INTRODUCTION

1

1.1 GENERAL CONSIDERATION

We are in the era of communication. The ever growing volume of communication makes different kinds of standards evolute exponentially. A summary of communication standards evolution over the past 3 decades has been plotted in Fig. 1.1. For example, Ethernet grows up steadily with a 10X speed improvement between generations, and Universal Serial Bus (USB) are moving toward data rate of 10 Gb/s in its newest standard USB 3.1. Huge amount of data is transmitted everyday

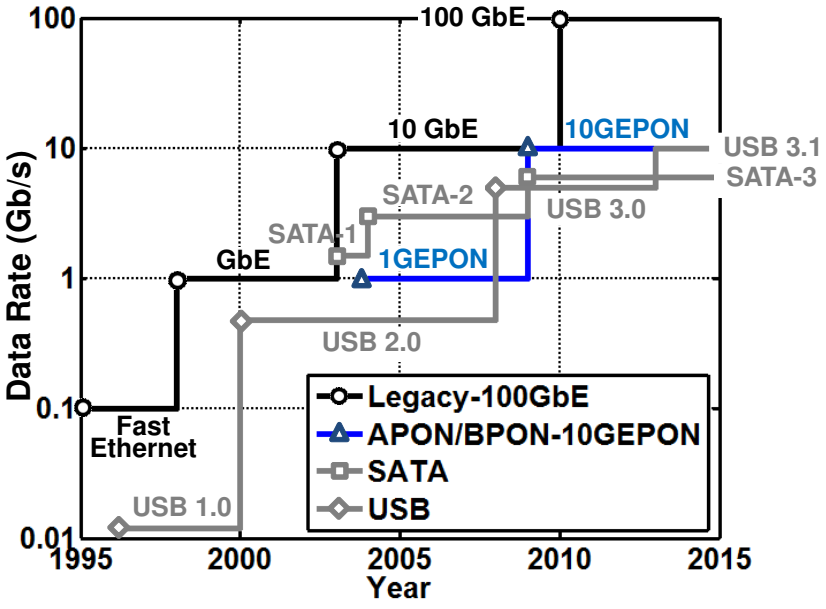


Fig. 1.1 Wireline communication upgrades.

around us, and trend of upgrades does not seem to slow down in the near future. It is estimated that more than 21 billion networked devices and connections will exist in our planet, and global IP traffic would be tripled from 2013 to 2018. The role of wireline communications, including backbone network, fibers, backplane, and chip-to-chip transceivers, are of great importance.

There are two key factors which have significant influence on the development of communication networks: device technology and circuit design. Figure 1.2 reveals the evolution for phase-locked loops (PLLs) ICs ever published since 1990's, revealing operation frequency improvement by 4 orders over the past 25 years [1]−[4]. However, the technology itself has been improved

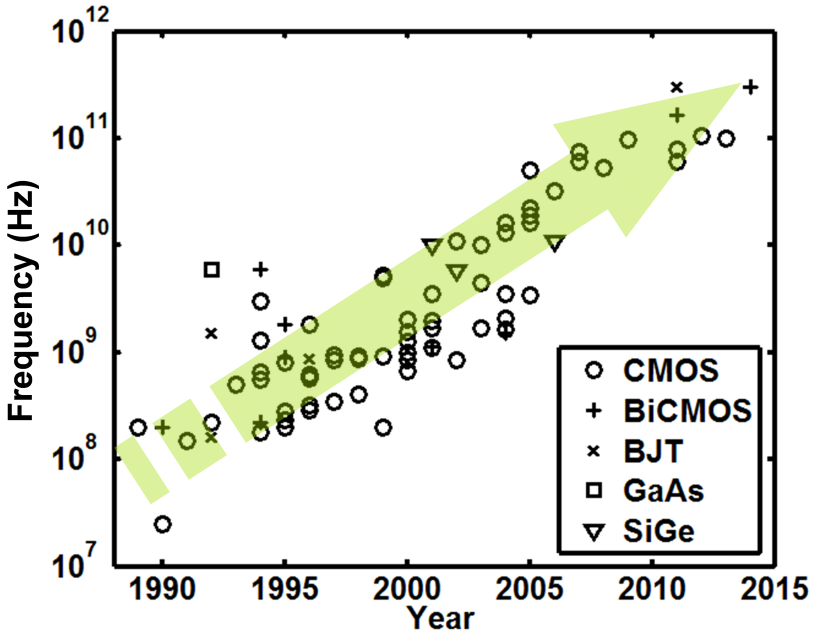


Fig. 1.2 Evolution of PLL circuits.

by only 100 times or less in terms of speed. For example, the mainstream CMOS in 90's is 0.8 μm, whose transit frequency (f_T) is about 12 GHz. In 2014, 40-nm CMOS in analog circuits has an f_T around 280 GHz. It implies that the intelligent work of IC designers contributes the other 100X development. Same situation can be found in other important blocks, such as clock and data recovery (CDR) circuits. Illustrated in Fig. 1.3(a) is an evolution plot for the CMOS CDRs, which presents 2.5X faster improvements than the technology nodes.

.

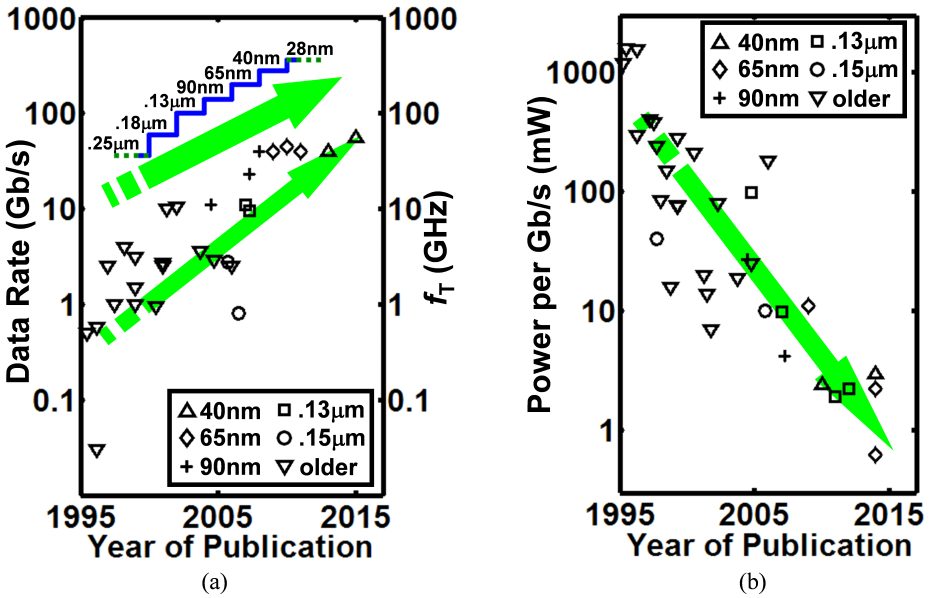


Fig. 1.3 CMOS circuits improvement trends: (a) CDR circuit, (b) I/O's power efficiency.

Other than speed, power consumption is of great concern. For example, a processor or a SOC chip needs to increase the power efficiency of the input-outputs (I/Os) in order to accommodate more communication channels. The I/O power efficiency improves in a rate approximately 1.4 times per year, arriving at about 0.63 mW per Gb/s in a data rate of 9 Gb/s in today's technology [Fig. 1.3(b)] [5].

Over the decades, supply voltages of CMOS technologies have been reduced from 5 V to 0.9 V (Fig. 1.4). Analog circuit designers must reform the circuit architecture from time to time to adopt new supply voltages. In most CMOS technologies, the shrinking of threshold voltages is slower than that of the supply voltages. It is an important factor for the exponentially growing up gate counts, otherwise the sub-threshold current would soon dominate the overall power consumption. However, it also makes analog circuits harder and harder to stack devices on top of each other. Consider a k-stage Gilbert cell in current-mode logic (CML) whose switching pairs are equally sized as $(\frac{W}{L})$ [Fig. 1.5(a)]. Stacking k stages is equivalent to a single switching pair with size $(\frac{W}{kL})$, as all inputs are in CML [Fig. 1.5(b)]. In other words, the circuit's current driving capability

is weakened by a factor of k. The required overdrive voltage would soon kill the tail current if k is large. With a supply of 0.9 V, k = 2 is barely acceptable.

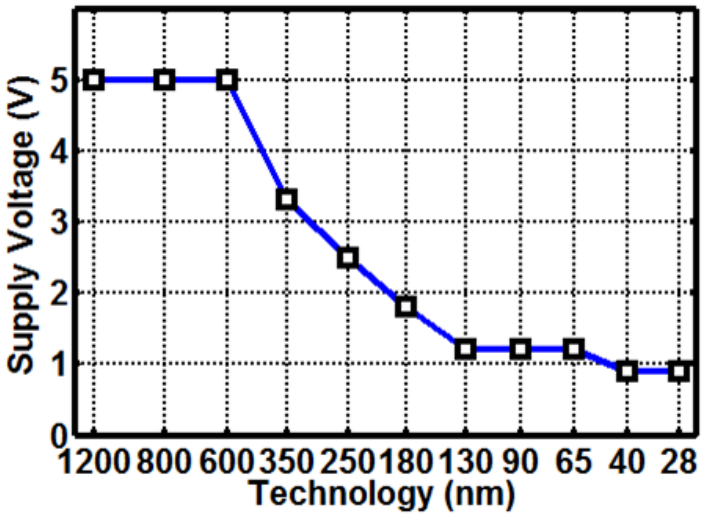


Fig. 1.4 Supply-voltage migration.

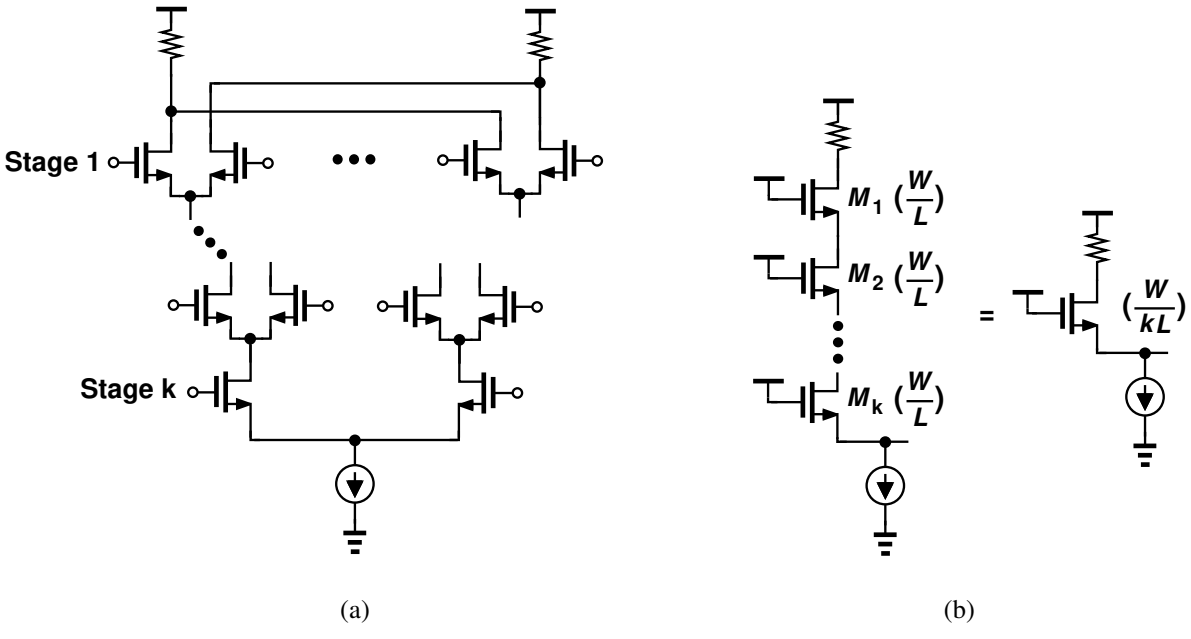


Fig. 1.5 (a) Gilbert-cell with k stages, (b) equivalent circuit in operation.

Example 1.1

Design a 3-input half adder in CML mode. Flatten the circuit structure as much as possible.

Solution:

An half adder produces an output ONE when there are odd number inputs of logic ONE. In low supply environments, CML swing can be set to $V_{DD}/2 \sim V_{DD}/3$. A straightforward realization is shown in Fig. 1.6(a). The over-stacking structure can be flattened as shown in Fig. 1.6(b). Here, V_A, V_B, and V_C drive three identical differential pairs with different polarities. One extra branch flows half amount of tail current to balance the output dc level, and peaking components are added to extend the bandwidth. As the thermometer code varies from 000 to 111, it forms an alternate output and $|V_{out}|$ is always equal to $I_{SS}R$. In other words, an LSB decoder with much faster operation is presented. Note that with I_{SS} = 2.5 mA and R = 2 kΩ, sufficient output swing would be obtained if the inputs are greater than 300 mV.

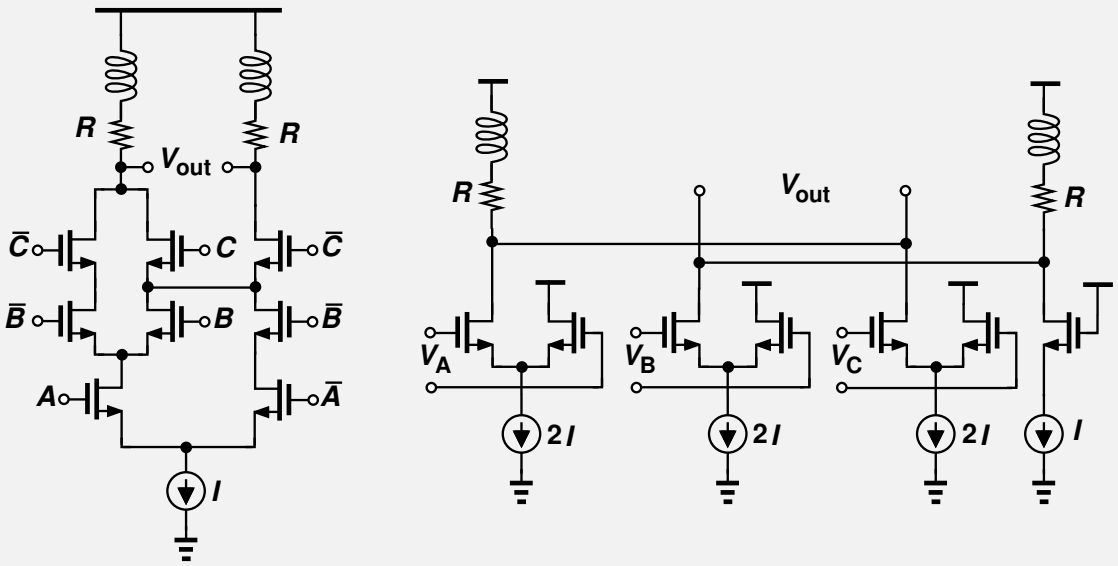


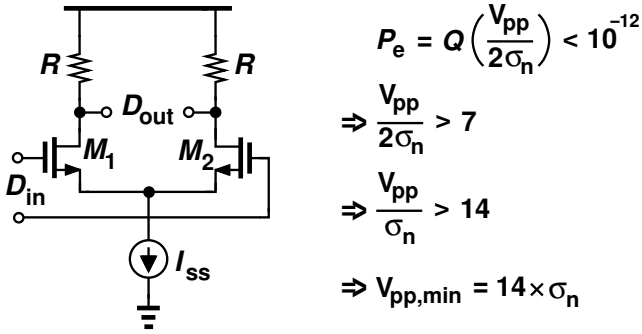
Fig. 1.6 High-speed half adder design: (a) conventional, (b) flattened.

How large a swing should we have for a CML block? Ideally, a larger swing is always preferable as it leads to better signal to noise ratio (SNR). However, in order to keep all devices in (or

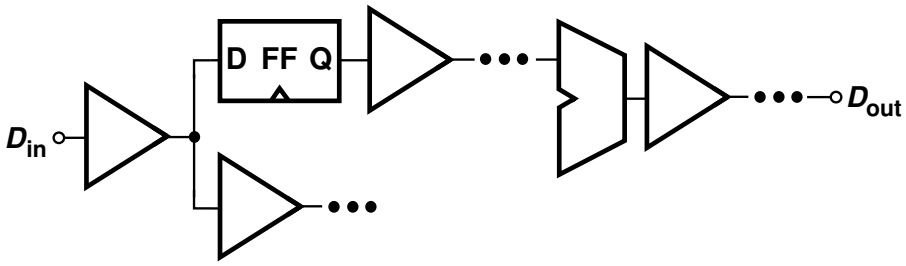
close to) saturation region, we usually set a single-ended peak-to-peak swing to be $400 \sim 600$ mV for low-supply high-speed circuits. The bottomline is to maintain an acceptable SNR in the worst case scenario. As we know, the additive noise imposed on a signal waveform potentially causes errors. The normal distribution on vertical jitter give rise to error probability as

$$P_e = \int_{\frac{V_{pp}}{2\,\sigma_n}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\Big(\frac{-x^2}{2}\Big)\,dx = Q\Big(\frac{V_{pp}}{2\,\sigma_n}\Big), \tag{1.1}$$

where V_{pp} denotes the signal swing and σ_n the standard deviation of noise distribution (or equivalently, rms jitter). To achieve BER $< 10^{-12}$, the SNR (V_{pp}/σ_n) must be greater than 14. Figure 1.7(a) illustrates the calculation.



(a)



(b)

Fig. 1.7 (a) Calculation of minimum required swing in CML, (b) a typical data path.

The above analysis only stands for a single CML buffer. In practice, the signal may experience quite a few blocks with similar CML structure before arriving at the final output, where noise from all blocks is accumulated [Fig. 1.7(b)]. For example, the overall output noise may be 5 or 10 times larger than a single differential pair. On the other hand, the equalization in the transmit side

(e.g., FFE) would further reduce the effective signal magnitude. For example, if the feedforward equalizer (FFE) in a transmitter provides 9.5-dB compensation at Nyquist frequency, the signal swing is basically shrunk to 1/3 of the full scale in the receiver's input. In a SerDes design, the minimum input swing (or equivalently, power) that allows a transceiver to correctly deliver data is known as the input sensitivity to the receiver, which is one of the key specifications. Nonetheless, if magnitude degradation is taken into consideration, the CML swing (from the TX's output) must be several times larger than the minimum requirement.

What is the ultimate supply voltage that a high-speed circuit can tolerate? Let us check the differential pair in Fig. 1.7(a) again. Based on our discussion, a simple differential pair may need a peak-to-peak swing of at least 250 mV to maintain signal integrity. We need one overdrive ($V_{GS} - V_{TH}$) for switching pair $M_{1,2}$ and one overdrive for I_{SS}. Overall speaking, the supply voltage has to be greater than 750 mV, given that one overdrive is roughly equal to 250 mV. It implies the supply voltage shrinking for analog/mixed-mode circuits would stop at 0.8 $\sim$ 0.9 V, unless better circuit structures can be invented.

It is worth noting that some processes have special devices with lower threshold voltages. These low-V_{TH} devices provide more current driving capability. That is, for a given current, the device size could be reduced. As a result, the parasitic capacitance decreases and the bandwidth increases. Experiments show that 20 $\sim$ 30% bandwidth improvement can be observed if $M_{1,2}$ in Fig. 1.7(a) are made of low-V_{TH} devices. However, the low-V_{TH} devices do not provide additional voltage headroom. The reader can prove that the minimum headroom does not change even though the tail current is replaced with a low-V_{TH} device.

It is also important to realize that, despite many merits, CMOS devices still suffer from bandwidth disadvantage if compared with their bipolar counterparts. For example, the transit frequency of an NMOS transistor with L = 65 nm, W = 1 μm, and $V_{GS} - V_{TH}$ = 250 mV is approximately equal to 180 GHz. Using this device in Fig. 1.7(a), we need 450 $\mathrm{mV_{pp}}$ single-ended input to ensure complete switching of the tail current I_{SS} (= 2 mA). This value is about 4 $\sim$ 5 times larger than that in bipolar devices with similar f_T. In other words, a bipolar transistor with the same f_T

allows much faster operation. Owing to this issue, some high-speed CMOS circuits are prone to be realized in sub-rate or parallelized structures.

1.2 PRBS

To fully use the available bandwidth, wireline data links usually deal with raw digital data without modulation. A random data sequence toggling between 0 and 1 randomly with a bit period T_b reveals a time domain expression [6]

$$x_1(t) = \sum_k b_k \, p(t - kT_b), \tag{1.2}$$

where $b_k \in \{0, 1\}$ and $p(t)$ is an ideal pulse with unity magnitude and pulsewidth T_b [Fig. 1.8(a)]. In general, such a random sequence possesses spectrum as [7]

$$S(\omega) = \frac{\sigma^2}{T_b}|P(\omega)|^2 + \frac{m^2}{T_b^2}\sum_k \left|P\Big(\frac{2\pi k}{T_b}\Big)\right|^2 \delta\Big(\omega - \frac{2\pi k}{T_b}\Big), \tag{1.3}$$

where σ^2 denotes the pulse variance, $P(w)$ the Fourier transform of $p(t)$, and m the mean amplitude of it. Thus, $\sigma^2 = 1/4$, $m = 1/2$, and

$$S_{x_1}(\omega) = \frac{1}{4T_b}\left[\frac{\sin(\omega T_b/2)}{\omega/2}\right]^2 + \frac{1}{4}\,\delta(\omega), \tag{1.4}$$

as shown in Fig. 1.8(a). Called "sinc" function, the first term presents nulls at data rate and its higher-order harmonics. The main lobe peaks at dc with a value of $T_b/4$, whereas the second lobe reaches a maximum of $T_b/(9\pi^2)$ at $w = 3\pi/T_b$. The 13.3-dB difference between the two implies that most power is concentrated in the main lobe. Integration of the power spectrum density gives rise to the total power

$$\int_{-\infty}^{\infty} S_{x_1}(\omega = 2\pi f)df = \int_{-\infty}^{\infty} \frac{1}{4T_b}\left[\frac{\sin(\pi f\, T_b)}{\pi f}\right]^2 df + \int_{-\infty}^{\infty} \frac{1}{4}\,\delta(f)\, df \tag{1.5}$$

$$= \frac{1}{4} + \frac{1}{4}, \tag{1.6}$$

where the first term represents the data power and the second the dc power.

Focusing on data power, we calculate the main lobe power and obtain

$$\int_{-\frac{1}{T_b}}^{\frac{1}{T_b}} \frac{1}{4T_b}\left[\frac{\sin(\pi f\, T_b)}{\pi f}\right]^2 df = \frac{1}{4}\cdot 0.9. \tag{1.7}$$

In other words, the main lobe contains 90% of signal power. It can be easily proven that 48.6% of signal power is contained from dc to Nyquist frequency [$f = 1/(2T_b)$].

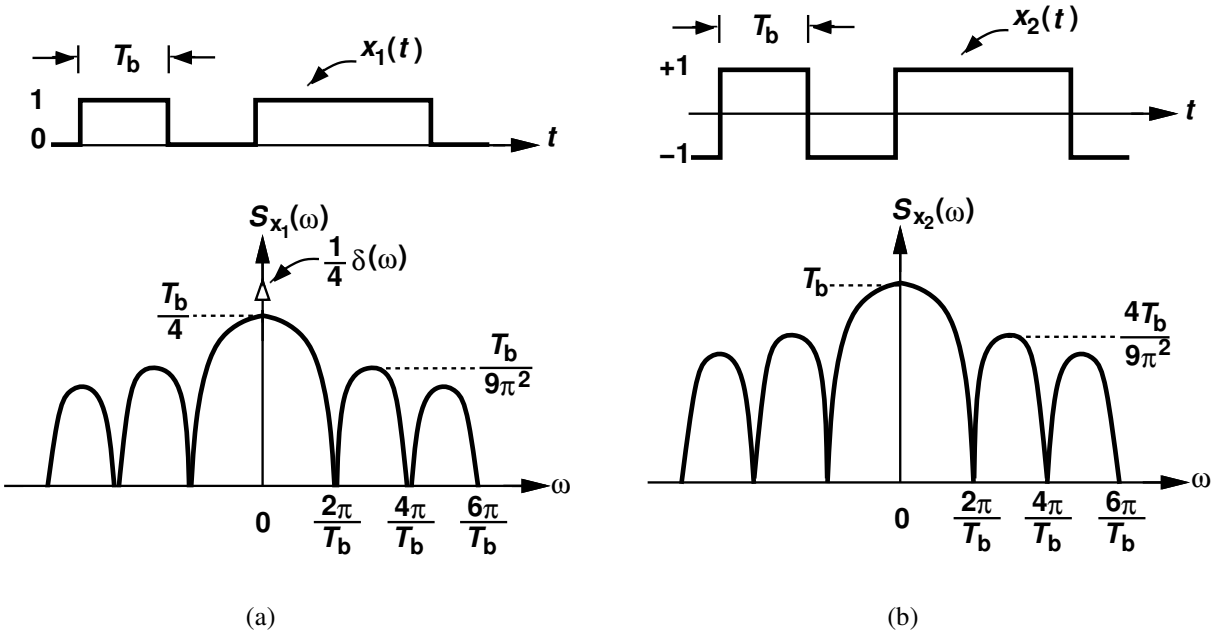


Fig. 1.8 Random sequence and spectrum (a) with, (b) without dc offset.

The reader should not be confused by the dc term of Eq. (1.6). For a balanced random data sequence with $m = 0$, the impulse of Eq. (1.4) is gone and the dc power of Eq. (1.6) disappears. Similarly, for a zero-dc random sequence $x_2(t)$ with $\{+1, -1\}$ magnitude, its power spectral density is equal to

$$S_{x_2}(\omega) = \frac{1}{T_b}\left[\frac{\sin(\omega T_b/2)}{\omega/2}\right]^2, \tag{1.8}$$

which is 4 times the first term (data power) of $S_{x_1}(w)$.

Since it is quite difficult to generate a true random data sequence, we instead create pseudo random binary sequence (PRBS) for testing, which is implemented by means of a linear feedback shift register that produces randomized (but still periodic) data sequence. Depending on the length,

it can provide PRBS with different randomness. A linear feedback shift register is actually characterized by the so-called "feedback polynomial". Consider a n-degree polynomial with only 1 or 0 coefficients. If it can not be further decomposed as a product of lower degree polynomials, we call it a primitive. For example, $p(x) = x^4+x^3+1$ is primitive, whereas x^4+x^3+x+1 is non-primitive because $x^4 + x^3 + x + 1 = (x^2 + x + 1)(x^2 + 1)$. Note that the arithmetic conducting here is modulo-2 operation, i.e., $x^n + x^n = x^n - x^n = 0$. Figure 1.9(a) illustrates examples of primitive polynomials with different degrees. We also define reciprocal polynomial $p^*(x) = x^n \cdot p(1/x)$. For example, if $p_4(x) = x^4+x^3+1$, then $p_4^*(x) = x^4+x+1$. For a given degree n, it is possible to find more than one primitive polynomial. Moreover, if a polynomial is primitive, then its reciprocal is also primitive.

The polynomial can be used to form a linear feedback shift register, which produces PRBS. Shown in Fig. 1.9(b) is an example of n = 4. Here, bit sequences are shifted from the very left $(x^0 = 1)$ to the very right (x^n), and the non-zero terms are taken out and XOR'd in the feedback loop. Driven by CK_{in}, the output x_4 presents $\{1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0\}$ and repeats itself every 15 bits. Actually, all terms $x_0, x_1, \cdots, x_n$ produce the same sequences with different shifting. The reader can find that the sequence is almost balanced, i.e., the difference between number of ones and zeros is always 1. If the shift register is realized based on a primitive polynomial with degree n, then the output would present length of $2^n - 1$ pseudo random data (called PRBSn), and the maximum number of consecutive bits is n. Note that a shift register with non-primitive polynomial would lead to bit sequence length less than $2^n - 1$.

The spectrum of a PRBS is slightly different from that of a real random data sequence. The periodicity of a PRBS suggests a spectrum with impulses. Since each pulse bit repeats every 2^n-1 bits, it is nothing more than one unit sequence of length $2^n - 1$ (bit period = T_b) convoluting with time-domain impulses [separated by $(2^n - 1)T_b$]. As a result, by convolution theory, the PRBS spectrum is the product of the spectrum of purely random data and a spectrum with frequency domain impulses separated by $2\pi/[(2^n-1)T_b]$.[1] In our case of n = 4, the PRBS spectrum is illustrated

[1]Here we assume one unit of data sequence $(2^n - 1)$ is long enough such that it have very similar spectrum as a real random data sequence. For $n \geq 7$, it is indeed true.

in Fig. 1.9(c). For a larger n, the sequence becomes more random and its spectral lines get closer to each other.

It is also instructive to observe the waveform with limited bandwidth. Consider that an ideal PRBS is fed into a filter that cuts off some side lobes. The output data eye gets round (rise/fall time increases) as high frequency side lobes are removed. With the main lobe only, the $20 \sim 80\%$ rise/fall time become $0.44\ T_b$ [Fig. 1.9(d)]. Such a high frequency loss makes the data defective and prone to error. Unfortunately, all high-speed wireline communication systems suffer from high frequency loss. We address channel loss issues in Chapter 5.

The PRBS generator in Fig. 1.9(b) suffers from speed limitation due to clock-to-Q delay and gate delay. For example, if $2^{16} - 1$ PRBS is to be generated, we resort to the feedback polynomial

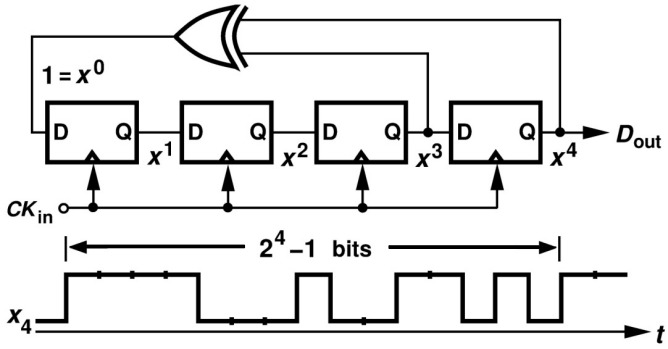
$$P(x) = x^{16} + x^{14} + x^{13} + x^{11} + 1 \tag{1.9}$$

and obtain the circuit implementation as depicted in Fig. 1.10(a). Due to the 3 XOR gates in serial, the clock cycle must be greater than (FF setup time + FF clock-to-Q delay + 3 XOR gate delay). An alternate structure (called Galois Configuration) splits the XOR chain into individual gates in conjunction with FFs, arriving at a structure as shown in Fig. 1.10(b). Here, the order of the taps are flipped to generate the same output stream.

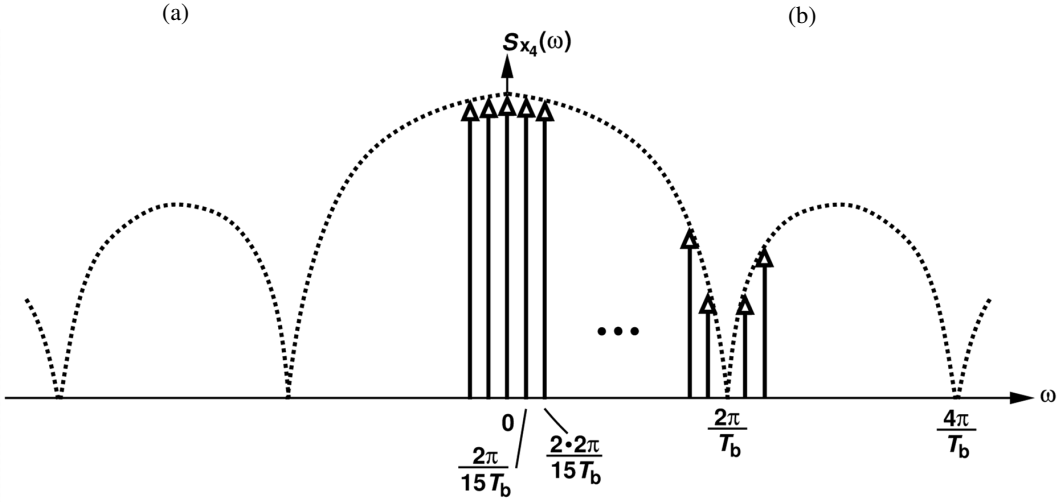
To generate PRBS with even higher data rate, we have to interleave the shift register and serialize the subrate outputs. One can multiplex the outputs of a lower-speed PRBS (with proper arrangement) to create a higher speed data sequence with the same pattern. One thing we need to pay attention to while combining the low-speed sequence into a high-speed output is to ensure proper delay between sub-rate data. To realize a PRBS of $2^n - 1$ with sub-rate data ratio 2^m (e.g., $m = 1$ for half rate, $m = 2$ for quarter rate, etc), the 2^m sub-rate data streams must be separated by 2^{n-m} bits (in terms of sub-rate bits). The following example illustrates how it works.

n	Polynomial	n	Polynomial
3	$x^3 + x^2 + 1$	15	$x^{15} + x^{14} + 1$
4	$x^4 + x^3 + 1$	23	$x^{23} + x^{18} + 1$
7	$x^7 + x^6 + 1$	31	$x^{31} + x^{28} + 1$
9	$x^9 + x^5 + 1$		

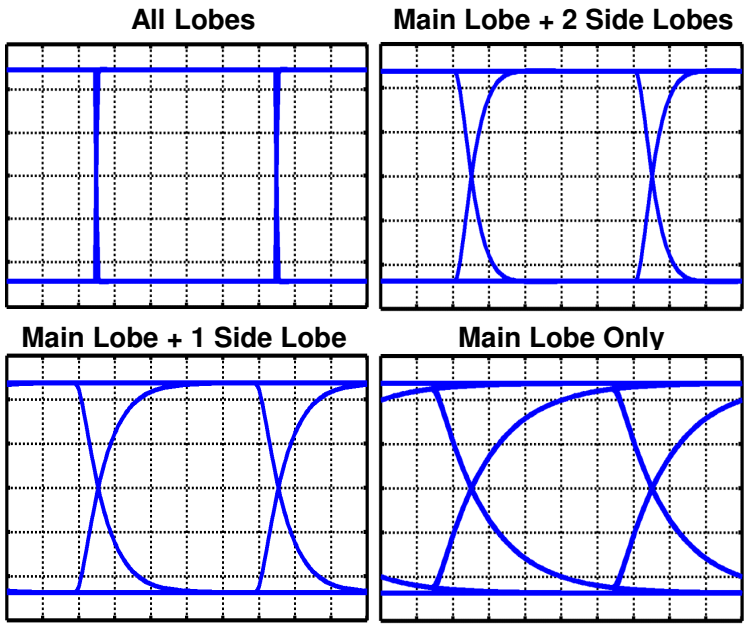
(a)



(b)

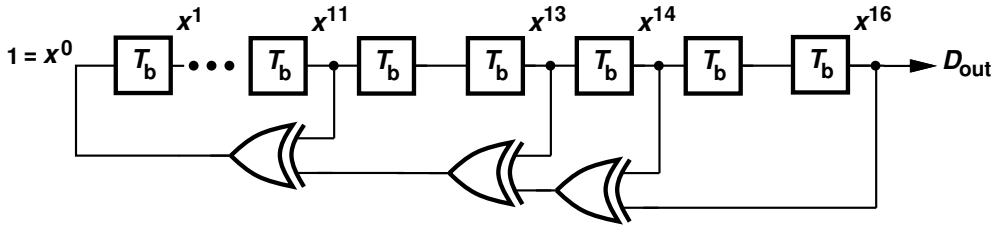


(c)

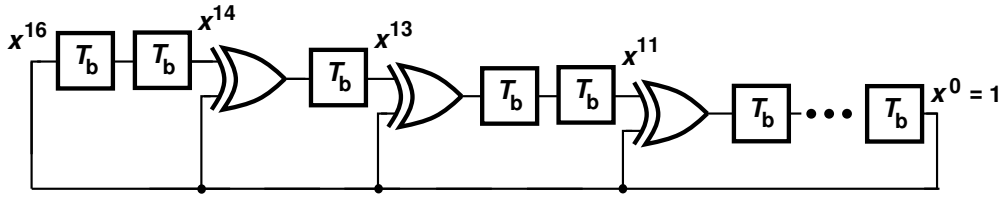


(d)

Fig. 1.9 (a) Primitive polynomial, (b) generating PRBS4, (c) spectrum of PRBS4, (d) waveforms anality with different number of side lobes (arbitrary units).



(a)



(b)

Fig. 1.10 $2^{16} - 1$ PRBS: (a) conventional (Fibonacci), (b) Galois.

Example 1.3

Determine the data sequence of a quarter-rate $2^4 - 1$ PRBS.

Solution:

The 4 quarter-rate data streams (D_0, D_1, D_2, and D_3) must be separated by 2^{4-2} bits. As illustrated in Fig. 1.11, identical PRBS4 patterns are obtained if the multiplexing order is $D_3 \rightarrow D_2 \rightarrow D_1 \rightarrow D_0$.

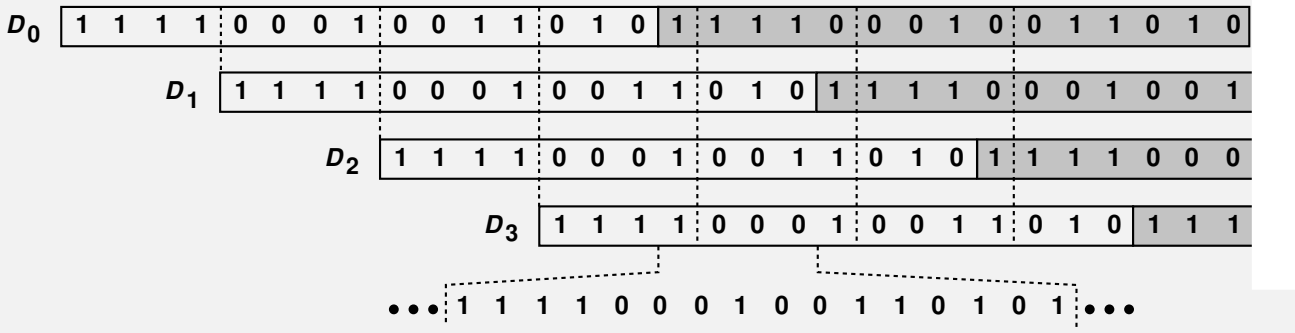


Fig. 1.11 Quarter-rate PRBS of $2^4 - 1$.

