncooled lasers is:

$$I_M = 5 \ldots 100\,\text{mA}. \tag{8.1}$$

Similarly, the bias-current range must be large enough to include the threshold current I_{TH} of the desired laser types under high-temperature and end-of-life conditions. A typical range seen in commercial 2.5 and 10 Gb/s laser drivers for uncooled lasers is:

$$I_B = 0 \ldots 100\,\text{mA}. \tag{8.2}$$

8.1.2 Voltage Swing and Bias Voltage Range

Definition. The peak-to-peak *Voltage Swing* V_S is the difference between the on-state and off-state voltage supplied by the modulator driver. In the case of the EAM driver, the *Bias Voltage* V_B (a.k.a. *DC Offset Voltage*) is the voltage supplied by the driver when the optical signal is on. In the case of the MZM driver, the *Bias Voltage* V_B is the average voltage (DC component) supplied by the driver. See Figure 8.2 for an illustration of V_S and V_B together with the switching curves of the EAM and MZM.

For an EA modulator, the voltage swing must be equal or larger than the modulator's switching voltage V_{SW}. The bias voltage is either zero or a small positive value (diode reverse voltage) used to optimize the chirp parameter α.

For an MZ modulator, the voltage swing must be set very close to the switching voltage V_π (or $V_\pi/2$ on each arm for a dual-drive MZM operated in push-pull mode). Due to the sinusoidal switching curve (see Fig. 8.2(b)), the ER degrades if the voltage swing is smaller or larger than V_π; these conditions are known as under- or over-modulation, respectively.

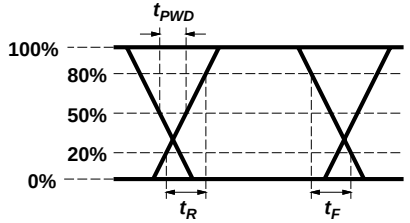

Figure 8.3: Eye diagram and AC parameters of laser/modulator driver.

However, a small amount of over-modulation is sometimes used to improve the rise/fall time of the optical signal. The optimum bias voltage is at the so-called *Quadrature Point* of the switching curve. Due to path mismatch and drift this voltage is not known a priori and the bias voltage range has to span at least one full period, i.e., it has to be at least $2 \cdot V_\pi$. Note that if the drift is larger than one period, the bias voltage can be reduced by $2 \cdot V_\pi$ because the switching curve is periodic with $2 \cdot V_\pi$. Usually, an automatic bias controller is used to generate V_B such that it is located at the quadrature point and tracks drift automatically.

Typical Values. Typical single-ended, peak-to-peak output swings seen in commercial 2.5 and 10 Gb/s modulator drivers are:

$$\text{EAM Driver:} \quad V_S = 0.2 \ldots 3\,\text{V} \tag{8.3}$$
$$\text{MZM Driver:} \quad V_S = 0.5 \ldots 5\,\text{V}. \tag{8.4}$$

Typical bias voltage ranges are:

$$\text{EAM Driver:} \quad V_B = 0 \ldots 1\,\text{V} \tag{8.5}$$
$$\text{MZM Driver:} \quad V_B = 0 \ldots 10\,\text{V}. \tag{8.6}$$

8.1.3 Rise and Fall Time

Definition. The *Rise Time* and *Fall Time* of a laser/modulator driver output signal can be measured in the electrical or optical domain. To display the signal waveform in the optical domain, an optical-to-electrical (O/E) converter is used in front of the oscilloscope. Usually, rise time t_R is measured from the point where the signal has reached 20 % to the point where it reaches 80 % of its full value. Fall time t_F is measured similarly going from 80 % down to 20 %. However, some manufacturers use 10 % and 90 % as measurement conditions and one has to be careful when comparing specifications of different products. In case the signal exhibits over- or undershoot, the 0 % and 100 % values correspond to the steady state values, *not* the peak values. See Fig. 8.3 for an illustration of rise and fall time in the eye diagram.

The rise and fall times must both be shorter than one bit interval, or else the full value of the transmitted signal is not reached for a "01010101..." pattern resulting in ISI and vertical eye closure. It is recommended for an NRZ system that the total system rise-time is kept below 0.7 UI [Agr97]. The driver rise-time t_R is just one component of the system *Rise-Time Budget* which also includes the fiber rise-time and the receiver rise-time (added in the square sense). Therefore, the driver rise-time must be made significantly shorter than 0.7 UI. However, in laser drivers the rise time should not be made shorter than necessary, because the optical chirp (Δf in Eq. (7.10)) increases for very short rise times [SA86].

Typical Values. Typical values of electrical rise/fall times seen in commercial 2.5 and 10 Gb/s laser/modulator drivers are:

$$2.5\,\text{Gb/s}: \quad t_R, t_F < 100\,\text{ps}\ (< 0.25\,\text{UI}) \tag{8.7}$$
$$10\,\text{Gb/s}: \quad t_R, t_F < 40\,\text{ps}\ (< 0.40\,\text{UI}). \tag{8.8}$$

8.1.4 Pulse-Width Distortion

Definition. An offset or threshold error in the driver circuit may lengthen or shorten the electrical output pulses relative to the bit interval. Furthermore, turn-on delay in the laser may shorten the optical pulses. This kind of pulse lengthening or shortening is known as *Pulse-Width Distortion* (PWD) and can be measured in the electrical as well as the optical domain. The amount of PWD, t_{PWD}, is defined as the difference between the wider pulse and the narrower pulse divided by two. Figure 8.3 shows how t_{PWD} can be determined from the eye diagram. If the crossing point is vertically centered, t_{PWD} is zero. Under this condition the horizontal eye opening is maximized as well.

Many laser and modulator drivers contain a so-called *Pulse-Width Control* (PWC) circuit connected to an external trim pot to compensate for PWDs. The trim pot must be adjusted, with the desired laser/modulator connected to the driver, until the crossing point of the optical eye is centered.

Low PWD is desirable because it improves the horizontal eye opening. Furthermore, some CDRs use both edges for phase detection and therefore require them to be precisely aligned with the bit intervals to get an accurate sampling phase.

Typical Values. Typical values of electrical PWDs seen in commercial 2.5 and 10 Gb/s laser/modulator are:

$$2.5\,\text{Gb/s}: \quad t_{PWD} < 20\,\text{ps}\ (< 0.05\,\text{UI}) \tag{8.9}$$
$$10\,\text{Gb/s}: \quad t_{PWD} < 5\,\text{ps}\ (< 0.05\,\text{UI}). \tag{8.10}$$

The above numbers are for drivers without a pulse-width control circuit or with that feature disabled. If a PWC circuit is present, the adjustment range for t_{PWD} is specified. A typical range is ± 0.20 UI ($\pm 20\,\%$).

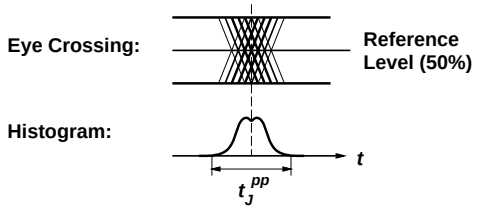

Figure 8.4: Jitter histogram with deterministic and random jitter.

8.1.5 Jitter Generation

One may think that in the absence of PWD the edges of the data signal are precisely aligned with the bit intervals. However, there is still a statistical uncertainty in the edge position known as *Timing Jitter*. As shown in Fig. 8.4 jitter can be measured in the (electrical or optical) eye diagram by computing a histogram of the time points when the signal crosses a reference voltage. This voltage is set to the eye-crossing point where the histogram has the tightest distribution; in the absence of PWD this is at the 50 % level. Many sampling oscilloscopes have the capability to calculate and display such histograms.

Jitter in the electrical output signal is caused by noise and ISI introduced by the driver circuit. Jitter already present in the driver's data or clock input signals will also appear at the output and must be subtracted out to obtain the driver's jitter performance. The optical output signal contains additional jitter produced by the laser (modulator), such as turn-on delay jitter.

Low output jitter is desirable because it improves the horizontal eye opening and improves the sampling precision in the CDR. Furthermore, in some types of *Regenerators* the clock signal recovered from the received optical signal is used to re-transmit the data. When cascading several such regenerators, jitter increases with every regenerator and to prevent excessive jitter accumulation at the end of the chain, very tough jitter specifications are imposed on each regenerator.

Definition. The jitter amount t_J can be specified in several ways. If the histogram is bounded, the peak-to-peak (t_J^{pp}) value is a good measure. If the histogram is Gaussian or has Gaussian tails, as in Fig. 8.4, the peak-to-peak measure is ambiguous because it depends on how many edges we measure. To be precise we need to specify a probability, such as 10^{-12}, that the jitter will be larger than t_J^{pp}. A BERT scan can be used to measure jitter relative to an error probability. If the histogram is purely Gaussian, the rms value (t_J^{rms}) is usually given. The corresponding peak-to-peak value is about six to eight times larger than the rms value. This rule-of-thumb relies on the fact that we "see" the $3-4\sigma$ value on a scope as the peak value. The peak-to-peak value for a probability 10^{-12} is $14 \cdot t_J^{rms}$ for reasons analogous to those given in Section 4.2.

We can distinguish two types of jitter: *Deterministic Jitter* and *Random Jitter*. A typical wideband jitter histogram, as shown in Fig. 8.4, contains both types of jitter. The inner part of the histogram is due to deterministic jitter, the Gaussian tails are due

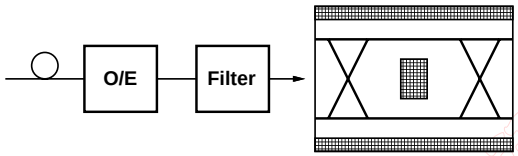

Figure 8.5: Eye-diagram mask for SONET OC-48.

to random jitter. Besides the *Wideband Jitter* discussed so far, *Narrowband Jitter* is measured in a restricted bandwidth, the so-called *Jitter Bandwidth*. A narrowband jitter measurement is required by some standards such as SONET.

For a more detailed discussion of jitter and its measurement see Section 4.9 and Appendix **??**.

Typical Values. The jitter generation limits prescribed by the SONET standard are:

$$2.5\,\text{Gb/s:} \qquad t_J^{rms} < 4\,\text{ps}\ (< 0.01\,\text{UI}) \tag{8.11}$$

$$10\,\text{Gb/s:} \qquad t_J^{rms} < 1\,\text{ps}\ (< 0.01\,\text{UI}). \tag{8.12}$$

These values are defined for a jitter bandwidth from 12 kHz to 20 MHz for 2.5 Gb/s and 50 kHz to 80 MHz for 10 Gb/s. This bandwidth is relevant, because high-frequency jitter outside this bandwidth is not passed on to the output of the regenerator (jitter-transfer specification) and therefore does not get accumulated in a chain of regenerators. Because the SONET jitter bandwidth is much lower than the bit rate, random jitter is the main contributor to the SONET jitter-generation numbers.

The laser/modulator driver's jitter specification must be much lower than the system limits given above because the driver is only one of several components contributing to the total jitter generation.

8.1.6 Eye-Diagram Mask Test

The so-called *Eye-Diagram Mask Test* checks the output signal for many impairments simultaneously such as slow rise/fall time, pulse-width distortions, jitter, ISI, ringing, noise, etc. In this test the (electrical or optical) eye diagram is compared to a mask which specifies regions inside and outside the eye which are off-limits to the signal. For example in Fig. 8.5 the signal must stay out of the hatched regions. The rectangle inside the eye diagram defines the required *Eye Opening*. The regions outside the eye diagram limit overshoot and undershoot.

If the signal contains Gaussian noise and/or Gaussian jitter, the mask will always be violated eventually, we just have to collect enough data samples. Therefore it is necessary to specify the time over which the eye diagram must be accumulated. Or, more precisely, we must specify a small probability, such as 10^{-12}, for which the hatched regions may be entered.

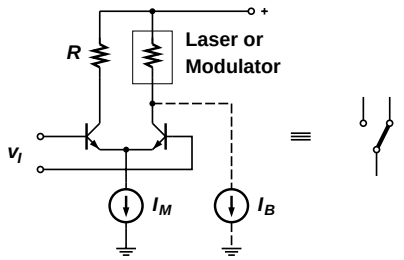

Figure 8.6: Current-steering stage for driving a laser or modulator.

Electrical eye-diagram mask tests are usually performed directly on the signal while optical eye-diagram mask tests require that the signal is first filtered to suppress effects which would not affect the receiver (e.g., relaxation oscillations from the laser will be attenuated). For example, the SONET standard requires that the optical signal is first passed through a 4th-order Bessel-Thomson filter with a 3-dB bandwidth equal to $0.75\,B$ before it is tested against the eye mask (cf. Fig. 8.5). The combination of O/E converter and filter is known as a *Reference Receiver*.

8.2 Driver Circuit Principles

In the following section, we will have a look at the design principles for laser and modulator drivers: How can we generate the necessary drive currents and voltages; how can we avoid reflections on transmission lines; how can we reduce PWD and jitter; how can we control the optical output power, and so on? Then in the next section we will examine concrete circuit implementations to illustrate these principles.