Figure 1.12 depicts examples for implementing sub-rate PRBS7. More details can be found in [8], [9].

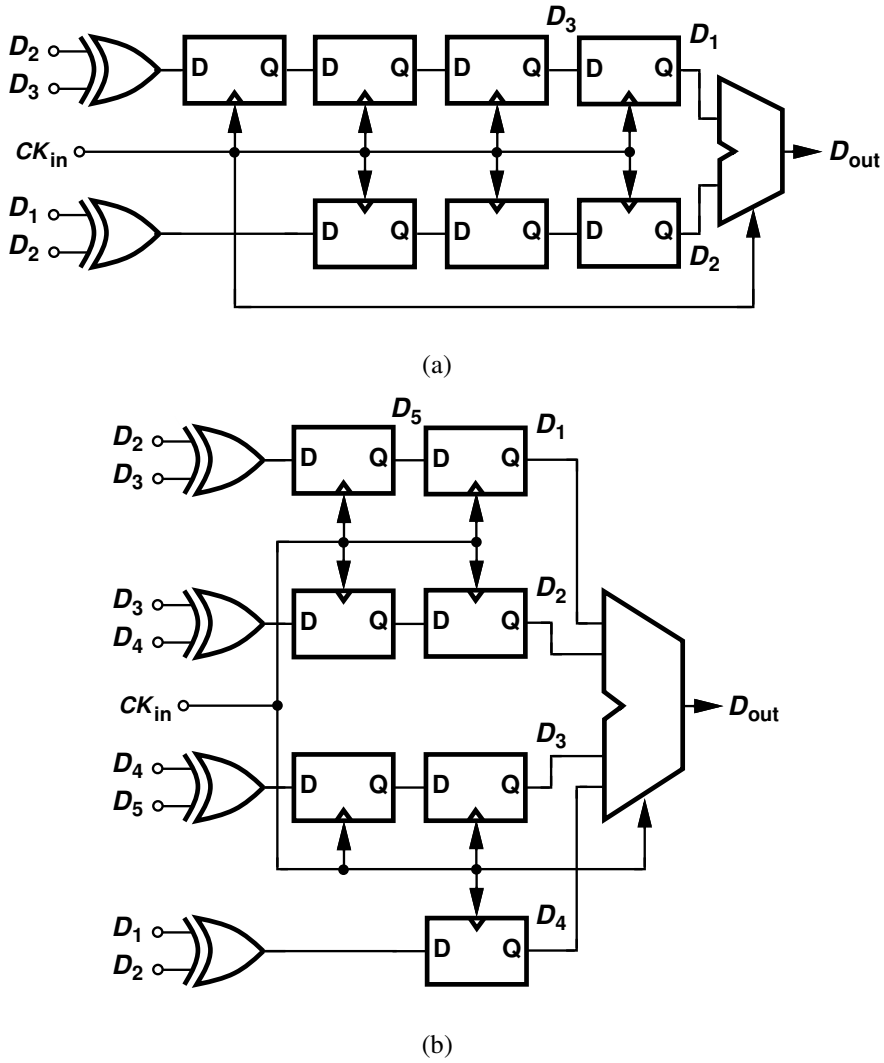


Fig. 1.12 Realization of sub-rate PRBS7: (a) half-rate, (b) quarter rate.

1.3 TRANSCEIVER ARCHITECTURE

A serializer/deserializer (SerDes) can possibly be implemented in many ways. We illustrate a generic architecture in Fig. 1.13. In general, low-speed, sub-rate data inputs are presented at the input ports in parallel. They are retimed and serialized into higher speed data streams by a

multiplexer (MUX), which is most likely made in a tree structure. A clock multiplication unit (basically a phase-locked loop) provides clocks with different frequencies and phases for the retimers and selectors in the MUX. In high data-rate SerDes, it may be necessary to incorporate a phase aligner (i.e., a delay-locked loop or equivalent circuit whose output phase is under control) so as to compensate for the skew and imbalanced delay. An output driver is responsible for delivering the data to the channel. In electrical domain, 50 Ω termination is usually required to minimize reflection. For optical applications, laser drivers are employed to emitted laser into the fiber, which may introduce distortion, dispersion, and other nonidealities. High frequency signal power tends to be attenuated more severely in the channel, so the transmitter usually includes a pre-emphasis device to neutralize the effect. The FFE is part of the equalization blocks.

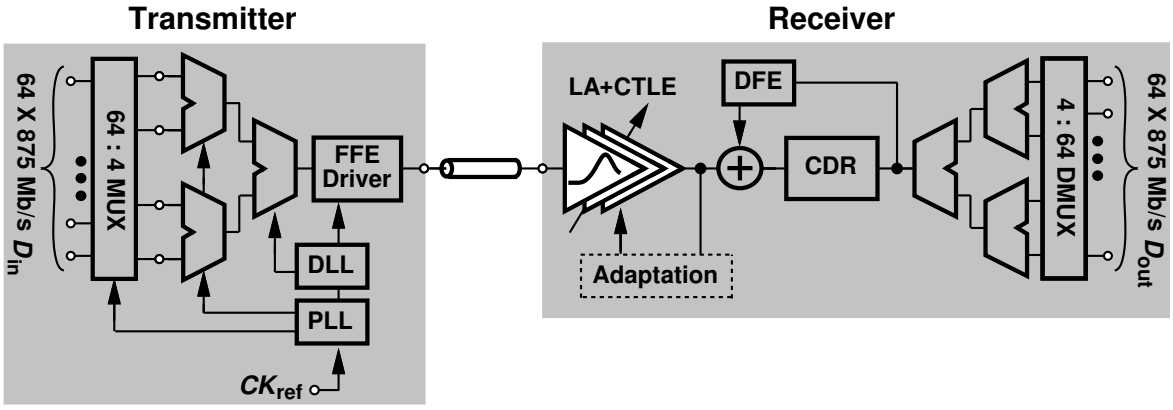


Fig. 1.13 General transceiver architecture.

In the receive side, data must be amplified and equalized before further processing. A so-called limiting amplifier (LA) co-working with a continuous time linear equalizer (CTLE) does the job. In optical, a photo detector must be used to convert the light back to electrical current, and such a tiny current subsequently gets converted to voltage with certain transimpedance gain. Similar to the low noise amplifier in wireless, this transimpedance amplifier (TIA) must be designed with very low additive noise as it locates in the very front end. After being equalized and amplified to normal logic level (e.g., 500 mV for CML), the input data must be retimed and demultiplexed. Except for some special systems in which system clock is embedded in the data or is transmitted in another line, the receiver has no information about the data rate. In other words, the system clock must be extracted

directly from the data stream, whose spectrum presents a null at the frequency of data rate! This task is taken care of by a circuit named clock and data recovery (CDR), which recovers the clock, retimes and demultiplexes the data. In modern SerDes architecture, a decision feedback equalizer (DFE) is usually adopted in the receiver to help equalize the data. Co-designed with the CDR circuit, this equalizer typically cooperates with FFE and CTLE to achieve the best compensation for loss and reflection. Since the receiver can monitor the signal quality after equalization in real time, the equalizers in the receive side can be implemented with adaptation. Finally, high-speed serial data is demultiplexed into low-speed outputs in parallel for further processing.

It is worth noting that there may exist frequency offset between the transmitter and the receiver. In short-range system, a reference clock can be provided from the transmitter, synchronizing the receiver.[2] The CDR here is only responsible for lining up the phases of clock and data, as the recovered clock frequency is exactly the same as the data rate (or the sub rates). In long distance applications, on the other hand, the CDR may need to recover both the phase and the frequency simultaneously. For example, the repeaters in a long-haul system are away from each other by tens of kilometers. It is impossible to transmit reference clock signal unless an additional pair of fibers are included. The CDR circuits in such cases must conduct frequency acquisition before phase locking.

1.4 PULSE AMPLITUDE MODULATION (PAM) SIGNAL

As the required data rate continuously goes up, the channel bandwidth becomes a bottleneck. To squeeze more data into a given bandwidth, data format itself needs to be modified. The binary NRZ data can be reformed as multiple-level signal to carry more bit per unit bandwidth. Shown in Fig. 1.14(a) is an example. Here, we combine two NRZ data with 2:1 weighting, resulting in a 4-level data. Recognized as pulse amplitude modulation with 4 levels (PAM4), this signaling carries twice as much information as NRZ does at the cost of 9.5-dB SNR degradation. It can be represented as

$$x_{PAM4}(t) = \sum_k b_k \, p(t - k \cdot 2T_b), \tag{1.10}$$

[2] Alternatively, a global reference clock could be created independently and sent to the transmitter and the receiver.

if the symbol rate is $1/(2T_b)$. Here, $b_k = \{-3, -1, +1, +3\}$, and $p(t)$ is still an ideal pulse with unity magnitude. For simplicity, we take off the dc port. Since the two inputs are independent, the PAM4 output should appear in the 4 levels with equal probability. The spectral density function of such a PAM4 signal is thus given by

$$S_{PAM4}(\omega) = \frac{5}{2T_b} \cdot \left[\frac{\sin(\omega T_b)}{\omega/2}\right]^2. \tag{1.11}$$

As expected, it still presents a sinc function but with half width as compared with an NRZ data

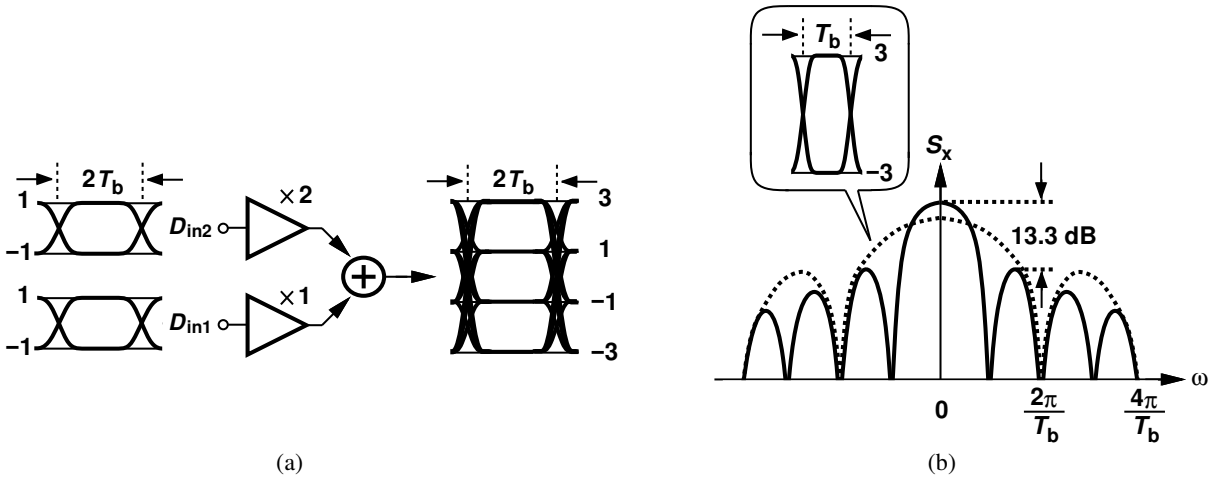


Fig. 1.14 (a) PAM4 signal, (b) its spectrum (bold dash line: spectrum of NRZ with the same data rate and magnitude).

with the same data rate $(1/T_b)$. That is, the nulls occurs at $w = \pi/T_b$ and its harmonics. The two-fold bandwidth efficiency makes it attractive for high-speed applications. Figure 1.14(b) illustrates the spectrum of PAM4 (solid) and NRZ (dotted) signals with the same data rate and data swing (i.e., ± 3). As will be shown in Chapter 2, this assumption is very realistic as the maximum current a differential pair can handle is almost constant for a given technology node. Note that the main lobe and the first side lobe of PAM4 still have 13.3 dB in difference. The reader can prove that the near-dc spectral density of PAM4 is slightly higher (i.e., 0.45 dB) than that of NRZ. Meanwhile, it can be easily shown that for a PAM signal with N levels (PAM-N), the first null locates at the frequency of data rate / $\log_2 N$.

The realization of a PAM4 signal is not difficult. As can be shown in Fig. 1.15, it is preferable and easier to add up two signals in current mode. The output driver converts the result back to voltage by loading (terminating) resistor and deliver the signal to the channel. The receiver is actually nothing more than a 2-bit analog-to-digital converter (ADC), which decodes the 2 bit/symbol data back to parallel NRZ format as MSB and LSB. In reality, the circuit implementation would be much more complicated. We leave circuit details to Chapter 11.

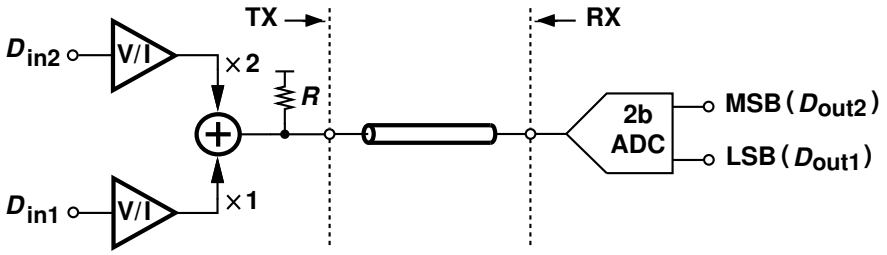


Fig. 1.15 Simplified PAM4 architecture.

It is instructive to calculate the probability of error in PAM signal (Fig 1.16). Taking PAM4 as an example, the 4 levels has equal probability of 1/4. With the same total swing V_{pp}, we calculate the error probability as

$$P_{e,\,PAM4} = (1+2+2+1) \times \frac{1}{4} \times \int_{V_{pp}/(6\,\sigma_n)}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(\frac{-x^2}{2})\,dx \tag{1.12}$$

$$= 1.5\,Q\Big(\frac{V_{pp}}{6\,\sigma_n}\Big). \tag{1.13}$$

Note that the 2 outmost levels have only one side for error to occur. In general, for a PAM-N signal, the probability of error becomes

$$P_{e,\,PAM\text{-}N} = \frac{2(N-1)}{N}\;Q\left[\frac{V_{pp}}{(N-1)\,\cdot 2\,\sigma_n}\right]. \tag{1.14}$$

Under what condition should we consider using PAM signaling to replace NRZ? This question is difficult to answer as it involves complicate tradeoffs among signal integrity, bandwidth, power consumption, circuit complexity, and so on. However, we can provide a simple yet useful way to estimate which data format is more advantage. It is to compare the channel loss at Nyquist

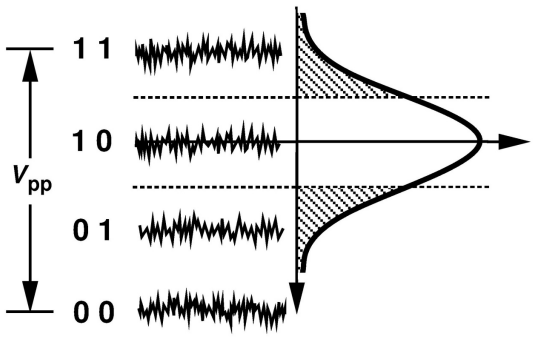


Fig. 1.16 Calculation of error probability in PAM4.

frequency. If a 56-Gb/s SerDes is evaluated, for example we check the 14-GHz point for PAM4 and the 28-GHz point for NRZ. Suppose circuit noise and other conditions are similar in both cases. If the channel loss difference P is greater than 9.5 dB, PAM4 is a better choice. Otherwise, NRZ should be used (Fig. 1.17). It is because the PAM4 is inherently inferior in signal power by 9.5 dB, and equalizations are to compensate for the channel loss within the Nyquist frequencies. In other words, we compare the expected eye opening after equalization. Certainly other considerations such as power, complexity, and area must be taken into account for a more accurate evaluation, but this quick check provides first-order estimation with minimum effort.

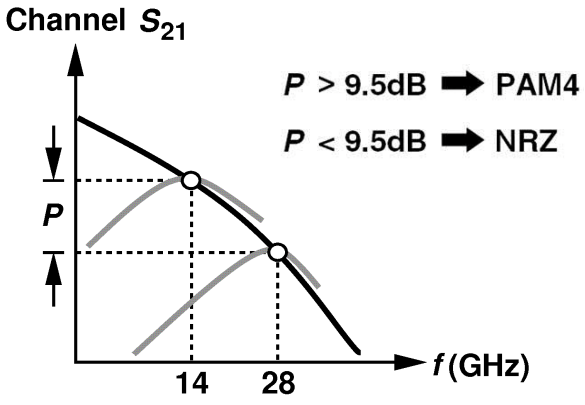


Fig. 1.17 Determine data format for a 56 Gb/s system [10].

1.5 DUOBINARY SIGNAL

In addition to PAM signals, the duobinary signal is often adopted as a substitute for NRZ. Having been used in optical communications and recently moving into electrical systems [11]−[13], duobinary modulation can also achieve a data rate theoretically twice as much as the channel bandwidth. In addition, intersymbol interference (ISI) is introduced in a controlled manner such that it can be cancelled out to recover the original signal. Unlike PAM4 or NRZ, duobinary signal incorporates the channel loss as part of the overall response [14], substantially reducing the required boost and relaxing the equalizer design. We introduce duobinary signal in this session.

A duobinary signal can be best described as the sum of the present bit and the previous bit of a binary (NRZ) data sequence

$$w[n] = x[n] + x[n-1]. \tag{1.15}$$

As shown in Fig 1.18(a), it correlates two adjacent bits to introduce the desired ISI. The transfer function of $H_1(z)$ is expressed in z-domain as

$$H_1[z] = \frac{1}{2}(1 + z^{-1}), \tag{1.16}$$

where the attenuating factor 1/2 is used to keep the signal swing constant before and after the conversion. Transforming it to continuous mode, we have

$$H_1(s) = \frac{1}{2}\,[\,1 + \exp(\,-j\omega\,T_b)], \tag{1.17}$$

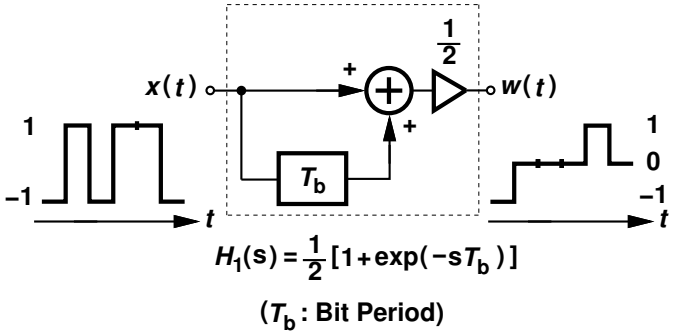
where T_b denotes the bit period. Since in an LTI system, the output spectrum is given by the product of the input spectrum and the magnitude square of the transfer function, we have

$$S_{duo}(\omega) = |H_1(\omega)|^2 \cdot S_x(\omega) \tag{1.18}$$

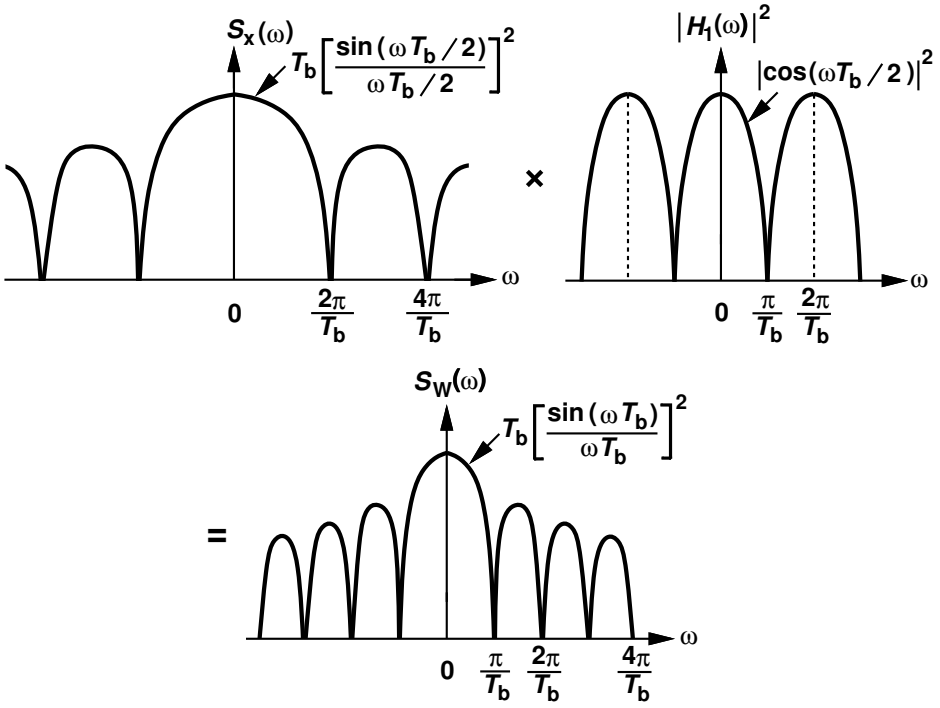
$$= \cos^2\left(\frac{\omega T_b}{2}\right) \cdot T_b \cdot \left[\frac{\sin(\frac{\omega T_b}{2})}{\frac{\omega T_b}{2}}\right]^2 \tag{1.19}$$

$$= \frac{1}{T_b} \cdot \left[\frac{\sin(\omega T_b)}{\omega}\right]^2. \tag{1.20}$$

As illustrated in Fig. 1.18(b), $S_{duo}(w)$ is still a sinc function with half the bandwidth as compared with $S_x(w)$. Just like PAM4, duobinary signaling reduces the required channel bandwidth by a factor of 2.



(a)



(b)

Fig. 1.18 (a) Linear model of duobinary signaling, (b) composition of duobinary spectrum [15].

It is worth noting that although the PAM4 signal possesses the same spectral efficiency as the duobinary does, the latter can further take advantage of the channel response as part of the transfer function. Fig. 1.19 illustrates the operation of duobinary signaling, where the transmit preemphasis and receive equalizer cooperate to reshape the low-pass response of the channel so that the overall transfer function approximates the first lobe of $H_1(w)$. In other words, a duobinary transceiver "absorb" significant amount of channel loss and makes it useful in the overall response, allowing more relaxed preemphasis and equalizer design.

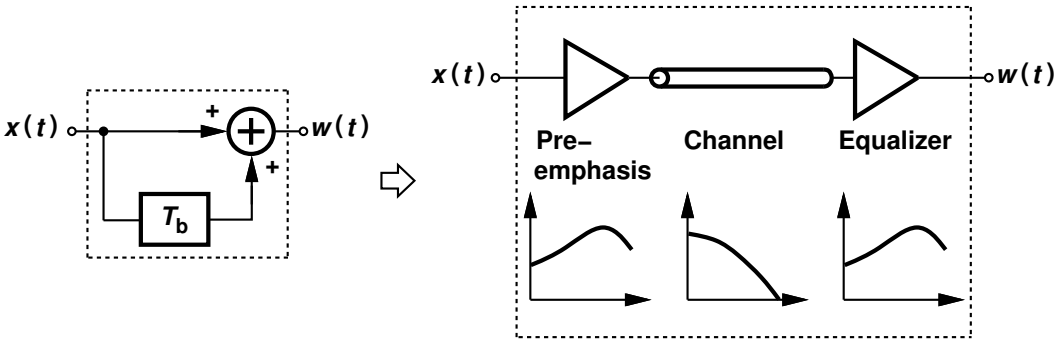


Fig. 1.19 Concept of duobinary signal formation [15].

In reality, a precoder $H_2(z) = 1/(1 + z^{-1})$ must be implemented in the transmit side. Here, we follow the design of [16], and the complete duobinary transceiver is shown in Fig. 1.20. The reshaped duobinary data gets decoded by an LSB distiller that takes the LSB as the output, recovering the binary NRZ data as $y[n]$. The waveforms of important nodes are also depicted in Fig. 1.20.

Although it looks attractive, duobinary signal has several issues. First, the precoder is difficult to implement in high speed unless an open-loop structure is adopted. The channel loss must be carefully shaped so as to mimic the main lobe of $|H_1(w)|^2$. It is not trivial at all if PVT variations are concerned. The CDR circuit for duobinary circuit is challenging as well.

Finally, to recover the duobinary data back to binary is another hurdle. The undesired ripple and time-domain jitter due to the imperfect response and finite rise/fall time may degrade the signal integrity considerably. We address practical circuit issue in Chapter 12.

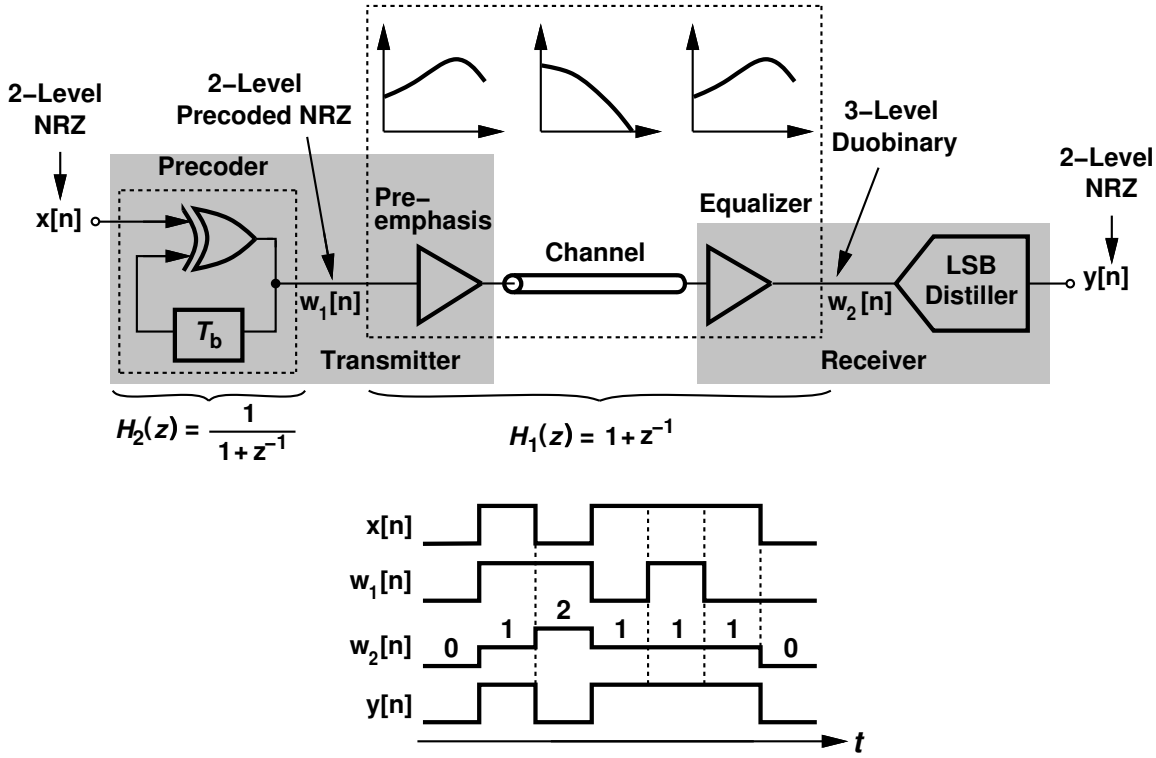


Fig. 1.20 Complete transceiver design and timing diagram of important nodes [15].

REFERENCES

[1] K. Tsai *et al.*, "A 43.7 mW 96 GHz PLL in 65 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers,* Feb. 2009, pp. 276-278.

[2] K. Tsai *et al.*, "A 104 GHz phase-locked loop using a VCO at second pole frequency," *IEEE Trans. on VLSI Systems,* vol. 20, pp. 80-88, Jan. 2012.

[3] M. Seo *et al.*, "A 300 GHz PLL in an InP HBT Technology," *IEEE MTT-S Int. Microw. Symp. Dig.,* pp. 1-4, June 2011.

[4] P. Chiang *et al.*, "A 300 GHz Frequency Synthesizer with 7.9% Locking Range in 90nm SiGe BiCMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers,* Feb. 2014, pp. 260-262.

[5] D. Baek *et al.*, "A 5.67 mW 9 Gb/s DLL-Based Reference-less CDR with Pattern-Dependent Clock-Embedded Signaling for Intra-Panel Interface," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers,* Feb. 2014, pp. 48-50.

[6] B. Razavi, "Design of Integrated Circuits for Optical Communications," NewYork: McGraw-Hill, 2002.

[7] B. Razavi, "RF Microelectronics," Upper Saddle River, NJ: Prentice-Hall, 1998.

[8] E. Laskin *et al.*, "A 60-mW per lane, $4 \times$ 23-Gb/s $2^7 - 1$ PRBS generator," *IEEE J. Solid-State Circuits,* vol. 41, no. 10, pp. 2198-2208, Oct. 2006.

[9] M. Chen *et al.*, "A low-power highly multiplexed parallel PRBS generator," in *Proc. IEEE Custom Integrated Circuits Conf. (CICC),* 2012, pp. 1-4.

[10] Jri Lee *et al.*, "Design of 56 Gb/s NRZ and PAM4 SerDes Transceivers in CMOS Technologies," *IEEE J. Solid-State Circuits,* vol. 50, pp. 2061-2073, Sept. 2015.

[11] A. Lender, "The duobinary technique for high-speed data transmission," *IEEE Trans. Commun. Electron.,* vol. 82, pp. 214-218, May. 1963.

[12] J. H. Sinsky *et al.*, "High-speed electrical backplane transmission using duobinary signaling," *IEEE Trans. Microw. Theory Tech.* vol. 53, no. 1, pp. 152-160, Jan. 2005.

[13] K. Yamaguchi *et al.*, "12 Gb/s duobinary signaling with 2 oversampled edge equalization," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers,* 2005, pp. 70-71.

[14] J. Sinsky *et al.*, "10 Gb/s duobinary signaling over electrical backplanes-Experimental results and discussion," Lucent Technologies, Bell Labs [Online]. Available: http://www.ieee802.org/3/ap /public/jul04/sinsky_01_0704.pdf

[15] Jri Lee *et al.*, "Design and Comparison of Three 20-Gb/s Backplane Transceivers for Duobinary, PAM4, and NRZ Data," *IEEE J. Solid-State Circuits,* vol. 50, pp. 2120-2133, Sept. 2008.

[16] M. Tomlinson, "New automatic equalizer employing modulo arithmetic," *Electron. Lett.,* vol. 7, pp. 138-139, Mar. 1971.

TERMINATION AND DRIVERS

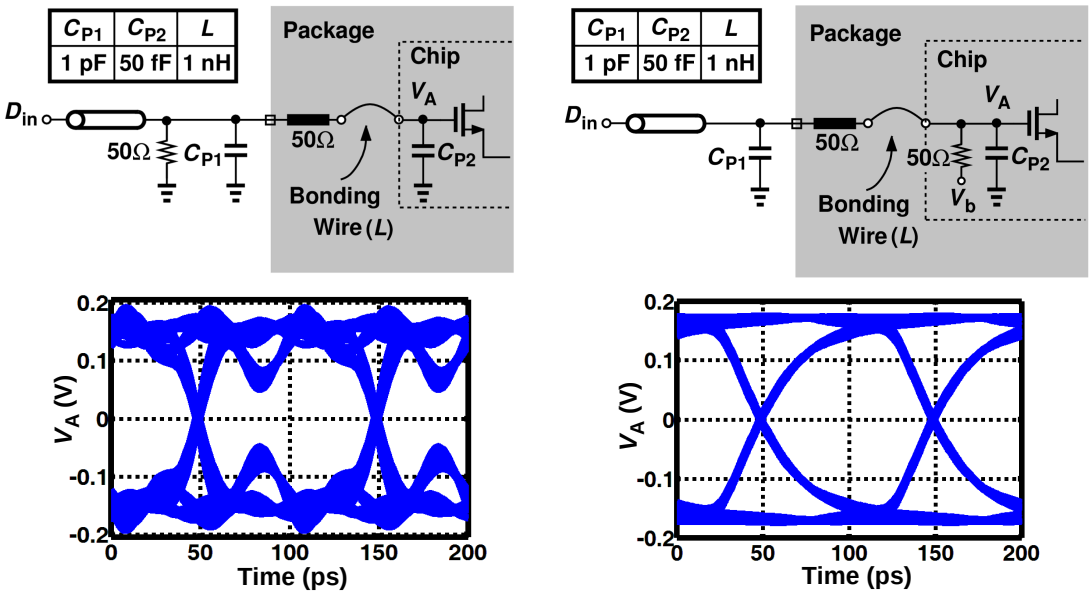
2

In this chapter, we study the termination technique and output drivers. High-speed data links necessitate well-behaved channels with good matching to ensure signal integrity. In addition, a robust and reliable driver is the key to high-quality signal delivery. We look at the output driver's properties and implementations in both electrical and optical domains.

2.1 TERMINATION

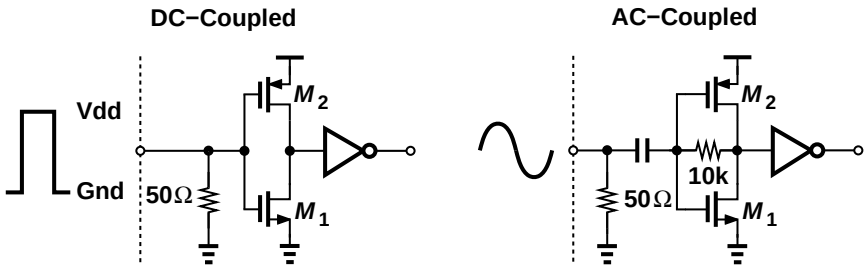
Termination is one of the fundamental techniques built up to maintain signal integrity, especially at high frequencies. Modern electrical devices require 50-Ω impedance matching along the signal path to ensure proper signal delivery with minimum reflection. For high-speed applications, it is always desirable to put termination on chip. As illustrated in Fig. 2.1(a), an off-chip terminator can only keep signal integrity to the soldering point of the package. The package parasitic capacitance C_P ($\approx$ 1 pF), internal trace (a few mm), bonding wire inductance ($\approx$ 1 nH), and pad capacitance ($\approx$ 50 fF) would cause significant distortion. The on-chip terminator, on the other hand, absorbs a significant amount of parasitics and arrives at a much better result [Fig. 2.1(b)]. 10-Gb/s input data eyes after external and internal terminations are depicted here to demonstrate the difference (C_{P1} = 1 pF, bond wire = 1 nH, C_{P2} = 50 fF).

There are several ways to perform on-chip termination. For digital input, the rail-to-rail signal can be dc-coupled to the input port with a 50-Ω terminator connecting to V_{DD} or ground. In advanced CMOS technologies, the input buffer (i.e., inverter M_1-M_2) can experience full swing input data up to several Gb/s. For clock input with smaller swing ($\leq V_{DD}/2$), it is also popular to incorporate internal coupling capacitors. Here, the self-biased inverter M_1-M_2 acquires the input



(a) (b)

Fig. 2.1 (a) External, (b) on-chip terminations.

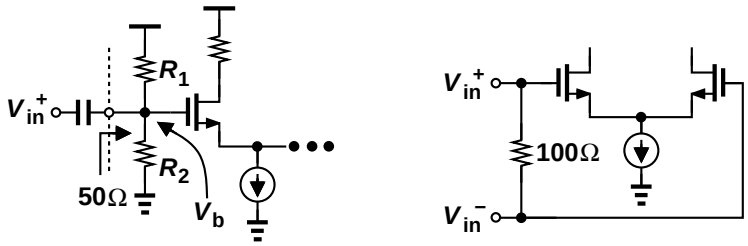


(a) (b)

Fig. 2.2 Input termination for digital circuits: (a) dc-, (b) ac-coupled.

signal (clock) by ac-coupling. The inverter M_1-M_2 is self-biased at the region of maximum gain. Depending on the input frequency, the coupling capacitor can be as small as several hundred fF. Note this structure is not recommended for broadband data. (Fig. 2.2)

What happens if a broadband data with small swing (e.g., 500 mV) is applied into the input port? Generally, the coupled capacitor must be placed externally as a high-quality, discrete device



(a) (b)

Fig. 2.3 Input termination for analog circuits: (a) dc-, (b) ac-coupled.

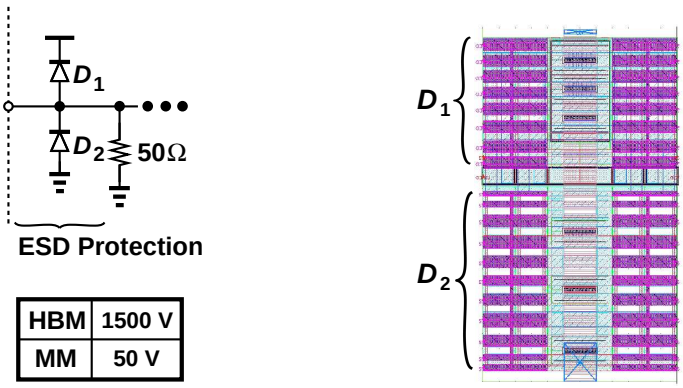
(i.e., a dc block); The internal bias circuit has to be codesigned with the input buffer, which is usually a differential pair. As can be shown in Fig. 2.3(a), two resistors R_1 and R_2 establish a proper input dc level V_b and maintain a 50-Ω impedance simultaneously:

$$R_1||R_2 = 50\ \Omega \tag{2.1}$$

$$V_b = V_{DD} \cdot \frac{R_2}{R_1 + R_2}. \tag{2.2}$$

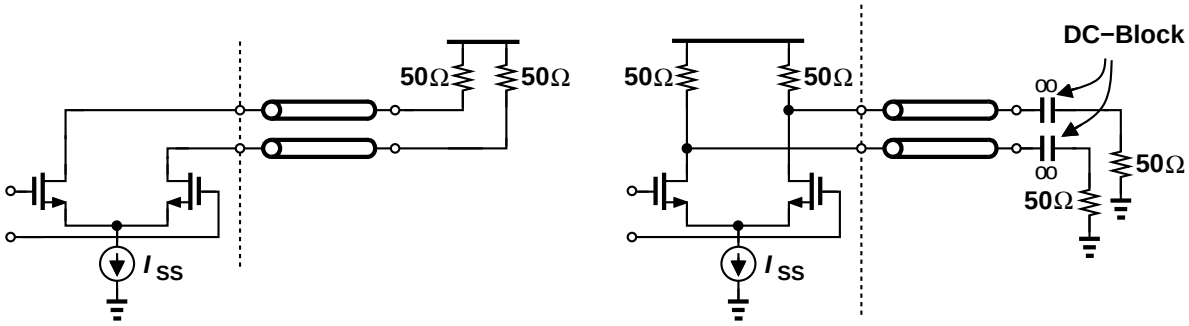
While providing good performance, the terminator structure in Fig. 2.3(a) suffers from larger power consumption. Assuming V_{DD}= 1.2 V and V_b = $0.5{\cdot}V_{DD}$, we have R_1 = R_2 = 100 Ω and the dc current flowing through R_1 and R_2 is 6 mA. For a differential buffer, 12 mA is dissipated in the input terminator, which is not acceptable in low power applications. An alternative approach is to dc-couple the input directly. Revealed in Fig. 2.3(b), the input dc level is determined by the transmitter's output common-mode level, and a differential terminator of 100 Ω is introduced between the two inputs. Note that this setup is suitable for local systems whose supply voltages for both Tx and Rx are well-defined. Nonetheless, ESD protection circuits should be added to avoid damage from electrostatic discharge. ESD circuits could be laid underneath the pads to save area (Fig. 2.4).

In addition to input ports, output ports must be terminated to prevent reflection as well. Figure 2.5 illustrates examples for ac- and dc-coupled structures. Regular testing would be quite similar to the ac-coupled cases, where the CML output is delivered to the cables + bias Tees and finally into the instruments (e.g., oscilloscope). The 50-Ω terminators at both far- and near-ends form



(a) (b)

Fig. 2.4 ESD protection circuit.



(a) (b)

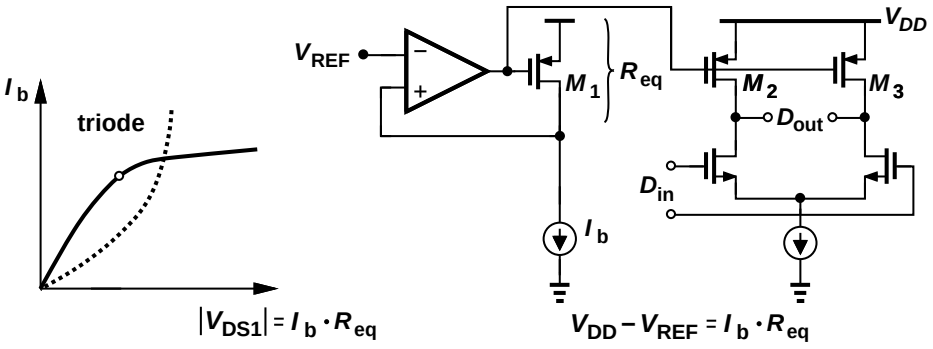
Fig. 2.5 Output termination: (a) dc-, (b) ac-coupled.

an equivalent 25-Ω loading for ac signals, leading to a smaller output swing (= I_{SS}·25 Ω). A compromised version is to terminate only one side (far-end) of the channel. Known as a “open drain” driver, this topology provides twice the output swing as compared with standard CML drivers. The dc current of it is directly provided from the far-end side. Undesired reflection may cause ringing to some extent, since only one side of the channel is properly terminated.

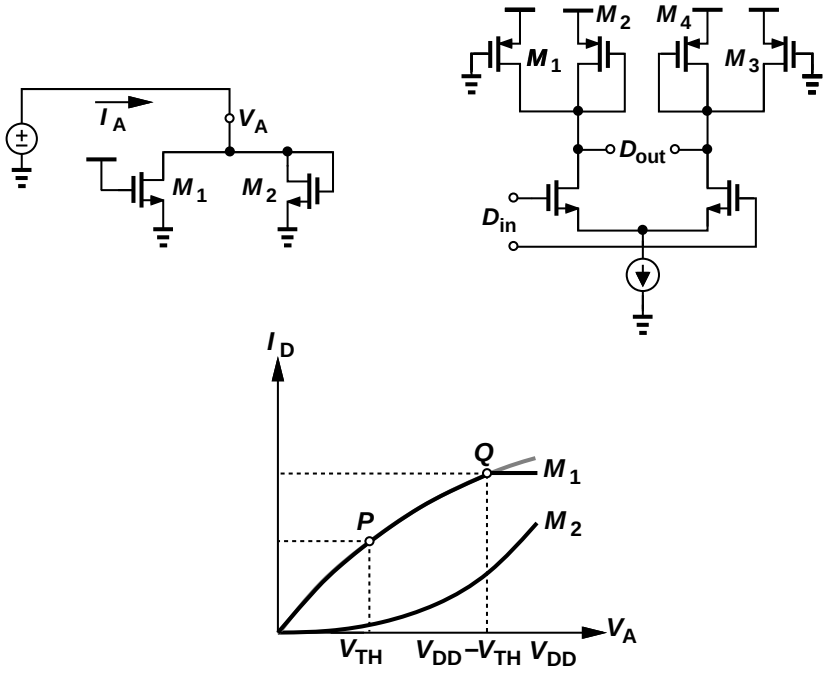
Resistors may not be avoidable in some CMOS processes. A device in triode (linear) region could serve as a substitute. Shown in Fig. 2.6(a) is an active resistor regulated by a servo controller.

Here, $M_1 - M_3$ are in triode region such that the equivalent resistance R_{eq} is defined by the negative feedback loop:

$$V_{DD} - V_{REF} = I_b \cdot R_{eq}, \tag{2.3}$$



(a)



(b)

Fig. 2.6 Resistorless termination: (a) servo-controlled, (b) active compensated, (c) active compensated in differential pair.

Where bias current I_b and reference voltage V_{REF} could be made accurately from bandgap. Mirroring the bias voltage to M_2 and M_3, we realize a well-controlled loading for the driver. A simpler approach without Opamp can be found in the Fig. 2.6(b), where a triode device (M_1) and a diode-connected device (M_2) are placed in parallel to realize a relatively constant resistance for a wide range. Here, M_1 stay in triode region until driving voltage $V_A \geq V_{DD} - V_{TH}$, and its current gradually decreases as V_A goes up. On the contrary, M_2 turns on as $V_A \geq V_{TH}$, carrying current in proportion to the square of its overdrive. If the sizes of $M_{1,}$, M_2 are properly chosen, it reveals a quite linear relationship between the total current I_A and V_A. A differential buffer with loading of this structure is also depicted in Fig. 2.6(c). An equivalent impedance of 50 Ω can be easily obtained by checking the overall I-V curve.

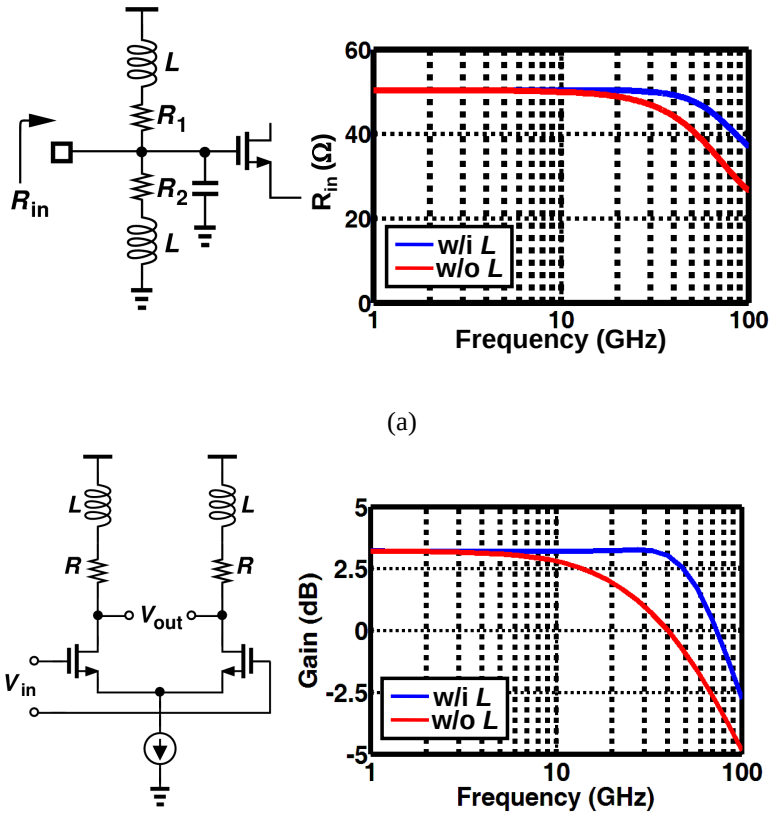


Fig. 2.7 Using inductive peaking to improve (a) input impedance matching, (b) output driver's bandwidth.

Example 2.1

Determine the device ratio of M_1 and M_2 in Fig. 2.6(b).

Solution: Assume second-order rule holds and M_2 is k times larger than M_1. We have the current of M_1 at point P as $V_A = V_{TH}$

$$I_{D1} \mid_{V_A = V_{TH}} = \frac{1}{2}\mu_n C_{ox}(\frac{W}{L})_1[2(V_{DD} - V_{TH})V_{TH} - V_{TH}^2] \tag{2.4}$$

$$\approx \mu_n C_{ox}(\frac{W}{L})_1[(V_{DD} - V_{TH}) \cdot V_{TH}], \tag{2.5}$$

where V_{TH} denotes the threshold voltage of $M_{1,2}$. In proper setting, the I-V curve extends linearly to $V_A = V_{DD}$. That is,

$$I_A \mid_{V_A = V_{DD}} = I_{D1} \mid_{V_A = V_{TH}} \cdot \frac{V_{DD}}{V_{TH}} \tag{2.6}$$

$$= \mu_n C_{ox}(\frac{W}{L})_1[(V_{DD} - V_{TH}) \cdot V_{DD}]. \tag{2.7}$$

On the other hand, I_{D1} saturates at point Q:

$$I_{D1} \mid_{V_A = V_{DD} - V_{TH}} = I_{D1} \mid_{V_A = V_{DD}} = \frac{1}{2}\mu_n C_{ox}(\frac{W}{L})_1(V_{DD} - V_{TH})^2. \tag{2.8}$$

Here channel-length modulation is neglected. Thus, M_2 is responsible for providing the difference current at $V_A = V_{DD}$:

$$I_{D2} \mid_{V_A = V_{DD}} = I_A \mid_{V_A = V_{DD}} - I_{D1} \mid_{V_A = V_{DD}} \tag{2.9}$$

$$= \frac{1}{2}\mu_n C_{ox}(\frac{W}{L})_2(V_{DD} - V_{TH})^2 \tag{2.10}$$

$$= \frac{1}{2}\mu_n C_{ox} k(\frac{W}{L})_1(V_{DD} - V_{TH})^2. \tag{2.11}$$

It follows that

$$k = \frac{V_{DD} + V_{TH}}{V_{DD} - V_{TH}}. \tag{2.12}$$

Note that k may slightly deviate from Eq. (2.12) if channel-length modulation is taken into consideration.

At high data rate, it is always desirable to extend the impedance matching to higher frequencies. Inductive peaking can help to enlarge the bandwidth by significant amount. Figure. 2.7 illustrates examples for bandwidth extension for input and output buffers. We addresses inductive peaking in chapter 3.

2.2 ELECTRICAL DRIVERS

There are many electrical signaling available in today's wireline communication systems, and some of these standards can be tracked back to 1960's. Among them, three technologies are especially popular: low-voltage differential signaling (LVDS), emitter-coupled logic (ECL), and current-mode logic (CML). We introduce these interfaces in this section.

2.2.1 Low-Voltage Differential Signaling (LVDS)

The LVDS is a versatile interface achieving low power consumption and high-speed operation. It can be beat described as a differential driver with push-pull currents. One typical example is illustrated in Fig. 2.8(a), where a constant tail current of 3.5 mA flows through the far-end terminator 100-Ω differentially with positive or negative polarity. As a result, a ± 350 mV differential swing is presented in the RX input. Since M_1-M_4 are switches, the output common-mode can be determined in either TX or RX. In Fig. 2.8(a), we have V_{CM} in the RX to setup the common-mode level, which can possibly vary by a significant amount. In an environment with V_{DD} = $2.5 \sim 3.3$ V, V_{CM} is usually set to be around 1.2 V. It is possible to determine the common-mode level in the TX side as well. The reader can find that the driver in Fig. 2.8(a) has only been terminated in the far end. Terminators can be also added at the near end to minimize reflection at the cost of reducing the swing by a factor of 2. We can apply ac coupling as long as common mode level is properly present.

In modern SerDes design, high supply voltage may not be available. For instance, CMOS technologies ask for a supply as low as 1 V. In such a case, a driver based on CMOS inverters can be used to perform push-pull of current [Fig. 2.8(b)]. Suppose the inverters are appropriately sized such that the large-signal equivalent on resistance ($R_{on,N}$ for $M_{1,3}$ and $R_{on,P}$ for $M_{2,4}$) is equal to

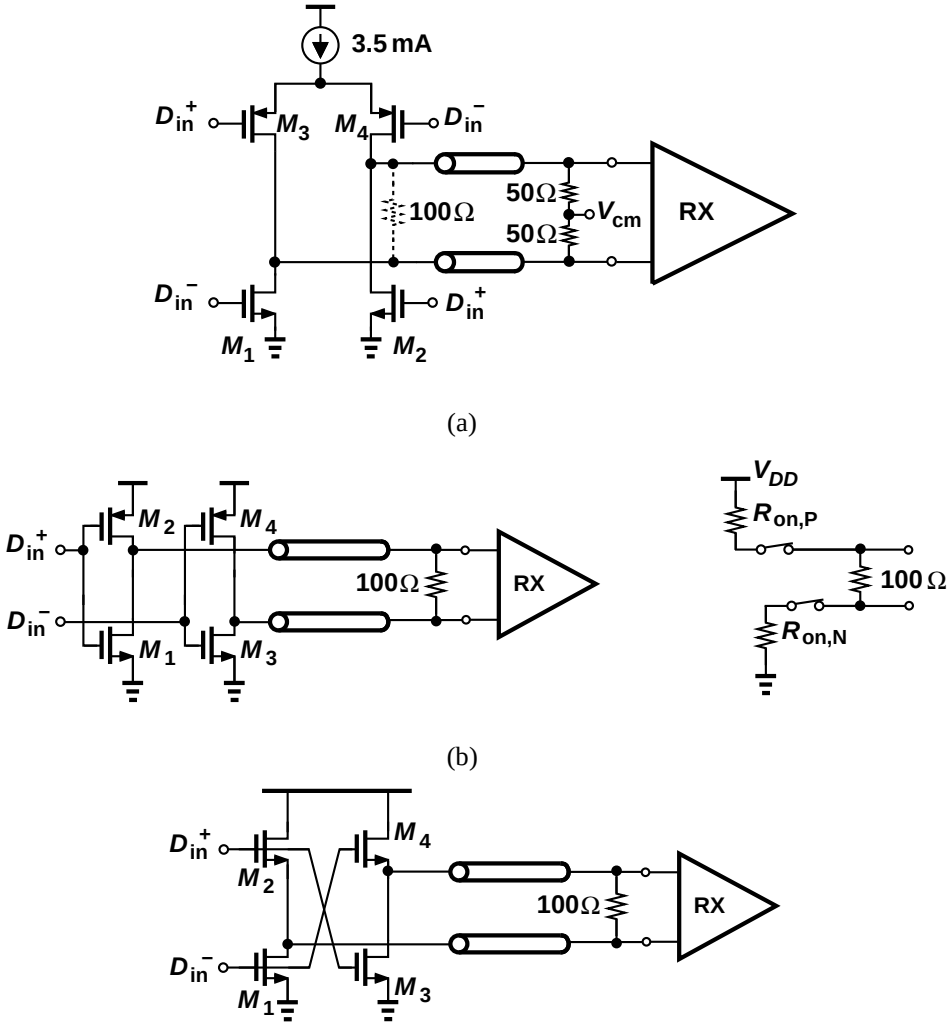
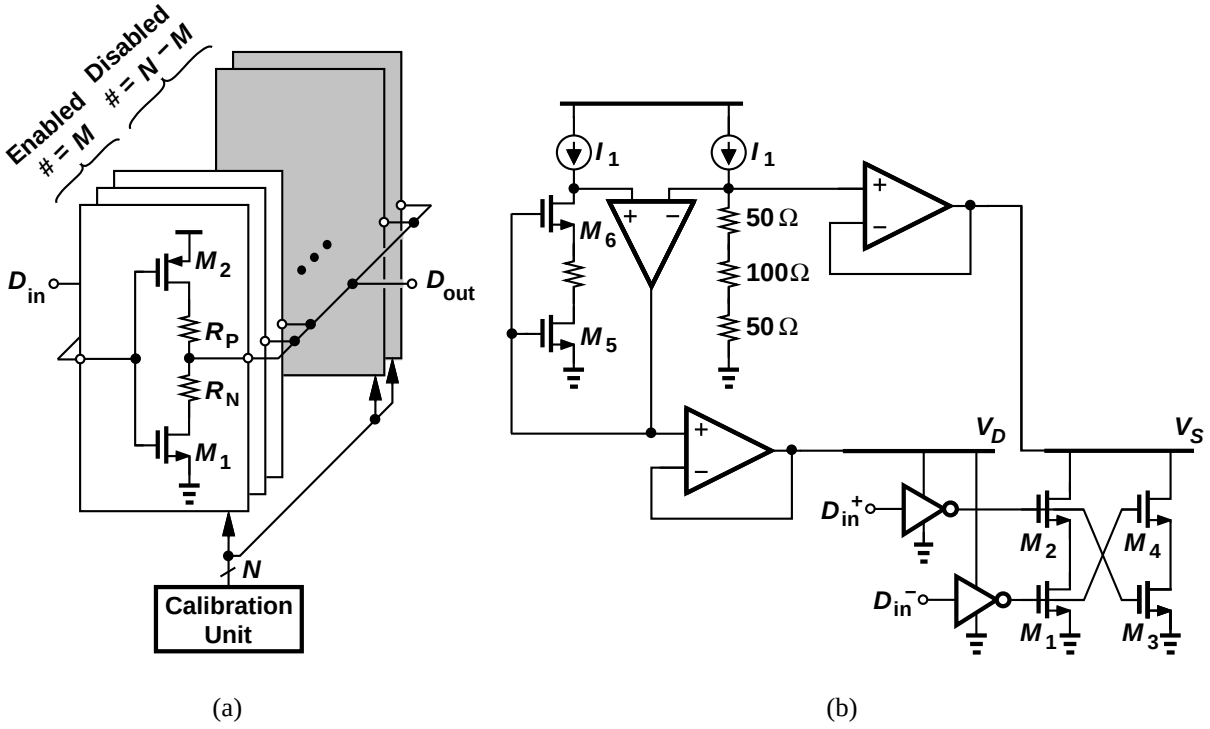


Fig. 2.8 LVDS driver designs: (a) standard, (b) low-supply, (c) low-supply low-swing.

$50\ \Omega$. The RX experiences a swing of $\pm\ V_{DD}/2$, which is quite sufficient in most applications. However, it would be difficult to maintain exactly $50\ \Omega$ on resistance over PVT variations without calibration. Similarly, if lower swing is acceptable, we can even realize the driver by NMOS solely. As demonstrated in Fig. 2.8(c), we have 4 NMOS devices M_1-M_4 to fulfill push-pull operation. With proper design, swing of several hundred mV can be achieved. Note that it is possible to realize impedance matching by using NMOS devices only. For example, if D_{in}^+ is high,



(a) (b)

Fig. 2.9 Impedance calibration of SST driver: (a) low-supply, (b) low-supply low-swing.

M_2 is in saturation and M_3 is in triode. That is, the impedance seen looking into source of M_2 is $1/g_{m2}$. A good matching could be obtained if both $1/g_{m2}$ and the equivalent on resistance of M_3 are close to $50\ \Omega$. The impedance matching here, however, is prone to degrade due to PVT variations, necessitating delicate calibration techniques. Recognized as source-series terminated (SST) drivers, Fig. 2.8(b) and (c) are widely used in low power transmitter design.

Several techniques can be adopted to overcome the above difficulties. For example, multiple drivers can be placed in parallel to achieve an equivalent output resistance approximately equal to $50\ \Omega$. The parallelism makes calibration much easier. As shown in Fig. 2.9(a), only M out of N identical buffer cells are tuned on (based on calibration result), arriving at an accurate impedance matching. Note that the calibration can be done either at power up or in background. Another example can be found in low-supply low-swing SST drivers. From the discussion of Fig. 2.8(c), we realize that it is difficult to simultaneously manage the equivalent impedance of devices in

saturation and in linear region. However, we could lower the driver's supply to put all devices of M_1-M_4 in Fig. 2.8(c) in triode region. Their equivalent on resistance would be much easier to control. As illustrated in Fig. 2.9(b), M_5 and M_6 (replica of $M_{1,3}$ and $M_{2,4}$) in serial with a 100-Ω resistor should present the same voltage drop as a 200-Ω resistor chain if they are carrying identical current of I_1. Thus, a negative feedback by means of the error amplifier establishes a proper supply voltage V_D for the pre-driver, and the desired voltage drop V_S can be applied to power the driver. As a result, good impedance matching is obtained. Unity-gain buffer is used here to minimize interference. Note that many other approaches can be found in the literature[xx][xx].

2.2.2 Emitter-Coupled Logic

ECL circuits were first invented in 1950's and later on widely used in high-speed bipolar circuits. Originally called current-steering logic, ECL is actually the predecessor of current mode logic. An ECL driver is shown in Fig. 2.10. Emitter followers M_3 and M_4 drive the 50-Ω channel and terminate at far-end side. The outstanding performance of bipolar emitter followers provides significant driving force and operation bandwidth. Note that in standard ECL operation $V_{CC} = 0$ V and $V_{EE} = -5.2$ V. Since no device enters saturation here, the switching time of ECL drivers is quite small. In other words, it is suitable for high-speed operation. The far-end side is usually terminated at $V_{CC} = -2$ V so as to achieve a $\pm 700 \sim 800$ mV swing in normal operation.

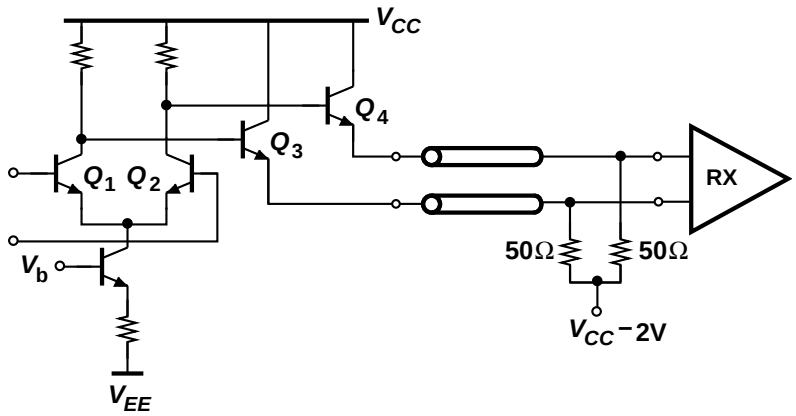


Fig. 2.10 ECL driver.

In modern communications, negative supply becomes rare and inconvenient. Setting $V_{CC} = 5.2$ V and $V_{EE} = 0$ V, we arrive at positive emitter-coupled logic (PECL) which is dedicated to positive supply environments. Similarly, if $V_{CC} = 3.3$ V and $V_{EE} = 0$ V, we call it low-voltage positive emitter-coupled logic (LVPECL).

2.2.3 Current-Mode Logic (CML)

The above ECL drivers can be further simplified to adopt even lower supply voltage. Figure. 2.11(a) illustrates one typical implementation, where the driving emitter followers are removed. The loading of differential pair can be made 50 Ω to achieve matching. Widely used in CMOS drivers, this simple differential pairs are used internally as high-speed data buffers, too. The output of CML drivers can also be ac-coupled in the far-end side, allowing the RX determining its input dc level. Since both ends are terminated, the data swing is relatively small. To generate $\pm$500 mV swing, I_{SS} needs to be 20 mA.

Several techniques have been proposed to increase the bandwidth of CML drivers/buffers. The most efficient one is inductive peaking [Fig. 2.11(b)]. With an inductor L in series the loading resistor R, the model of it is illustrated as well. Since the " rising " edge of data at nodes X and Y is nothing more than the process of charging C (parasitic capacitance) through L and R from supply, we model the rising edge as a step function driving the L-R-C network.

The transfer function is giving by

$$\frac{V_{out}}{V_{in}} = \frac{\frac{1}{LC}}{s^2 + \frac{R}{L}s + \frac{1}{LC}} \tag{2.13}$$

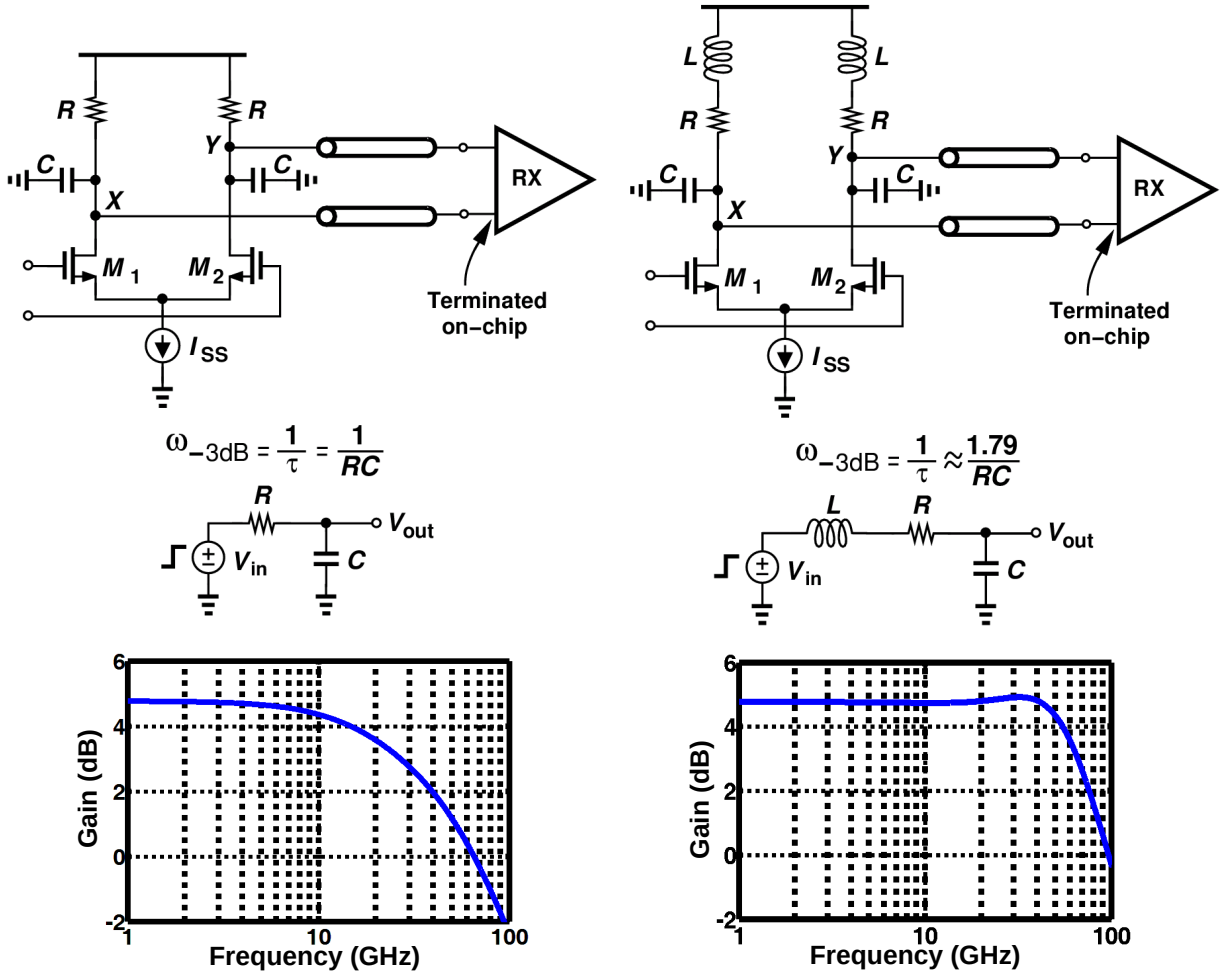
$$= \frac{s^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}, \tag{2.14}$$

where $\omega_n = 1/\sqrt{LC}$ and $\zeta = (R/2)(\sqrt{C/L})$. Note that without the peaking inductor L [Fig. 2.11(a)], the transfer function is simply a low-pass RC responses with bandwidth $\omega_{-3dB} = 1/\tau = 1/(RC)$. The second-order transfer function of Eq. (2.13) reaches a maximum flat response as $\zeta = 1/\sqrt{2}$. This is quite normal in a typical setup.For example,if C = 100 fF, R = 50 Ω, then L =

130 pH. In such a case, the−3-dB bandwidth can be obtained by

$$\left|\frac{V_{out}}{V_{in}}\right|(\omega_{-3dB}) = \frac{1}{\sqrt{2}} \tag{2.15}$$

$$\omega_{-3dB} = \frac{\sqrt{2}}{\tau} = \frac{\sqrt{2}}{RC}. \tag{2.16}$$



(a) (b)

Fig. 2.11 CML driver : (a) typical realization, (b) with inductive peaking.

That is, the bandwidth is extended by 41% with the help of inductor. The following example addresses the details of large-signal operation.

Example 2.2

The driver usually deal with large signal rather than small signal. Determine the 10% $\sim$ 90% rising time of Fig. 2.11(b) and compare with that of Fig. 2.11(a).

Solution: Applying a step function with unity magnitude gives rise to an output as

$$V_{out}(s) = \frac{1}{s} \cdot \frac{\omega_n^2}{s^2 + 2\zeta\omega_n + \omega_n^2} \tag{2.17}$$

$$= \frac{1}{s} - \frac{(s + \frac{\omega_n}{\sqrt{2}})}{(s + \frac{\omega_n}{\sqrt{2}})^2 + (\frac{\omega_n}{\sqrt{2}})^2} - \frac{(\frac{\omega_n}{\sqrt{2}})}{(s + \frac{\omega_n}{\sqrt{2}})^2 + (\frac{\omega_n}{\sqrt{2}})^2} \tag{2.18}$$

where we have $\xi = 1/\sqrt{2}$. The corresponding $V_{out}(t)$ becomes

$$V_{out}(t) = 1 - \sqrt{2} \cdot \exp\frac{-\omega_n t}{\sqrt{2}} \cdot \cos(\frac{\omega_n t}{\sqrt{2}} - \frac{\pi}{4}) \tag{2.19}$$

The responses for Fig. 2.11(a) and (b) are illustrated in Fig. 2.12. The 10% $\sim$ 90% rising time is equal to 1.52τ. As a comparison, the RC network in Fig. 2.11(a) takes 2.2τ to pull up the output from 10% $\sim$ 90%.

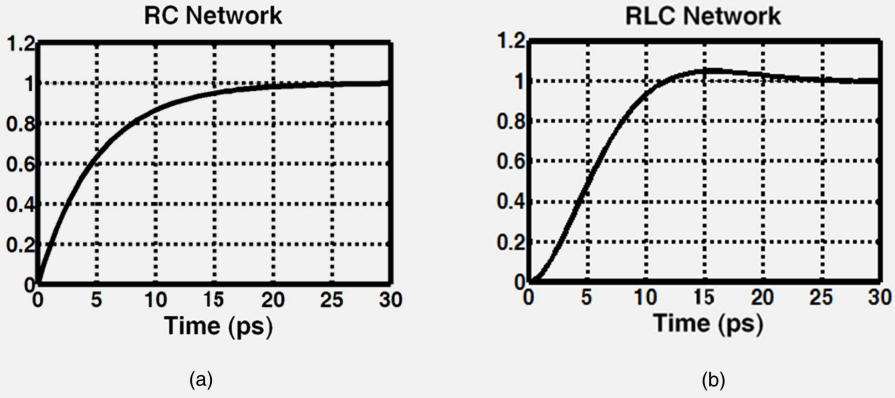


Fig. 2.12 Step response for pull-up network in Fig. 2.11 with and without inductive peaking.

2.3 OPTICAL DRIVERS

Two modulation schemes for optical lasers have been developed to achieve high-speed data transmission. For long haul optical links repeaters are separated by tens of kilometers, mandating higher

power lasers. External modulators based on Mach-Zehnder modulation are usually adopted in such cases, requiring several Volts of swing. Short-distance applications, on the other hand, target low power solutions. Since the impedance of different type of laser diodes varies significantly (i.e., 5 $\Omega \sim 50\ \Omega$), some laser drivers are required to drive current as large as 100 mA. Modern optical communication relies on direct modulated laser diodes for light sources. Recent development on vertical-cavity surface-emitting laser (VCSEL) makes it very suitable for short-range optical communication, such as Ethernet and fiber channel. Compared with laser diode having other geometry structures, VCSEL has remarkable advantages. A VCSEL is actually heterostructure laser diode with active region covered by distributed Bragg reflectors (DBRs) on top and bottom. As shown in Fig. 2.13, light comes out in the direction perpendicular to the surface, facilitating 20 array realization. The easy in-wafer testing and circular beam makes VCSEL superior to its edge emitting counterpart. More specifically, VCSEL consumes less current, achieves higher operation bandwidth, and reveals purer spectrum. The better stability over temperature and lower cost also make VCSELs attractive in different applications.

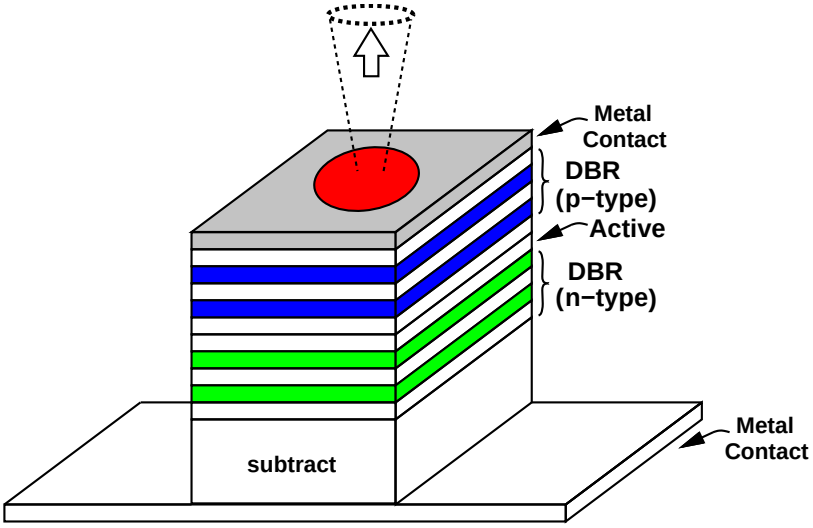


Fig. 2.13 VCSEL.

Today's high-speed VCSELs are designed to emit light with wavelength ranging from 650 nm to 1300 nm. Fig. 2.14(a) reveals the photo of a 25-Gb/s VCSEL with dimensions 150 by 150 um^2. The cathode is connected to chip for current pulling due to its smaller parasitic capacitance [1].