8.2.1 Current Steering

The output stage of most laser and modulator drivers are based on the *Current Steering* circuit shown in Fig. 8.6 for the case of a bipolar technology. Although this stage looks like a differential amplifier, it behaves more like a buffer from the *Current-Mode Logic* (CML) family. As indicated by the switch to the right of the circuit, the tail current I_M is completely switched either to the dummy load R on the left or the laser/modulator load on the right. For this reason this circuit is also known as a *Differential Current Switch*. The differential input-voltage swing (v_I) must be large enough to ensure full switching. If the driver needs to output a bias current or bias voltage, a current source I_B can be connected to the output of the stage as shown by the dashed lines in Fig. 8.6.

The current-steering stage of Fig. 8.6 can drive a variety of laser/modulator load configurations as illustrated in Fig. 8.7. For example, the load can be (a) a laser diode with or without a series resistor, (b) an EA modulator with a parallel resistor to convert the drive current into a voltage, (c) an AC-coupled laser diode with relaxed headroom

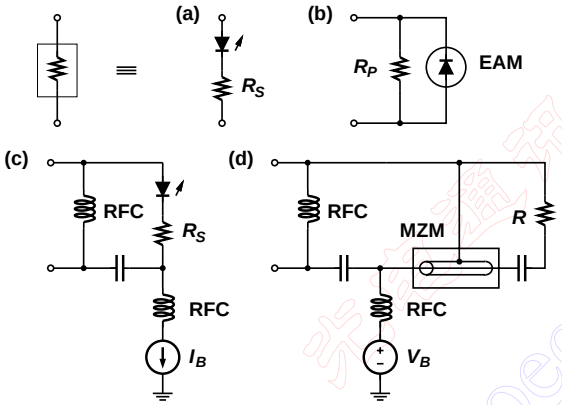

Figure 8.7: Load configurations for the current-steering stage: (a) laser, (b) EAM, (c) AC-coupled laser, and (d) MZM.

constraints compared to the configuration in (a), or (d) an AC-coupled MZ modulator. Separate biasing must be provided to the laser/modulator when AC coupling is used.

The advantages of the current-steering stage are similar to those well known from current-mode logic (CML):

- The differential design is insensitive to input common-mode noise and power/ground bounce. This is an important prerequisite to achieve low jitter. The differential design further ensures a low input offset voltage and prevents the associated pulse-width distortions (no reference voltage is required).

- Ideally, the power-supply current remains constant, i.e., it is always I_M ($+I_B$) regardless of the bit value transmitted. The tail current is either routed through the laser/modulator or dumped into the dummy load R. As a result, the generation of power and ground bounce is minimized. On the down side, power-dissipation is twice that necessary to drive the laser/modulator.

- The modulation current (for lasers) and the voltage swing (for modulators) can be controlled conveniently with the tail-current source I_M.

The purpose of the dummy load R is to improve the symmetry of the driver stage. An asymmetric load configuration can cause an input offset voltage and an undesirable modulation of the voltage across the tail-current source. This voltage modulation together with a parasitic tail-current source capacitance causes the drive current to over- and undershoot. Furthermore, a finite tail-current source resistance translates the voltage modulation into a current modulation enhancing the power-supply noise. In laser drivers

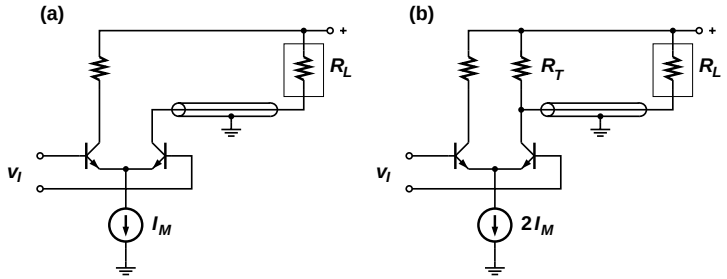

Figure 8.8: Current-steering stage for driving a laser or modulator through a transmission line: (a) without back termination (b) with passive back termination.

the dummy load can be a forward biased diode with a series resistor to match the laser load but often a simple resistor or sometimes just a short to the positive supply is used.

The current-steering stage connected directly to the laser/modulator load, as shown in Fig. 8.6, is useful for low-speed applications and in situations where "copackaging" can be used. The latter means that the laser/modulator is located in close proximity to the driver and that the two components are interconnected by means of short wire bonds or flip-chip bonds. In all other cases a transmission line between the driver and the laser/modulator is required.

8.2.2 Back Termination

When using a transmission line to connect the driver to the load undesirable reflections may occur at either end of the transmission line due to impedance mismatch. To avoid reflections from the load end of the transmission line back into the driver, the laser/modulator must be matched to the characteristic impedance of the transmission line. EA modulators can be matched to a 50 Ω transmission line with a 50 Ω parallel resistor. Laser diodes, which have a typical resistance of 5 Ω, can be matched to a 25 Ω transmission line with a 20 Ω series resistor, or to a 50 Ω transmission line with a 45 Ω resistor. For power reasons 25 Ω transmission lines are generally preferred for laser drivers: Driving 100 mA into a 25 Ω load dissipates 0.25 W (ignoring the laser's nonlinear I/V characteristics), whereas driving the same current into a 50 Ω load costs 0.5 W in power. Furthermore, a lower supply voltage can be used to drive 100 mA into a 25 Ω load than a 50 Ω load.

Open Collector/Drain. Figure 8.8(a) shows an open-collector current-steering stage which drives a load R_L (laser/modulator) matched to the impedance of the transmission line. The drawback of this simple arrangement is that it lacks a back termination. If the match between the load R_L and the transmission line is not perfect (e.g., in the case of an EAM load where R_L is bias dependent and cannot be matched well under all conditions),

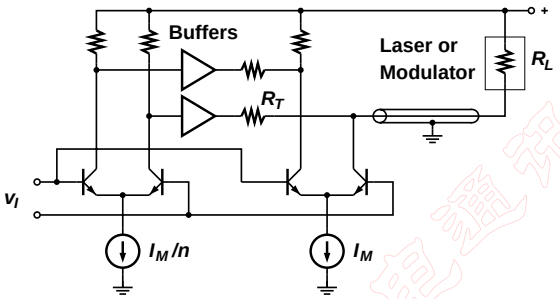

Figure 8.9: Current-steering stage with active back termination for driving a laser or modulator.

some part of the signal will be reflected back into the driver which is unable to absorb it. As a result, there will be a second reflection back to the load where it may degrade the extinction ratio and jitter performance.

Passive Back Termination. Figure 8.8(b) shows a simple extension of the previous circuit which incorporates a termination resistor R_T on the driver side. This modification will fix the problem with double reflections, but now the driver has to supply twice as much current as before: About half of the current reaches the load and performs a useful function, the other half gets "burned up" in the back termination resistor. When the load is a laser diode, even less than half of the current reaches the laser because of its nonlinear I/V characteristics.

Active Back Termination. Figure 8.9 shows a laser/modulator driver stage with *Active Back Termination* protecting against reflections without wasting an excessive amount of power [RSDG01]. Here the back termination resistor R_T is not connected to the positive supply, but to a *replica* of the intended driving signal voltage. The replica signal is generated with a scaled-down $(1/n)$ version of the current-steering stage followed by a voltage buffer (unity gain and zero output impedance assumed in Fig. 8.9) to drive the resistor R_T. Now, in the case of normal operation without reflections, the voltage drop across R_T is zero and no power is wasted. In case of reflections, the reflected signal appears only at the driver output (right side of R_T) but not at the replica output (left side of R_T) and thus gets absorbed by R_T.

When adding a bias current source I_B to the driver stage in Fig. 8.9 we have to draw a corresponding (scaled) current from the replica stage to avoid some of the bias current from getting dissipated in R_T.

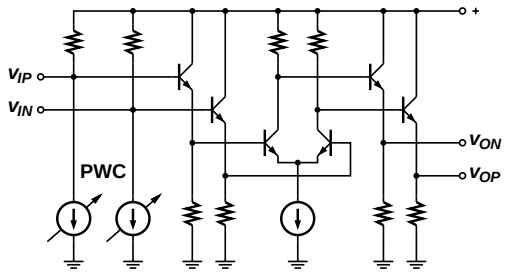

Figure 8.10: Pre-driver stage with pulse-width control.

8.2.3 Pre-Driver Stage

The transistors in the driver's output stage have to switch a current of around 100 mA and are therefore quite large. A so-called *Pre-Driver* stage is generally used to drive this heavy load. The pre-driver's input impedance is much higher than that of the output stage and can be driven easily from off-chip or another on-chip circuit block (e.g., a retiming flip-flop). The pre-driver has to provide enough voltage swing to the output stage to completely switch it but not too much to avoid distortions in the output signal. The pre-driver's output common-mode voltage and timing must be such that always one of the transistors in the output stage is on. If both transistors were to switch off momentarily, the voltage at the common-emitter node would drop and cause a current overshoot at the beginning of the next pulse.

The pre-driver can be implemented with another current-steering stage having smaller transistors and operating at a smaller tail current followed by emitter followers, as shown in Fig. 8.10. Alternatively, any one of the wideband amplifier stages discussed in Section 6.3.2 could be used as a pre-driver. Often a pulse-width controller to adjust the cross-over point in the eye diagram is integrated into the pre-driver. We will discuss this feature next.

8.2.4 Pulse-Width Control

Several effects lead to pulse-width distortions in the optical output signal (cf. Section 8.1.4). For example, the time to turn the laser current on may be different from the time to turn it off. Although the driver circuits are usually differential, the laser and EAM are single-ended devices breaking the symmetry. Furthermore, the laser may exhibit a turn-on delay or in the case of an EAM the nonlinear voltage-to-light characteristics may add distortions to the optical pulses.

These pulse-width distortions can be compensated by pre-distorting the signal in the pre-driver. Figure 8.11 shows how the pulse width can be controlled with a *voltage offset*. The top traces show a perfect signal without PWD. By offsetting the inverting and non-inverting signals the cross-over points can be shifted in time. Finally, the signal is regenerated with a limiter to produce a clean signal with the desired PWD.

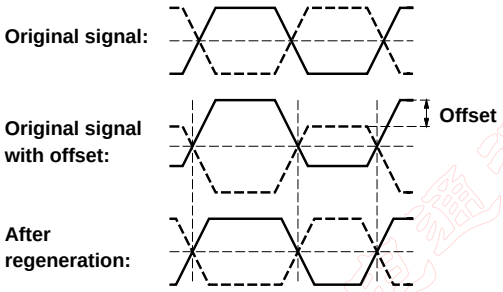

Figure 8.11: Operation of the pulse-width control circuit.

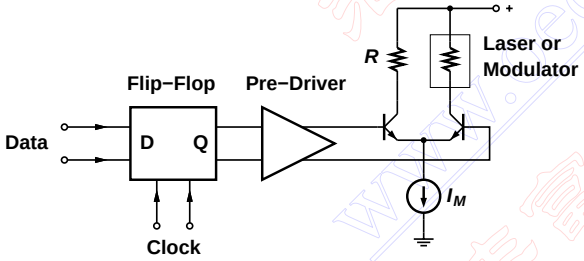

Figure 8.12: Block diagram of a laser/modulator driver with retiming flip-flop.

Figure 8.10 shows a simple way to implement this scheme with two adjustable current sources, labeled PWC, at the input of the pre-driver. These current sources produce the offset voltage and the following current-steering pair in the pre-driver acts as the limiter. Note, that the offset cannot be made too large in order for the current-steering pair to obtain enough signal to switch completely. A drawback of this scheme is that the PWD depends on the edge-rate of the signal at the input of the pre-driver which may not be well defined.

8.2.5 Data Retiming

High-speed laser/modulator drivers often contain a flip-flop located between the data input and the pre-driver, as shown in Fig. 8.12. The purpose of this flip-flop is to retime (resynchronize) the data signal with a precise clock signal. In this manner, pulse-width distortions and jitter in the input data signal are eliminated. However, jitter in the clock signal will be transferred undiminished to the output of the driver. It is therefore important that the clock source has very low jitter. For example, SONET compliant laser/modulator drivers require a clock source with less than 0.01 UI rms jitter.

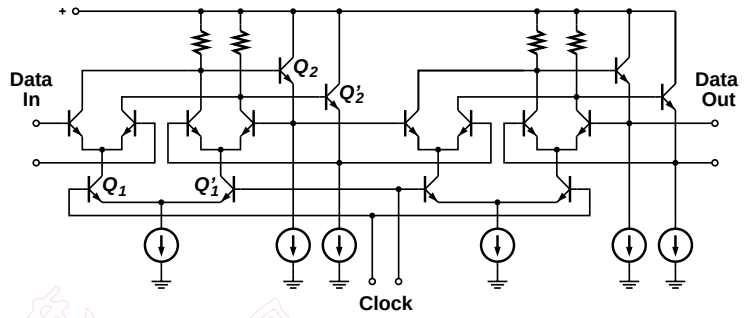

Figure 8.13: BJT/HBT implementation of the retiming flip-flop.