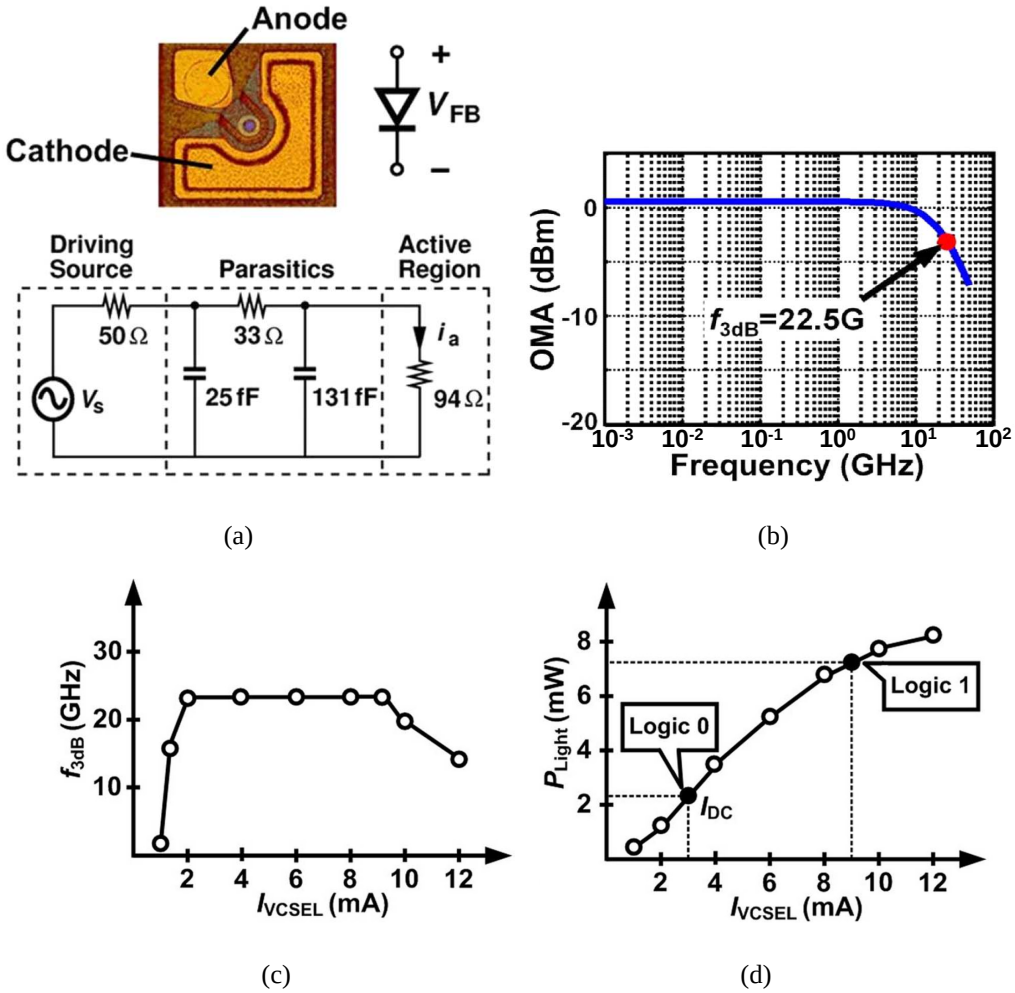


Fig. 2.14 (a)Typical VCSEL 850 nm and its small-signal model, (b)measured frequency reponse as bias current = 6 mA, (c)−3-dB bandwidth as a function of bias current, (d)VCSEL transfer function.

A small-signal model is established for transient simulation, which tightly matches the measured frequency response as shown in Fig. 2.14(b). With bias current of 6 mA, the bandwidth is barely enough for 25-Gb/s operation. In addition, the −3-dB bandwidth of VCSEL increases as bias current increases and get saturated as it becomes larger than 2 mA. We plot the −3-dB bandwidth for VCSEL as a function of bias current in Fig. 2.14(c). A threshold current if 1 ∼ 2 mA is usually required to turn on a VCSEL. High-speed VCSEL may need dc current as large as 3 mA to ensure

fast switching time between on and off [Fig. 2.14(d)]. Otherwise, the VCSEL deviates from linear operation and begins to cause errors. A constant pulling current over PVT variations becomes essential here. Typical slope efficiency can be as high as 0.8 $\sim$ 1.0 W/A at 850 nm.

In many systems, it is preferable to integrate optical devices and electrical drivers in one set as a receptacle module. The most popular assembly is the so called transmitter optical sub-assembly (TOSA), which includes laser diodes, filter, lens, ceramic tube, and driver IC in one module. It allows standardized connection and easy further integration. Similarly, a receiver optical sub-assembly (ROSA) can be found as a reciprocal module.

Due to the low impedance, driving a laser diode may need to pull very large amount of current (10 $\sim$ 100 mA). Figure. 2.15 illustrates an example realized in bipolar devices. Here, a laser diode of 25-Ω impedance serves as a loading device and, V_b, R_1, R_2 determine the reference current I_{REF}. The modulation and bias currents I_m and I_B are set by mirroring ratio (i.e., both are 100 in this case). The driving pair $Q_{1,2}$ are designed to pull current through the 25-Ω transmission lines, which are biased to V_{CC} by means of large external inductors (L = 10 μH). As a result, the current flowing through the laser diode would be $2I_B$ or 0, depending on the input. Note that a large inductor is necessary to put in the cathode side of the diode to block the parasitic capacitance of I_B pin.

The CMOS drivers usually suffer from poor current driving force. To pull a large amount of current, we break up the differential pair into 2-3 identical slices to avoid overlong routing. Fig. 2.15(b) illustrates one example. Each of the two identical pairs ($M_{1,2}$ and $M_{3,4}$) carries 30 $\sim$ 40 mA of current, achieving twice in total. The inter-connection can be implemented as 50-Ω transmission line, which also match the loading resistors. As a result, the output impedance seen looking into the driver is 25-Ω single-ended. The driver together with the external 25-Ω transmission lines form a differential driving on the 50-Ω laser diode.

The lower power VCSEL drivers encounter different issues. The driver needs to overcome the VCSEL's relaxation oscillation phenomenon at large signal. As illustrated in Fig. 2.16(a), a high-speed VCSEL present quite significant ringing effect at rising and falling edges due to the exchange of energy between photons and electrons [2]. Unlike regular signal distortion cause by

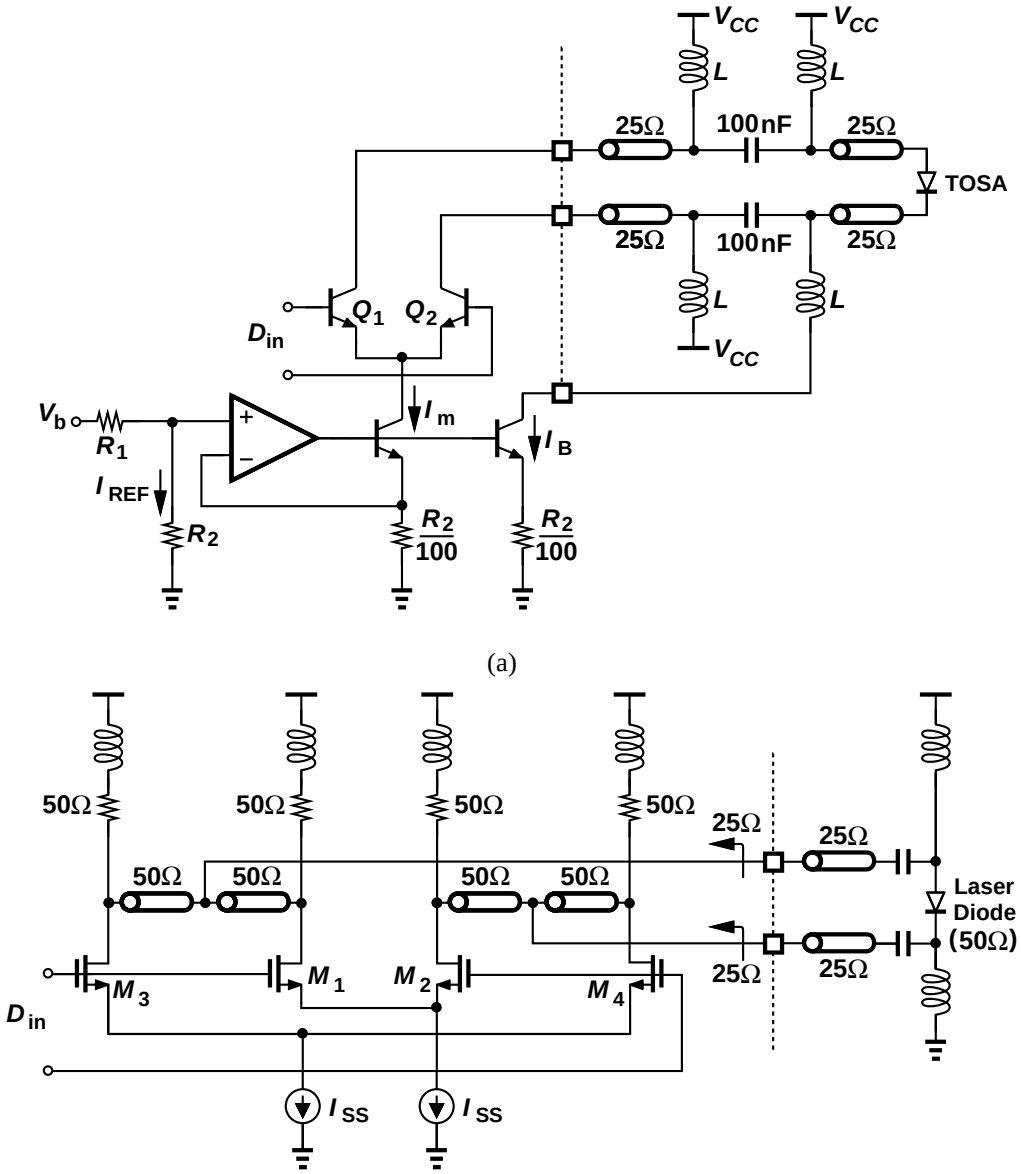


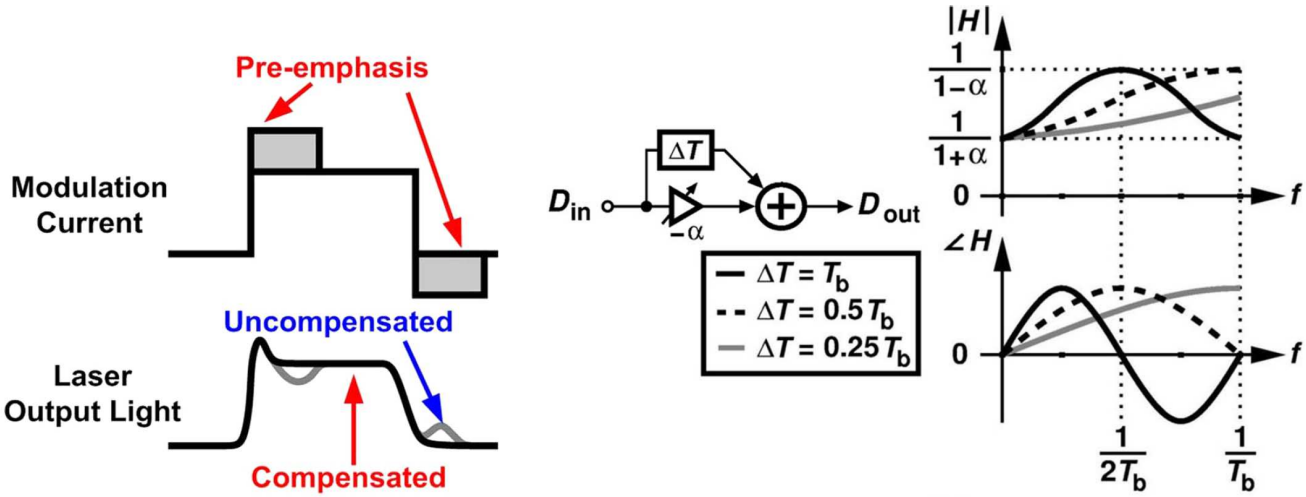
Fig. 2.15 High current laser drivers: (a)bipolar, (b)CMOS.

channel loss or reflection, these sharp humps and dents need fractional-bit pre-emphasis. For a two-tap fractional FFE with tunable delay $\triangle T$ and pre-emphasis factor α [Fig. 2.16(b)], we arrive

at magnitude and phase response as

$$|H(j\omega)| = \sqrt{1 + \alpha^2 - 2\alpha\cos(\omega\triangle T)} \tag{2.20}$$

$$\angle H(j\omega) = \tan^{-1}\left[\frac{\alpha\sin(\omega\triangle T)}{1 - \cos(\omega\triangle T)}\right] \tag{2.21}$$



(a) (b)

Fig. 2.16 (a) VCSEL relaxation oscillation, (b)two-tap fractional-bit boosting.

Typical VCSEL requires a compensation of approximately 2 dB, i.e., α = 0.25. For high-speed operation, e.g., 25 Gb/s, we choose $\triangle T \cong 0.5\ T_b$ as a compromise between boosting efficiency and phase concordance. More details about pre-emphasis techniques will be discussed in Chapter 5.

REFERENCES

[1] N. Li *et al*., "High-performance 850 nm VCSEL and photodetector arrays for 25 Gb/s parallel optical interconnects," in *Proc. Optical Fiber Commun. Conf. (OFC)*, Mar. 2010, paper OTuP2.

[2] B. Razavi, *Design of Integrated Circuits for Optical Communications.*, New york, NY, USA: McGraw-Hill, 2002.

BROADBAND AMPLIFIERS

3

3.1 GENERAL CONSIDERATION

Front-end circuits for high-speed data links necessitate broadband amplifiers to enlarge input data. In this chapter, we discuss two main-stream broadband amplifiers: transimpedance amplifiers (TIAs) and limiting amplifiers (LAs). The former is dedicated to optical links, whereas the latter can be found in both electrical and optical receivers. Figure 3.1 illustrates a typical optical receiver frontend, which includes a transimpedance amplifier followed by the subsequent equalizer and limiting amplifier. The TIA converts the tiny current coming from the photodiode into voltage with some gain, and the equalizer restores the input data from channel distortion. The limiting amplifier increases the data seing until it saturates as a typical logic level, which is a few hundred mV in CMOS. The equalizer in front-end usually refers to a continuous-time linear equalizer (CTLE, see chapter 5), which is generally codesigned with the subsequent limiting amplifier.

In most optical cases, the TIA serves as the only single-ended device along the data path, requiring single-ended to differential converter between TIA and equalizer/limiting amplifier combination. It is to protect the subsequent circuits from common-mode noise or coupling. For an electrical front-end, on the other hand, no TIA is required as the input signal (whether in voltage or current) directly gets amplified by the limiting amplifier. Depending on the applications and system-level requirements, the CTLE is either put in front of the LA, or the two blocks are placed alternately. We focus our discussion on TIAs and LAs in this chapter, and address the issues of equalizers in chapter 5.

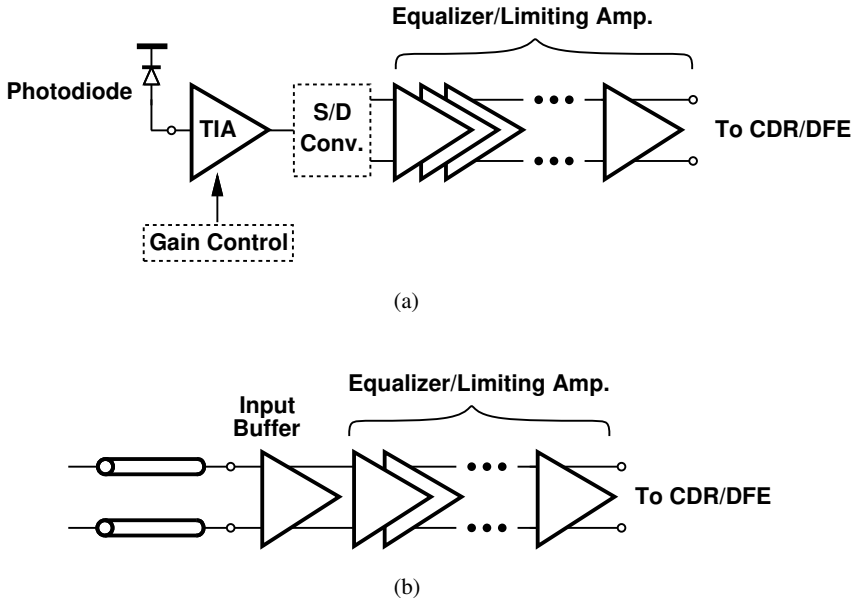


Fig. 3.1 Receiver frontend of (a)optical, (b)electrical systems.

Before getting into details of TIAs and LAs, we need to understand the fundamental properties of photodiodes. The most commonly used photodiode is realized as a P-intrinsic-N(PIN) structure of semiconductor[1]. As depicted in Fig. 3.2, such a PIN diode is usually reversely biased, conducting current whenever light (i.e., a photon with sufficient energy) enters the depletion region of the diode. The reverse-biased field sweeps the carriers and creates a current.The N-type and P-type regions are heavily doped to form ohmic contacts.

Figure 3.2 also illustrates an example of small-signal model for high-speed photodiodes and a picture showing how it looks like. Similar to other discrete components, a photodiode inevitably introduce parasitic capacitance. Since the quantum efficiency is above 90%, modern photodiodes usually present good responsivity R (defined as output laser current per unit input power). Typical responsivity is around 0.5 A/W for 850 nm laser and 0.9 A/W for 1.55 μm laser, respectively. The bandwidth of a photodiode actually depends on the reverse-biased voltage V_{RB}, as shown in Fig. 3.2. Note that the breakdown voltage could be as low as -5V. Bandgap references would be mandatory for TIAs to provide stable input common-mode levels.

[1]Intrinsic means undoped here.

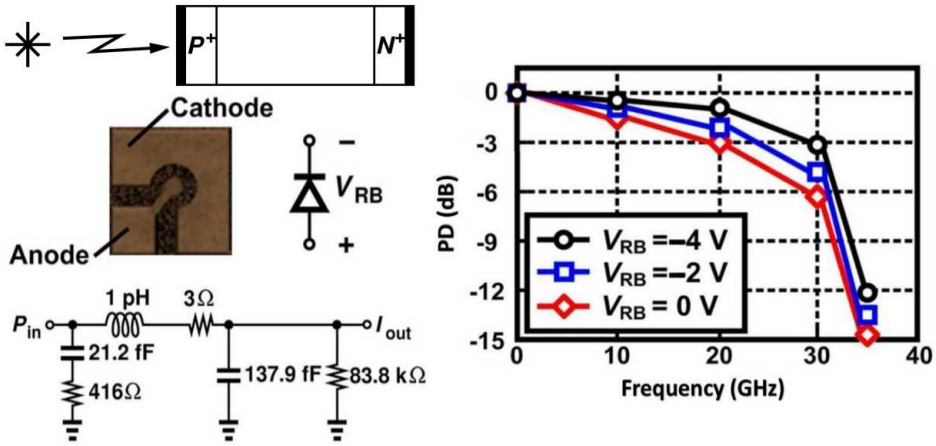


Fig. 3.2 PIN photodiode and typical responsibility as a function of frequncy.

Another important issue related to optical receiver front-ends is the difference between on (logic "ONE") and off (logic "ZERO") signals. Modern laser diode does not 100% turn off while transmitting a "0" signal in order to reduce reaction time. Extinction ratio (ER) is therefore defined to express the power ratio for the light source as on and off. It is also important to look at the average input power of light. These parameters help us to evaluate the signal-to-noise ratio (SNR), required conversion gain (i.e., transimpedance gain R_{Tran}), and link budget at the receive side.

Example 3.1

Consider an optical frontend shown in Fig.3.3, where the average optical input power is -12 dBm with ER = 6 dB . Equalization is neglected here. The photodiode has responsivity of 0.9 A/W, and the final data output must be as large al 600 mV_{PP}. (a) Determine the overall gain. (b) Estimate the maximum tolerable input-referred noise for BER $< 10^{-12}$.

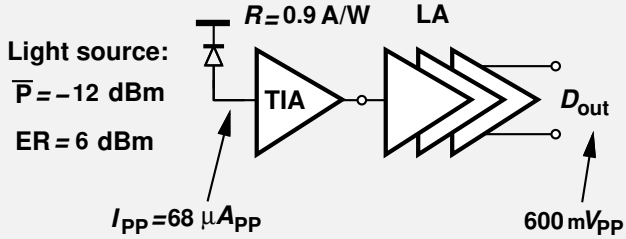


Fig. 3.3 Example of link budget of optical frontend.

Example 3.1 ***(Continued)***

Solution: (a) Denoting input power for "1" and "0" as P_1 and P_0, respectively, we have

$$P_1 = 4P_0$$
$$\frac{1}{2} \cdot (P_1 + P_0) = -12\ dBm = 63\ \mu W.$$

It follows P_1 = 100.8 μW and P_0 = 25.2 μW, and the corresponding current from photodiode becomes

$$I_1 = 90.7\ \mu A$$
$$I_0 = 22.7\ \mu A.$$

For small signal analysis, the peak-to-peak current input is given by 68μA. The total gain from the input of TIA to the output of LA is

$$Total \quad Gain = \frac{600\ mV}{68\ \mu A} = 8.8\ k\Omega = 79\ dB\Omega.$$

In practice, we may leave some margin for PVT variations. For example, we can choose TIA gain = 46 dBΩ , LA gain = 40 dB. (b) From the BER discussion in chapter 1, we need

$$\frac{I_{PP}}{2I_{n,RMS}} \geq 7$$

to ensure BER $< 10^{-12}$, where $I_{n,RMS}$ represents the square root of the input-referred noise power

$$I_{n,RMS} \triangleq \sqrt{\overline{I_{n,in}^2}}.$$

That is, the maximum allowable noise current is 4.8 μA,rms.

The term ”input-referred noise” needs explanation here. Different from low-frequency amplifier whose (thermal) noise flat within the band of interest, broadband amplifiers such as TIAs and LAs coners much wider bandwidth. Some components may contribute noise only at high frequencies. Consequently, to fairly estimate the noise performance, we integrate the output noise across the whole spectrum. The input referred noise power is defined as the overall output noise power divided by the square of (low-frequency) transimpedance gain:

$$\overline{I_{n,\,in}^2} = \frac{\int_0^\infty \overline{V_{n,\,out}^2}\, df}{R_{Tran,DC}^2}\,. \tag{3.1}$$

We use $I_{n,RMS} = \sqrt{\overline{I_{n,in}^2}}$ to describe RMS noise current.

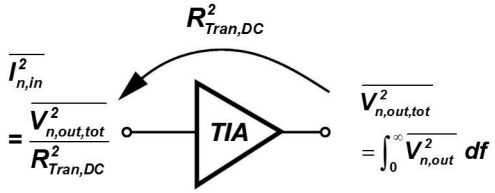


Fig. 3.4 input-referred noise.

In reality, the photodiode itself contribute noise too. Since a diode's shot noise is given by $\overline{I_n^2} = 2qI$, where q denotes electron charge and I the carrying current, we have the RMS noise attributed to photodiode as

$$I_{n,\,shot,\,1}^2 = \sqrt{2qI_1BW_n} \tag{3.2}$$

$$I_{n,\,shot,\,0}^2 = \sqrt{2qI_0BW_n}\,, \tag{3.3}$$

where $\mathrm{BW_n}$ represents the equivalent noise bandwidth. If $\mathrm{BW_n}$ = 10 GHz, for instance, we arrive at $I_{n,\,shot,\,1}$ = 0.54 (μA,rms) and $I_{n,\,shot,\,0}$ = 0.27 (μA,rms), respectively. The shot noise from photodiode is usually small as compared with the TIA/LA noise.

The single-end operation of TIA makes itself vulnerable to common-mode noise or unwanted coupling. Several ways can be adopted to do the single-ended to differential conversion. A straightforward approach can found in Fig.3.5(a), where an RC low-pass filter takes out the dc value of the single-ended output voltage from TIA.The current steering pair $M_{1,2}$ thus creates a differential output. Some frontend designs may have dummy TIAs to provide reference power level

for automatic gain control. In such case, we can take the outputs from both the real and dummy TIAs and adjust the intrinsic offset through the loop across the buffer.As shown in Fig.3.5(b), the output of TIA_2 stays at "0" level all the time. The $M_{1,2}$ pair together with R_S and imbalanced tail currents I_{ss1} and I_{ss2} counter-balance the tilted input to the first order.Taking the average dc level of D_{out} by RC low-pass filter, we utilize an error amplifier along with an auxiliary current source M_3 to tune the residual offset. Due to the error amplifiers high gain, the negative feedback loop forces the output data to be fully differential.

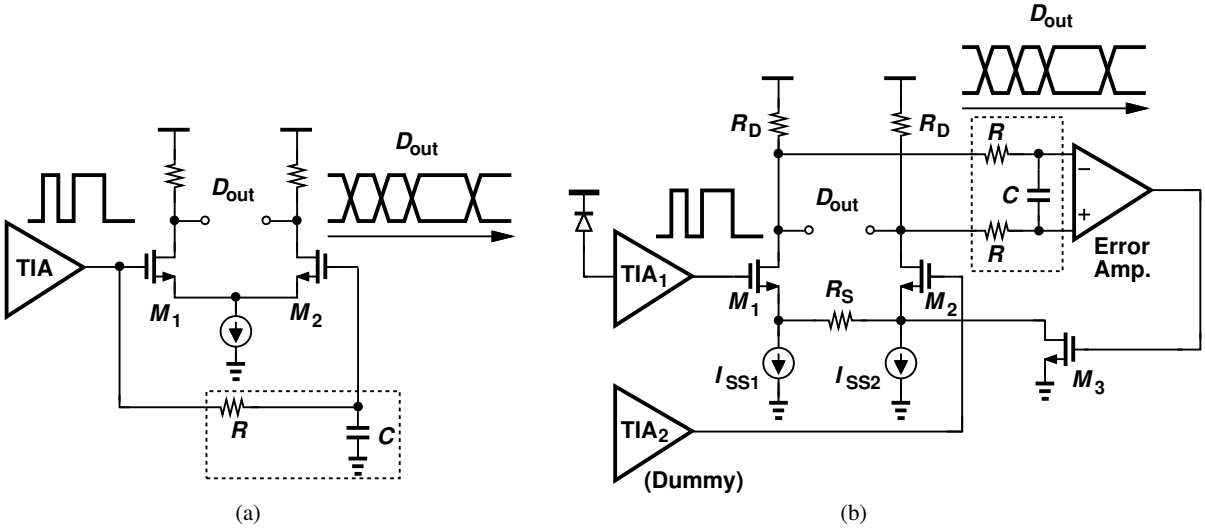


Fig. 3.5 Single-end to differential conversion with (a)RC low-pass filter, (b)error amplifier.

The TIA and LA still have other issues in design. For example, to avoid saturation, automatic gain control can be introduced to TIAs so as to cover a longer dynamic range. We address these issues when getting into circuit details.

3.2 FEEDBACK TIA

A conventional feedback TIA usually employs a low-noise operational amplifier (Opamp) with a resistive feedback. As shown in Fig.3.6, the injected current I_{in} (from photo diode) is transferred to voltage by means of the feedback resistor R_F. At low frequencies, the transimpedance gain

R_{Tran} is given by

$$R_{Tran} = \frac{V_{out}}{I_{in}} = -R_F \,. \tag{3.4}$$

For example, if R_F = 1kΩ, we have R_{Tran} = 50 dBΩ.

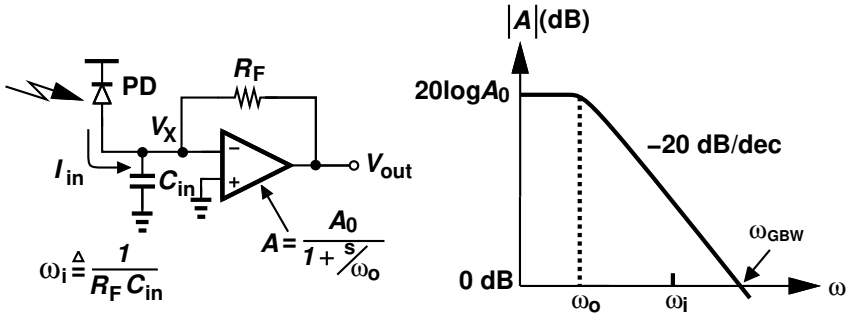


Fig. 3.6 Conventional feedback TIA.

One important issue of such an implement is the parasitic capacitance of the photo diode and Opamp input port. The former is on the order of hundreds of fF, and the latter may be as large as tens of pF. We lump it as C_{in}. If the open-loop response of Opamp is represented as a first order transfer function (which is true in most cases), we obtain R_{Tran} as

$$R_{Tran} = -\frac{R_F \cdot A_0 \omega_o \omega_i}{[s^2 + (\omega_o + \omega_i)s + (A_0 + 1)\omega_o \omega_i]} \,. \tag{3.5}$$

Here, A_0 and ω_o denote the open loop gain and bandwidth of the Opamp. We also define

$\omega_i \triangleq (R_F C_{in})^{-1}$. The $-$20-dB/dec slope also suggests the gain-bandwidth product of the Opamp is equal to $A_0\,\omega_o$. If ω_{GBW} denote the frequency at which the open-loop gain intersects 0 dB, we have

$$\omega_{GBW} = A_0\,\omega_o \,. \tag{3.6}$$

As expected, R_{Tran} approaches R_F as S→0.

The second-order transfer function of Eq(3.4) can be studied in standard form:

$$R_{Tran} \triangleq \frac{K_1}{s^2 + \left(\frac{\omega_n}{Q}\right)s + \omega_n^2} \,. \tag{3.7}$$

Where

$$\omega_n^2 = (A_0 + 1)\omega_o\omega_i \tag{3.8}$$

$$Q = \frac{\sqrt{(A_0 + 1)\omega_o\omega_i}}{\omega_o + \omega_i} \tag{3.9}$$

$$K_1 = -R_F A_0 \omega_o \omega_i \,. \tag{3.10}$$

Since $A_0 \gg 1$ and $\omega_o \ll \omega_i$, we have $Q = [A_0\omega_o/omega_i]^{1/2}$. In practical realization, $omega_i$ is very likely to be less than or much less than the unity-gain bandwidth (ω_{GBW}) of the Opamp. This especially true for discrete implementation targeting high speed and high gain simultaneously. For example, if gain-bandwidth product = 300 MHz, R_F = 1 kΩ and C_{in} = 5pF, we have $\omega_{GBW} = 9.4\omega_i$ and Q = 3.1. Such a high Q leads to severe peaking on the response of transimpedance gain. We study the peaking effect in the following example.

Example 3.2

Determine the peaking of R_{Tran} for (a) $\omega_{GBW} = 10\omega_i$, (b) $\omega_{GBW} = 100\omega_i$.

Solution:

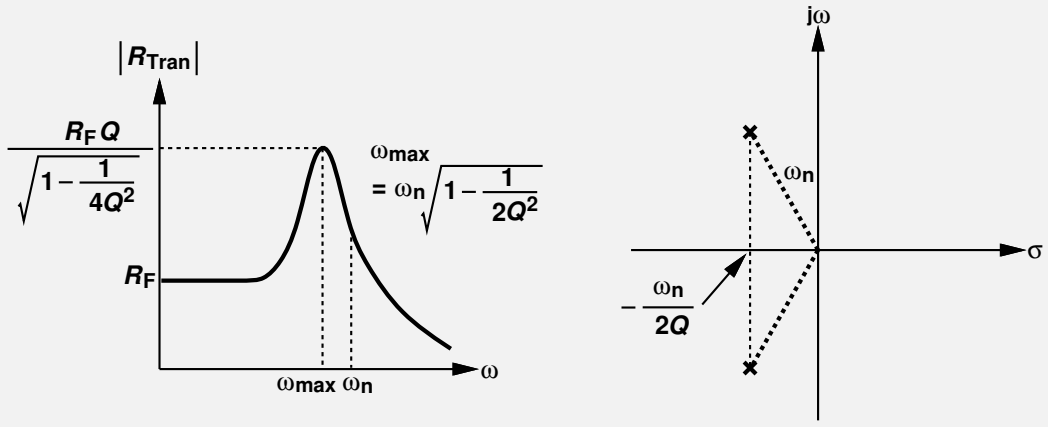


Fig. 3.7 Analysis of peaking due to different Q.

Example 3.2 *(Continued)*

Based on standard 2nd-order transfer function analysis, we plot $|R_{Tran}|$ as a function of ω in Fig.3.7.It is well-known that the peaking appears for $Q >$ 1/2:

$$Peaking = 20\ log_{10}\left(\frac{Q}{\sqrt{1-\frac{1}{4Q^2}}}\right) = \begin{cases} 10.1dB,\ for\ \ Q = 3.16 \\ 20dB,\ for\ \ Q = 10\,. \end{cases}$$

Meanwhile, we have

$$\omega_n = \begin{cases} 3.16\omega_i,\ for\ \ Q = 3.16 \\ 10\omega_i,\ for\ \ Q = 10\,. \end{cases}$$

The poles of the denominator of R_{Tan} is also plotted here. The larger the Q is, the closer the Conjugate poles approach imaginary axis.

Example 3.2 implies that the circuit in Fig.3.6 may be prove to instability or even oscillation. Since ω_i and ω_{GBW} are quite restricted by specifications, they form a severe tradeoff and a significant peaking seems inevitable. Fortunately, a simple modification can provide efficient rescue. As illustrated in Fig.3.8, a capacitor C_F is introduced in parallel with R_F in the feedback loop. Denoting $\omega_F = (R_F C_F)^{-1}$, we re-calculate the transimpedance gain R_{Tran}. Omitting the tedious derivation, we obtain

$$R_{Tran} = \frac{K_2}{s^2 + \left(\frac{\omega_n}{Q}\right)s + \omega_n^2}\,. \tag{3.11}$$

Here,

$$\omega_n^2 = (1 + A_0)\frac{\omega_o\omega_i\omega_F}{\omega_F + \omega_i} \tag{3.12}$$

$$Q = \left[\frac{1}{\sqrt{1+A_0}} \cdot \left(\frac{1}{\omega_o} + \frac{1}{\omega_i} + \frac{A_0}{\omega_F}\right) \cdot \sqrt{\frac{\omega_o\omega_i\omega_F}{\omega_F + \omega_i}}\ \right]^{-1} \tag{3.13}$$

$$K_2 = -A_0\, R_F \cdot \frac{\omega_o\omega_i\omega_F}{\omega_F + \omega_i}\,. \tag{3.14}$$

The response of R_{Tran} is still in second order. However, we have one more parameter (i.e., ω_F) to moderate Q. For most cases, it is preferable to put ω_F somewhere between ω_i and ω_{DBW} to dramatically reduce Q. We study the following example to gain more insight into this compensation technique.

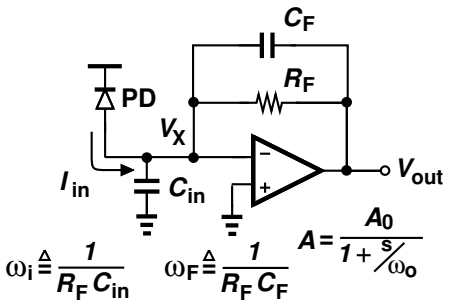


Fig. 3.8 Modified feedback TIA and C_F.

Example 3.3

In Example 3.2(b), if we choose ω_F = 10ω_i, calculate the peaking of R_{Tran}.

Solution:

If ω_{GBW} = $A_0\omega_o$ = 100ω_i and ω_F = 10ω_i, we obtain

$$Q = 0.95$$

$$\omega_n = \omega_F = 10\omega_i.$$

Figure 3.9 illustrates locations of poles. The peaking now reduces to 0.97 dB, well acceptable in most applications. Note that sn does not change at all. In other words, the introduction of C_F neither sacrifices bandwidth nor dissipates more power.

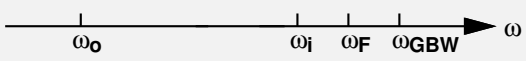


Fig. 3.9 Pole arrangement.

In reality, the choice of C_F may require iterative calculation or even simulation to achieve the optimum performance. However, an easy estimation can be obtained if we put ω_F as the geometric of ω_i and ω_{GBW}, i.e., $\omega_F = \omega_F = \sqrt{\omega_i \omega_{GBW}}$. The reader can prove that Q is not sensitive as C_F varies.

It is instructive to examine the input impedance of a feedback TIA. For simplicity we assume $\omega_{GBW} \gg \omega_i$ and $\omega_F = \omega_F = \sqrt{\omega_i \omega_{GBW}}$. To clarify the effect, let us take C_{in} away from the rest of the circuit and consider their impedance separately. Placing a testing current source. It with voltage Vt into TIA [Fig.3.10(a)], we obtain the equivalent input impedance

$$Z_1 = \frac{V_t}{I_t} \cong \frac{R_F(1 + s/\omega_o)}{(1 + A_0)(1 + s/\omega_{GBW})(1 + s/\omega_F)}. \tag{3.15}$$

Here we assume $A_0 \gg 1$. At low frequencies, Z_1 degenerates to $R_F/(1 + A_0)$.

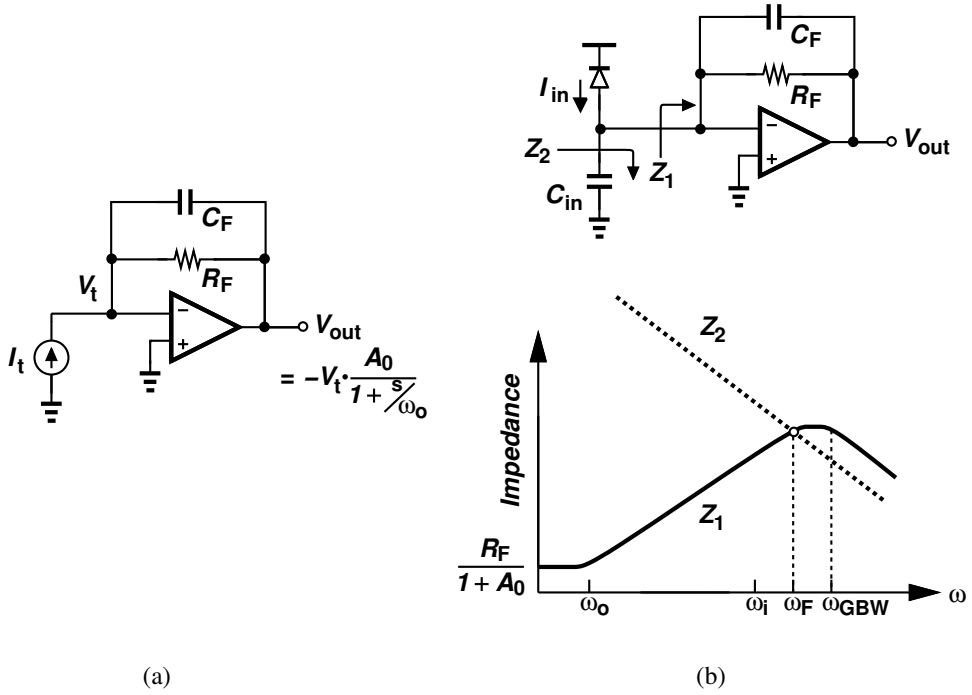


Fig. 3.10 (a)Input impedance calculation, (b)effect of input impedance.

The zero pushes Z_1 to climb up with a slope of 20 dB/dec until ω_F, where it encounters the first pole. Z_1 falls down after ω_{GBW} again, as the effect of second pole occurs, on the other hand,

the impedance of C_{in} (defined as Z_2) falls at -20 dB/dec. That is,

$$Z_2 = \frac{1}{sC_{in}}. \tag{3.16}$$

Interestingly, it intersects with Z_1 at $\omega_F = \sqrt{\omega_i \omega_{GBW}} = \omega_F$. That is, as frequency approaches ω_F, half of the input current from photo diode no longer flows into the TIA but rather C_{in}. Such a high input impedance issue would become worse if A_0 is not large enough (which is true for monolithic implementation). We introduce TIA architecture low input impedance in 3.XX

How do we implement a high-speed feedback TIA in CMOS? Apparently we can not build up an Opamp, as it would be too slow, noisy, and power hungry. A simple common source amplifier could be a good choice, which provides sufficient bandwidth.

We intuitively think of a source follower as shown in Fig.3.11(a). At low frequencies, the transimpedance gain and input, output impedance are given by

$$R_{Tran,DC} = \frac{g_{m1} R_D R_F}{1 + g_{m1} R_D} \tag{3.17}$$

$$R_{in} = \frac{R_F}{1 + g_{m1} R_D} \tag{3.18}$$

$$R_{out} = \frac{1 \,/ g_{m2}}{1 + g_{m1} R_D}. \tag{3.19}$$

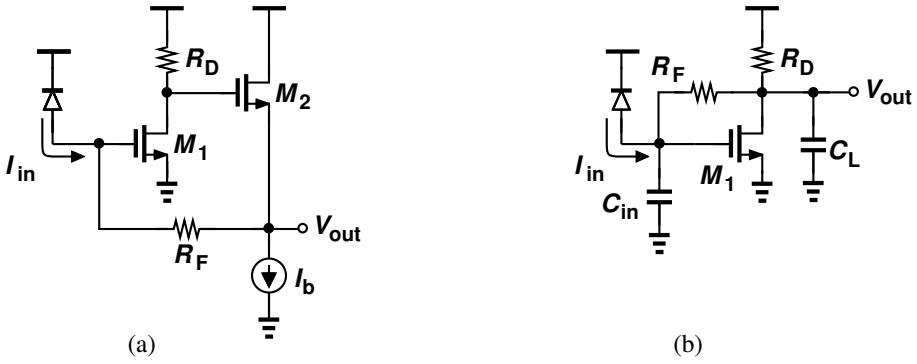


Fig. 3.11 Monolithic feedback TIA in CMOS (a)with, (b)without source follower.

As expected, R_{Tran} approaches R_F as $g_{m1} R_o \gg 1$. Meanwhile, the shunt-shunt feedback lowers the input/output impedance significantly. However, the source follower introduces a series of

issues. First, the parasitic capacitance introduced by the I_b severely degrades the operation speed. The source follower itself also presents inductive output impedance, potentially causing ringing if the loading capacitor is heavy. The supply must be large enough to accommodate voltage headroom, including one overdrive for I_b, one V_{GS} for M_2, and IR drop for R_D. Typically a supply voltage equal to or greater than 1.8V is a better choice. As a result, it is preferable to get rid of the source follower. Shown in Fig.3.11(b) is TIA with direct feedback. Here, the input and output capacitances are denoted as C_{in} and C_L respectively. At first glance, we neglect the effect of C_{in} of C_{in} and C_L and chuck the low-frequency properties. The transimpedance gain and input/output impedance now become

$$R_{Tran,DC} = -\frac{g_{m1}R_F - 1}{g_{m1}R_D + 1}R_D \tag{3.20}$$

$$R_{in} = \frac{R_F + R_D}{1 + g_{m1}R_D} \tag{3.21}$$

$$R_{out} = R_D \| (1/g_{m1})\,. \tag{3.22}$$

As $g_{m1}R_D \gg 1$ and $g_{m1}R_F \gg 1$, R_{Tran} approaches to $-R_F$. The input and output impedances are greater than Eq.3.18 and Eq.3.19 due to the lack of isolation in feedback loop. A supply voltage as low as 1V is sufficient for the TIA in Fig.3.11(b), as it only has to cover one V_{GS} for M_1 and one IR drop for R_D. Note that R_F carries no dc current.

The major advantage of such a direct feedback TIA is that it needs no additional capacitor along the feedback path. To gain more insight, we express the transimpedance gain including the capacitance

$$\begin{aligned} R_{Tran} &= \frac{(1 - g_{m1}R_F)R_D}{R_F R_D C_L C_{in} s^2 + [R_D C_L + (R_F + R_D)C_{in}]s + 1 + g_{m1}R_D} \\ &\cong \frac{-R_F A_0 \omega_o \omega_i}{s^2 + \left[\omega_i + \dfrac{1}{(R_F \| R_D)C_L}\right] s + (1 + A_0)\omega_o \omega_i}\,. \end{aligned} \tag{3.23}$$

Again, we lump resistors and capacitors as $\omega_i = (R_F C_{in})^{-1}$, $\omega_o = (R_D C_L)^{-1}$ and $A_0 = g_{m1}R_D$. We also assume $g_{m1}R_F \gg 1$, which is reasonable in most cases. In fact, Eq.3.23 becomes exactly the same as Eq.3.5 if $R_F \gg R_D$.

The key point here is that ω_o is now much higher than ω_i. It is because the feedback resistor is usually greater than the loading resistor (to achieve high transimpedance gain and save voltage headroom), and the capacitance from photodiode is typically larger than the output loading. As a result, we arrive at

$$Q = \frac{\sqrt{(A_0+1)}\omega_o\omega_i}{\omega_o+\omega_i} \cong \sqrt{(A_0+1)\frac{\omega_i}{\omega_o}}\,. \tag{3.24}$$

Certainly, $\omega_n^2 = (1+A_0)\omega_o\omega_i$.

Since A_0 is quite low in this single-stage structure (e.g. $A_0 \approx 10$ in low-supply CMOS design), Q is usually a small number. As we know, for $Q \approx 1$ the peaking phenomenon is negligible. That is, the direct feedback TIA as illustrated in Fig.3.11(b) needs no feedback capacitor.

Example 3.4

Determine the peaking of R_{Tran} for the circuit in Fig.3.11(b), where R_D = 250Ω, R_F = 1kΩ, C_{in} = 300 fF, C_L = 100 fF, g_{m1} = 0.04 A/V.

Solution:

With the given condition we have

$$\omega_i = 2\pi \times 0.53\ \text{GHz}$$

$$\omega_o = 2\pi \times 6.37\ \text{GHz}$$

$$A_0 = 10\,.$$

Taking into Eq.(3.23), we have ω_n = $2\pi \times 6.1$ GHz, Q = 0.72, and R_{Tran} at low frequencies as -909 dBΩ. It almost presents a maximum flat response and the peaking is negligible.

Let us consider the noise performance of the direct feedback TIA. Denoting the current noise sources of R_D, R_F and M_1 as $\overline{I_{n,R_D}^2}$, $\overline{I_{n,R_F}^2}$, and $\overline{I_{n,M_1}^2}$, respectively, we draw the small-signal model in Fig.3.12 and obtain

$$\frac{-V_{n,\,out}}{R_D} + I_{n,R_D} = I_{n,R_F} + \frac{V_{n,\,out}-V_x}{R_F} + V_x\, g_{m1} + I_{n,M_1} + V_{n,\,out}\cdot sC_L \tag{3.25}$$

$$I_{n,R_F} + \frac{V_{n,\,out}-V_x}{R_F} = V_x \cdot sC_{in}, \tag{3.26}$$

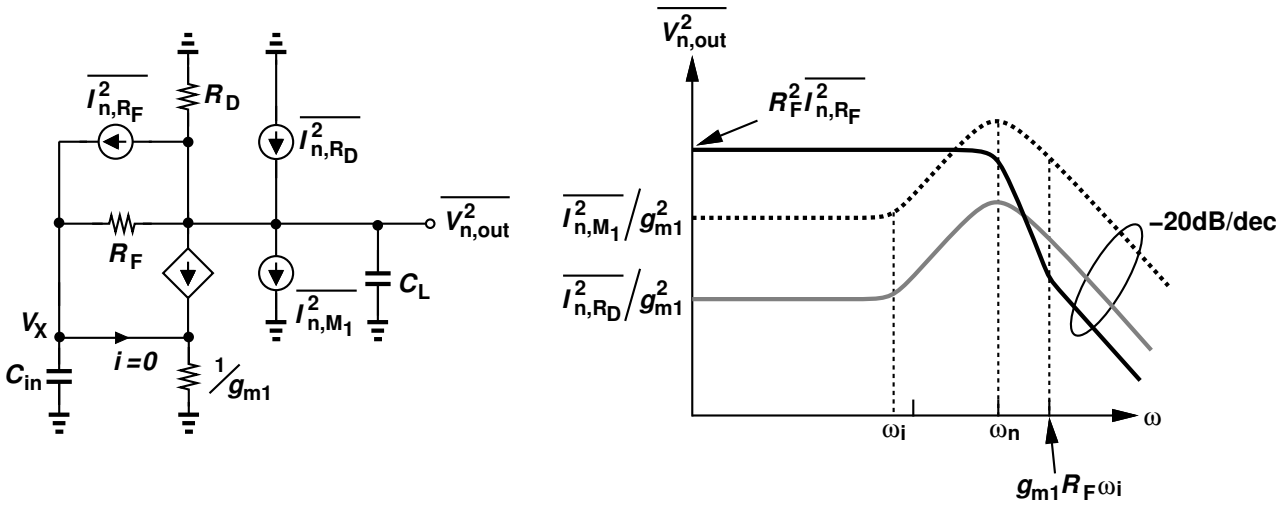


Fig. 3.12 Noise calculation of direct feedback TIA in Fig.11(b).

where V_x represents the gate voltage in small signal. Here we assume $Q \leq 1$ so that -3 dB bandwidth of the circuit is in the vicinity of ω_n. It is true for regular designs. After reorganizing the equations, we obtain

$$\begin{aligned}\overline{V^2_{n,\,out}} = &\frac{\overline{I^2_{n,R_F}}(1+g_{m1}R_F)^2R_D^2\,\omega_i^2\,\omega_o^2}{|s^2+(\frac{\omega_n}{Q})s+\,\omega_n^2|^2}\cdot|1+\frac{s}{\omega_i(1+g_{m1}R_F)}|^2\\ &+\frac{\overline{I^2_{n,M_1}}R_D^2\,\omega_i^2\,\omega_o^2}{|s^2+(\frac{\omega_n}{Q})s+\,\omega_n^2|^2}\cdot|1+\frac{s}{\omega_i}|^2\\ &+\frac{\overline{I^2_{n,R_D}}R_D^2\,\omega_i^2\,\omega_o^2}{|s^2+(\frac{\omega_n}{Q})s+\,\omega_n^2|^2}\cdot|1+\frac{s}{\omega_i}|^2. \qquad (3.27)\end{aligned}$$

Although it looks complicated, $\overline{V^2_{n,uot}}$ can be easily explained by observing the spectra of its 3 components (Fig.3.12). The first tern, R_F noise (solid line), starts as approximately $R_F^2\overline{I^2_{n,R_F}}$(assume $g_{m1}R_F \gg 1$) at dc and keeps flat until ω_n, at which it bends down to a sharp slope of -40 dB/dec. It soon turns back at $g_{m1}R_F\omega_i$ and reduces the slope to -20 db/dec since then. The second term, M_1 noise (dash line), also starts as a flat line of $\overline{I^2_{n,M_1}}/g^2_{m1}$. However, it rises up at the zero ω_i and falls down around ω_n. The third term, R_D noise (gray line), has the same shape. Since $g_{m1}R_D \gg 1$

(at least on the order of 10), M_1 contributes much more noise than R_D does. Similarly, R_F reveals the most noise at low frequencies as $g_{m1}R_F \gg \gamma$. In other words, R_F presents a tradeoff between conversion gain and noise. All three noise components roll off at -20 dB/dec for $\omega > g_{m1}R_F\omega_i$. In practice, the noise contribution highly depends on design parameters, and simulation is mandatory for performance optimization. Nonetheless, integrating the noise spectrum leads to the overall noise voltage at output:

$$\overline{V_{n,not,tot}^2} = \int_0^\infty \overline{V_{n,out}^2}\, df\,. \tag{3.28}$$

The input-referred noise is therefore obtained as

$$\overline{I_{n,in}^2} = \frac{\int_0^\infty \overline{V_{n,out}^2}\, df}{R_{Tran,dc}^2}\,. \tag{3.29}$$

Where R_{Tran}, dc denotes the transimpedance gain at dc.

It is instructive to examine the noise performance of our previous example. Figure 3.13 illustrates the simulated noise performance of the TIA in Fig.3.11(b) with the same device parameters of Eample 3.4. γ is set to 3 in this case. The integrated output noise $\overline{V_{n,out,tot}^2}$ is given by xx V^2, where R_F, M_1 and R_D contribute xx, xx, and xx V^2, respectively. Since R_{Tran} = -909 dBΩ, the input-referred noise $I_{n,in}$ is equal to xx $\mu A, rms$.

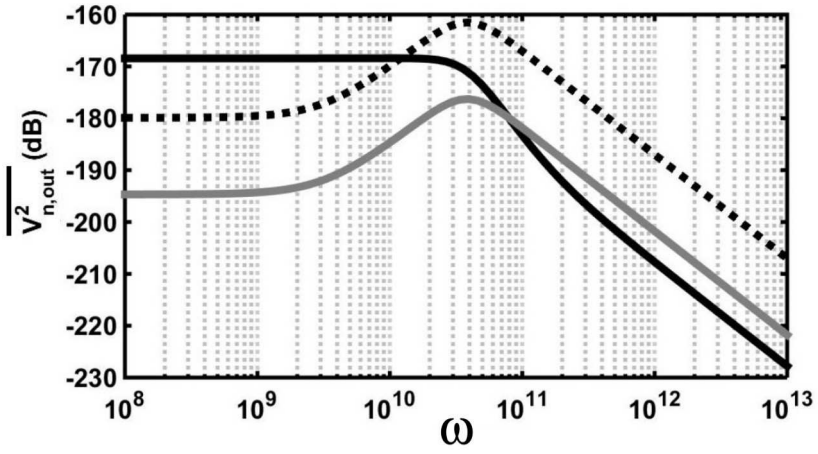


Fig. 3.13 Simulated noise profile of circuit in Fig.11(b) (with R_D = 250Ω, R_F = 1 kΩ, C_{in} = 300fF, C_L = 100fF, g_{m1} = 0.04 A/W and γ = 3).

We investigate a transformed version of direct feedback TIA to close this section. Shown in Fig.3.14 is a self-biased inverter, which is potentially suitable for converting current into voltage. Here g_{mN}, g_{mP} and r_{oN}, r_{oP} denote the transconductance and output resistance of M_N and M_P, respectively. Indeed, if we look at its small signal model, we realize that it is identical to that of a direct feedback TIA in Fig.3.14(b) except g_{m1} becomes $(g_{mN} + g_{mP})$ and R_D becomes $r_{oN} \| r_{oP}$. The low-frequency gain and input/output impedance are given by

$$R_{Tran,DC} \cong -R_F \tag{3.30}$$

$$R_{in} \cong (g_{mN} + g_{mP})^{-1} \tag{3.31}$$

$$R_{out} \approx (g_{mN} + g_{mP})^{-1}\,, \tag{3.32}$$

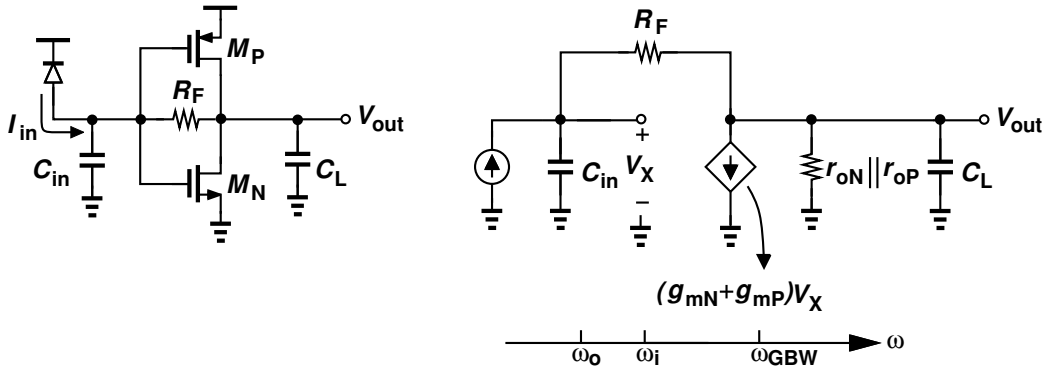


Fig. 3.14 Inverter based TIA.

if $A_0 = (g_{mN} + g_{mP})(r_{oN} \| r_{oP}) \gg 1$, $(g_{mN} + g_{mP})R_F \gg 1$ and $(r_{oN} \| r_{oP}) \gg R_F$. The complete R_{Tran} as a function of ω is readily available as well:

$$R_{Tran} \cong \frac{-R_F\ A_0\ \omega_o\ \omega_i}{s^2 + \left(\omega_i + \dfrac{1}{R_F\ C_L}\right) s + (1 + A_0)\ \omega_o\ \omega_i}\,, \tag{3.33}$$

where $\omega_i = (R_F C_{in})^{-1}$, $\omega_o = [(r_{oN} \| r_{oP}) \cdot C_L]^{-1}$. Since the open loop gain A_0 becomes much larger now and ω_o be significantly lower than ω_i, this inverter-based TIA might be subject to instability. The noise would become higher because of the introduction of M_P.

3.3 COMMON-GATE TIA

Perhaps the simplest structure to realize a TIA is to use a common-gate amplifier [Fig.3.15(a)]. Here, input current from photodiode injects into the source of M_1 and converts to output voltage by means of R_D. Here, M_2 serves as a constant current source. At low frequency, R_{Tran} = R_D, $R_{in} \approx 1/g_{m1}$, and R_{out} R_D. As frequency goes up, parasitic capacitors C_{in} and C_L come into the picture and the transimpedance gain becomes

$$R_{Tran} = \frac{R_D}{(1 + s/\omega_{in})(1 + s/\omega_{out})}. \tag{3.34}$$

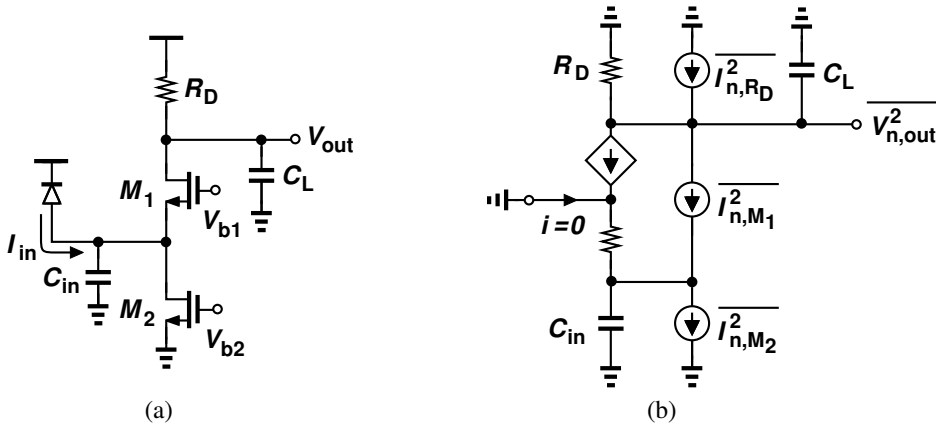


Fig. 3.15 (a)Common-gate TIA,(b)its noise modal.

Following the same notation, we define ω_{in} = g_{m1}/C_{in} and ω_{out} = $(R_D C_L)^{-1}$. Note that now we are dealing with two real poles rather than conjugate ones. For high-speed design, it is desirable to push all poles as high as possible. Typically, ω_{in} and ω_{out} have the same order of magnitude in most cases.

Let us look at the noise performance of a common-gate TIA. For simplicity, we assume $\omega_{in} \approx \omega_{out}$. With the noise model shown in Fig.3.15(b), the output noise can be calculated as

$$\begin{aligned}\overline{V_{n,\,out}^2} &= \overline{I_{n,R_D}^2} \cdot \left|\frac{R_D}{1+s/\omega_{out}}\right|^2 \\ &+ \overline{I_{n,M_1}^2} \cdot \left|\frac{R_D}{1+s/\omega_{out}}\right|^2 \cdot \left|\frac{s/\omega_{out}}{1+s/\omega_{out}}\right|^2 \\ &+ \overline{I_{n,M_2}^2} \cdot \left|\frac{R_D}{1+s/\omega_{out}}\right|^2 \cdot \left|\frac{1}{1+s/\omega_{out}}\right|^2 . \end{aligned} \tag{3.35}$$

The spectrum components are depicted in Fig.3.16. The noise from R_D rolls off beyond ω_{out} at a rate of -20 dB/dec, but the noise from M_2 decays beyond the same point at a steeper rate of -40 dB/dec. Since $g_{m2}R_o\gamma$ is greater than 1 in regular cases, the M_2 noise has higher dc value. The noise from M_1 experience first-order low-pass and high-pass response at the same corner frequency ω_{out}, resulting in a hill shape spectrum.

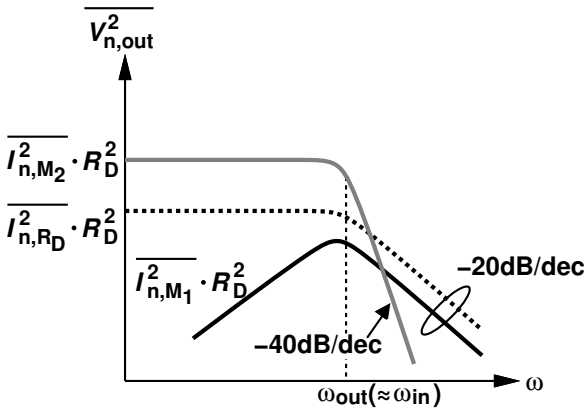


Fig. 3.16 Noise spectrum of common-gate TIA for $\omega_{in} \approx \omega_{out}$: noise contributed by M_1(solid), R_D(dash), M_1(gray).

To estimate the overall noise at output port, we integrate the noise spectrum across the whole bandwidth:

$$\overline{V_{n,\,out,\,tot}^2} = \int_0^\infty \overline{V_{n,\,out}^2}\,(\omega = 2\pi \mathrm{f})\ df. \tag{3.36}$$

The three can be separately calculated. Namely,

$$\overline{V^2_{n,\,out,\,R_D}} = \int_0^\infty \frac{\overline{I^2_{n,\,R_D}} \cdot R_D^2}{1+(\mathrm{f}/\mathrm{f_{out}})^2} df = \overline{I^2_{n,\,R_D}} \cdot R_D^2 \cdot \mathrm{f_{out}} \cdot \frac{\pi}{2} \tag{3.37}$$

$$\begin{aligned}\overline{V^2_{n,\,out,\,M_1}} &= \int_0^\infty \frac{\overline{I^2_{n,\,M_1}} \cdot R_D^2}{1+(\mathrm{f}/\mathrm{f_{out}})^2} \cdot \frac{1+(\mathrm{f}/\mathrm{f_{out}})^2}{1+(\mathrm{f}/\mathrm{f_{out}})^2}\, df \\ &= \overline{I^2_{n,\,M_1}} \cdot R_D^2 \cdot \mathrm{f_{out}} \cdot \int_0^\infty \frac{1}{1+\mathrm{u}^2} - \frac{1}{(1+\mathrm{u}^2)^2}\, du \\ &= \overline{I^2_{n,\,M_1}} \cdot R_D^2 \cdot \mathrm{f_{out}} \cdot \frac{\pi}{4}\end{aligned} \tag{3.38}$$

$$\begin{aligned}\overline{V^2_{n,\,out,\,M_2}} &= \int_0^\infty \frac{\overline{I^2_{n,\,M_2}} \cdot R_D^2}{[1+(\mathrm{f}/\mathrm{f_{out}})^2]^2}\, df \\ &= \overline{I^2_{n,\,M_2}} \cdot R_D^2 \cdot \mathrm{f_{out}} \cdot \frac{\pi}{4}.\end{aligned} \tag{3.39}$$

As a result, we arrive at

$$\overline{V^2_{n,\,out,\,tot}} = \frac{\pi}{4} \cdot R_D^2 \cdot f_{out} \left[2\overline{I^2_{n,\,R_D}} + \overline{I^2_{n,\,M_1}} + \overline{I^2_{n,\,M_2}}\right]. \tag{3.40}$$

The input-referred noise power is defined as the overall output noise power divided by the square of conversion gain at dc. That is,

$$\overline{I^2_{n,\,in}} = \frac{\overline{V^2_{n,\,out,\,tot}}}{R_D^2} = \frac{\pi}{4} \cdot f_{out} \left[2\overline{I^2_{n,\,R_D}} + \overline{I^2_{n,\,M_1}} + \overline{I^2_{n,\,M_2}}\right]. \tag{3.41}$$

Since g_{m1} and g_{m2} are on the same order of magnitude, M_1 actually contributes commeasurable amount of noise as M_2 does.

It is instructive to investigate the noise performance for the case $\omega_{out} \gg \omega_{in}$. Following the same noise calculation, we obtain the output noise as

$$\begin{aligned}\overline{V^2_{n,\,out}} = \overline{I^2_{n,R_D}} &\cdot \left|\frac{R_D}{1+s/\omega_{out}}\right|^2 \\ &+ \overline{I^2_{n,M_1}} \cdot \left|\frac{\omega_{in}}{1+s/\omega_{in}}\right|^2 \cdot \left|\frac{R_D}{1+s/\omega_{out}}\right|^2 \\ &+ \overline{I^2_{n,M_2}} \cdot \left|\frac{1}{1+s/\omega_{in}}\right|^2 \cdot \left|\frac{R_D}{1+s/\omega_{out}}\right|^2.\end{aligned} \tag{3.42}$$

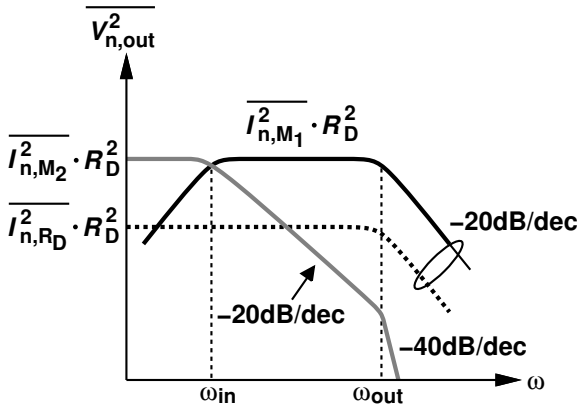


Fig. 3.17 Noise spectrum of common-gate TIA for $\omega_{in} \gg \omega_{out}$: noise contributed by M_1(solid), R_D(dash), M_1(gray).

Since ω_{in} and ω_{out} are apart from each other, it is straight forward to plot the noise spectrum as illustrated in Fig.3.17. The R_D noise (dash) keeps flat as $\overline{I^2_{n,R_D}} \cdot R^2_D$ and presents a first-order rolling off beyond ω_{out}. The M_1 noise (solid) reveals a high-pass response with pass band from ω_{in} to ω_{out}. The M_2 noise (gray) exhibits low-pass response with two poles of ω_{in} and ω_{out}, respectively. Since $\omega_{out} \gg \omega_{in}$, we integrate $\overline{V^2_{n,out}}$ across the whole spectrum and obtain $\overline{V^2_{n,out,tot}}$ as

$$\begin{aligned}\overline{V^2_{n,\,out,\,tot}} &= \int_0^\infty \overline{V^2_{n,\,out}}\,(\omega = 2\pi \mathrm{f})\ df \\ &\cong R^2_D \cdot \left(\overline{I^2_{n,\,R_D}} + \overline{I^2_{n,\,M_1}}\right) \int_0^\infty \frac{1}{1 + f^2/f^2_{out}}\, df \\ &+ R^2_D \cdot I^2_{n,\,M_2} \int_0^\infty \frac{1}{1 + f^2/f^2_{in}}\, df. \end{aligned} \tag{3.43}$$

Note that the integration variable is $f(=\frac{\omega}{2\pi})$. Owing to the fact that $R_D\gamma g_{m1} \gg 1$ and g_{m1} is on the same order as g_{m2}, we can further simplify the total output noise

$$\overline{V^2_{n,\,out,\,tot}} \approx R^2_D \cdot \overline{I^2_{n,\,M_1}} \cdot f_{out} \cdot \frac{\pi}{2}\,. \tag{3.44}$$

The input referred noise is thus given by

$$\overline{I^2_{n,\,in}} = \frac{\overline{V^2_{n,\,out,\,tot}}}{R^2_D} \cong \overline{I^2_{n,\,M_1}} \cdot f_{out} \cdot \frac{\pi}{2}\,. \tag{3.45}$$

which implies the noise performance is dominated by M_1 noise. In reality, $\omega_{out} \gg \omega_{in}$ means the R_{Tran} bandwidth is limited to ω_{in}, which contradicts the requirement for high-speed operation. In order words, this kind of situation rarely happens.

3.4 REGULATED-CASCODE TIA

The above two TIA structures encounter the same difficulty−the input resistance is too high. Recall from section 3.1 that the photodiode presents a significant capacitance, whose equivalent impedance might be smaller than the input resistance of TIA at high frequencies. As a result, input current from the photodiode gets harder and harder to be injected into the TIA as data rate goes up. To improve the bandwidth, a so-called regulated cascode (RGC) TIA has been introduced in Fig.3.18(a). Applying the feedback source follower M_2 directly to the input port without a resistor, this architecture is well known for its low input impedance. The output is no longer taken from the source of M_2, but instead the drain of it. To further speed up the circuit, sometimes a resistor can be used to replace the tail current I_b. Inductor peaking can be added on top of R_{D1} and R_{D2} as well.

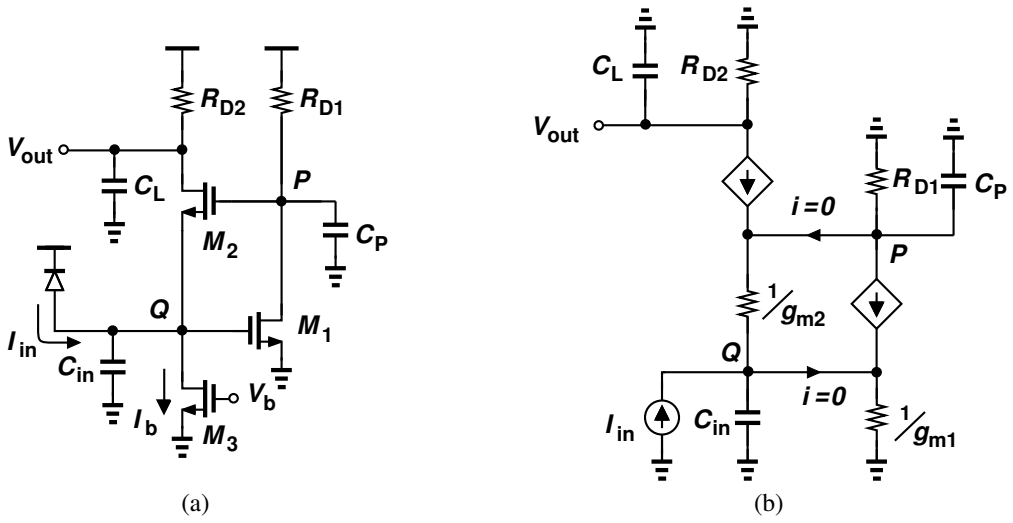


Fig. 3.18 (a)Regulated Cascode TIA, (b)its small signal model.

Let us consider the frequency response of transimpedance gain. Lumping the capacitance at input, output and node P as C_{in}, C_L and C_P, respectively, we draw the small-signal model in Fig.3.18(b) investigate R_{Tran}. At first glance, we neglect these capacitances for the time being and check the dc gain. Since the input current flows all the way up to R_{D2}, we have

$$R_{Tran,DC} = R_{D2}\,. \tag{3.46}$$

Now we look at the frequency response. Unfortunately, the three capacitors would make direct calculation too messy. We then calculate their associated poles independently. The reader can easily prove the three poles of V_{out}/I_{in} are given by

$$\omega_{in} = \frac{1}{C_{in}\left(\frac{1/g_{m2}}{1+g_{m1}R_{D1}}\right)} \tag{3.47}$$

$$\omega_{out} = \frac{1}{R_{D2}C_L} \tag{3.48}$$

$$\omega_P = \frac{1}{C_P\left(\frac{R_{D1}}{1+g_{m1}R_{D1}}\right)}\,. \tag{3.49}$$

We see that the input resistance here becomes $[g_{m2}\cdot(1+g_{m1}R_{D1})]^{-1}$. Compared with common-gate TIAs, RGC TIA's input resistance is reduced by a factor of $(1+g_{m1}R_{D1})$. The coupled cascode structure also lowers the equivalent resistance at node P. In regular designs, ω_{out} may probably serve as the dominant pole of R_{Tran}(s) with ω_{in} and ω_P not far away from it. Since ω_{out} is commensurate with the bandwidth of a typical differential pair, we expect RGC TIAs to operate at high speed.

Example 3.5

Determine the finite zero of the RGC TIA in Fig.3.18(a).

Solution:

Example 3.5 ***(Continued)***

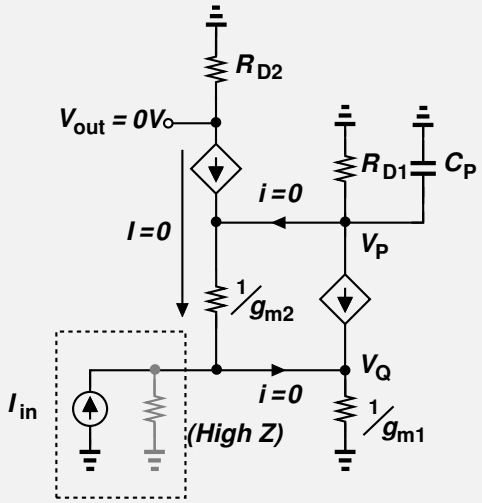


Fig. 3.19 Calculating zero associated with C_P.

We calculate the zero associated with C_P (Fig.3.19). Since V_{out} = $0V$, the current flowing through M_2 is also zero. Therefore V_P = V_Q. The current flowing through M_1 branch is isolated, we have

$$R_{D1} \| \frac{1}{s_z C_P} = -1/g_{m1} \, .$$

It follows that

$$s_z = \frac{1 + g_{m1} R_{D1}}{R_{D1} C_P} \, .$$

which is identical to its pole. The other two zeros caused by C_{in} and C_L are infinite.

The above example allows as to describe the complete R_{Tran}:

$$R_{Tran} \approx \frac{R_{D2}}{(1 + s/\omega_{in})(1 + s/\omega_{out})} \, . \tag{3.50}$$

In practical design, the three poles may not be easily separable, and the pole and zero of C_P could deviate from each other to same extent. Anyhow for simplicity, we still take the result from individual pole/zero analysis and preserve the approximation symbol to avoid inaccuracy.

Now we examine the noise performance. To make the analysis tolerable in hand calculation, we neglect the effect of C_P and assume $\omega_{in} \approx \omega_{out}$. These conditions are quite normal in high-speed

RGC TIAs. The 5 noise sources are drawn in Fig.3.20 as a small signal model, and the direction of noise currents are defined as shown. KCL suggests that

$$\frac{-V_P}{R_{D1}} + I_{n,R_{D1}} = I_{n,M_1} + V_Q g_{m1} \tag{3.51}$$

$$I_{n,R_{D2}} - \frac{V_{n,out}}{R_{D2} \| 1/sC_L} = I_{n,M_2} + (V_P - V_Q) \cdot g_{m2} = V_Q \cdot + sC_{in} + I_{n,M_3}\,. \tag{3.52}$$

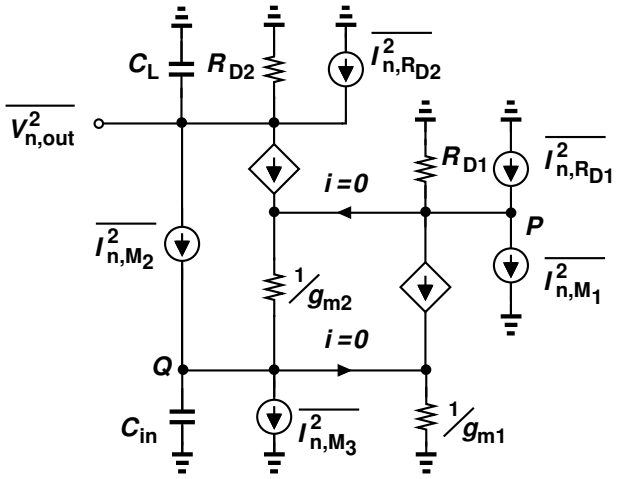


Fig. 3.20 Noise calculation.