Retiming flip-flops are usually based on *Current-Mode Logic* (CML) because it is fast, insensitive to common-mode and power-supply noise, and produces little power/ground bounce. In bipolar technologies CML is also known as *Emitter-Coupled Logic* (ECL), in FET technologies it is known as *Source-Coupled FET Logic* (SCFL). Figure 8.13 shows a typical bipolar implementation of a CML (or ECL) flip-flop. This master-slave flip-flop consists of a cascade of two identical D latches. Using the left D latch as an example, its operation can be explained as follows: If Q_1 is turned on and Q_1' is turned off by the clock signal, the D latch acts like an amplifier passing the input logic state directly to the output. Conversely, If Q_1' is turned on and Q_1 is turned off, the D latch acts as a regenerator storing the previous logic state by means of positive feedback through Q_2 and Q_2'. In the latter case the output state is independent of the input state. The D-latch output is level-shifted and buffered with transistors Q_2, Q_2' such that the second D latch can be driven. The second D latch is clocked from the inverted clock such that when the first latch is in amplification mode, the second one is in regeneration mode and vice versa.

The flip-flop in Figure 8.13 requires a full-rate clock and only one clock edge is used for retiming. For very high bit-rate systems a retiming circuit operating from a half-rate clock while using both clock edges can be an interesting alternative [LTN+98].

8.2.6 Inductive Load

The rise/fall time of a current-steering stage can be improved by inserting small inductors in series with the load resistors. This technique can be applied to an output-driver stage with back termination (cf. Section 8.2.2), a pre-diver stage (cf. Section 8.2.3), or any CML circuit (e.g., the retiming flip-flop of Section 8.2.5). Since one end of the inductor is connected to the supply voltage it is possible to use a bond wire to realize it; this approach is frequently used for output-driver stages.

Figure 8.14 shows the principle of operation. At first we assume $L = 0$ and thus the load consists of the resistor R and the parasitic load capacitance C only. The tail current-source is switched by a step-like input signal. The $20-80\,\%$ rise/fall time of the

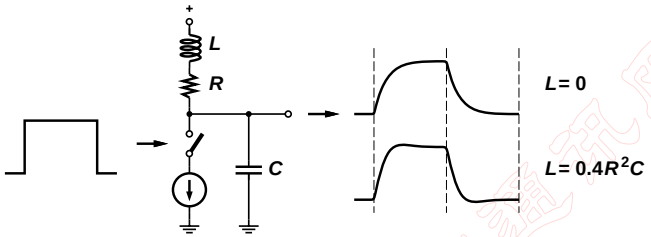

Figure 8.14: Improvement of the rise/fall time with an inductive load.

output pulse can be calculated as

$$t_R = t_F = \ln(0.8/0.2) \cdot RC = 1.39 \cdot RC \quad \text{for} \quad L = 0. \tag{8.13}$$

The output pulse for this situation is shown in the upper trace of Fig. 8.14. Next, if we insert an inductor with the optimum value $L = 0.4 \cdot R^2C$ in series with the load resistor, the rise/fall time improves by about 40 % to

$$t_R = t_F = 0.85 \cdot RC \quad \text{for} \quad L = 0.4 \cdot R^2C. \tag{8.14}$$

The output pulse for the latter situation is shown in the lower trace of Fig. 8.14. It can be seen that besides the improvement in rise/fall time there is also a slight over- and under-shoot (about 2.6 %) which is usually harmless.

An intuitive explanation of the edge-rate improvement is as follows: When the current source is switched on (falling edge at output) the inductor will at first act like an open and all of the tail current is used to discharge C, rather than some of it flowing into R. When the current source is switched off (rising edge) the inductor, which is "charged up" with current, charges C more rapidly than R alone. By now you have probably noticed that this circuit is equivalent to the amplifier stage with shunt peaking discussed in Section 6.3.2. The difference is just that we discussed shunt peaking in the frequency domain, while we are now working in the time domain.

8.2.7 Automatic Power Control (Lasers)

Since the laser's L/I characteristics is strongly temperature and age dependent, an *Automatic Power Control* (APC) mechanism is usually required to stabilize the output power. Many laser drivers contain the APC circuit integrated with the driver on the same chip. Figure 8.15 shows the circuit principle for a continuous-mode APC. A monitor photodiode, with good temperature, age, and coupling stability, produces a current proportional to the transmitted optical power. This current is low-pass filtered and converted to a voltage at node x with R and C. This voltage, which is proportional to the average optical power, is compared to the desired reference value V_{REF} by means of an op amp. The op amp output voltage, V_B, controls the laser's bias current, I_B such that the desired

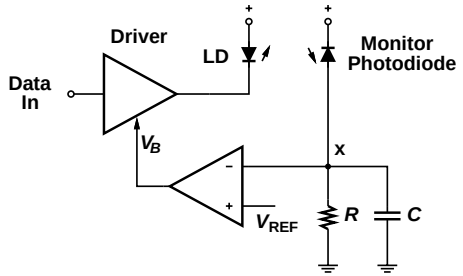

Figure 8.15: Automatic power control for continuous-mode laser drivers.

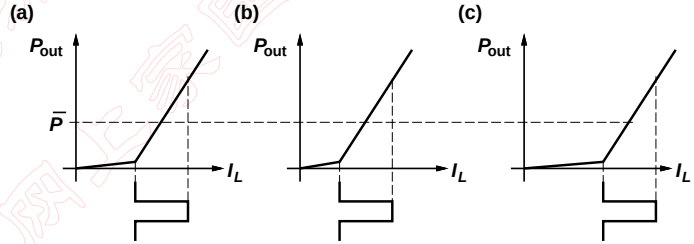

Figure 8.16: Single-loop APC: (a) nominal threshold, (b) reduced threshold, and (c) increased threshold.

average optical power is obtained. Because only the *average* photodiode current matters, a slow and low-cost monitor photodiode is sufficient.

The power-control mechanism just described is slightly inaccurate for long strings of zeros or ones. For example, given a long string of ones, the voltage at node x will rise above the average value causing the laser power to drop below its nominal value as a result of the APC loop. The power penalty associated with this impairment depends on the low-pass filter bandwidth $f_{APC} = 1/(2\pi \cdot RC)$ and is identical to that derived in Section 6.2.5 for the LF-cutoff in an MA:

$$PP = 1 + r \cdot \frac{2\pi f_{APC}}{B}. \tag{8.15}$$

To minimize this power penalty, the APC filter bandwidth must be made as small as possible. For an implementation of an APC loop see [Ølg94].

Dual-Loop APC. The circuit shown in Fig. 8.15 is also known as a *Single-Loop APC*, because only the bias current, and not the modulation current, is controlled with a feedback loop. If the threshold current of the laser is the *only* parameter that changes with temperature and age, the single-loop APC works perfectly well (see Fig. 8.16). However,

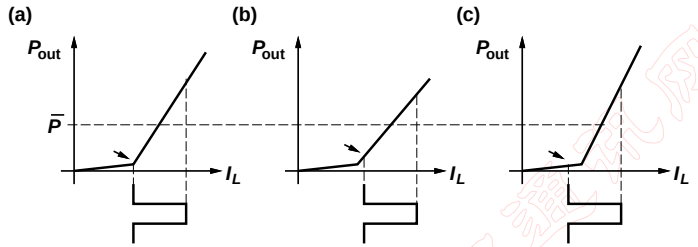

Figure 8.17: Single-loop APC: (a) nominal slope, (b) reduced slope, and (c) increased slope.

if the laser's slope efficiency changes as well, we experience the undesirable effects illustrated in Fig. 8.17. The single-loop APC will always keep the average output power $\overline{P}$ constant. Thus, if the slop reduces, the power for zeros increases and the power for ones decreases (Fig. 8.17(b)). In other words, the extinction ratio degrades. Alternatively, if the slope increases, the current for zeros drops below the laser's threshold current causing PWD and jitter (Fig. 8.17(c)).

These problems can be alleviated by increasing the modulation current in proportion to the bias current. This method helps if the slope efficiency degrades in proportion to the threshold current. Alternatively, the modulation current can be made a function of the laser temperature, a major factor in determining the slope efficiency. A simple circuit with a thermistor can do this job. However, laser aging is still not accounted for and a more robust solution is desirable. The so-called *Dual-Loop APC* uses two feedback loops to achieve this goal [Shu88]: (i) The bias current is controlled by the average optical power, just like in the single-loop APC. (ii) The modulation current is controlled by the laser's slope efficiency. The latter can for example be determined with the monitor photodiode by taking the difference between the peak and the average power. A peak-detector circuit similar to that shown in Fig. 8.21 can be used to measure the optical peak power. A drawback of this approach is that it requires a fast monitor photodiode which can follow the bit pattern. An alternative approach is to modulate the laser current with a low-frequency tone at a low modulation index and infer the slope efficiency from the response at the photodiode.

In practice, a single-loop APC is often sufficient, because the slope efficiency is less dependent on temperature and age than the threshold current. This is especially true for cooled lasers which are operated at a controlled temperature for reasons of wavelength stabilization or reliability.

Mark-Density Compensation. The circuit shown in Fig. 8.15 works well for DC-balanced data signals, or more generally, for signals with a constant mark density. (Remember, a DC-balanced signal has an average mark density of 1/2.) But imagine what happens if we send nothing but zeros to the driver. Will the transmitter shut off? No, an optical power meter connected to the output of the transmitter will always show the same

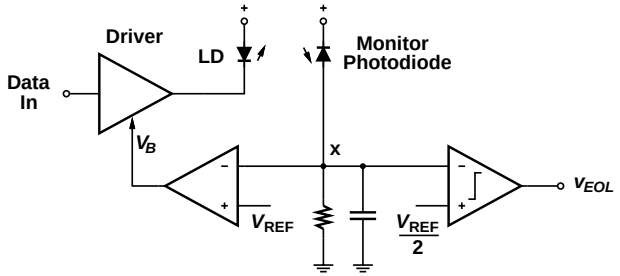

Figure 8.18: End-of-life detection circuit for continuous-mode laser drivers.

value no matter if we transmit all zeros, all ones, or a PRBS! This is, of course, because the APC circuit keeps the *average* output power constant. So, for an all-zero sequence, the transmitter outputs a constant power half-way in between that for a one and a zero. To deal with varying mark densities, special circuits for *Mark Density Compensation* have been proposed [Shu88]. These circuits generate a voltage proportional to the mark density and take it into account when comparing the voltage at node x with V_{REF}. We will discuss a related circuit when we examine APC for burst-mode transmitters.

Fortunately, all major transmission standards (SONET/SDH, Gigabit-Ethernet) do have DC-balanced data signals and the simple APC circuit in Fig. 8.15 is usually adequate.

8.2.8 End-of-Life Detection (Lasers)

As we have discussed in Section 7.2, the MTTF of a continuously operated laser at 70 °C is limited to about $1-10$ years. For this reason it is required by some standards, such as FSAN, that the laser's *End-Of-Life Condition* (EOL) is detected automatically. This feature can be implemented easily with the help of the monitor photodiode which is already present for the purpose of APC. Figure 8.18 shows a simple extension of the APC circuit of Fig. 8.15 to generate a binary EOL signal. A voltage comparator compares the voltage at node x with a reference voltage which is lower than V_{REF}, for example $V_{\mathrm{REF}}/2$ as shown in the figure. The comparator detects when the APC mechanism fails to keep the output power at the desired level and activates the EOL alarm when the power drops below a set value (e.g., 1/2 nominal value for $V_{\mathrm{REF}}/2$).

8.2.9 Automatic Bias Control (MZ Modulators)

Since MZ modulator suffer from voltage drift over temperature and age, an *Automatic Bias Control* (ABC) mechanism is usually required [PP97, HKV97].

The ABC circuit shown in Fig. 8.19 [Luc98] modulates the driver's output voltage with a low-frequency tone (e.g., 1 kHz) at a low modulation index. An optical splitter at the output of the modulator directs a small amount of the signal (e.g., 10 %) to a monitor photodiode. The signal from the photodiode is amplified and mixed with the original

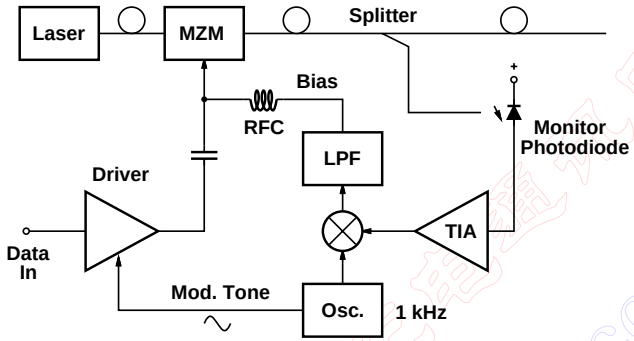

Figure 8.19: Automatic bias controller for MZ modulator.

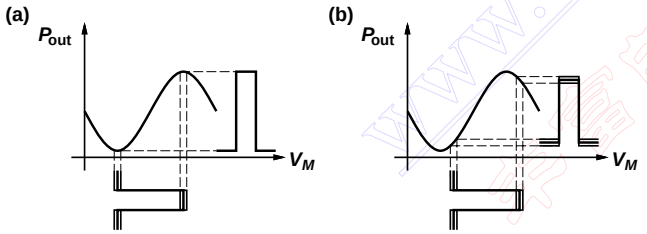

Figure 8.20: Automatic bias control for MZ modulator: (a) correct bias, (b) incorrect bias.

modulation tone. The mixer output is low-pass filtered to produce the bias voltage which is applied to the MZM by means of a bias tee. The ABC circuit adjusts the bias voltage until the amplitude of the detected tone is minimized. The direction of the bias adjustment is given by the phase relationship between the detected tone and the modulation tone.

Figure 8.20(a) shows that for a correctly adjusted bias voltage the tone does not appear in the optical output signal. However, for an incorrect bias, the on and off voltages move into the steep part of the sinusoidal switching curve and the tone appears in the optical output signal (see Fig. 8.20(b)).

Older MZ modulators exhibited so much drift that after an extended period of operation the upper or lower limit of the bias voltage generator was reached. Under these circumstances the bias generator needed to be reset ($\pm 2n \cdot V_\pi$). Fortunately, much progress in reducing the MZM drift has been made in recent years [HKV97].

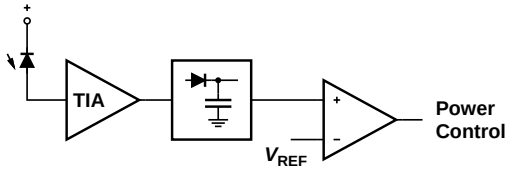

Figure 8.21: Burst-mode APC based on a peak detector.

8.2.10 Burst-Mode Laser Drivers

The main differences between a burst-mode laser driver and its continuous-mode cousin are: (i) A very high inter-burst extinction ratio is required and (ii) the APC must operate correctly for a bursty data signal.

Extinction Ratio. Burst-mode transmitters which are used in a multiple-access network are required to keep the undesired light output in between bursts at a very low level. For example, in a PON system with 32 subscribers, 31 subscribers are "polluting" the shared medium with residual light output. This background light reduces the received ER of the one subscriber that is transmitting. Typically, the inter-burst ER for a burst-mode laser driver is required to be over 30 dB. The ER requirement within a burst is around 10 dB, similar to that of a continuous-mode transmitter.

To achieve this high ER, the laser bias current in between bursts must be chosen very low (i.e., well below the laser's threshold current) or zero. During transmission of a burst, the bias current can be increased to reduce the turn-on delay and jitter. Furthermore, burst-mode laser drivers often feature a shunt transistor across the laser diode rapidly removing carriers at the end of the burst with the purpose to reach the 30 dB extinction ratio before another subscriber starts to transmit the next burst.

Automatic Power Control. A power control circuit as shown in Fig. 8.15 does not work correctly under burst-mode operation because the transmitted signal is not DC balanced and keeping the average output power constant is not sufficient. One way to implement a burst-mode APC is shown in Fig. 8.21 [MS96, INA+97]. The current from the monitor photodiode is converted to a voltage with a broadband TIA, then a peak detector finds the *peak* value which corresponds to the power transmitted during a one bit. The laser output power is adjusted until the peak detector output voltage equals the reference voltage V_{REF}.

However, the approach shown in Fig. 8.21 has some shortcomings: (i) During long idle periods, when no bursts are transmitted, the analog peak detector drifts away from the actual peak value causing an output power error at the beginning of the next burst. (ii) A fast monitor photodiode that can operate at bit-rate speed is required. (iii) The power dissipation in the TIA and peak detector, which have to operate at bit-rate speed, can be quite significant.

Another approach, based on an integrate-and-dump circuit and digital storage, avoids

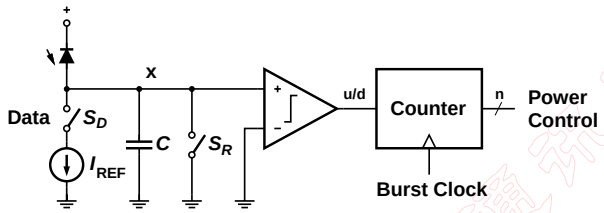

Figure 8.22: Burst-mode APC based on an integrate-and-dump circuit and digital storage.

these shortcomings. The principle of this APC is shown in Fig. 8.22 [SOGF00, SO01a]. The power level is stored in a digital up/down counter which can hold the precise power level as long as the power supply is up, therefore output power errors at the beginning of the bursts are avoided. The power level is increased or decreased in between bursts with the "Burst Clock" signal. The count direction is given by the up/down signal "u/d" which is generated as follows: Before the burst starts, capacitor C (node x) is discharged by briefly closing the reset switch S_R (dump). During the burst, the photodiode current $i_{PD}(t)$ charges the capacitor C (integrate). Simultaneously, the desired peak current I_{REF} is modulated with the transmitted data by switch S_D producing $i_{\mathrm{REF}}(t)$ discharging the same capacitor C. At the end of the burst, the capacitor is charged to the following voltage:

$$\frac{1}{C}\int_{\mathrm{Burst}} i_{PD}(t) - i_{\mathrm{REF}}(t)\,dt = \frac{1}{C}\left(I_{PD}\cdot\frac{n_1}{B} - I_{\mathrm{REF}}\cdot\frac{n_1}{B}\right) \tag{8.16}$$

where I_{PD} is the peak value of the photodiode current, n_1 is the number of ones in the burst, and B is the bit rate. The comparator compares this voltage (at node x) to 0 V, the reset voltage. As we can see from Eq. (8.16), the outcome of this comparison is independent of C, B, or n_1 and is only affected by $I_{PD} - I_{\mathrm{REF}}$. If this difference is positive, the counter is stepped down, if it is negative, the counter is stepped up.

End-of-Life Detection. Just like in the continuous-mode case, EOL detection can be implemented with a circuit very similar to that for APC. But now, because the transmitted data consists of discrete bursts, it is easily possible to time multiplex a single circuit for both APC and EOL detection. During one burst, the reference current is set to I_{REF} and the circuit performs APC, during another burst the reference current is set to $I_{\mathrm{REF}}/2$ and the same circuit performs EOL detection. The advantage of this multiplexing scheme is that it performs both functions at the power consumption of one [SOGF00, SO01b].

8.3 Driver Circuit Implementations

In the following section we will examine some representative transistor-level laser- and modulator-driver circuits taken from the literature and designed for a variety of technologies (cf. Appendix **??**). These circuits will illustrate how the design principles discussed in the previous section are implemented in practice.

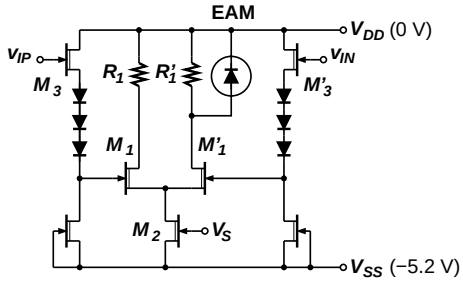

Figure 8.23: MESFET/HFET implementation of an EAM output-driver stage.

8.3.1 MESFET & HFET Technology

Modulator Driver. Figure 8.23 shows an EAM output-driver stage implemented in a depletion-mode MESFET or HFET technology. For instance the modulator drivers reported in [SSO+92, MYK+97] are based on such a topology and realized in GaAs HFET technology.

The circuit consists of a FET current-steering stage M_1, M_1' and load resistors R_1, R_1'. The EAM in parallel to R_1' is connected to the driver directly or through a transmission line without back termination. The circuit drives the EAM single-endedly, which means that one electrode of the EAM (cathode) remains at a DC level. This is a necessity for standard EAMs especially if they are integrated with the laser on the same substrate and therefore share one electrode (the substrate) with the laser. The output voltage swing of the modulator driver can be adjusted with the tail current source M_2 which is controlled by the voltage V_S. In the circuit shown, the output bias voltage of the modulator driver is 0 V, but it could be made non-zero by applying a DC current to the output node. The current steering stage, which consists of very large transistors (e.g., $W = 400\,\mu$m), is driven by the source followers M_3, M_3'.

Typically, the output stage shown in Fig. 8.23 is driven by a pre-driver stage which consists either of another smaller current-steering stage or one of the wideband stages discussed for the MAs, such as the FET stage covered in Section 6.4.1 [LTN+98]. The output voltage swing of the pre-driver and the transistor sizes in the output drivers must be carefully optimized to give the best rise and fall times.

Modulator/Laser Driver with Back Termination. Figure 8.24 shows an output-driver stage with back termination for driving an EAM or laser through a transmission line. For instance the modulator driver reported in [LTN+98] is based on such a topology and realized in GaAs HFET technology.

To avoid double reflections on the transmission line, both sides must be terminated. The termination on the modulator side is implemented by R_P, the termination on the driver side is implemented by R_1'. As a result, the driver transistor M_1' is presented with two parallel load resistors (R_1' and R_P) and *twice* the tail current is needed to achieve

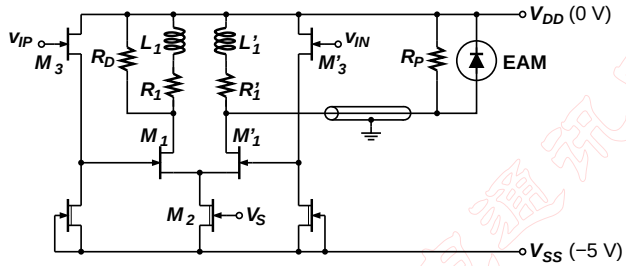

Figure 8.24: MESFET/HFET implementation of an EAM output-driver stage with back termination.

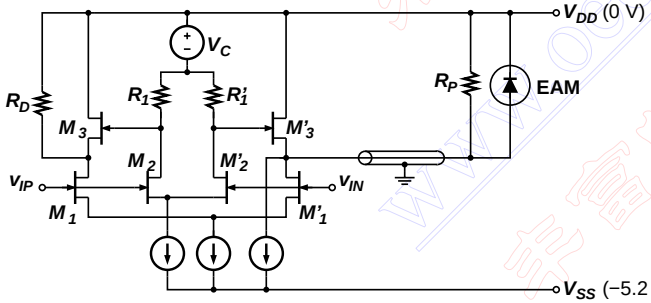

Figure 8.25: MESFET/HFET implementation of an EAM output-driver stage with active back termination.

the same output-voltage swing as the driver circuit of Fig. 8.23. To reduce the power dissipation, i.e., to get a better output swing for a given tail current, the resistors R_1, R'_1, are often made larger than what is the optimum for matching (e.g., 100 Ω for a 50 Ω transmission line). This degrades the matching quality at DC, but may not have a big impact at high frequencies where parasitic capacitances degrade the matching quality anyway.

The peaking inductors, L_1 and L'_1, improve the rise/fall time of the output signal as discussed in Section 8.2.6. The modulator driver circuit in Fig. 8.24 can also operate as a laser driver if a bias current source, similar to Fig. 8.27, is added.

Modulator/Laser Driver with Active Back Termination. Figure 8.25 shows an output-driver stage with *active* back termination for driving an EAM or laser through a transmission line. For instance the laser/modulator driver reported in [RSDG01] is based on such a topology and realized in GaAs HFET technology.

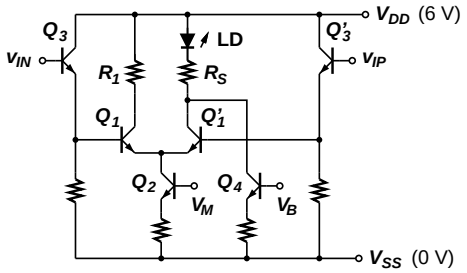

Figure 8.26: BJT/HBT implementation of a laser output-driver stage.

As usual the driver generates the output signal with a current-steering stage M_1 and M'_1. A scaled-down replica (e.g., 1:8) of this stage M_2 and M'_2 generates a copy of the intended output voltage (without reflections). Source followers M_3 and M'_3 buffer the replica signal and are sized such that their output impedance ($\approx 1/g_m$) matches the transmission line. The output impedance of the source follower M'_3 acts as the back termination and absorbs possible reflections. The current through M'_3, is kept at a constant value I_3 by means of a feedback circuit (not shown) controlling V_C. Therefore the power dissipated in the active back termination circuit is $P = V_{DD} \cdot I_3$ where I_3 is typically around 10 mA.[1] This compares to the much larger power of a resistive back termination $P = V_{DD} \cdot I_T/2$, where I_T is the tail current of the current steering pair (e.g., 200 mA). The feedback circuit keeping the current in M'_3 constant is also making sure that a bias current which may be applied to the output node doesn't flow back into M'_3.

8.3.2 BJT & HBT Technology

Laser Driver. Figure 8.26 shows a simple laser output-driver stage implemented in a BJT or HBT technology. It can be used to drive FP, DFB, or VCSEL lasers directly or through a transmission line without back termination. For instance the laser drivers reported in [Rei88] and [RDS+92] are based on this topology and realized in Si BJT and GaAs HBT technology, respectively.