Since $\omega_{in} \approx \omega_{out}$, $C_{in}/C_L \approx R_{D2}(1 + g_{m1}R_{D1})$. After re-arrangement, $\overline{V_{n,out}^2}$ can be obtained as

$$\overline{V_{n,\,out}^2} = \overline{V_{n,\,out,\,R_{D1}}^2} + \overline{V_{n,\,out,\,R_{D2}}^2} + \overline{V_{n,\,out,\,M_1}^2} + \overline{V_{n,\,out,\,M_2}^2} + \overline{V_{n,\,out,\,M_3}^2}\,, \tag{3.53}$$

where

$$\overline{V_{n,\,out,\,R_{D1}}^2} = g_{m2}^2 R_{D2}^2 R_{D1}^2 \frac{|s/\omega_{out}|^2}{|1 + s/\omega_{out}|^4} \cdot \overline{I_{n,,\,R_{D1}}^2} \tag{3.54}$$

$$\overline{V_{n,\,out,\,R_{D2}}^2} = R_{D2}^2 \frac{1}{|1 + s/\omega_{out}|^2} \cdot \overline{I_{n,,\,R_{D2}}^2} \tag{3.55}$$

$$\overline{V_{n,\,out,\,M_1}^2} = g_{m2}^2 R_{D2}^2 R_{D1}^2 \frac{|s/\omega_{out}|^2}{|1 + s/\omega_{out}|^4} \cdot \overline{I_{n,,\,M_1}^2} \tag{3.56}$$

$$\overline{V_{n,\,out,\,M_2}^2} = R_{D2}^2 \frac{|s/\omega_{out}|^2}{|1 + s/\omega_{out}|^4} \cdot \overline{I_{n,,\,M_2}^2} \tag{3.57}$$

$$\overline{V_{n,\,out,\,M_3}^2} = R_{D2}^2 \frac{1}{|1 + s/\omega_{out}|^4}\,. \tag{3.58}$$

The spectrum of the 5 components are shown in Fig.3.21. The cascode branch devices R_{D2}, M_1 and M_3 reveal the shapes of noise spectrum as their counterparts in a common-gate TIA. It can be clearly shown that R_D2 and M_3 present the same amount of noise as compared with a common-gate TIA (Fig.3.16), and M1(cascode device) contributes $g_{m2}^2 R_{D1}^2$ times more noise. In addition, both R_{D1} and M_2 present hill-shape noise spectrum. RGC TIAs inevitably present more noise than simple common gate TIAs. A careful simulation is therefore mandatory to optimize the performance of gain, noise and power consumption.

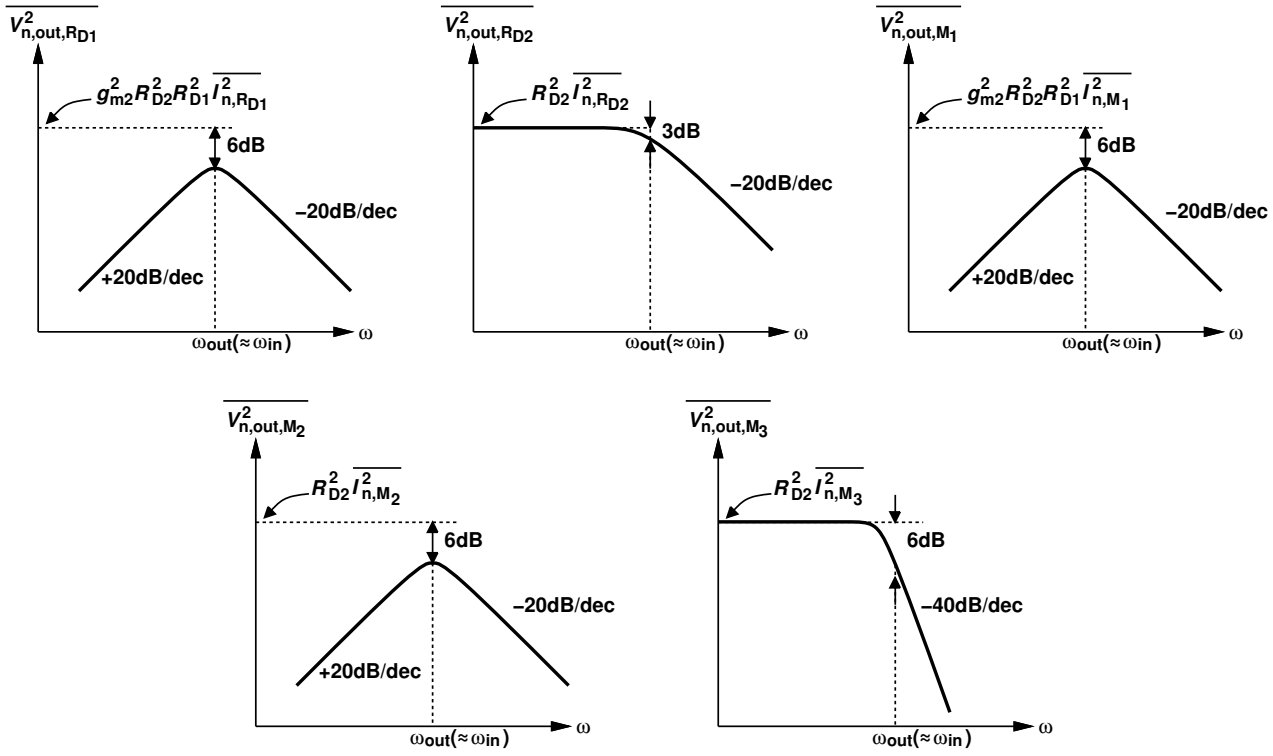


Fig. 3.21 RGC TIA noise componemts.

The RGC TIA introduced in Fig.3.18 suffers from voltage headroom issue. Letting all active devices in saturation, we must have supply higher than the lower bound:

$$V_{DD,min} = V_{GS1} + V_{GS2} - V_{TH} + I_b R_{D2}\,. \tag{3.59}$$

To ensure high-speed operation, the active device in Fig.3.18 must be biased with sufficient overdrive. As a result, it is difficult to accommodate a conventional RGC TIA into a 1.2-V supply.

A modified version of RGC TIA can relax the voltage headroom issue. As depicted in Fig.3.22, an additional stage M_2 is inserted between M_1 and M_3 stages. Adapting the input current and converting it into voltage by another common-gate structure, M_2 prevents the input from being connected to a common-source directly. In other words, the required voltage headroom is reduced. The minimum acceptable supply now becomes

$$V_{DD,min} = V_{DS4} + V_{GS1} - V_{TH} + I_b R_{D1} \,. \tag{3.60}$$

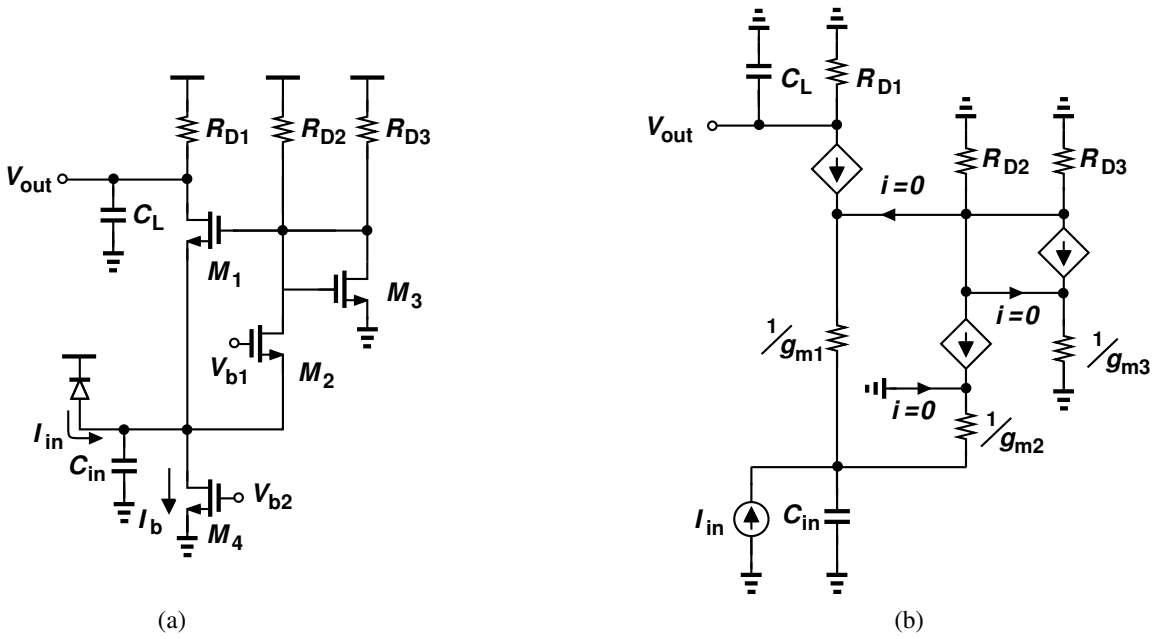


Fig. 3.22 (a)Low-supply RGC TIA, (b)its small signal model.

Saving several hundred mV of headroom. Note that M_4 serves as a current source, which might be replaced by a simple resistor.

The circuit in Fig.3.22 preserves similar characteristic of conventional RGC TIAs. We redraw the small-signal model in Fig.22(b). Neglecting the effect of capacitors, we obtain the low-frequency gain as

$$\begin{aligned} R_{Tran,DC} &= \frac{g_{m1}R_{D1}(1+g_{m2}R_{D2}g_{m3}R_{D3})}{(g_{m1}+g_{m2}+g_{m1}g_{m2}R_{D2}g_{m3}R_{D3})} \\ &\cong R_{D1}, \end{aligned} \tag{3.61}$$

the approximation holds as $g_{m2}R_{D2}g_{m3}R_{D3} \gg 1$. The input resistance is also readily available

$$R_{in} = (g_{m1}+g_{m2}+g_{m1}g_{m2}R_{D2}g_{m3}R_{D3})^{-1}, \tag{3.62}$$

which is approximately equal to $(g_{m1}+g_{m2}+g_{m1}g_{m2}R_{D2}g_{m3}R_{D3})^{-1}$. The additional stage brings down the input resistance even further. Using the same small-signal model, we investigate the poles associated with input and output. It can be shown that the equivalent resistance seen looking into the output port is equal to R_{D1}. Thus, the two poles are

$$\omega_{in} = (R_{in}C_{in})^{-1} \tag{3.63}$$

$$\omega_{out} = (R_{D1}C_L)^{-1}, \tag{3.64}$$

similar to those of a conventional RGC TIA. The reader can demonstrate that the circuit in Fig.3.22 exhibits more noise as compared with that in Fig.3.18. Overall speaking, the modified version improves the voltage headroom and bandwidth at a cost of higher noise and power consumption.

3.5 LIMITING AMPLIFIER FUNDAMENTALS

Limiting amplifiers with large bandwidth have been used extensively in various wireline systems. We study the fundamental techniques of limiting amplifiers in this section.

3.5.1 Bandwidth Extension

A limiting amplifier in modern technologies must provide voltage gain of at least 30dB with tens of GHz bandwidth. Almost all broadband amplifiers rely on the cascading technique to achieve

wide bandwidth with reasonable gain. Shown in Fig.3.23 is a general illustration, where n stages of identical amplifiers (could be as simple as a differential pair) line up as a cascade structure. Assuming each stage has a dc gain of A_0(= $g_{m1,2}R_D$) and a pole of ω_o[= $(R_DC_L)^{-1}$], we obtain the overall transfer function as

$$H(s) = \frac{V_{out}}{V_{in}} = [\frac{A_0}{1 + s/\omega_o}]^n. \tag{3.65}$$

Meanwhile, the overall bandwidth is given by

$$\omega_{-3dB} = \sqrt{2^{1/n} - 1} \cdot \omega_o \,. \tag{3.66}$$

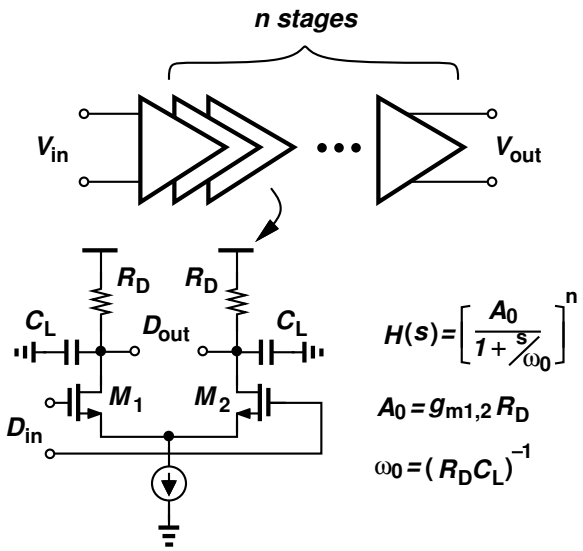
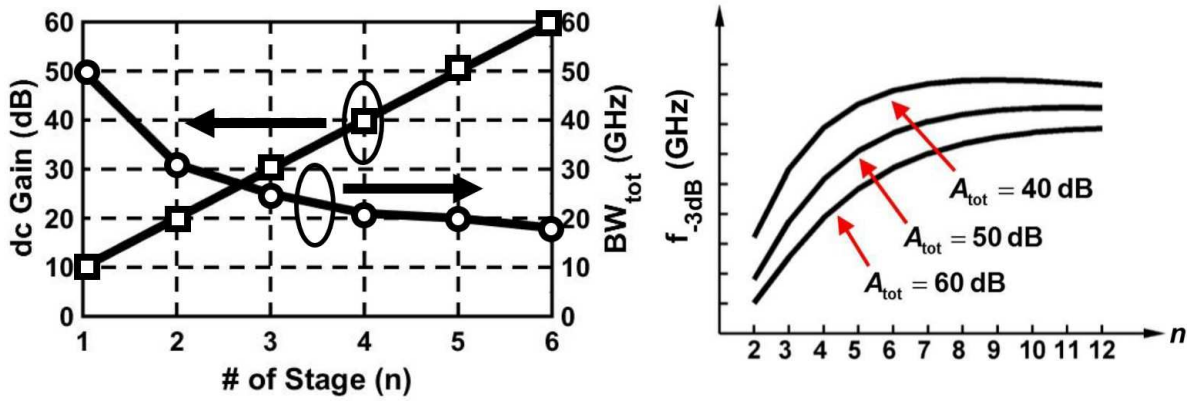


Fig. 3.23 General limiting amplifier architecture.

The secret behind it is that, as n increases, the overall gain accumulates in a way faster than the rate overall bandwidth decreases. Figure 3.24(a) reveals an example, where A_0 = 3.16 (=10dB) and ω_o = $2\pi\times$50 GHz. It can be easily shown that for n = 5, the gain becomes 5 times larger (i.e., 50dB) whereas the $-$3-dB bandwidth only drops from 50 to 19.3 GHz. In other words, we trade power dissipation for gain and bandwidth. For a first-order amplifier stage, the gain-bandwidth product (or equivalent, unity-gain bandwidth ω_{GBW}) is relatively constant for a given



(a) (b)

Fig. 3.24 (a)Gain and bandwidth variation for different n, (b)−3-dB bandwidth for GBW=185 GHz.

technology. Indeed, since $\omega_{GBW} = A_0\omega_o = g_{m1,2}/C_L$ transit frequency is C_L is primarily composed of the gate capacitance of the nest stage. Defining the required dc gain as A_{tot},
we arrive at

$$\omega_{-3dB} = \omega_{GBW} \cdot \frac{\sqrt{2^{1/n}-1}}{A_{tot}^{1/n}}. \tag{3.67}$$

For the same condition as Fig.3.24(a) (i.e., $\omega_{GBW} = 2\pi\times 158$ GHz), we plot the -3-dB bandwidth for different A_{tot} as a number of n. it approaches the maximum bandwidth as n increases. However, we limit the maximum number of stages in order to save power and reduce noise [XX].

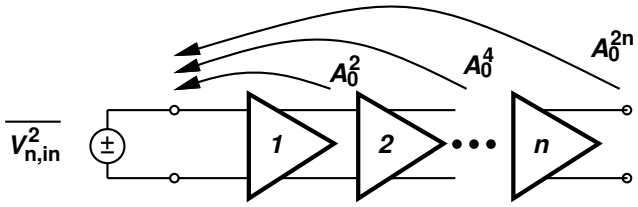


Fig. 3.25 Calculating input-referred noise of LA.

To determine n specifically, let us consider the noise performance of a limiting amplifier. Each stage contributes a voltage noise power at its output as

$$\overline{V_{n,out,tot}^2} = 2R_D^2(\overline{I_{n,M_1}^2} + \overline{I_{n,R_D}^2}) \cdot BW_n \,, \tag{3.68}$$

where BW_n denotes the equivalent noise bandwidth. The input-referred noise becomes

$$\overline{V_{n,in}^2} = 2R_D^2(\overline{I_{n,M_1}^2} + \overline{I_{n,R_D}^2}) \cdot BW_n \cdot \sum_{i=1}^{n} A_0^{-2i} \,. \tag{3.69}$$

Since all the stages are identical, one stage contributes only 1/10 of noise power as compared with the precedent one if A_0 = 10 dB. It shows the importance of sufficient gain for each stage. Meanwhile, since the tail current of a CML amplifier is relatively constant, the number of stage should be limited to around 5.

3.5.2 Tapered LAs

In many applications, a limiting amplifier may need to drive heavy loading in the output, e.g., 50Ω loading, significant around of capacitance and so on. Like clock buffer chains driving large capacitors in digital circuit, a tapered structure would be suitable here as well. The key point is to achieve a bandwidth as wide as possible with a given power budget. We look at the following example first.

Example 3.6

Consider a two-stage amplifier shown in Fig.3.26, where both amplifying units are in first-order response. The dc-gain and corner frequency are $(A_0,\ \omega_o)$ and $(A_0/\alpha,\ \alpha\omega_o)$, respectively. Determine α that maximize the total -3-dB bandwidth for a given A_{tot}.

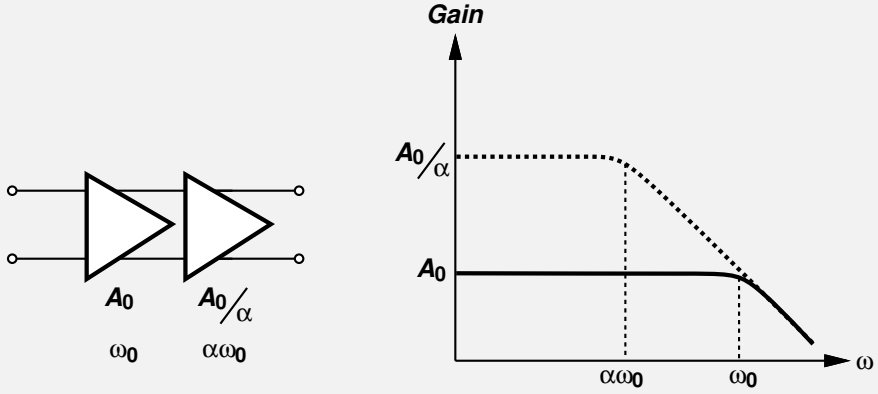


Fig. 3.26 Calculating optical sizing factor of stages.

Solution:

The transfer function is given by

$$H(s) = \frac{A_0^2/\alpha}{(1+\frac{s}{\omega_o})(1+\frac{s}{\alpha\omega_o})},$$

where $A_0^2/\alpha = A_{tot}$. The -3-dB bandwidth is calculated as

$$(1+\frac{\omega_{-3dB}^2}{\omega_o^2})(1+\frac{\omega_{-3dB}^2}{\alpha^2\omega_o^2}) = 2.$$

Since $A_0\omega_o = \omega_{GBW}$, we have

$$\omega_{-3dB}^4 + (\alpha^2+1)\frac{\omega_{GBW}^2}{\alpha A_{tot}} \cdot \omega_{-3dB}^2 - \frac{\omega_{GBW}^4}{A_{tot}^2} = 0.$$

Taking $\partial\omega_{-3dB}/\partial\alpha$, we obtain

$$\alpha = 1.$$

That is, stage in cascade amplifiers are preferable to have equal gain and corner-frequency if we want to optimize the overall bandwidth.

The foregoing example reveals the fact that the best strategy for tapered limiting amplifiers is to balance the gain and bandwidth for each stage. Figure 3.27 illustrates a 5-stage design example based on this principle. Suppose the final capacitor to drive is C5 and with a factor k, we get

$$k = \sqrt[4]{\frac{C_5}{C_1}}\,. \tag{3.70}$$

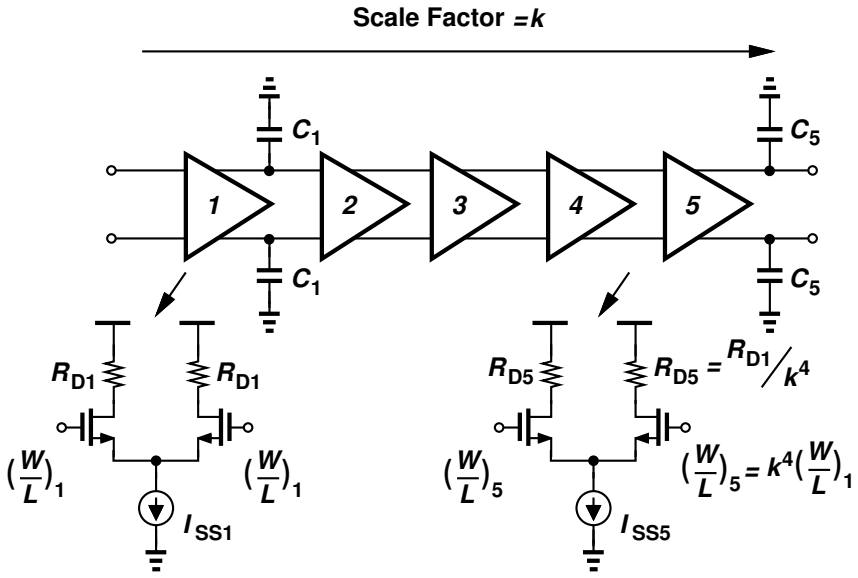


Fig. 3.27 Optical scaling of a 5-stage limiting amplifier driving heavy-loading.

Denoting the loading resistor, device size and tail current of stage 1 as R_{D1}, $(\frac{\mathrm{W}}{\mathrm{L}})_1$ and I_{SS1}, respectively, we can arrange the sizing from stage 1 to 5 as

$$\text{Loading Resistor} = R_{D1}, R_{D1}/k, \cdots, R_{D1}/k^4 \tag{3.71}$$

$$\text{Device Size} = (\frac{W}{L})_1, k(\frac{W}{L})_1, \cdots, k^4(\frac{W}{L})_1 \tag{3.72}$$

$$\text{Tail current} = I_{SS1}, kI_{SS1}, \cdots, k^4 I_{SS1}\,. \tag{3.73}$$

Since the resistor scales down by the same factor as capacitor scales up, each stage maintains identical corner frequency. Similarly, the same gain is achieved for all stages as $g_{mi} \propto \sqrt{(\frac{W}{L})_i I_i}$. The reader can see the IR drop (or common-mode level) is a constant, too.

3.5.3 Offset Cancellation

Just like other high gain amplifiers, a limiting amplifier also suffers from offset issues. It gets more and more serious as data rate goes up, where advance technologies with small device size become mandatory. As we can see in chapter 1, a typical differential pair would present input-referred offset as large as tens of mV. The input data (from TIA) would be buried if its magnitude is less than the input-referred offset of the LA.

A remarkable way to remove the offset is to adopt a (negative) feedback loop around the amplifier. By proper setting, the feedback would neutralize the imbalance by means of the high loop gain. Figure 3.28 depicts such a technique. The n-stage main amplifier is surrounded by a feedback, which distills the output offset by a low-pass filter with using low corner frequency. Here, we take n=5 as an example. The main amplifier presents an open loop gain $A = A_{tot}/(1 + s/\omega_o)^5$, whose -3-dB bandwidth is ω_{-3dB}. The subtraction between input and feedback signals is accomplished in current mode. To investigate how much offset can be reduced, we define the input-referred offset of the (open-loop) main amplifier as $V_{OS,in}$. (In open-loop mode, the output offset is $A_{tot} \cdot V_{OS,in}$.)

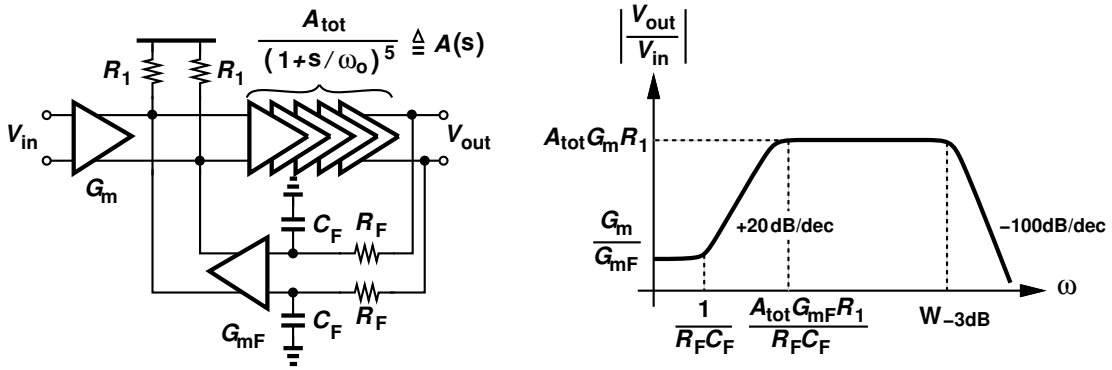


Fig. 3.28 Offset cancelation technique using feedback.

Now we close the loop and check the output offset. By setting V_{in}=0 and putting an imaginary $V_{OS,in}$ at the input of main amplifier, we obtain

$$(V_{OS,in} - V_{OS,out} G_{mF} R_1) A_{tot} = V_{OS,out}\,, \tag{3.74}$$

where $V_{OS,out}$ denotes the output offset in closed-loop mode. It turns out

$$V_{OS,out} \cong \frac{V_{OS,in}}{G_{mF}R_1} \approx V_{OS,in}\,. \tag{3.75}$$

Here the offset associated with G_m,$G_{mF}$and R_1 is neglected. Note that G_m and G_{mF} are simply for V/I conversion. It is fair to assume $G_{m1}R_1 \approx 1$ and $G_{mF}R_1 \approx 1$. As will be shown below, the closed-loop amplifier still presents a midband gain of approximately A_{tot}. We thus obtain the new input-referred noise $V^*_{OS,in}$ under closed-loop condition as

$$V^*_{OS,in} \cong \frac{V_{OS,in}}{A_{tot}}\,. \tag{3.76}$$

In other words, the loop reduces the input-referred offset by a factor of midband gain.

What does the closed-loop transfer function look like? Considering the loop, the reader can easily show that

$$\frac{V_{out}}{V_{in}} = \frac{AG_mR_1(1+sR_FC_F)}{1+AG_mR_1+sR_FC_F} \tag{3.77}$$

At low frequencies, $A = A_{tot}$, we arrive at

$$\frac{V_{out}}{V_{in}} \approx \frac{G_m}{G_{mF}} = \frac{1+sR_FC_F}{1+sR_FC_F/(A_{tot}G_{mF}R_1)}\,. \tag{3.78}$$

That means the gain begins to climb up at $\omega = (R_FC_F)^{-1}$ at a rate of $+20$ dB/dec and saturates to $A_{tot}G_mR_1$ at $\omega = AG_{mF}R_1/(R_FC_F)$. At high frequencies, $A = A_{tot}/(1+s/\omega_o)^5$. We therefore obtain

$$\frac{V_{out}}{V_{in}} \approx G_mR_1 \cdot \frac{A_{tot}}{(1+s/\omega_o)^5}\,. \tag{3.79}$$

Which rolls off at rate of -100 dB/dec beyond $\omega = \omega_{-3dB}$. The offset cancelation loop does not affect the high-frequency response. Figure 3.28 depicts the transfer function for n = 5.

Now that we understand the operation principle of limiting amplifiers, the remaining issue is to design broadband gain stages. We introduce bandwidth extension techniques in the next section.

3.6 BROADBAND TECHNIQUES

3.6.1 Inductive Peaking

Perhaps the most powerful broadband technique is inductive peaking. As we describe in chapter xx, adding an inductor can substantially improve the bandwidth. To be more specific, we redraw the equivalent circuit in Fig.3.29(a) and define $\omega \triangleq (R_D C_L)^{-1}$. The transfer function now becomes

$$\frac{V_{out}}{V_{in}} = -g_m R_D \frac{\omega_n Q s + \omega_n^2}{s^2 + \frac{\omega_n}{Q} s + \omega_n^2}\,, \tag{3.80}$$

where

$$\omega_n^2 = \frac{1}{L_P C_L} \tag{3.81}$$

$$Q = \frac{1}{R_D}\sqrt{\frac{L_P}{C_L}}\,. \tag{3.82}$$

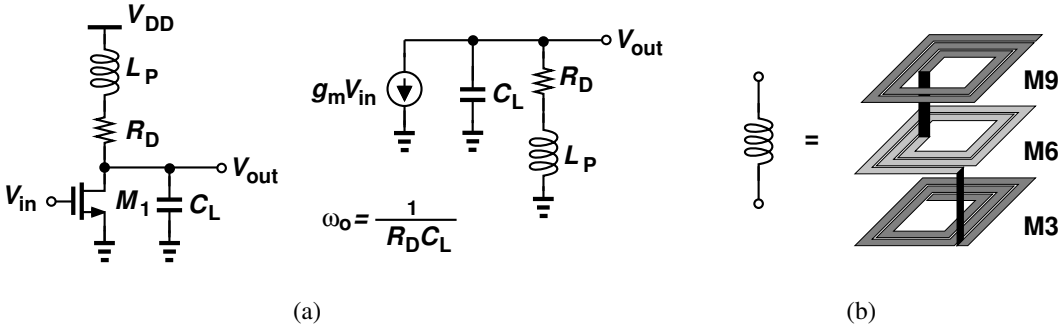


Fig. 3.29 (a)Inductive peaking technique, (b)multiple-layer inductor.

The first-order term in the number makes the analysis complicate. Instead of hand calculation, we plot the bandwidth extension and peaking effect as a function of Q in Fig.3.30.Generally speaking, significant ringing would begin to appear in the output data eye as the LA presents a peaking exceeding 5 dB. That is, for a 5-stages LA, each stage can allow only 1dB of peaking, which corresponds to XX-times bandwidth improvement. Actual design would present less enhancement in bandwidth due to the parasitic capacitance of the inductor itself. To minimize the

area occupied by the peaking inductor, stacked spirals can be used here. Putting N identical spirals on top of each other creates N^2 times inductance in theory. For example, a 0.5 nH 3-layer stacked inductor as shown in Fig.3.29(b) occupies only $14\times14\mu m^2$. Note that the inductor's quality factor is not an issue here, as it has to be put in series with a physical resistor R_D anyway.

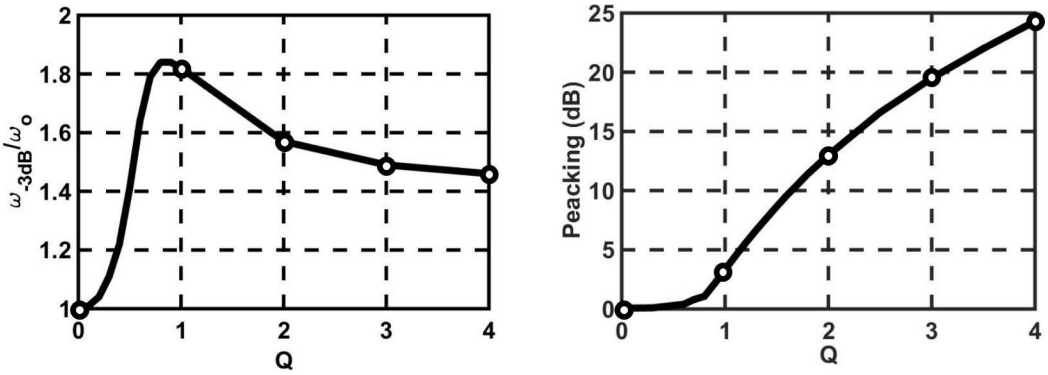
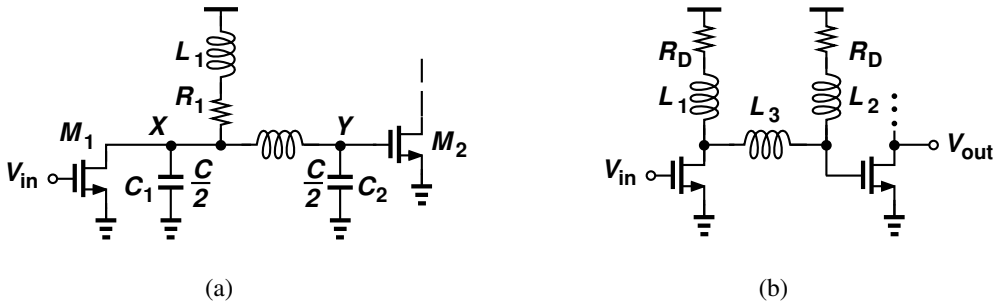


Fig. 3.30 Inductive peaking performance as a function of Q.

Inductors could be places in both series and parallel directions as the loading capacitance. Figure 3.31(a) illustrates an example, where the second inductor L_2 is inserted between stages [xx]. Since the loading capacitance has been split into two portions, L_2 creates a second resonance network to further extend the bandwidth.



(a) (b)

Fig. 3.31 Multiple-resonance peaking.

Figure 3.31(b) reveals another example [xx]. Not only does an inter-stage inductor L_3 split the parasitic, but both ends of it are terminated with inductive peak. Similar approach can be found in

[XX], where the peaking inductors are split in series. While looking attractive, the multiple resonance technique must be used conservatively. For example, if we choose $L_2 = 2L_1$ in Fig.3.31(a), a peaking of 1.8 dB appears in transfer function of a single stage. Cascading 5 identical stages would be tough as the ringing effect becomes significant. The peaking per stage in [XX] would be as high as 3dB if the parasitic capacitance are evenly split.

3.6.2 Cherry-Hooper Amplifiers

It is well known that a shunt-shunt feedback presents greater bandwidth by reducing both the input and output resistance. The feedback TIA is an example. If a trans-admittance device is coupled with the trans-impedance amplifier, we arrive at a voltage-in voltage-out amplifier which still preserves the voltage gain as

$$\frac{V_{out}}{V_{in}} = g_{m1}R_F - \frac{g_{m1}}{g_{m2}}\,. \tag{3.83}$$

It is not difficult to realize a reasonable gain (say, 10dB per stage). The key point here is that both node X and Y reveal a low equivalent resistance ($\approx 1/g_{m2}$) when looking into it. The poles associated with these nodes are

$$\omega_X \approx g_{m1}/C_X \tag{3.84}$$

$$\omega_Y \approx g_{m2}/C_Y\,, \tag{3.85}$$

where C_X and C_Y denote the parasitic capacitances. Obviously such a combination has potential to achieve high bandwidth. Indeed, quite a few bipolar broadband amplifiers over the past decades were realized in Cherry-Hooper topology. A typical structure is illustrated in Fig.3.33(a), where emitter followers Q_3 and Q_4 provide feedback paths and output ports simultaneously. The structure, however, requires large voltage headroom, i.e., tail current + V_{BE1} + V_{BE5} + IR drop for R_C and R_F. Although good performance can be achieved in bipolar devices [xx] [xx], a CMOS realization would be extremely difficult. To realize a Cherry-Hooper limiting amplifier in CMOS, we first need to make the circuit in Fig.3.32 a differential structure, i.e., adding the other half and placing tail currents at bottom. The current sources on top must be removed, as they introduce significant capacitance and mandate common-mode feedback. As a result, using resistive loads (with

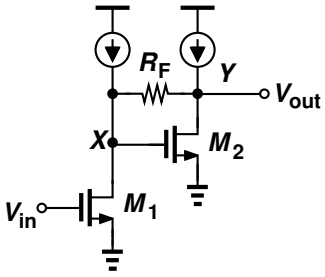


Fig. 3.32 Coupling trans-impedance and trans-admittance stages.

inductive peaking perhaps) becomes the only applicable solution. The gain would be degraded to some extent, as expected. Other than the gain issue, such a CMOS topology still suffers from high voltage headroom and output swing issue. We study the following example for more details.

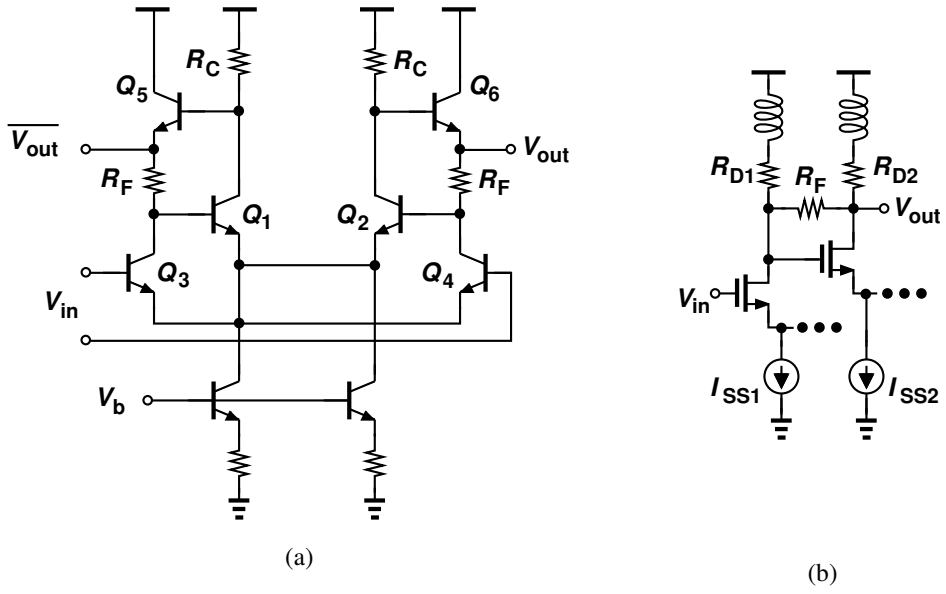


Fig. 3.33 Cherry-Hooper amplifier (a)bipolar, (b)CMOS.