The circuit consists of a bipolar current-steering stage Q_1, Q'_1 with a dummy resistor on one side and the laser with a series resistor R_S on the other side. The series resistor dampens the ringing caused by parasitic inductances and capacitances (e.g., package, bond wire, etc.) and may also serve to move some of the power dissipation out of the driver IC. The modulation current I_M is supplied by the tail current source, consisting of Q_2 and an emitter resistor, which is controlled by the voltage V_M. The laser bias current I_B is supplied by the current source consisting of Q_4 and an emitter resistor. The value of the bias current is controlled by the voltage V_B. The output stage is driven by the emitter followers Q_3, Q'_3; in practice a cascade of several emitter followers may be used

[1] More precisely, $P = V_{DD} \cdot (I_3 + I_M/n)$ where I_M/n is the current in the scaled down replica.

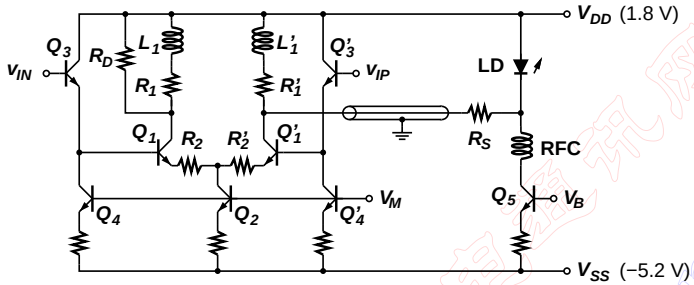

Figure 8.27: BJT/HBT implementation of a laser output-driver stage with back termination.

to improve the drive capability and to increase the collector-emitter voltage of Q_1 and Q_1' boosting their f_T (cf. Section 6.3.2).

Typically, the output stage shown in Fig. 8.26 is driven by a pre-driver stage which consists either of another smaller current-steering stage or one of the wideband stages discussed for the MAs, such as the Cherry-Hooper stage covered in Section 6.4.2 [RDS+92].

Laser/Modulator Driver with Back Termination. Figure 8.27 shows an output-driver stage with passive back termination for driving a laser or EAM through a transmission line. For instance the laser/modulator driver reported in [RSW+94] and [KBA+01] are based on this topology and are realized in Si BJT and InP HBT technology, respectively.

To avoid double reflections on the transmission line, both sides must be terminated. As a result, the driver transistor Q_1' is presented with two parallel load resistors (R_1' and R_S plus the laser) and *twice* the tail current is needed to supply the same modulation current to the laser as the driver circuit of Fig. 8.26. In addition to this, the bias current from Q_5 must be increased as well because the back-termination resistor R_1' absorbs some of this current, further increasing the power dissipation. The power dissipated in the termination resistor R_1' can be reduced by increasing its value at the expense of the matching quality.

The peaking inductors, L_1 and L_1', improve the rise/fall time of the output signal as discussed in Section 8.2.6. The emitter resistors R_2 and R_2' are distributed among the emitter fingers to ensure an even current distribution within the transistors Q_1 and Q_1'. To obtain good performance over a wide range of modulation currents, the bias currents of the emitter followers Q_3 and Q_3' are varied with the modulation current. Usually, the bias currents have a part which varies with the modulation current (Q_4, Q_4') as well as a constant part (not shown in Fig. 8.27). The laser bias current is applied to the laser through an RF choke (RFC) to avoid an increased voltage drop over R_S and minimize capacitive loading of the output. Note that the driver circuit in Fig. 8.27 can also be used to drive a properly terminated modulator at the other end of the transmission line.

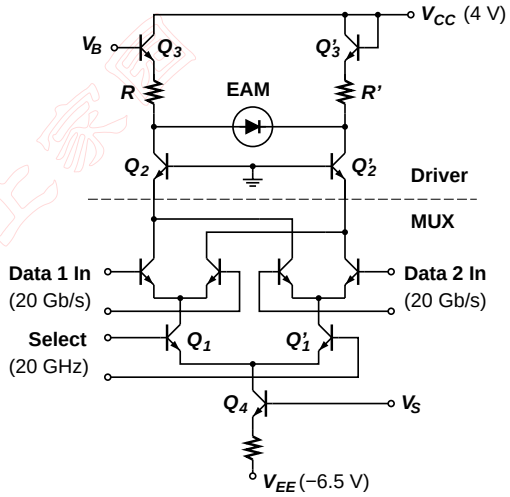

Figure 8.28: BJT/HBT implementation of a differential 40 Gb/s EAM-driver stage with built-in MUX.

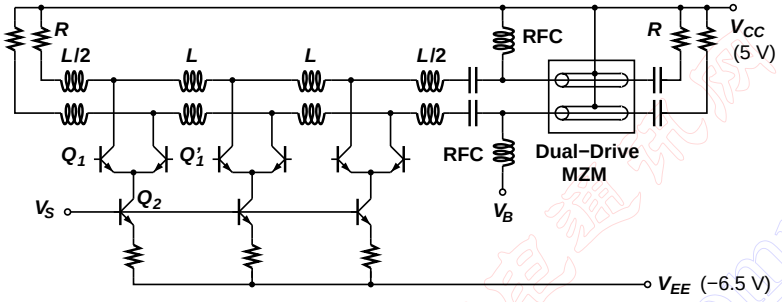

Figure 8.29: BJT/HBT implementation of a distributed MZM output-driver stage.

Modulator Driver with Built-In MUX. Figure 8.28 shows a high-speed EAM driver stage with a built-in data multiplexer implemented in a BJT or HBT technology. The modulator driver reported in [MSR+98] is based on this topology and realized in SiGe HBT technology. The SiGe modulator driver reported in [SMRR98] uses a similar topology, except that it does not have a built-in multiplexer.

The circuit shown in Fig. 8.28 combines the data-multiplexer and modulator-driver function into a single high-speed circuit and is called "Power MUX" for this reason. The multiplexer combines, for example, two 20 Gb/s data streams into a single 40 Gb/s data stream which is then used to drive the modulator. The current from current source Q_4 ($40 - 50$ mA) is directed to either the right or left output of the MUX depending on the select and data signals. The MUX is implemented with three current-steering differential pairs. A first pair, Q_1, Q_1', controls which data input is selected. The other two pairs steer the current according to the input data. The output currents from the MUX are routed through the cascode transistors Q_2 and Q_2' to prevent breakdown-voltage violations and to reduce the capacitance at the driver's output. The output currents drop over the loads R, R', Q_3, and Q_3' where they produce the driving voltage for the EAM. The output voltage swing can be adjusted with V_S and the output bias voltage is controlled by $V_{CC} - V_B$.

This circuit drives the EAM differentially and thus requires a symmetrical modulator, i.e., it must be possible to drive both electrodes of the EAM independently (no shared connection with the laser) and they must have about equal capacitance. The advantage of driving the modulator differentially is that only half the voltage swing is required at each output. For example, a $1\,V_{pp}$ signal at each output produces a $2\,V_{pp}$ signal across the EAM. However, at the time of writing such a modulator does not appear to be commercially available.

Modulator Driver with Distributed Output Stage. Figure 8.29 shows a distributed MZM output-driver stage implemented in a BJT or HBT technology. The modulator driver reported in [WFBS96] is based on this topology and realized in GaAs HBT technology.

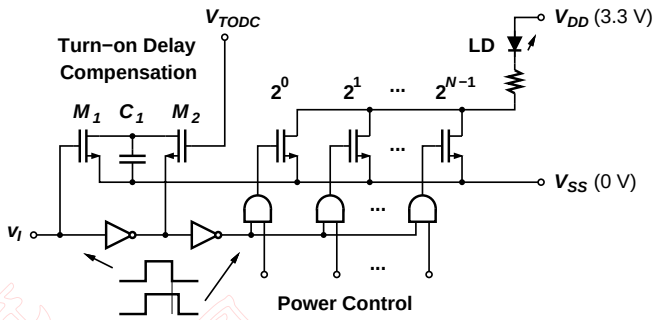

Figure 8.30: CMOS implementation of a low-power burst-mode laser driver.

A major speed limitation of the current-steering stage is the time constant formed by the output capacitance (e.g., collector-base capacitance of the driver transistors) and the load resistor. Just like in the distributed amplifier introduced in Section 6.3.2, the output capacitance can be absorbed into an artificial (discrete) transmission line by splitting the driver transistors into n smaller ones, Q_1, Q_1', ..., and connect them as shown in Fig. 8.29. With the appropriate value for the inductors L, an artificial transmission line matched to the load resistors R and the impedance of the MZM is formed. Now the speed is limited by the cutoff frequency of the artificial transmission line which can be made high by choosing a large n ($n = 5$ in [WFBS96]). The distributed output stage is driven from the taps of a second transmission line which is not shown in Fig. 8.29.

The distributed output-driver stage can be connected to a dual-drive MZ modulator operating in a push-pull fashion as shown in Fig. 8.29. In this way only half the switching voltage V_π is needed per arm. The output swing can be adjusted with the tail current sources, Q_2, ..., which are all controlled by the same voltage V_S. The bias voltage V_B is supplied to the MZ modulator through a bias tee (AC-coupling capacitor and RF choke). The low-frequency modulation tone necessary for the ABC (cf. Section 8.2.9) can be superimposed on V_S.

8.3.3 CMOS Technology

Figure 8.30 shows a low-power laser-driver stage implemented in CMOS technology. The burst-mode laser driver reported in [SOGF00] is based on this topology.

In contrast to the drivers discussed previously, this driver uses *Current Switching* rather than current steering. This means that when transmitting a zero, the current is shut off completely and thus the average power dissipation is reduced by a factor two. This scheme is particularly suitable for burst-mode laser drivers because additional power is saved during idle periods when no bursts are transmitted. For example, if the average burst activity is 10 % the total power savings are 20×. Furthermore, the driver stage in Fig. 8.30 operates with *zero* laser bias current. This ensures the high inter-burst extinction

ratio required for burst-mode drivers (cf. Section 8.2.10) and saves power as well.

As a result of operating the laser without bias current we are affected by turn-on delay and turn-on delay jitter (cf. Section 7.2). Turn-on delay can be compensated by predistorting the input data signal. The simple *Turn-On Delay Compensation* circuit shown in Fig. 8.30 delays the falling data edge by the laser's TOD. This delay is implemented by loading the inverter output with M_2 and C_1. The amount of delay can be controlled with the gate voltage V_{TODC}. On the rising edge, M_1 discharges C_1 rapidly to prevent a delay. The turn-on delay jitter cannot be compensated, but is most likely not a problem in low-speed applications up to 155 Mb/s.

The laser driver in Fig. 8.30 also features digital power control. N parallel driver transistors M_n with widths proportional to 2^n form a driver with built-in D/A converter. The N-bit word controlling these transistors determines the output power. The counter-based APC circuit, discussed in Section 8.2.10, is well suited to control this driver.

8.4 Product Examples

Tables 8.1 and 8.2 summarize the main parameters of some commercially available laser and modulator drivers. The numbers have been taken from data sheets of the manufacturer which were available to me at the time of writing. For up-to-date product information please contact the manufacturer directly. Rise and fall times are measured from $20-80\,\%$. The power dissipation numbers quoted for the laser drivers are under the condition of *zero* modulation and bias current. Similarly, the power dissipation numbers quoted for the modulator drivers are for *zero* voltage swing and bias voltage. Exceptions are the power numbers for the Giga parts (in parenthesis), which are quoted at maximum voltage swing. The column labeled "Output" indicates if the modulator driver has single-ended (s.e.) or differential (diff.) outputs. For consistency, the tabulated output swing $V_{S,\max}$ is the *single-ended* swing, even for modulator drivers which have differential outputs.

From these tables, it can be seen that the power consumption of current 10 Gb/s laser and modulator drivers is quite high. Even with zero output signal it is in the range $0.5-1.5\,\mathrm{W}$, for a typical output current/swing it is typically well above 1 W. Most modulator drivers listed are implemented in GaAs technology. GaAs FETs are preferred over other high-speed devices such as SiGe BJTs because of their high breakdown voltage. The difference between the similar Agere laser drivers is that the LG1627BXC and TCLD0110G contain a retiming flip-flop whereas the LG1625AXF and TLAD0110G do not.

8.5 Research Directions

The research effort can be divided roughly into three categories: higher speed, higher integration, and lower cost.

Higher Speed. It has already been pointed out in Section 5.5 that many research groups are now aiming at the 40 Gb/s speed and beyond. To this end fast modulator

Company & Product	Speed	$I_{M,\max}$	$I_{B,\max}$	$t_{R,F}$	Power	Technology
Agere LG1625AXF	2.5 Gb/s	65 mA	40 mA	90 ps	520 mW	GaAs HFET
Agere LG1627BXC	2.5 Gb/s	85 mA	60 mA	100 ps	730 mW	GaAs HFET
Infineon S896A006	2.5 Gb/s	60 mA	60 mA		165 mW	Si BJT
Maxim MAX3867	2.5 Gb/s	60 mA	100 mA	90 ps	205 mW	
Nortel YA08	2.5 Gb/s	80 mA	100 mA	150 ps	350 mW	Si BJT
Philips OQ2545HP	2.5 Gb/s	60 mA	100 mA	120 ps	350 mW	Si BJT
Agere TLAD0110G	10 Gb/s	100 mA	120 mA	30 ps	780 mW	GaAs HFET
Agere TCLD0110G	10 Gb/s	100 mA	100 mA	25 ps	780 mW	GaAs HFET
Maxim MAX3930	10 Gb/s	100 mA	100 mA	27 ps	540 mW	SiGe HBT

Table 8.1: Examples for 2.5 and 10 Gb/s laser driver products.

Company & Product	Speed	Output	$V_{S,\max}$	$t_{R,F}$	Power	Technology
Agere LG1626DXC	2.5 Gb/s	s.e.	3.0 V	90 ps	730 mW	GaAs HFET
Agere TMOD0110G	10 Gb/s	diff.	2.7 V	30 ps	1400 mW	GaAs HFET
Giga GD19901	10 Gb/s	diff.	3.0 V	45 ps	(3200 mW)	GaAs HFET
Giga GD19903	10 Gb/s	diff.	5.0 V	45 ps	(8600 mW)	GaAs HFET
Maxim MAX3935	10 Gb/s	s.e.	3.0 V	34 ps	550 mW	SiGe HBT
OKI KGL4115F	10 Gb/s	s.e.	2.7 V	40 ps	1300 mW	GaAs HFET

Table 8.2: Examples for 2.5 and 10 Gb/s modulator driver products.

drivers, possibly with integrated MUX, must be designed. Usually GaAs, InP, and SiGe technologies combined with heterostructure devices such as HFETs and HBTs are used to reach this goal.

Here are some examples from the literature:

- In SiGe-HBT technology a 23 Gb/s modulator driver with 3.5 V_{pp} single-ended swing, a 40 Gb/s EAM driver with 1.0 V_{pp} single-ended swing, and a 50 Gb/s EAM driver with the same swing have been reported in [SMRR99], [SMRR98], and [MSR+98], respectively.

- In GaAs-HFET technology a 20 Gb/s laser driver has been reported in [WBN+93]; a 40 Gb/s modulator driver with a single-ended swing of 2.9 V_{pp} has been reported in [LTN+98].

- In InP-HBT technology a 40 Gb/s EAM-driver with 2.2 V_{pp} single-ended swing and a 20 Gb/s modulator driver with 4.0 V_{pp} single-ended swing have been reported in [KBA+01] and [MBK98], respectively.

Higher Integration. Another interesting research direction, aiming at higher integration, is to combine the laser diode, monitor photodiode, and the driver circuit on the same substrate, so-called *Optoelectronic Integrated Circuits* (OEIC) [Kob88]. Complete systems consisting of a 1.3 μm DFB laser, monitor photodiode, and driver circuit have been integrated on a single InP substrate. However, it is a challenge to effectively combine laser and circuit technologies into a single one because of the significant structural differences between lasers and transistors, for example lasers require mirrors or gratings while transistors don't. As a result transmitter OEICs are not as far advanced as receiver OEICs.

Another OEIC approach is to use a flip-chip laser on top of the driver chip. An important advantage of this approach is that the laser chip and the driver chip can each be fabricated in their most suitable technologies, avoiding the compromises of monolithic OEICs.

Lower Cost. Another area of research is focusing on the design of high-performance laser and modulator drivers in low-cost, mainstream technologies, in particular digital CMOS.

For example, a laser driver for a fiber-to-the-home system (PON) must be very low cost, to be competitive with traditional telecom services, and low power, to minimize the size and cost of the back-up battery. A low-power (15 mW) CMOS laser driver for 155 Mb/s has been reported in [SOGF00].

8.6 Summary

The main specifications for laser and modulator drivers are:

- The modulation and bias current ranges for laser drivers which must be large enough to operate the desired laser under worst-case conditions. In particular, uncooled lasers require large current ranges.

- The voltage swing and bias voltage ranges for modulator drivers which must be large enough to operate the desired modulator under worst-case conditions. In particular, high-speed MZ modulators require large voltage swings.

- The rise and fall times which must be short compared to the bit period but not too short to avoid excessive chirping in lasers.

- The pulse-width distortion which is usually compensated with an adjustable pulse-width control circuit.

- The jitter generation which must be very low for SONET compliant equipment.

In addition, some standards such as SONET require that the filtered optical signal fits into a given eye mask.

Most drivers are based on the current-steering principle. Some drivers are directly connected to the laser or modulator (e.g., through a short bond wire), others are connected trough a matched transmission line (with or without back termination) permitting a larger distance between the driver and the laser or modulator. Pulse-width control is usually implemented by introducing an adjustable offset voltage to the pre-driver. A flip-flop for data retiming can be used to reduce jitter and pulse-width distortions in the output signal. An automatic power control (APC) circuit uses negative feedback from the monitor photodiode to keep the optical output power, and optionally the ER, of the laser constant. Similarly, an automatic bias control (ABC) circuit is required to stabilize the operating point of a MZ modulator. Some laser drivers feature an end-of-life detector alerting to the fact that the laser must be replaced soon. Burst-mode laser drivers require a very high inter-burst ER and a special APC circuit that can deal with a bursty data signal.

Laser and modulator drivers have been implemented in a variety of technologies including MESFET, HFET, BJT, HBT, BiCMOS, and CMOS.

Currently, researchers are working on 40 Gb/s modulator drivers, drivers integrated with the laser or modulator on the same chip, as well as laser and modulator drivers in low-cost technologies such as CMOS.

Bibliography

[Agr97] Govind P. Agrawal. *Fiber-Optic Communication Systems*. John Wiley & Sons, New York, 2nd edition, 1997.

[AH87] Phillip E. Allen and Douglas R. Holberg. *CMOS Analog Circuit Design*. Holt, Rinehart and Winston, New York, 1987.

[AHK+02] Kamran Azadet, Erich F. Haratsch, Helen Kim, Fadi Saibi, Jeffrey H. Saunders, Michael Shaffer, Leilei Song, and Meng-Lin Yu. Equalization and FEC techniques for optical transceivers. *IEEE J. Solid-State Circuits*, SC-37(3):317–327, March 2002.

[AT&95] AT&T. The relationship between chirp and voltage for the AT&T mach-zehnder lithium niobate modulators. Agere Systems, Technical Note, October 1995.

[Bel95] Bellcore. SONET transport systems: Common criteria, GR-253-CORE – Issue 2. Bellcore Documentation, Morristown, N.J. (now Telcordia Technologies), December 1995.

[Bel98] Bellcore. SONET OC-192 transport systems generic criteria, GR-1377-CORE – Issue 4. Bellcore Documentation, Morristown, N.J. (now Telcordia Technologies), March 1998.

[Ben83] M. J. Bennett. Dispersion characteristics of monomode optical-fiber systems. *IEE Proceedings, Pt. H*, 130(5):309–314, August 1983.

[BM95] Aaron Buchwald and Ken Martin. *Integrated Fiber-Optic Receivers*. Kluwer Academic Publisher, 1995.

[CFL99] Walter Ciciora, James Farmer, and David Large. *Modern Cable Television Technology – Video, Voice, and Data Communications*. Morgan Kaufmann, San Francisco, 1999.

[CH63] E. M. Cherry and D. E. Hooper. The design of wide-band transistor feedback amplifiers. *Proceedings IEE*, 110(2):375–389, February 1963.

[dSBB+99] Valeria L. da Silva, Yvonne L. Barberio, Olen T. Blash, Leslie J. Button, Karin Ennser, Laura L. Hluck, Alan J. Lucero, Margarita Rukosueva, Sergio

Tsuda, and Robert J. Whitman. Capacity upgrade for non-zero dispersion-shifted fiber based systems. National Fiber Optics Engineers Conference (NFOEC), 1999.

[Ein96] Göran Einarsson. *Principles of Lightwave Communications*. John Wiley & Sons, 1996.

[Est95] Donald Estreich. Wideband amplifiers. In Ravender Goyal, editor, *High-Frequency Analog Integrated Circuit Design*, pages 170–240. John Wiley & Sons, Inc., 1995.

[Feu90] Dennis L. Feucht. *Handbook of Analog Circuit Design*. Academic Press, San Diego, 1990.

[FJ97] Daniel A. Fishman and B. Scott Jackson. Transmitter and receiver design for amplified lightwave systems. In Ivan P. Kaminow and Thomas L. Koch, editors, *Optical Fiber Telecommunications IIIB*, pages 69–114. Academic Press, San Diego, 1997.

[FSA01] FSAN. Common technical specifications of ATM subscriber system, version 3.021, May 2001. `http://www.fsanet.net`.

[GHW92] Richard D. Gitlin, Jeremiah F. Hayes, and Stephen B. Weinstein. *Data Communications Principles*. Plenum Press, 1992.

[GM77] Paul R. Gray and Robert G. Meyer. *Analysis and Design of Analog Integrated Circuits*. John Wiley & Sons, New York, 1977.

[Gre84] Alan B. Grebene. *Bipolar and MOS Analog Integrated Circuit Design*. John Wiley & Sons, New York, 1984.

[Gre01] Yuriy M. Greshishchev. Front-end circuits for optical communications, February 2001. ISSCC'2001 Tutorial.

[GS99] Yuriy M. Greshishchev and Peter Schvan. A 60-dB gain, 55-dB dynamic range, 10-Gb/s broad-band SiGe HBT limiting amplifier. *IEEE J. Solid-State Circuits*, SC-34(12):1914–1920, December 1999.

[Hec99] Jeff Hecht. *City of Light – The Story of Fiber Optics*. Oxford University Press, 1999.

[HG93] Timothy H. Hu and Paul R. Gray. A monolithic 480Mb/s parallel AGC/decision/clock-recovery circuit in 1.2-μm CMOS. *IEEE J. Solid-State Circuits*, SC-28(12):1314–1320, December 1993.

[HKV97] Fred Heismann, Steven K. Korotky, and John J. Veselka. Lithium niobate integrated optics: Selected contemporary devices and system applications. In Ivan P. Kaminow and Thomas L. Koch, editors, *Optical Fiber Telecommunications IIIB*, pages 377–462. Academic Press, San Diego, 1997.

[HLG88] Paul S. Henry, R. A. Linke, and A. H. Gnauck. Introduction to lightwave systems. In Stewart E. Miller and Ivan P. Kaminow, editors, *Optical Fiber Telecommunications II*, pages 781–831. Academic Press, San Diego, 1988.

[IEE01] IEEE. Ethernet in the first mile, task force IEEE 802.3ah, 2001. `http://www.ieee802.org/3/efm/`.

[INA+97] Noboru Ishihara, Makoto Nakamura, Yukio Akazawe, Naoto Uchida, and Yhuji Akahori. 3.3V, 50Mb/s CMOS transceiver for optical burst-mode communication. In *ISSCC Dig. Tech. Papers*, pages 244–245, 1997.

[IT94] ITU-T. Digital line systems based on the synchronous digital hierarchy for use on optical fibre cables, recommendation G.958. International Telecommunication Union, November 1994.

[IT98] ITU-T. Broadband optical access systems based on passive optical networks (PON), recommendation G.983.1. International Telecommunication Union, October 1998.

[IT00] ITU-T. Forward error correction for submarine systems, recommendation G.975. International Telecommunication Union, October 2000.

[IVCS94] Mark Ingels, Geert Van der Plas, Jan Crols, and Michel Steyaert. A CMOS 18THzΩ 240Mb/s transimpedance amplifier and 155Mb/s LED-driver for low cost optical fiber links. *IEEE J. Solid-State Circuits*, SC-29(12):1552–1559, December 1994.

[Jin87] Renuka P. Jindal. Gigahertz-band high-gain low-noise AGC amplifiers in fine-line NMOS. *IEEE J. Solid-State Circuits*, SC-22(4):512–521, August 1987.

[Jin90] Renuka P. Jindal. Silicon MOS amplifier operation in the integrate and dump mode for gigahertz band lightwave communication systems. *Journal of Lightwave Technology*, LT-8(7):1023–1026, July 1990.

[JM97] David Johns and Ken Martin. *Analog Integrated Circuit Design*. John Wiley & Sons, New York, 1997.

[Kas88] Bryon L. Kasper. Receiver design. In Stewart E. Miller and Ivan P. Kaminow, editors, *Optical Fiber Telecommunications II*, pages 689–722. Academic Press, San Diego, 1988.

[KB00] Helen Kim and Jonathan Bauman. A 12GHz, 30dB modular BiCMOS limiting amplifier for 10Gb/s SONET receiver. In *ISSCC Dig. Tech. Papers*, pages 160–161, February 2000.

[KBA+01] Nicolas Kauffmann, Sylvain Blayac, Miloud Abboun, Philippe André, Frédéric Aniel, Muriel Riet, Jean-Louis Benchimol, Jean Godin, and Agnieszka Konczykowska. InP HBT driver circuit optimization for high-speed

ETDM transmission. *IEEE J. Solid-State Circuits*, SC-36(4):639–647, April 2001.

[KCBB01] Helen H. Kim, S. Chandrasekhar, Charles A. Burrus, and Jon Bauman. A Si BiCMOS transimpedance amplifier for 10Gb/s SONET receiver. *IEEE J. Solid-State Circuits*, SC-36(5):769–776, May 2001.