Example 3.7

Consider a CMOS cherry-Hooper amplifier stage shown in Fig.3.34(a). For simplicity we have R_D as loading resistor for all arms and R_D = kR_F. The two tail currents are identical (i.e., I_{SS}). (a) Calculate the voltage gain. (b) Determine the saturated output swing.

Example 3.7 ***(Continued)***

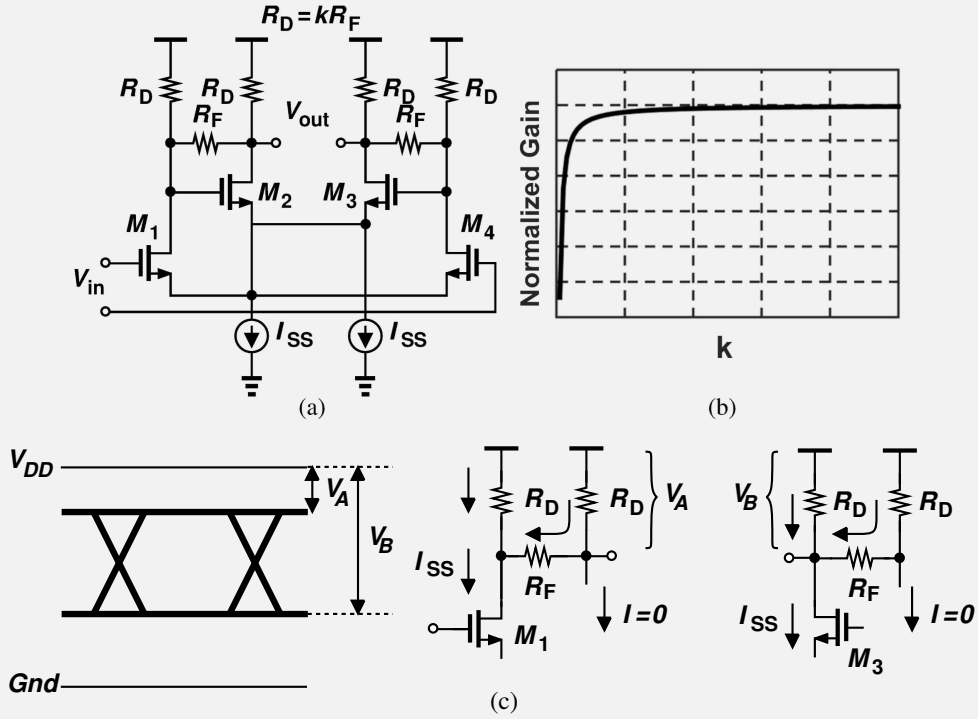


Fig. 3.34 Cherry Hooper amplifier with resistive loads:(a)circuit, (b)gain degradation as a function of k, (c)output data level calculation.

Solution:

using small-signal analysis we obtain the gain as

$$\frac{V_{out}}{V_{in}} = \frac{g_{m1}R_F(1 - g_{m2}R_F)}{1 - g_{m2}R_F - (k+1)^2/k^2},$$

which degenerates to $g_{m1}R_F - g_{m1}/g_{m2}$ as k approaches infinity. The gain degradation as a function of k is depicted in Fig.3.34(b)

Unlike a typical CML gain stage, the output level of Cherry-Hooper amplifiers would not simply locate from V_{DD} to V_{DD}-IR. In saturation, both differential pairs $M_{1,4}$ and $M_{2,3}$ are tilted completely. As illustrated in Fig.3.34(c), the higher and lower levels are V_A and V_B away from V_{DD}. For instance, the extreme levels are obtained by flowing no current through M_2 and M_4, but all I_{SS} through M_1 and M_3. That is, current flowing through R_F are from right to left in both side.

Example 3.7 ***(Continued)***

Thus, V_A and V_B are readily available:

$$V_A = \frac{k^2}{2k+1} I_{SS} \cdot R_F$$
$$V_B = \frac{k(k+1)}{2k+1} I_{SS} \cdot R_F \,.$$

The final output data differential swing is therefore given by

$$V_{PP} = 2(V_B - V_A) = \frac{2k}{2k+1} I_{SS} \cdot R_F \,.$$

The reader can imagine the situation would become much more complicated if loading resistors and tail currents are different in the two differential pairs. The unusual output levels necessitate larger voltage headroom, as all devices must stay in saturation region under any circumstance. It is sort of challenging to implement such a topology in supply voltage as low as 1.2V.

3.6.3 Darlington Amplifiers

Darlington pair has been extensively used in various applications. An important feature is that poles of a Darlington pair are relatively high. It is naturally possible to use this structure to physicalize a limiting amplifier. Figure 35 shows a simplified version of a Darlington amplifier.

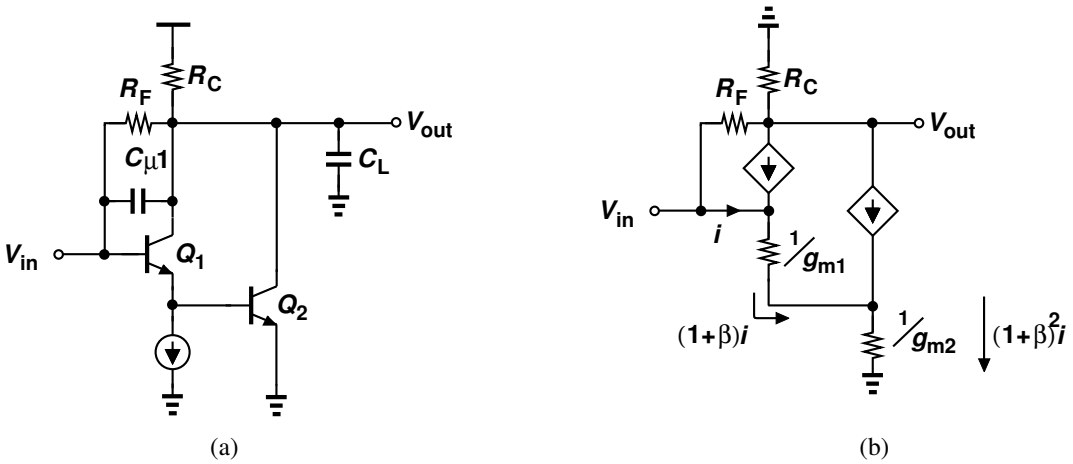


Fig. 3.35 Gain stage based on Darlington pair: (a)circuit, (b)small-signal model.

Serving as an emitter follower, Q_1 presents negligible C_π to input node as V_{BE1} is constant. It is expected to have broader bandwidth. Taking the small-signal model into calculation and neglecting the parasitic capacitances for the time being, we arrive at the voltage gain as

$$\frac{V_{out}}{V_{in}} = -g_{m2}(R_F \| R_C)\,. \tag{3.86}$$

Again, such a moderate gain is well-suitable for LAs.

The dc gain implies $C_{\mu 1}$ can be replaced with two imaginary capacitors by Miller's Effect. That is, the equivalent capacitor associated with the input node is given by $C_{\mu 1}[1 + g_{m2}(R_F \| R_C)]$. The equivalent resistance at input can be estimated by the small-signal model in Fig.3.35(b) as well. Assuming $g_{m2}(R_F \| R_C) \gg 1$, the input resistance is roughly equal to

$$R_{in} = \frac{R_F}{1 + g_{m2}(R_F \| R_C)}\,, \tag{3.87}$$

which is relatively low. The output resistance is also easy to obtain:

$$R_{out} = R_F \| R_C\,. \tag{3.88}$$

As a result, for a single-stage Darlington amplifier, the poles are approximately equal to

$$\omega_{in} = (R_F C_{\mu 1})^{-1} \tag{3.89}$$

$$\omega_{out} = [(R_F \| R_C) \cdot C_L]^{-1}\,. \tag{3.90}$$

The above analysis demonstrates that cascading several Darlington amplifier is possible to achieve wide bandwidth with reasonable gain. Note that the tail current below Q_1 could be replaced by a resistor to minimize parasitic capacitance. A possible differential realization is revealed in Fig.3.36.

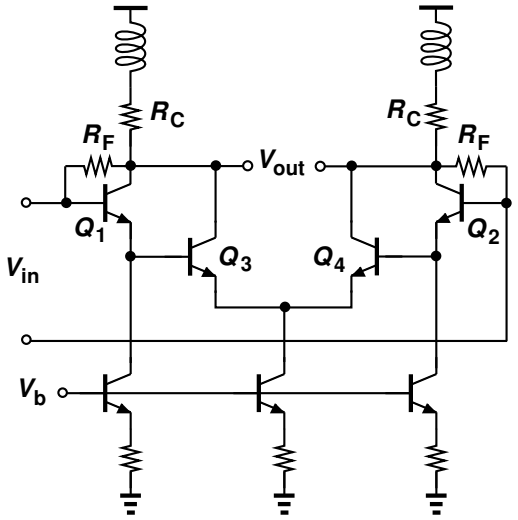


Fig. 3.36 Differential Darlington amplifier.

3.6.4 Distributed Amplifiers

HIGH-SPEED LOGICS AND CALIBRATION TECHNIQUES

Broadband data link relies on high-speed operation of mixed-mode logics. Unlike digital circuits which can significantly benefit from scaling, broadband building blocks necessitate more design techniques in architecture and circuit. Peripheral circuits for calibration and stabilization are of great importance, as the overall system performance would be highly determined by them.

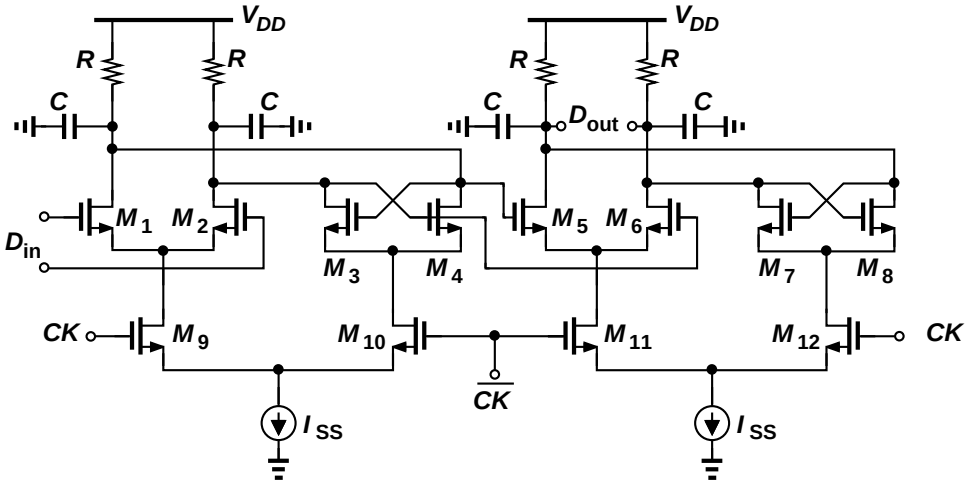
We study important circuits and techniques in this chapter, namely, flipflops (FFs), clock distribution, high-speed logic gates, multiplexers (MUXes), demultiplexers (DMUXes), and calibration skills.

4.1 FLIPFLOPS

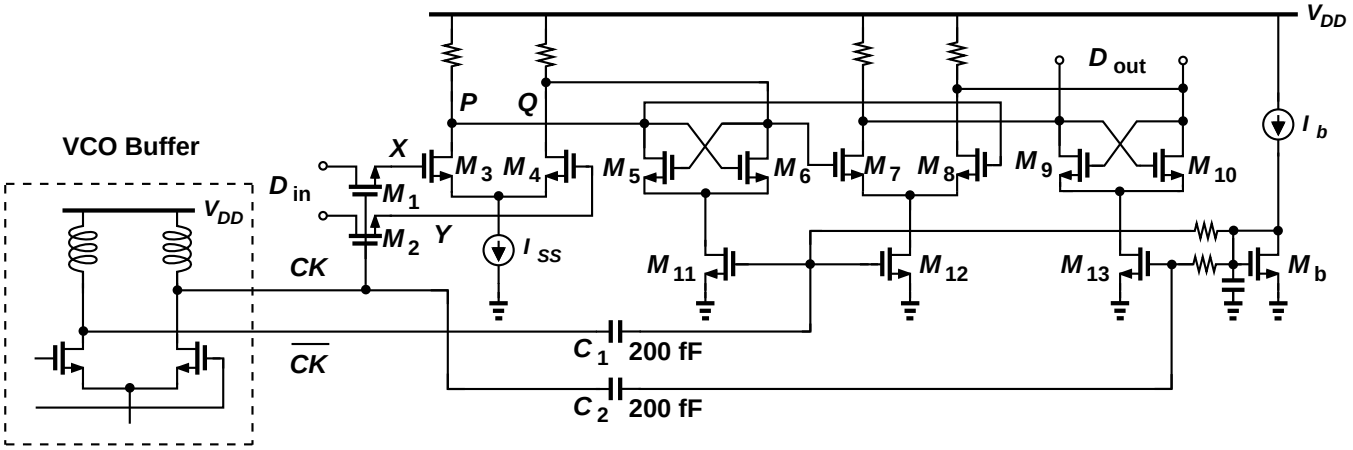
Perhaps the most commonly used block in wireline communication is the flipflop. By definition, a "flipflop" (or "D-flipflop") here means a bistable circuit, whose output states is solely determined by the rising or falling edge of the driving clock. It is usually accomplished by placing two latches (i.e., master and slave) in cascade. The most popular structure for high-speed operation is realized in CML. Figure 1(a) illustrates the circuit, where two identical latches driven by CK and $\overline{CK}$ are placed in series. Each latch has a differental pair for sampling (e.g., $M_{1,2}$), and a cross-coupled pair for regeneration (e.g., $M_{3,4}$). Starting from the falling edge of CK, the sampled data in the master latch is regenerated by the cross-coupled pair of $M_{3,4}$, and is transparently presented at the output. When CK goes high, new data is coming in while this data is preserved at the output port by the positive feedback of $M_{7,8}$. As a result, the output data updates itself once per cycle at the falling edge of CK.

Quite a few things must be paid attention to while designing a CML flipflop.First of all, in order to properly lock the data, the regeneration pair must be stronger (wider) than the sampling

pair by roughly a factor of 2 (or more). Otherwise, the data stored could be contaminated by the transition of input data. It is not difficult to check whether a FF is properly functioning: by applying a clock frequency slightly different from the data rate of D_{in}, we observe D_{out} in transient. If D_{out} always follows the falling edge of CK, the FF is properly functioning. If not, the locking behavior of it is not working and the FF degenerates to a buffer. Meanwhile, the current switched M_9-M_{12} need not stay in saturation all the time. The rule of thumb is that, as long as tail currents can be completely switched between two arms, the current switching stage (M_9-M_{12}) is doing fine.



(a)



(b)

Fig. 4.1 CML:(a) standard, (b) class-AB biasing.

The third issue is the finite sampling time. To understand more details, let us consider the small-signal model of $M_{3,4}$ pair for the regeneration mode [1]. As shown in Fig. 4.2, the output V_{out} (= $V_X - V_Y$) can be

$$g_{m3,4}R \cdot V_{out} \cdot \frac{\frac{1}{sC}}{R + \frac{1}{sC}} = V_{out}. \tag{4.1}$$

where A_O = $g_{m3,4}$R. Taking the inverse Laplace Transform,we obtain

$$V_{out} = V_{out,0} \cdot exp[(g_{m3,4}R - 1)t/RC] = V_{out,0} \cdot exp[t/\tau_0]. \tag{4.2}$$

Here, $\tau_0 \triangleq RC/(g_{m3,4}R - 1)$ and $V_{out,0}$ denotes the initial value of V_{out} at the beginning of regeneration. The two output nodes deviate from each other exponentially in the beginning of regeneration and saturate to dc levels afterwards. More derivation details can be found in [1].

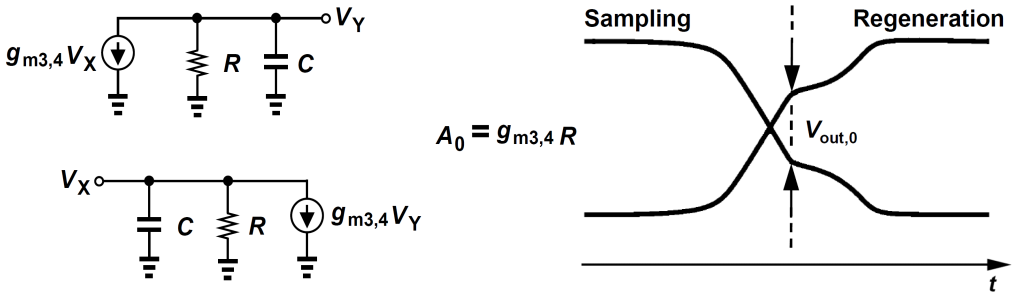
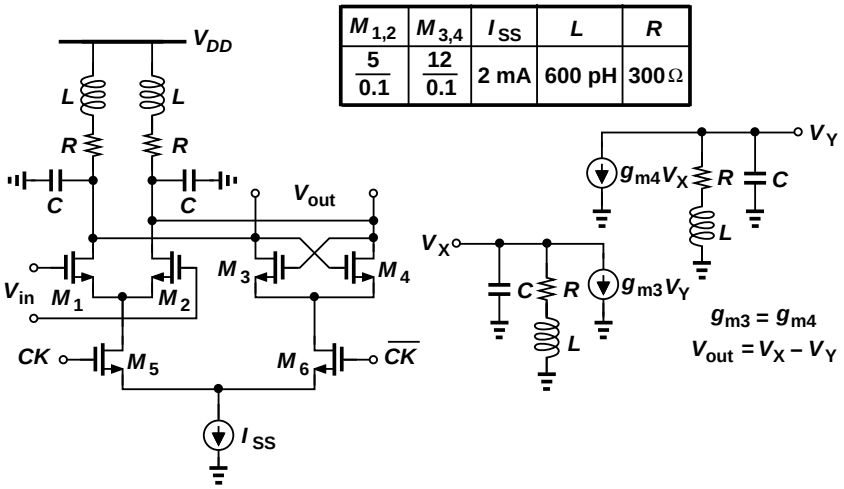


Fig. 4.2 Analysis of regeneration.

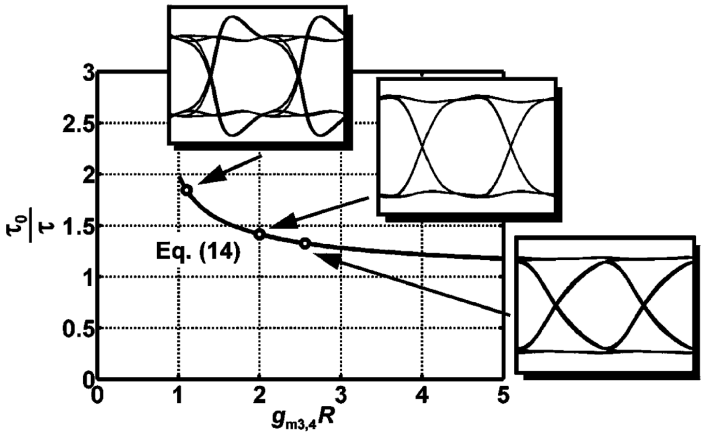
At high data rates, the loading resistor R could be reduced to less than 200 Ω in order to increase the bandwidth. That is, $g_{m3,4}R \gg 1$ may not hold any more. Meanwhile, for a given power budget, increasing $g_{m3,4}$ also means enlarging C (in a more rapid way). As a result, there exists an upper limit of operation speed.

The regeneration speed limitation can be alleviated by introducing inductive peaking into the loadings. Redrawing the latch with peaking inductor L and the equivalent model in regeneration mode [Fig. 4.3(a)], we calculate the output V_{out} (= $V_X - V_Y$) again:

$$LC\frac{\mathrm{d}^2V_{out}}{\mathrm{d}t^2} + (RC - g_{m3,4}L)\frac{dV_{out}}{dt} + (1 - g_{m3,4}R)V_{out} = 0. \tag{4.3}$$



(a)



(b)

Fig. 4.3 (a) CML latch with inductive peaking, (b) regeneration speed improvement.

For the most flat respone [i.e., Q = (1/R)$\sqrt{L/C}$ = 0.7], we obtain an explicit solution for $V_{out}(t)$, which grows up exponentially with a new time constant τ :[3]

$$\tau = \frac{2RC}{g_{m3,4}R - 2 + \sqrt{g_{m3,4}^2R^2 + 4g_{m3,4}R - 4}}. \tag{4.4}$$

As compared with τ_0, the positive-feedback process is accelerated by a factor of

$$\frac{\tau_0}{\tau} = \frac{g_{m3,4}R - 2 + \sqrt{g_{m3,4}^2R^2 + 4g_{m3,4}R - 4}}{2(g_{m3,4}R - 1)} \geq 1. \tag{4.5}$$

Note that $g_{m3,4}R$ must be greater than unity to guarantee positive feedback. Fig. 4.3(b) plots the speed improvement as a function of $g_{m3,4}R$, demonstrating that the inductive peaking unconditionally improves the regeneration. However, aggressive peaking not only risks the regeneration but leads to significant ringing on the output data. Three cases of time domain waveforms with $g_{m3,4}R$ = 1.1 and 2 are shown in the insets of Fig. 4.3(b) to illustrate such a trade-off. As a result, an improving factor of 1.4 as $g_{m3,4}R$ = 2 has been chosen as an optimal point in this design. Note that in actual design, the speed may be boosted to a lesser extent due to some other considerations such as power consumption and routing convenience. Device sizes of an design example in 90nm CMOS process are listed ni Fig. 4.3(a) as well. Note that sampling speed is also improved by using inductive peaking, which has been explained in chapter 3.

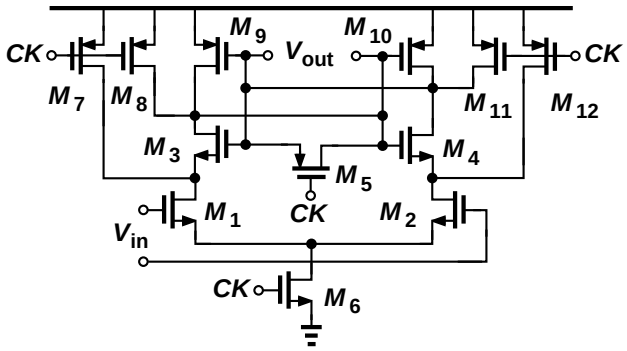
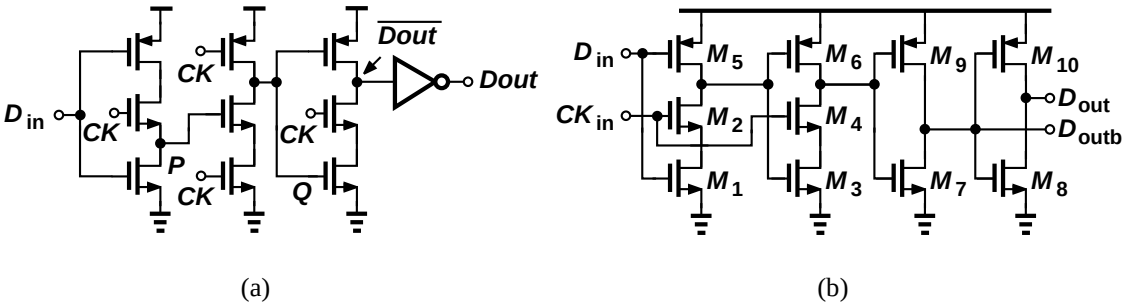


Fig. 4.4 DCVS latch.

Other than CML topology, another popular latch architecture can be found in Fig. 4.4. Known as differential cascode voltage switch (DCVS) latched, this type of circuits utilizes the techniques of some amplifiers. Driven by a single-phase, rail-to-rail clock, the circuit operates as follows. In sampling mode, the latch resets the output. As CK is low, differential pair $M_{1,2}$ senses the input data but no active current flowing through them. The back-to-back inverters (M_3, M_4, M_9, and M_{10}) are both pull up to V_{DD}. Switch M_5 is inserted here to equalize voltages at both ends more rapidly. In regeneration, CK is high, and the input static (in the last moment of sampling) immediately determine which arm carries more current. The positive feedback formed by the two inverters soon regenerates the output data to rail-to rail format. Owing to the reset, the data output

must be taken subsequently by a non-resetting latch. The heavy clock loading would be an issue if significant amount of DCVS latches are used in a system.



(a) (b)

Fig. 4.5 TSPC latch example.

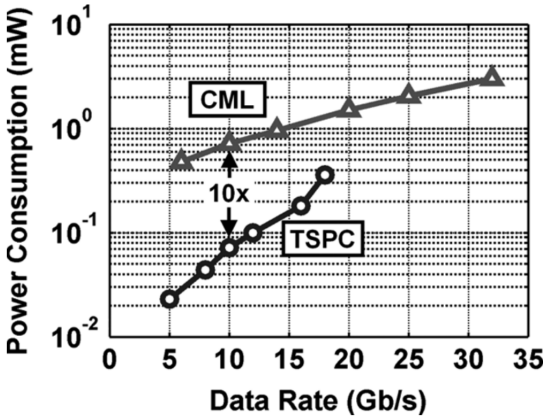


Fig. 4.6 Power consumption of CML and TSPC FFs.

The digital operation of DCVS latches benefits from low power consumption. For many applications, even single-ended data path (usually with rail-to-rail swing) is sufficient. That allows us to further simplify the latch structure. A popular logic family named true single phase clock (TSPC) provides useful latch architecture. Again, a flipflop is achieved by cascading two latched and driving them with complementary clocks. Figure 5 illustrates two possible structures for TSPC

latches. For the standard structure shown in Fig. 4.5(a), the operation can be expressed as

- $CK = 0\ (Locking),\ P = 0,\ Q = 1,\ \overline{D_{out}} = High\ Z,$
- $CK = 1\ (Sampling),\ P = \overline{D_{in}}$:

$$\Longrightarrow\ if\ D_{in} = 1,\ P = 0,\ Q = 1,\ \overline{D_{out}} = 0$$

$$\Longrightarrow\ if\ D_{in} = 0,\ P = 1,\ Q = 0,\ \overline{D_{out}} = 1.$$

Here, 4 transistors need to be clocked, creating significant loading. Figure. 5(b) reveals a modified version, where CK_{in} only needs to drive 2 devices. The reader can demonstrate a similar operation for this circuit.

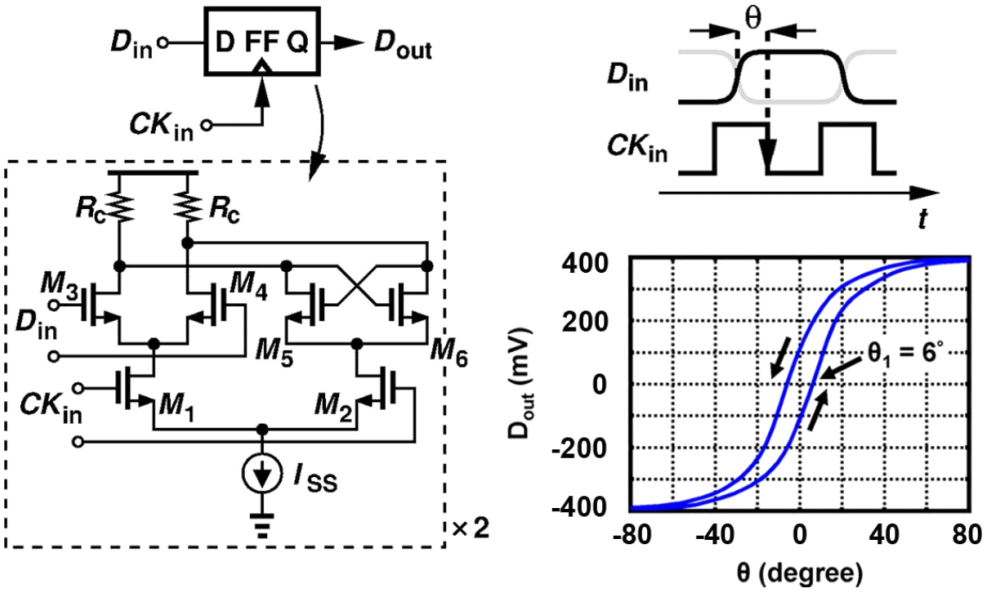


Fig. 4.7 Hysteresis sample of D-flipflop.

The power efficiency for TSPC latches is remarkable. A comparison of power consumption between TSPC and CML latches is shown in Fig. 4.6. Here, both latches are designed in 65-nm CMOS. If operated at 10 Gb/s, the TSPC dissipates only 1/10 of power. D-flipflops are key blocks in lots of systems. However, one should pay close attention to it, if a flipflop is mainly desinated to deal with the edge of input data. The finite regeneration time and limited sampling bandwidth may cause uncertainty of the results. In 65-nm CMOS technology, for example a conventional

flipflop fails to operate beyond 24 Gb/s even with CML topology. As data rate increases, a serious issue may occur if we take a flipflop as a sampler. An illustration is shown in fig. 4.7, where a conventional CML flipflop is used. Here, the flipflop is operated in full-rate mode, i.e., each data bit is sampled once. If we gradually shift the clock edge of CK_{in} to the right, the output D_{out} will not flip immediately at $\theta = 0$, but will rather stay in its original state for a finite phase difference θ_1. It is because the cross-coupled pair M_3-M_4 needs large enough initial voltage to overcome the mismatch, finite bandwidth, and limited regeneration time. Similarly, as we shift CK_{in} to the left, the flipflop takes exceeding phase($-\theta_1$) to change state. As a result, a hysteresis characteristic appears. In 65-nm CMOS, θ_1 can be as large as 6° if D_{in} has a rate of 24 Gb/s. Such a phase uncertainty prohibits the use of a simple flipflop as a phase detector. We discuss more details about phase detectors in Chapter 8.

4.2 CLOCK BUFFERS

Delivering high-speed clocks becomes increasingly difficult as data rate goes up. For a SerDes system, it is preferable to use CML format for clocks above a few GHz to keep good manners of differential circuits. Clocks in CMOS logic with full-swing (rail-to-rail) magnitude are usually used for circuits below $3 \sim 5$GHz. It is of course a rough division. Overall optimization relies on proper choosing of clock buffers.

Let us first consider a general CML buffer with inductive peaking, where the loading and tail current are optimized. Design criteria here is to maintain at least the same clock magnitude (i.e., large-signal gain ≥ 1). Fig. 4.8 shows the simulates power dissipation as a function of bandwidth for such a differential pair in 90-nm CMOS with fanout-of-4 loading. The interconnect is also taken into consideration by extracting the parasitic capacitance from layout. Drawing a best-fit curve, we conclude that a good power efficiency can be maintained only up to 15 GHz.

In other words, to drive clock at higher frequencies, we may (a) use a more advanced technology node; (b) reduce the number of fan-out; (c) modify the buffer. It is always desirable to use more advanced processes, but cost may be a concern. Besides, scaling does not provide 100% speed improvement, as the distances between layers are getting smaller. Reducing the number

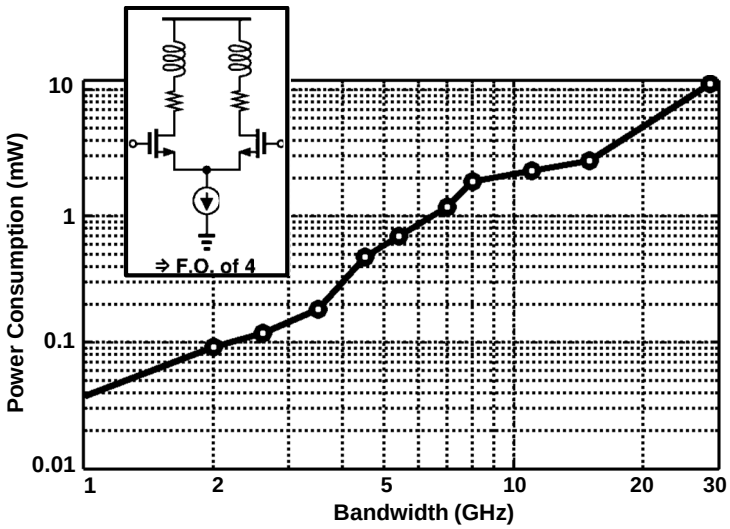
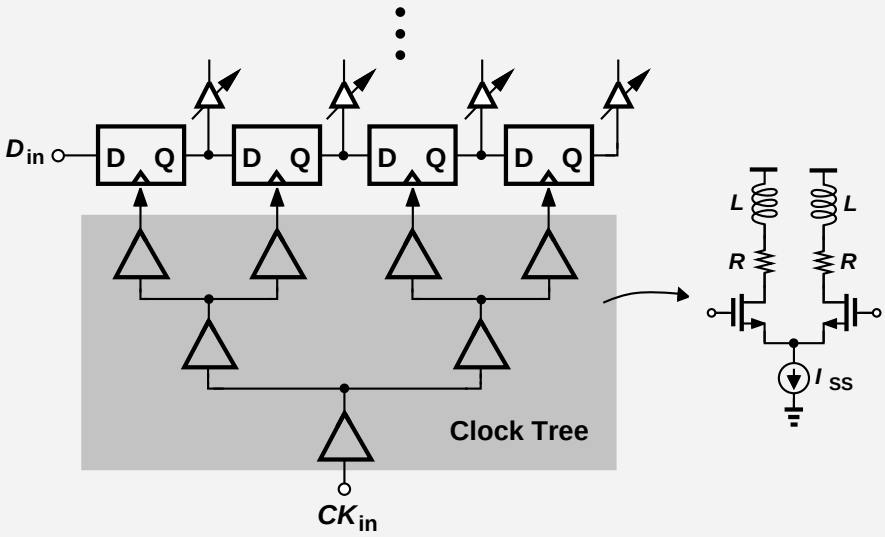


Fig. 4.8 Power efficiency of high-speed buffer in 90-nm CMOS technology.

of fan-out leads to a larger clock buffer tree, which not only consumes more power but increases layout difficulties.

Example 4.1

Consider a 40 Gb/s 4-tap feedforward equalizer shown in Fig. 4.9, which is driven by a clock buffer tree with fan-out of 2. The loading for each driver is equal to 50 fF. Estimate the total power consumption of the tree.



Example 4.1 ***(Continued)***

Fig. 4.9 Power estimation for high-speed clock tree.

Solution:

Assume the inductor can flatten the response so that the -3-dB loss at the original corner $[(RC)^{-1}]$ is diminished. Thus,

$$2\pi \cdot 40\ GHz = \frac{1}{R \cdot 50\ fF}, \tag{4.6}$$

$$R = 80\ \Omega. \tag{4.7}$$

To keep clock swings about 500mV, $I_{SS} \cong 6$ mA. The buffer tree itself consumes 42 mA from supply.

Reducing the number of fan-out helps to improve the bandwidth to some extent. However, the overall power dissipation of the buffer tree increases significantly, as more buffers are employed. We resort to circuit techniques to achieve a better performance.

As we learn form Chapter 3, inductive peaking extends the bandwidth but it also suffers from tradeoffs. Figure. 10(a) illustrates such a buffer. Generally speaking, we have Q $\approx$ 0.7 to reach a flat response, making the bandwidth approximately equal to ω_n. Similar to the small-signal analysis, the differential pair steers the tail current I completely under large input and presents a flat $|V_{out}|$ from dc to approximately the bandwidth ω_n. Note that the large signal behavior resembles the small-signal response and it can be verified by simulation. The key point is, since the filpflop and other gates require a swing of at least 500 mV, R must be 150 $\Omega \sim$ 200 Ω or larger. Otherwise I must be increased, which in turn leads to bigger device sizes and larger C. If we were to keep the optimal Q by increasing L, the bandwidth would be decreased. In other words, it is difficult to realize a larger swing output by using inductive peaking only. In fact, it is waste to keep the voltage gain all the way down to dc, because the clock buffer only operates at a narrow band of high frequencies (e.g., in the vicinity of ω_n).