[KDV+01] Benedik Kleveland, Carlos H. Diaz, Dieter Vook, Liam Madden, Thomas H. Lee, and S. Simon Wong. Exploiting CMOS reverse interconnect scaling in multigigahertz amplifier and oscillator design. *IEEE J. Solid-State Circuits*, SC-36(10):1480–1488, October 2001.

[Kil96] Ulrich Killat (editor). *Access to B-ISDN via PONs – ATM Communication in Practice*. John Wiley and B. G. Teubner, 1996.

[KMJT88] Bryon L. Kasper, Alfred R. McCormick, Charles A. Burrus Jr., and J. R. Talman. An optical-feedback transimpedance receiver for high sensitivity and wide dynamic range at low bit rates. *Journal of Lightwave Technology*, LT-6(2):329–338, February 1988.

[Kob88] Kohroh Kobayashi. Integrated optical and electronic devices. In Stewart E. Miller and Ivan P. Kaminow, editors, *Optical Fiber Telecommunications II*, pages 601–630. Academic Press, San Diego, 1988.

[Koc97] Thomas L. Koch. Laser sources for amplified and WDM lightwave systems. In Ivan P. Kaminow and Thomas L. Koch, editors, *Optical Fiber Telecommunications IIIB*, pages 115–162. Academic Press, San Diego, 1997.

[Kra88] John D. Kraus. *Antennas*. McGraw Hill, 2nd edition, 1988.

[KSC+98] P. I. Kuindersma, M. W. Snikkers, G. P. J. M. Cuypers, J. J. M. Binsma, E. Jansen, A. van Geelen, and T. van Dongen. Universality of the chirp-parameter of bulk active electro absorption modulators. European Conference on Optical Communication (ECOC), 1998.

[KTT95] Haideh Khorramabadi, Liang D. Tzeng, and Maurice J. Tarsia. A 1.06Gb/s, −31dBm to 0dBm BiCMOS optical preamplifier featuring adaptive transimpedance. In *ISSCC Dig. Tech. Papers*, pages 54–55, February 1995.

[KWOY89] M. Kawi, H. Watanabe, T. Ohtsuka, and K. Yamaguchi. Smart optical receiver with automatic decision threshold setting and retiming phase alignment. *Journal of Lightwave Technology*, LT-7(11):1634–1640, November 1989.

[LBH+97] Zhihao Lao, Manfred Berroth, Volker Hurm, Andreas Thiede, Roland Bosch, Peter Hofman, Alex Hülsmann, Canute Moglestue, and Klaus Köhler. 25Gb/s AGC amplifier, 22GHz transimpedance amplifier and 27.7GHz limiting amplifier ICs using AlGaAs/GaAs-HEMTs. In *ISSCC Dig. Tech. Papers*, pages 356–357, February 1997.

[Lee98] Thomas H. Lee. *The Design of CMOS Radio-Frequency Integrated Circuits*. Cambridge University Press, Cambridge, U.K., 1998.

[Liu96] Max Ming-Kang Liu. *Principles and Applications of Optical Communications*. Irwin, McGraw-Hill, Chicago, 1996.

[LM94] Edward A. Lee and David G. Messerschmitt. *Digital Communication*. Kluwer Academic Publishers, 2nd edition, 1994.

[LTN+98] Zhihao Lao, Andreas Thiede, Ulrich Nowotny, Hariolf Lienhart, Volker Hurm, Michael Schlechtweg, Jochen Hornung, Wolfgang Bronner, Klaus Köhler, Alex Hülsmann, Brian Raynor, and Theo Jakobus. 40-Gb/s high-power modulator driver IC for lightwave communication systems. *IEEE J. Solid-State Circuits*, SC-33(10):1520–1526, October 1998.

[Luc98] Lucent Technologies. Using the lithium niobate modulator: Electro-optical and mechanical connections. Agere Systems, Technical Note, April 1998.

[Luc99] Lucent Technologies. Low-cost, high-voltage APD bias circuit with temperature compensation. Agere Systems, Application Note, January 1999.

[Luc00] Lucent Technologies. Electroabsorptive modulated laser (EML): Setup and optimization. Agere Systems, Technical Note, May 2000.

[LWL+97] Manfred Lang, Zhi-Gong Wang, Zhihao Lao, Michael Schlechtweg, Andreas Thiede, Michaela Rieger-Motzer, Martin Sedler, Wolfgang Bronner, Gudrun Kaufel, Klaus Köhler, Axel Hülsmann, and Brian Raynor. 20-40Gb/s, 0.2-μm GaAs HEMT chip set for optical data receiver. *IEEE J. Solid-State Circuits*, SC-32(9):1384–1393, September 1997.

[MBK98] Mounir Meghelli, Michel Bouché, and Agnieszka Konczykowska. High power and high speed InP DHBT driver IC's for laser modulation. *IEEE J. Solid-State Circuits*, SC-33(9):1411–1416, September 1998.

[MHBL00] Sunderarajan S. Mohan, Maria del Mar Hershenson, Stephen P. Boyd, and Thomas H. Lee. Bandwidth extension in CMOS with optimized on-chip inductors. *IEEE J. Solid-State Circuits*, SC-35(3):346–355, March 2000.

[MKD97] Pablo V. Mena, Sung-Mo Kang, and Thomas A. DeTemple. Rate-equation-based laser models with a single solution regime. *Journal of Lightwave Technology*, LT-15(4):717–730, April 1997.

[MM96] Robert G. Meyer and William D. Mack. Monolithic AGC loop for a 160Mb/s transimpedance amplifier. *IEEE J. Solid-State Circuits*, SC-31(9):1331–1335, September 1996.

[MMR+98] J. Müllrich, T. F. Meister, M. Rest, W. Bogner, A. Schöpflin, and H.-M. Rein. 40Gb/s transimpedance amplifier in SiGe bipolar technology for receiver in optical-fibre TDM links. *Electronics Letters*, Vol. 34(5):452–453, March 1998.

[MOA+00] Toru Masuda, Ken-ichi Ohhata, Fumihiko Arakawa, Nobuhiro Shiramizu, Eiji Ohue, Katsuya Oda, Reiko Hayami, Masamitchi Tanabe, Hiromi Shimamoto, Masao Kondo, Takashi Harada, and Katsuyoshi Washio. 45GHz transimpedance, 32dB limiting amplifier, and 40Gb/s 1:4 high-sensitivity demultiplexer with decision circuit using SiGe HBTs for 40Gb/s optical receiver. In *ISSCC Dig. Tech. Papers*, pages 60–61, February 2000.

[MOO+98] Toru Masuda, Ken-ichi Ohhata, Eiji Ohue, Katsuya Oda, Masamitchi Tanabe, Hiromi Shimamoto, T. Onai, and Katsuyoshi Washio. 40Gb/s analog IC chipset for optical receiver using SiGe HBTs. In *ISSCC Dig. Tech. Papers*, pages 314–315, February 1998.

[MRW94] M. Möller, H.-M. Rein, and H. Wernz. 13Gb/s Si-bipolar AGC amplifier IC with high gain and wide dynamic range for optical-fiber receivers. *IEEE J. Solid-State Circuits*, SC-29(7):815–822, July 1994.

[MS96] Th. Mosch and P. Solina. Burst mode communication. In Ulrich Killat (editor), editor, *Access to B-ISDN via PONs – ATM Communication in Practice*, pages 157–175. John Wiley and B. G. Teubner, 1996.

[MSR+98] M. Möller, T. F. Meister R. Schmid, J. Rupeter, M. Rest, A. Schöpflin, and H.-M. Rein. SiGe retiming high-gain power MUX for direct driving an EAM up to 50Gb/s. *Electronics Letters*, Vol. 34(18):1782–1784, September 1998.

[MSW+97] Mehran Mokhtari, Thomas Swahn, Robert H. Walden, William E. Stanchina, Michael Kardos, Tarja Juhola, Gerd Schuppener, Hannu Tenhunen, and Thomas Lewin. InP-HBT chip-set for 40-Gb/s fiber optical communication systems operational at 3V. *IEEE J. Solid-State Circuits*, SC-32(9):1371–1383, September 1997.

[MTM+00] Jens Müllrich, Herbert Thurner, Ernst Müllner, Joseph F. Jensen, William E. Stanchina, M. Kardos, and Hans-Martin Rein. High-gain transimpedance amplifier in InP-based HBT technology for receiver in 40-Gb/s optical-fiber TDM links. *IEEE J. Solid-State Circuits*, SC-35(9):1260–1265, September 2000.

[MYK+97] Miyo Miyashita, Naohito Yoshida, Yoshiki Kojima, Toshiaki Kitano, Norio Higashisaka, Junichi Nakagawa, Tadashi Takagi, and Mutsuyuki Otsubo. An AlGaAs/InGaAs pseudomorphic HEMT modulator driver IC with low power dissipation for 10-Gb/s optical transmission systems. *IEEE Trans. on Microwave Theory and Techniques*, MTT-45(7):1058–1064, July 1997.

[NCI00] NCITS. Fibre channel – methodologies for jitter specification – 2, T11.2 / Project 1316-DT / Rev 0.0. National Committee for Information Technology Standardization, April 2000.

[NIA98] Makoto Nakamura, Noboru Ishihara, and Yukio Akazawa. A 156-Mb/s CMOS optical receiver for burst-mode transmission. *IEEE J. Solid-State Circuits*, SC-33(8):1179–1187, August 1998.

[Nor83] Ernst H. Nordholt. *The Design of High-Performance Negative-Feedback Amplifiers*. Elsevier Scientific Publishing Company, Amsterdam, NL, 1983.

[NRW96] Michael Neuhäuser, Hans-Martin Rein, and Horst Wernz. Low-noise, high-gain Si-bipolar preamplifiers for 10Gb/s optical-fiber links – design and realization. *IEEE J. Solid-State Circuits*, SC-31(1):24–29, January 1996.

[Ølg94] Christian Ølgaard. A laser control chip combining power regulator and a 622-MBit/s modulator. *IEEE J. Solid-State Circuits*, SC-29(8):947–951, August 1994.

[Ols89] N. A. Olsson. Lightwave systems with optical amplifiers. *Journal of Lightwave Technology*, LT-7(7):1071–1082, July 1989.

[OMI+99] Kenichi Ohhata, Toru Masuda, Kazuo Imai, Ryoji Takeyari, and Katsuyoshi Washio. A wide-dynamic-range, high-transimpedance Si bipolar preamplifier IC for 10-Gb/s optical fiber links. *IEEE J. Solid-State Circuits*, SC-34(1):18–24, January 1999.

[OMOW99] Kenichi Ohhata, Toru Masuda, Eiji Ohue, and Katsuyoshi Washio. Design of a 32.7-GHz bandwidth AGC amplifier IC with wide dynamic range implemented in SiGe HBT. *IEEE J. Solid-State Circuits*, SC-34(9):1290–1297, September 1999.

[OS90] Yusuke Ota and Robert G. Swartz. Burst-mode compatible optical receiver with a large dynamic range. *Journal of Lightwave Technology*, LT-8(12):1897–1903, December 1990.

[OSA+94] Yusuke Ota, Robert G. Swartz, Vance D. Archer III, Steven K. Korotky, Mihai Banu, and Alfred E. Dunlop. High-speed, burst-mode, packet-capable optical receiver and instantaneous clock recovery for optical bus operation. *Journal of Lightwave Technology*, 12(2):325–331, February 1994.

[PA94] Patrick K. D. Pai and Asad A. Abidi. A 40-mW 55Mb/s CMOS equalizer for use in magnetic storage read channels. *IEEE J. Solid-State Circuits*, SC-29(4):489–499, April 1994.

[PD97] Mary R. Phillips and Thomas E. Darcie. Lightwave analog video transmission. In Ivan P. Kaminow and Thomas L. Koch, editors, *Optical Fiber Telecommunications IIIA*, pages 523–559. Academic Press, San Diego, 1997.

[Per73] S. D. Personick. Receiver design for digital fiber optic communication systems. *The Bell System Technical Journal*, 52(6):843–886, July-August 1973.

[PP97] A. J. Price and K. D. Pedrotti. Optical transmitters. In Jerry D. Gibson, editor, *The Communications Handbook*, pages 774–788. CRC Press, 1997.

[PT00] Sung Min Park and C. Toumazou. A packaged low-noise high-speed regulated cascode transimpedance amplifier using a 0.6 μm N-well CMOS technology. In *Digest of European Solid-State Circuits Conference*, September 2000.

[Ran01] Hans Ransijn. Receiver and transmitter IC design, May 2001. CICC'2001 Ed. Session 3-2.

[Raz96] Behzad Razavi. A 1.5V 900MHz downconversion mixer. In *ISSCC Dig. Tech. Papers*, pages 48–49, February 1996.

[Raz00] Behzad Razavi. A 622Mb/s, 4.5pA/$\sqrt{\text{Hz}}$ CMOS transimpedance amplifier. In *ISSCC Dig. Tech. Papers*, pages 162–163, February 2000.

[RDR+01] Mario Reinhold, Claus Dorschky, Eduard Rose, Rajasekhar Pullela, Peter Mayer, Frank Kunz, Yves Baeyens, Thomas Link, and John-Paul Mattia. A fully integrated 40-Gb/s clock and data recovery IC with 1:4 DEMUX in SiGe technology. *IEEE J. Solid-State Circuits*, SC-36(12):1937–1945, December 2001.

[RDS+92] Klaus Runge, Detlef Daniel, R. D. Standley, James L. Gimlett, Randall B. Nubling, Richard L. Pierson, Steve M. Beccue, Keh-Chung Wang, Neng-Haung Sheng, Mau-Chung F. Chang, Dong Ming Chen, and Peter M. Asbeck. AlGaAs/GaAs HBT IC's for high-speed lightwave transmission systems. *IEEE J. Solid-State Circuits*, SC-27(10):1332–1341, October 1992.

[Rei88] Hans-Martin Rein. Multi-gigabit-per-second silicon bipolar IC's for future optical-fiber transmission systems. *IEEE J. Solid-State Circuits*, SC-23(3):664–675, June 1988.

[Rei01] Hans-Martin Rein. Design of high-speed Si/SiGe bipolar ICs for optical-fiber systems with data rates up to 40Gb/s, March 2001. Lecture Notes, MEAD Microelectronics.

[RM96] H.-M. Rein and M. Möller. Design considerations for very-high-speed Si-bipolar IC's operating up to 50Gb/s. *IEEE J. Solid-State Circuits*, SC-31(8):1076–1090, August 1996.

[RR87] Reinhard Reimann and Hans-Martin Rein. Bipolar high-gain limiting amplifier IC for optical-fiber receivers operating up to 4Gbit/s. *IEEE J. Solid-State Circuits*, SC-22(4):504–511, August 1987.

[RR89] Reinhard Reimann and Hans-Martin Rein. A single-chip bipolar AGC amplifier with large dynamic range for optical-fiber receivers operating up to 3Gbit/s. *IEEE J. Solid-State Circuits*, SC-24(6):1744–1748, December 1989.

[RS98] Rajiv Ramaswami and Kumar N. Sivarajan. *Optical Networks: A Practical Perspective*. Morgan Kaufmann Publishers, San Francisco, 1998.

[RSDG01] Hans Ransijn, Gregory Salvador, Dwight D. Daugherty, and Kenneth D. Gaynor. A 10-Gb/s laser/modulator driver IC with dual-mode actively matched output buffer. *IEEE J. Solid-State Circuits*, SC-36(9):1314–1320, September 2001.

[RSW+94] H.-M. Rein, R. Schmid, P. Wenger, T. Smith, T. Herzog, and R. Lachner. A versatile Si-bipolar driver circuit with high output voltage swing for external and direct laser modulation in 10Gb/s optical-fiber links. *IEEE J. Solid-State Circuits*, SC-29(9):1014–1021, September 1994.

[RZP+99] K. Runge, P. J. Zampardi, R. L. Pierson, R. Yu, P. B. Thomas, S. M. Beccue, and K. C. Wang. AlGaAs/GaAs HBT circuits for optical TDM communications. In Keh-Chung Wang, editor, *High-Speed Circuits for Lightwave Communications*, pages 161–191. World Scientific, Singapore, 1999.

[SA86] T. M. Shen and Govind P. Agrawal. Pulse-shape effects on frequency chirping in single-frequency semiconductor lasers under current modulation. *Journal of Lightwave Technology*, LT-4(5):497–503, May 1986.

[Säc89] Eduard Säckinger. Theory and monolithic CMOS integration of a differential difference amplifier. In W. Fichtner, W. Guggenbühl, H. Melchior, and G. S. Moschytz, editors, *Series in Microelectronics*. Hartung-Gorre Verlag, Konstanz, Germany, 1989.

[SBL91] Norman Scheinberg, Robert J. Bayruns, and Timothy M. Laverick. Monolithic GaAs transimpedance amplifiers for fiber-optic receivers. *IEEE J. Solid-State Circuits*, SC-26(12):1834–1839, December 1991.

[SD89] M. Sherif and P. A. Davies. Decision-point steering in optical fibre communication systems: Theory. *IEE Proceedings, Pt. J*, 136(3):169–176, June 1989.

[SDC+95] Michiel S. J. Steyaert, Wim Dehaene, Jan Craninckx, Máirtín Walsh, and Peter Real. A CMOS rectifier-integrator for amplitude detection in hard disk servo loops. *IEEE J. Solid-State Circuits*, SC-30(7):743–751, July 1995.

[Sen85] John M. Senior. *Optical Fiber Communications – Principles and Practice*. Prentice Hall, 1985.

[SF00] Eduard Säckinger and Wilhelm C. Fischer. A 3GHz, 32dB CMOS limiting amplifier for SONET OC-48 receivers. *IEEE J. Solid-State Circuits*, SC-35(12):1884–1888, December 2000.

[SG87] Eduard Säckinger and Walter Guggenbühl. A versatile building block: The CMOS differential difference amplifier. *IEEE J. Solid-State Circuits*, SC-22(2):287–294, April 1987.

[SH97] Stefanos Sidiropoulos and Mark Horowitz. A 700-Mb/s/pin CMOS signaling interface using current integrating receivers. *IEEE J. Solid-State Circuits*, SC-32(5):681–690, May 1997.

[Shu88] P. W. Shumate. Lightwave transmitters. In Stewart E. Miller and Ivan P. Kaminow, editors, *Optical Fiber Telecommunications II*, pages 723–757. Academic Press, San Diego, 1988.

[SMRR98] R. Schmid, T. F. Meister, M. Rest, and H.-M. Rein. 40Gb/s EAM driver IC in SiGe bipolar technology. *Electronics Letters*, Vol. 34(11):1095–1097, May 1998.

[SMRR99] R. Schmid, T. F. Meister, M. Rest, and H.-M. Rein. SiGe driver circuit with high output amplitude operating up to 23Gb/s. *IEEE J. Solid-State Circuits*, SC-34(6):886–891, June 1999.

[SO01a] Eduard Säckinger and Yusuke Ota. Burst-mode laser techniques. U.S. Patent No 6,229,830, May 2001.

[SO01b] Eduard Säckinger and Yusuke Ota. Burst-mode laser techniques. U.S. Patent No 6,219,165, April 2001.

[SOGF00] Eduard Säckinger, Yusuke Ota, Thaddeus J. Gabara, and Wilhelm C. Fischer. A 15mW, 155Mb/s CMOS burst-mode laser driver with automatic power control and end-of-life detection. *IEEE J. Solid-State Circuits*, SC-35(2):269–275, February 2000.

[SP82] R. G. Smith and S. D. Personick. Receiver design for optical fiber communication systems. In H. Kressel, editor, *Topics in Applied Physics Vol. 39 – Semiconductor Devices for Optical Communication*. Springer Verlag, Berlin, Germany, 1982.

[SR99] Jafar Savoj and Behzad Razavi. A CMOS interface circuit for detection of 1.2Gb/s RZ data. In *ISSCC Dig. Tech. Papers*, pages 278–279, February 1999.

[SSA+99] Yasuyuki Suzuki, Hidenori Shimawaki, Yasushi Amamiya, Nobuo Nagano, Hitoshi Yano, and Kazuhiko Honjo. A 40-Gb/s preamplifier using AlGaAs/InGaAs HBT's with regrown base contacts. *IEEE J. Solid-State Circuits*, SC-34(2):143–147, February 1999.

[SSO+92] Yasuyuki Suzuki, Tetsuyuki Suzaki, Yumi Ogawa, Sadao Fujita, Wendy Liu, and Akihiko Okamoto. Pseudomorphic 2DEG FET IC's for 10-Gb/s optical communication systems with external optical modulation. *IEEE J. Solid-State Circuits*, SC-27(10):1342–1346, October 1992.

[SSO+97] E. Sano, K. Sano, T. Otsuij, K. Kurishima, and S. Yamahata. Ultra-high speed, low power monolithic photoreceiver using InP/InGaAs double heterojunction bipolar transistors. *Electronics Letters*, Vol. 33(12):1047–1048, June 1997.

[SSS+01] Hisao Shigematsu, Masaru Sato, Toshihide Suzuki, Tsuyoshi Takahashi, Kenji Imanishi, Naoki Hara, Hiroaki Ohnishi, and Yuu Watanabe. A 49-GHz preamplifier with a transimpedance gain of 52dBΩ using InP HEMTs. *IEEE J. Solid-State Circuits*, SC-36(9):1309–1313, September 2001.

[Ste01] Michiel Steyaert. CMOS optical communication circuits, March 2001. Lecture Notes, MEAD Microelectronics.

[STS+94] Masaaki Soda, Hiroshi Tezuka, Fumihiko Sato, Takasuke Hashimoto, Satoshi Nakamura, Toru Tatsumi, Tetsuyuki Suzaki, and Tsutomu Tashiro. Si-analog IC's for 20Gb/s optical receiver. *IEEE J. Solid-State Circuits*, SC-29(12):1577–1582, December 1994.

[Sze81] S. M. Sze. *Physics of Semiconductor Devices*. John Wiley & Sons, New York, 2nd edition, 1981.

[Sze98] S. M. Sze (editor). *Modern Semiconductor Device Physics*. John Wiley & Sons, New York, 1998.

[TSN+98a] Akira Tanabe, Masaaki Soda, Yasushi Nakahara, Takao Tamura, Kazuyoshi Yoshida, and Akio Furukawa. A single-chip 2.4-Gb/s CMOS optical receiver IC with low substrate cross-talk preamplifier. *IEEE J. Solid-State Circuits*, SC-33(12):2148–2153, December 1998.

[TSN+98b] Akira Tanabe, Masayuki Soda, Yasushi Nakahara, Akio Furukawa, Takao Tamura, and Kazuyoshi Yoshida. A single chip 2.4Gb/s CMOS optical receiver IC with low substrate crosstalk preamplifier. In *ISSCC Dig. Tech. Papers*, pages 304–305, February 1998.

[Tuc85] Rodney S. Tucker. High-speed modulation of semiconductor lasers. *Journal of Lightwave Technology*, LT-3(6):1180–1192, December 1985.

[VT95] Tongtod Vanisri and Chris Toumazou. Integrated high frequency low-noise current-mode optical transimpedance preamplifiers: Theory and practice. *IEEE J. Solid-State Circuits*, SC-30(6):677–685, June 1995.

[Wal99] Robert H. Walden. A review of recent progress in InP-based optoelectronic integrated circuit receiver front-ends. In Keh-Chung Wang, editor, *High-Speed Circuits for Lightwave Communications*, pages 319–330. World Scientific, Singapore, 1999.

[Wan99] Keh-Chung Wang (editor). *High-Speed Circuits for Lightwave Communication*. World Scientific, Singapore, 1999.

[WBN+93] Zhi-Gong Wang, Manfred Berroth, Ulrich Nowotny, Manfred Ludwig, Peter Hofmann, Alex Hülsmann, Klaus Köhler, Brian Raynor, and Joachim Schneider. Integrated laser-diode voltage driver for 20-Gb/s optical systems using 0.3-μm gate length quantum-well HEMT's. *IEEE J. Solid-State Circuits*, SC-28(7):829–834, July 1993.

[WFBS96] Thomas Y. K. Wong, Al P. Freundorfer, Bruce C. Beggs, and John Sitch. A 10Gb/s AlGaAs/GaAs HBT high power fully differential limiting distributed amplifier for III-V Mach-Zehnder modulator. *IEEE J. Solid-State Circuits*, SC-31(10):1388–1393, October 1996.

[WG90] Jack H. Winters and Richard D. Gitlin. Electrical signal processing techniques in long-haul fiber-optic systems. *IEEE Trans. on Communications*, COM-38(9):1439–1453, September 1990.

[WHKY98] Richard C. Walker, Kuo-Chiang Hsieh, Thomas A. Knotts, and Chu-Sun Yen. A 10Gb/s Si-bipolar TX/RX chipset for computer data transmission. In *ISSCC Dig. Tech. Papers*, pages 302–303, February 1998.

[WK92] Jack H. Winters and Sanjay Kasturia. Adaptive nonlinear cancellation for high-speed fiber-optic systems. *Journal of Lightwave Technology*, LT-10(7):971–977, July 1992.

[WK98] T. K. Woodward and A. V. Krishnamoorthy. 1Gb/s CMOS photoreceiver with integrated detector operating at 850 nm. *Electronics Letters*, Vol. 34(12):1252–1253, June 1998.

[Won93] Thomas T. Y. Wong. *Fundamentals of Distributed Amplification*. Artech House, Boston, 1993.

[YKNS95] K. Yonenaga, S. Kuwano, S. Norimatsu, and N. Shibata. Optical duobinary transmission system with no receiver sensitivity degradation. *Electronics Letters*, Vol. 31(4):302–304, February 1995.

[Yod98] James Daniel Yoder. Optical receiver preamplifier dynamic range enhancing circuit and method. U.S. Patent No 5,734,300, March 1998.

[ZNK97] John L. Zyskind, Jonathan A. Nagel, and Howard D. Kidorf. Erbium-doped fiber amplifiers for optical communications. In Ivan P. Kaminow and Thomas L. Koch, editors, *Optical Fiber Telecommunications IIIB*, pages 13–68. Academic Press, San Diego, 1997.