Another possible approach attempting to deliver large swing is to employ pure inductive loads, which resonates out the parasitic capacitance C at the desire frequency [Fig. 4.10(b)]. In

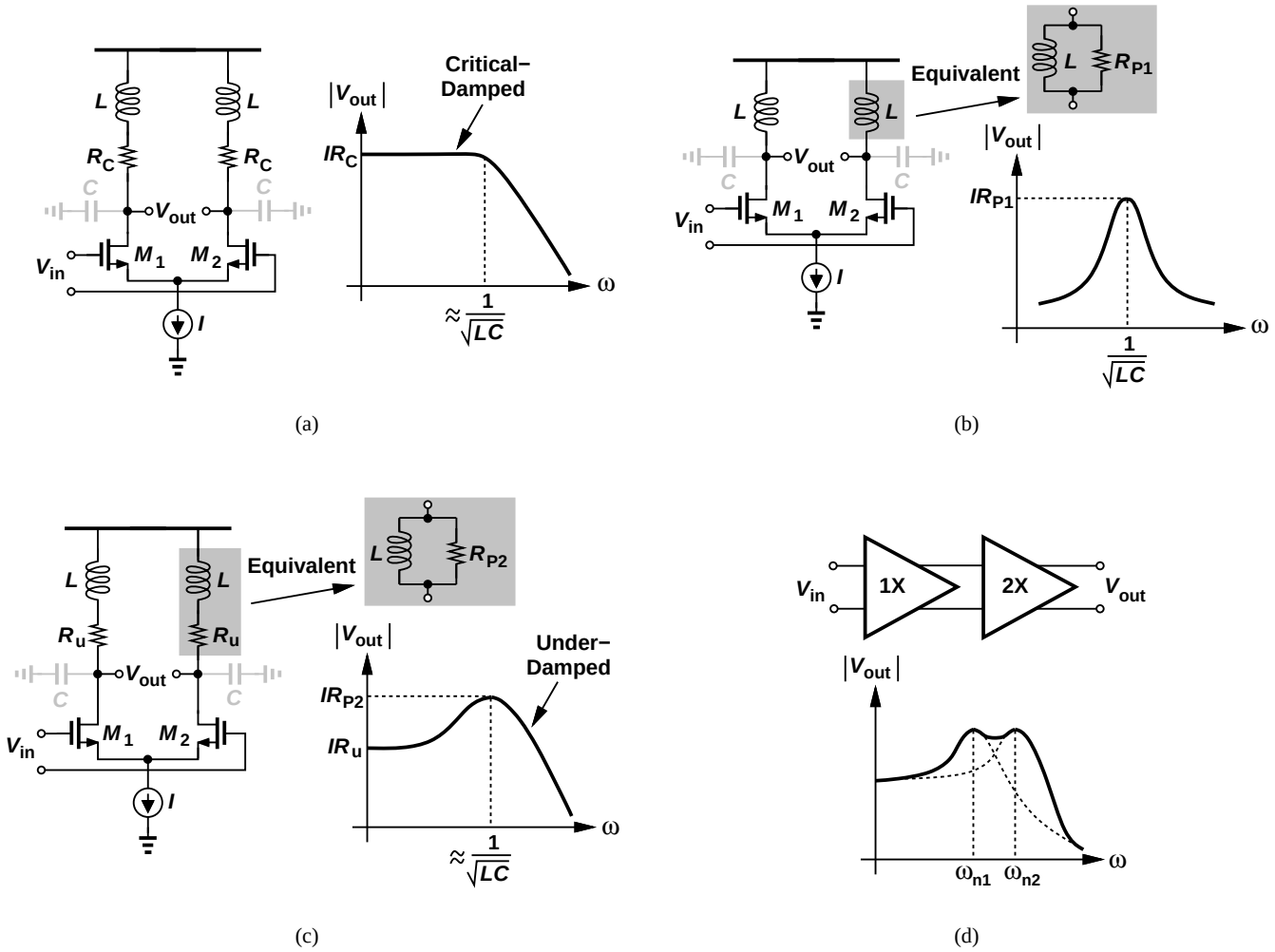


Fig. 4.10 Clock buffer with (a) critical inductive peaking, (b) pure inductive loads, and (c) underdamped peaking with realization [12].

deed, this method produces a swing of IR_{P1} in the vicinity of $1/\sqrt{LC}$, where R_{P1} represents the loss of L as an equivalent resistance in parallel. With a quality factor Q above 4, the buffer in Fig. 4.10(b) can create a large output swing easily. However, it is very challenging to precisely line up the resonance frequencies of the VCO and the buffer. For instance, if Q = 5, than a 50% magnitude degradation would occur if the two resonance frequencies deviate from each other by 17%. Practical application thus becomes very hard to implement due to PVT variations. The output swing is not predictable since the Q of the on-chip inductors is hard to control.

The above difficulties can be alleviated by introducing an underdamped peaking. As depicted in Fig. 4.10(c), we keep the loading resistors of Fig. 10(a) but reduce the value R_μ. The output swing starts with a lower value of IR_μ from dc and presents a gradual peaking of IR_{P2} at $1/\sqrt{LC}$. Here, we convert the series L-R network into an equivalent parallel combination L-R_{P2}. The difference between Fig. 4.10(b) and (c) is that the R_{P2} in Fig. 4.10(c) now becomes predictable, because the physical resistor R_μ is fully under our control. In other words, we degenerate the tuned amplifier of Fig. 4.10(b) in such a way that its peaking and bandwidth become well-behaved to accommodate the desired operation points. As compared with that in Fig. 4.10(a), this buffer allows more efficient optimization of the gain and bandwidth. For example, if we choose Q = 2, then R_{P2} = 4R_u. To be more specific, this method plays a compromising role between the resistive and inductive loadings, alleviating bandwidth limitation and providing accurate swing control. Note that this structure is totally different from the purely inductive designs such as [2] and [3].

To further increase the bandwidth, we cascade two stages with different peaking frequencies [Fig. 4.10(c)]. The split peaks enlarge the operation range significantly. For example, if a 20 GHz clock buffer is to be designed, in 90nm CMOS , we realize the peaking moderately to ensure a stable operation, i.e., the maximum peak only exceeds the dc gain by 2.7 dB. As a result, the -3-dB bandwidth of this clock buffer is about 24.6 GHz, which provides adequate margin for PVT variations. Note that the two-stage topology also achieves a good isolation for clock source, protecting it from being disturbed by the sampling flipflop and the frequency detector. A reverse isolation (S_{12}) of -74 dB is observed in simulation.

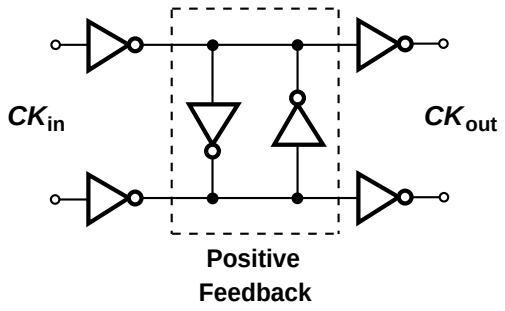
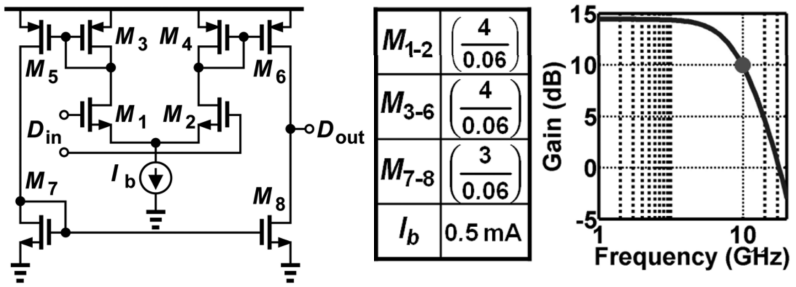
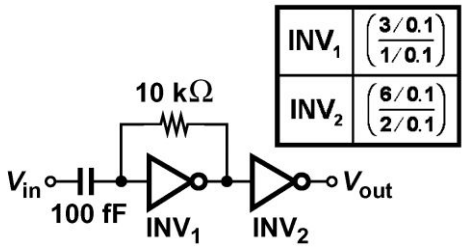


Fig. 4.11 Rail-to-Rail (CMOS) clock buffer.



(a)



(b)

Fig. 4.12 CML-to-CMOS Converters with (a) differential, (b)single-ended inputs.

Clock buffers for lower speed clocks in CMOS logic with rail-to-rail swings are relatively straightforward. CMOS inverters with proper sizing (i.e., tapered) are sufficient in most applications. If both CK and $\overline{CK}$ are delivered, deskew coupler can be introduced to ensure the two clocks are $180°$ out of phase. Figure. 11 illustrates an example, where back-to-back invers are used to couple the two outputs. Duty cycle correction is also possible to achieve by similar structure [xx].

Most systems have clocks at different frequencies, perhaps varying from tens of GHz to hundreds of MHz. A PLL is a good example. At certain point of frequency, the CML level needs to be converted into CMOS level. Figure 12 reveals two possible converters. In Fig. 4.12(a), we see a differential input converter, which contains a current steering adapter to translate the CML input into full-swing levels. Note that minimum channel length is used for all devices, which generates a conversion gain greater than 10 dB at 10 GHz with only 0.5 mA in 65 nm CMOS process. For

single-ended input, the converter in Fig. 4.12(b) can be adopted, where the inverter INV_1 is self-biased at its high-gain region. Such a structure ensures proper operation up to 10^{+}GHz for (90 nm or more advanced processes) with very low power dissipation.

4.3 MULTIPLEXERS

Multiplexing has been serving as a key function in data links for decades. Let us look at the transmitter architecture first. Fig. 4.13 illustrates a typical realization of a wireline transmitter, composing multiple ranks of 2:1 selectors and clock multiplication unit (CMU) providing the clocks. The last-stage MUX and the voltage-controlled oscillator (VCO) play critical roles simply due to the high-speed requirement. A tree structure is commonly used in high-speed transmitters, as it provides the highest bandwidth. As we know the two data inputs of a 2:1 selector must be shifted by 0.5 UI with each other in order to reach the best sampling. Each high-speed 2:1 selector has at least 5 latches in front to line up the data inputs.

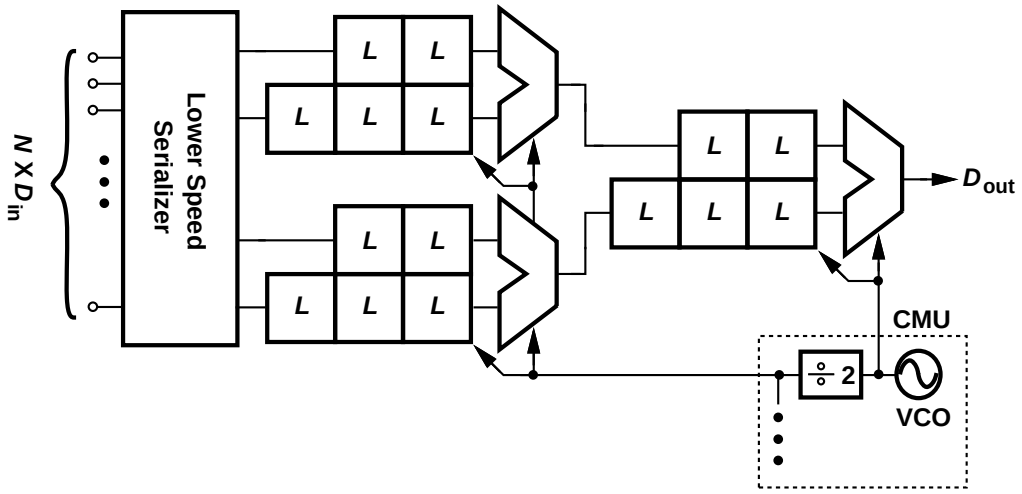
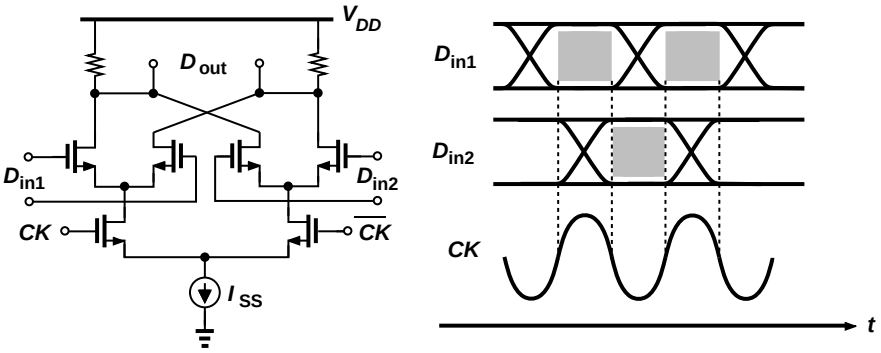


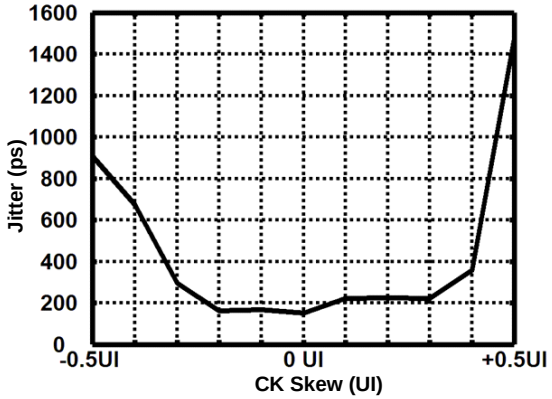
Fig. 4.13 Conventional transmitter architecture with tree-type MUXes.

One issue arises from the arrangement is that the 2:1 selector and its lineup latches are driven by the same clock. The intrinsic phase relationship between the 2 input data and the driving clock is not quite right, let along the uncertainty caused by clock-to-Q delay of the FFs and routing. At least a constant delay is needed between the two versions of clock to ensure proper phase relationship.

Figure 14(a) explains the phase requirement. With finite rese/fall times, it is desirable to place the peak of CK around the center of one data (e.g., D_{in1}) and the valley the center of the other (D_{in2}). As data rate goes up, eye opening gets smaller, making it harder to get a clean shot. Moreover, owing to the sinusoidal clock, both data paths would be turned on momentarily over a significant portion of a bit period. The output could be contaminated by the unselected path. Figure. 14(b) reveals the data jitter (peak-to-peak) of a 40 Gb/s 2:1 selector with $2^7 - 1$ PRBS inputs simulated in 40 nm CMOS technology. The output jitter increase dramatically as CK_{in} deviates from its optimal position by $\pm XXUI$.



(a)

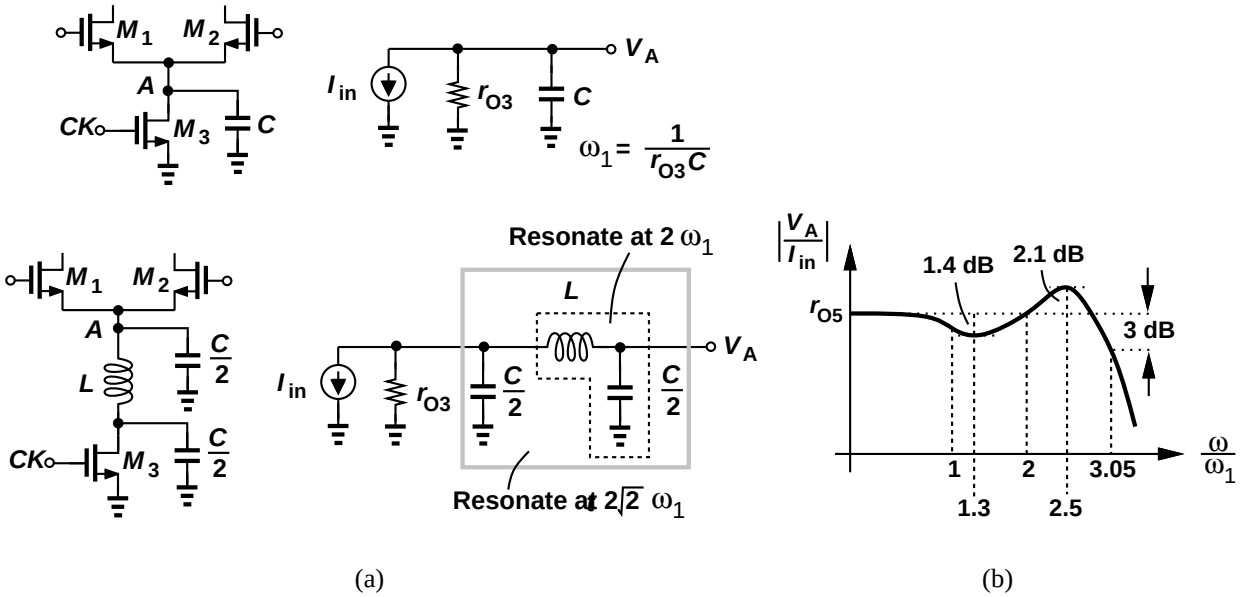


(b)

Fig. 4.14 (a) CML 2:1 selector and timing diagram, (b)jitter as a function of skew.

To speed up the operation of a 2:1 selector, we introduce inductive peaking to the output port. Interestingly, same technique can be used to accelerate the charging/discharging process at

internal node. As illustrated in Fig. 4.15(a), when the clock turns on, the parasitic capacitance C at node A must be discharged so as to lower V_A until either M_1 or M_2 is on. The -3-dB bandwidth ω_1 is thus given by $(r_{O3}C)^{-1}$, where r_{O3} denotes the output resistance of M_3. The relatively large capacitance C considerably degrades the performance at high speed.



(a) (b)

Fig. 4.15 Internal node behavior (a) small-signal model, (b) transfer function.

Now, a series inductor L is inserted between the clock and data stages as shown in Fig. 4.15(b) [4], [5], splitting C into two components [6]. Assuming the M_1-M_2 pair and M_3 contribute approximately equivalent capacitance (C/2), we choose L to resonate with C/2 at $2\omega_1$ to minimize peaking: at $\omega = 2\omega_1$, the L-C/2 network acts as a short, absorbing all of I_{in} and causing $|V_A/I_{in}|$ = $[2\omega_1(C/2)]^{-1}$ = r_{O3}; at $\omega = 2\sqrt{2}\omega_1$, the π network of C/2-L-C/2 resonates, forcing all of I_{in} to flow through r_{O3} and making $|V_A/I_{in}| = r_{O3}$. (The two capacitors in the π network carry equal and opposite currents.) Quantitative analysis reveals that

$$\left|\frac{V_A}{I_{in}}\right|(j\omega) = \frac{4r_{O3}}{\sqrt{[4-(\frac{\omega}{\omega_1})^2]^2 + [4(\frac{\omega}{\omega_1}) - \frac{1}{2}(\frac{\omega}{\omega_1})^3]^2}} \tag{4.8}$$

and the transfer function is plotted in Fig. 4.15. The peak (2.1 dB) and valley (-1.4 dB) occur at $2.5\omega_1$ and $1.3\omega_1$, respectively. The -3-dB bandwidth is approximately equal to $3.05\omega_1$. In other words, this technique extends the bandwidth associated with the internal node A by a factor of 3.

In practice, the inductor L introduces parasitic capacitance and loss, limiting the bandwidth improvement to a lesser extent. The large-signal behavior of a MUX restricts the bandwidth enhancement as well. The capacitance C may not be split evenly either. For example, if M_3 contributes $C/3$ and the M_1-M_2 pair $2C/3$ to node A, we could choose the L-$2C/3$ network to resonate at $1.5\omega_1$, arriving at a 2.3-times bandwidth improvement of the internal node with passband ripple of less than 0.2 dB. Note that unlike that of double resonance peaking in chapter 3.xx, the peaking in Fig. 4.15(b) is only associated with internal nodes and has negligible impact of output data ringing.

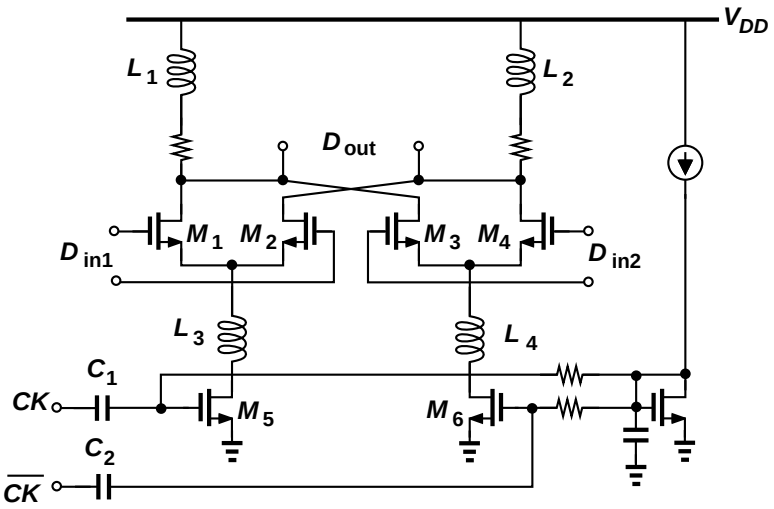


Fig. 4.16 2:1 selector with double peaking.

A modified selector design is thus depicted in Fig. 4.16, where the tail current source is eliminated to relax the voltage headroom requirement. Current switching in M_5-M_6 is accomplished by gate control or so-called "Class-AB" operation. Since the tail current source is removed, M_5-M_6 can be much narrower, presenting a smaller capacitance to the clock buffer. Such Class-AB current sources create a large peak current and provide greater voltage swings at the outputs.

In addition to the skew issue between a 2:1 selector and its preceding FFs, we ought to properly arrange the clock phase relationship at different frequencies. After all, the clock multiplication unit (i.e., a PLL) has very loose control over the timing of data paths. It is especially true for the very last stage of multiplexing at high data rate. For example, a 56-Gb/s transmitter

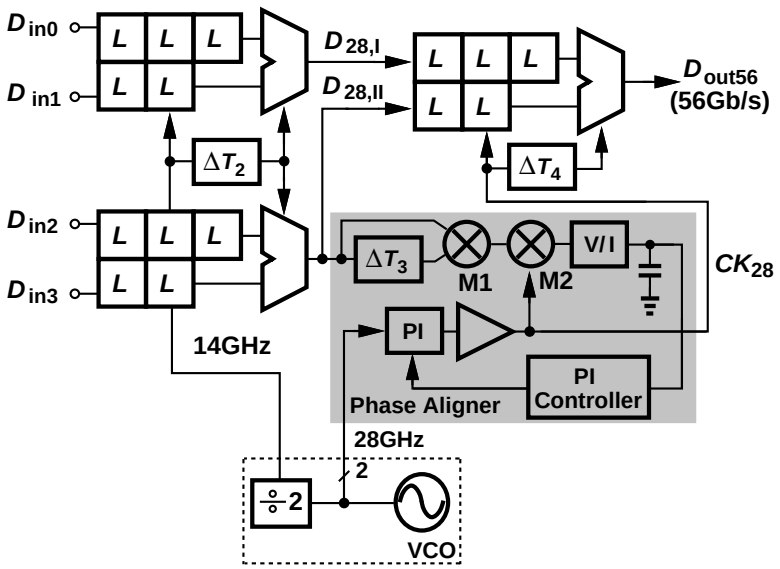


Fig. 4.17 ultra high-speed TX CMU with internal phase aligner.

design is illustrated in Fig. 4.17. In the first multiplexing stage, delays ΔT_1 and ΔT_2 are inserted to balance the sample timing. These delays are properly designed to match the internal skews over a wide temperature range. At 56 Gb/s, the phase alignment issue becomes so severe that a static delay can hardly work. For instance, the acceptable sampling window in the last stage (56-Gb/s output) is about 8 to 10 ps, but the phase drifting caused by PVT variations could be as large as 15 to 20 ps. That is, two 28-Gb/s data streams $D_{28,I}$ and $D_{28,II}$ in Fig. 4.17 created by the first multiplexing stage need to be retimed before entering the final 2:1 selector. However, the 28-Gb/s data is too fast to be sampled by a 28-GHz clock with arbitrary phase, no matter where it comes from. To accommodate random phase relationship, we put a phase aligner in front of the second multiplexing stage to dynamically track the optimal clock and data phases. The phase tracking operates as follows. First, the synchronization clock (wherever it comes from) is divided by two to generate quadrature clocks at 28 GHz. The data transition is examined by using a roughly 16.5-ps delay ΔT_3 with a mixer (M1) to detect the arrival of the internal 28-Gb/s data. With the help of the 28-GHz phase interpolator (PI) and the second mixer (M2), we arrive at a feedback loop that forces the PI to produce clock phase which aligns the data transition. To be more specific, mixer

M1 serves as a XOR gate a XOR to distill ”pulses” (actually as round as a sinusoid due to ultra high speed) upon occurrence of data transition of $D_{28,II}$. This pulse sequence gets mixed up with 28-GHz clock (after the phase interpolator) to create phase error information. Since high-frequency terms are filtered out, the phase error is presented as a cosine function and is applied to the V/I converter. The phase interpolator and its control unit therefore rotates the clock phase based on the control voltage until phase locking is accomplished. The gate delay of M_1, M_2, and clock buffer makes the falling edges of the locked clock (CK_{28}) locate right in the data eye center of $D_{28,I}$ and $D_{28,II}$, leading to perfect alignment. Finally, ΔT_4 provides the phase difference between the retiming latches and the final 56-Gb/s 2:1 selector. Note that the phase aligner here is purely linear and is unconditionally stable.

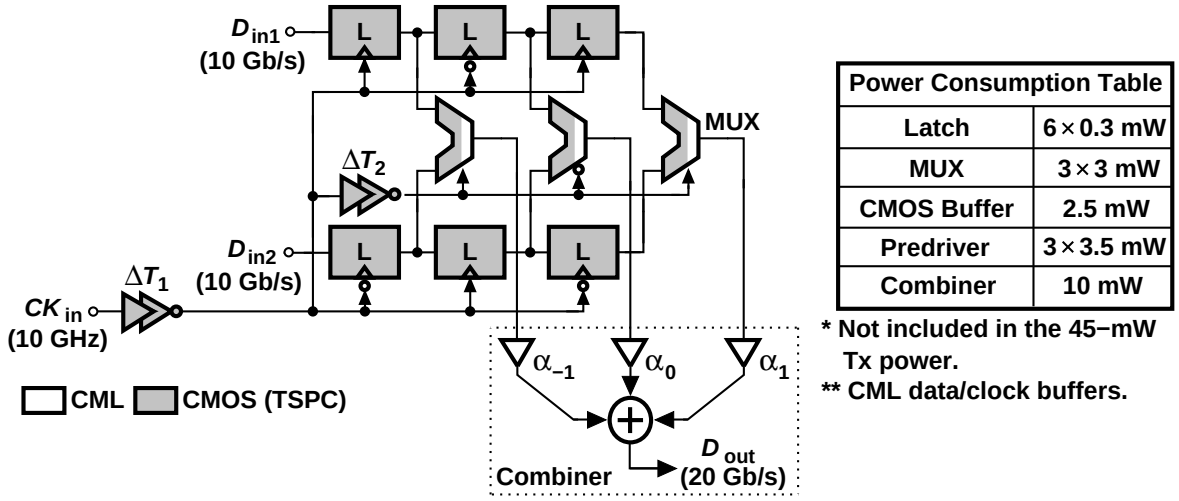


Fig. 4.18 Hybrid transmitter architecture.

At moderate speed around 20 Gb/s, the MUX design becomes more relaxed. Applications at this speed may require feedforward equalizers (FFEs) with $3 \sim 5$ taps in the transmitter, which must be codesigned with MUX. Full-rate structures inevitably dissipates significant power, because every single block in it has to be made in CML. Half-rate architecture, however, can leverage against the stringent speed requirement and save considerable power. It is primarily because in 65-nm CMOS, the half-rate data (10 Gb/s) and clock (10 GHz) can be handled purely in the digital domain, which, even with design margin, still consumes less power as compared with its

CML counterpart. To be more specific, we introduce a 20-Gb/s transmitter frontend with half-rate architecture and a 3-tap FFE (Fig. 4.18). Here, the two data inputs and deployed for the MUXes to pick up alternatively, producing appropriate bit sequence to be multiplied with the corresponding coefficients α_{-1}, α_0, α_1. The output driver thus combines the three and delivers the pre-emphasized output. The 10-GHz clock is buffered by delays ΔT_1 and ΔT_2 (both are made of CMOS inverters) to provide provide proper phase shifts for the DEMUX, the latches, and the MUXes. A table summarizing the power dissipation of each block is also demonstrated.

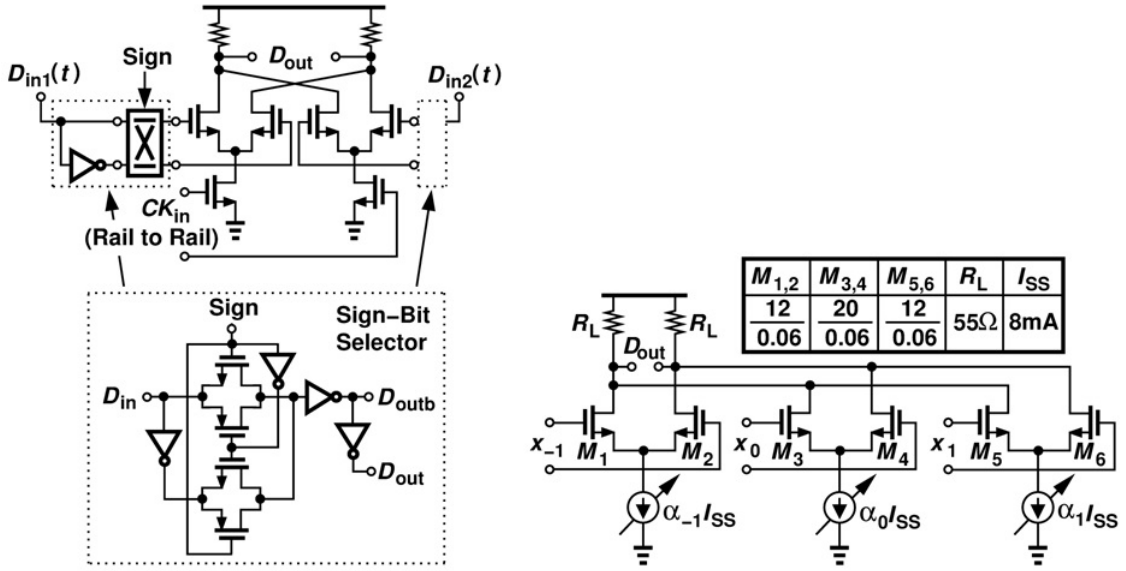


Fig. 4.19 (a) Hybrid MUX and (b) final output combiner.

The MUX design is shown in Fig. 4.19(a). With the help of rail-to-rail data and clock, it is possible to realize such a hybrid MUX at 20 Gb/s. Here, the sign-bit selection of the two data streams is accomplished by two-way switches made of transmission gates. Note that the MUX in Fig. 4.19(a) naturally restores the output signal back to CML levels. The output combiner (driver) follows conventional designs [7], [8] (Fig. 4.19). CML pairs with tunable tail currents are combined by means of the 55 Ω loading resistors. The three tail current sources have a constant total current of 8 mA, leading to a maximum swing (when no boosting) of 200 mV. Note that the devices in different taps are slightly scaled with current to further reduce the output capacitance.

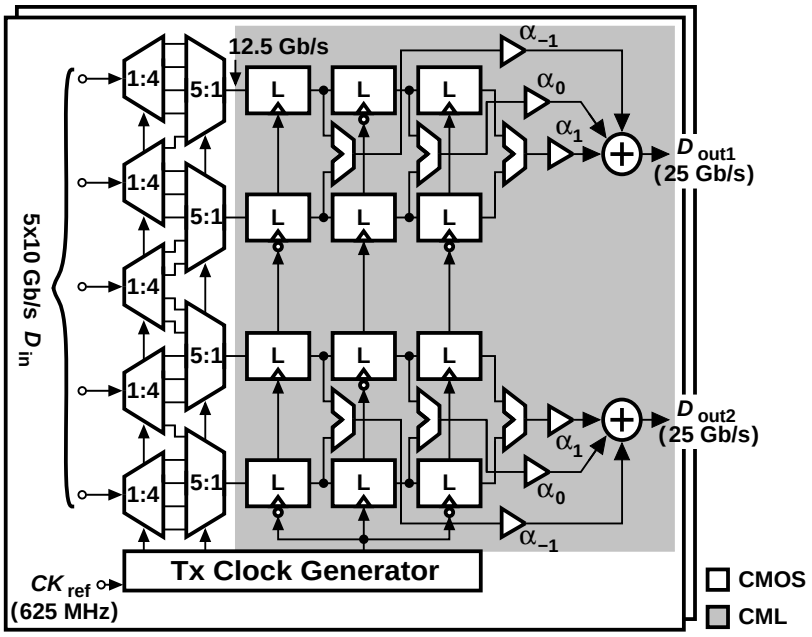
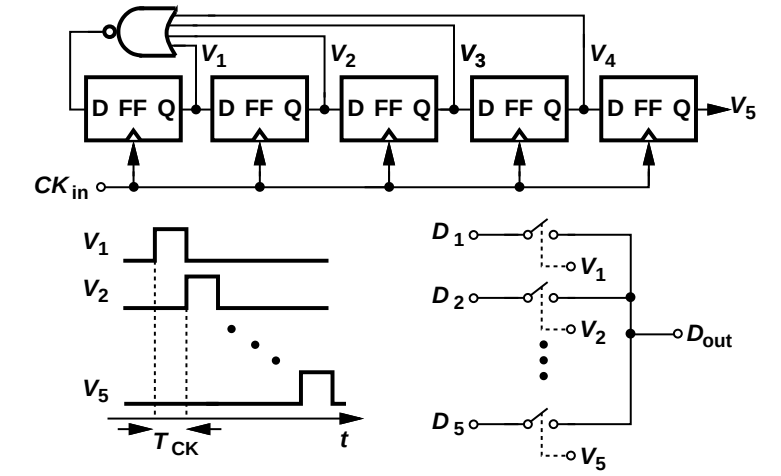
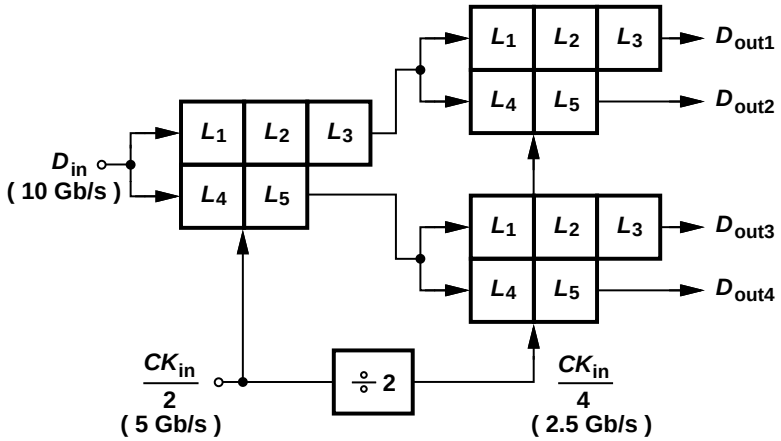


Fig. 4.20 100GbE gearbox with 5:2 multiplexing.

Some applications need MUXes with non-power-of-2 multiplexing ratio. For instance. 100 Gb/s Ethernet serializes 10 $\times$ 10 Gb/s input data into 4 $\times$ 25 Gb/s output data stream. The serializer and deserializer (also known as gearbox) must accomplish 5:2 and 2:5 data transforming. Figure 20 depicts such a transmitter. A complete 100Gb/s gearbox requires two identical 5:2 serializers. Each of them is responsible for converting 5 $\times$ 10 Gb/s inputs into 2 $\times$ 25 Gb/s outputs. It consists of a multi-frequency and multi-phase clock generator for different stages of multiplexing. Since combining 5 $\times$ 10 Gb/s input data directly may consume large amount of power, we realize the multiplexing by first speed down the data rate by a factor of 4. The 20 sub-rate data can be lumped as a group of 5 in digital circuits. Finally, 4 $\times$ 12.5Gb/s data streams are further serialized in half-rate operation with 3-tap FFEs. The 5:1 MUX circuit is illustrated in Fig. 4.21, which is realized as a 5-input transmission-gate sampler operated by rail-to-rail data and clocks. Five TSPC flipflops with a NOR gate feedback produce five 20% duty-cycle clocks $CK_{1\sim5}$ for proper sampling. The 1:4 DMUX is realized as a typical tree structure [Fig. 4.21], which employs TSPC latches to minimize power consumption.



(a)



(b)

Fig. 4.21 (a) 5:1MUX, (b) 1:4 DMUX in 100GbE gearbox.

4.4 DEMULTIPLEXERS

DMUXes are relatively easier to design in general, as the data rate goes down after demultiplexing. Generally speaking, two FFs driven by differential clocks can do the jobs. As illustrated in Fig. 4.22, we put a 40 Gb/s DMUX as an example. The input data is sampled, demultiplexed, and aligned by the $2\frac{1}{2}$ FFs directly. Note that no full-rate clock is required here. Some delay buffers may need to be inserted into the data paths, but the timing requirement is much more relaxed. At

lower speed, a direct 1:N structure can be adopted to save power, as depicted in Fig. 4.23. The low duty-cycle clocks can be generated in the same way as that in Fig. 4.21(a).

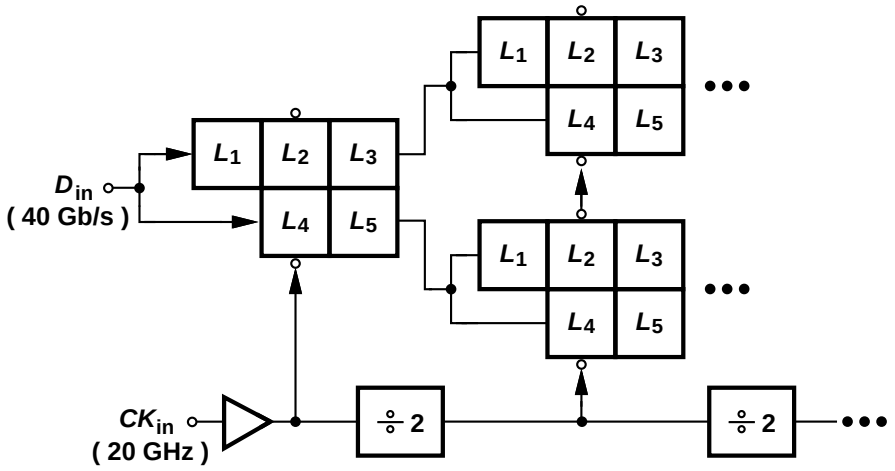


Fig. 4.22 DMUX in the structure.

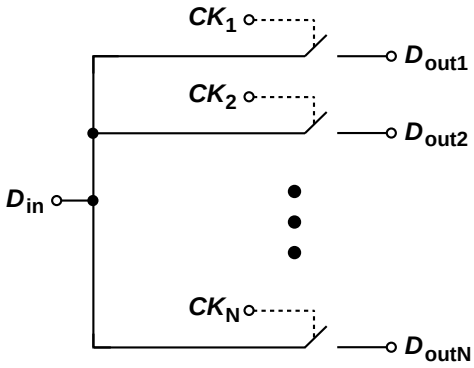


Fig. 4.23 Direct 1:N demultiplexing.

For non-power-of-2 DMUXes, circuits become much more complicated as deskew and alignment functions are now mandatory. Again, we take the 100GbE gearbox as an example. The 2:5 deserializer architecture is shown in Fig. 4.24. Two channels process the input data independently, presenting an aggregate data rate of 50 Gb/s. Each channel consists of a limiting amplifier with constant gain biasing, and a full-rate CDR circuit. The two retimed data streams are further demultiplexed into five 10-Gb/s lanes in parallel. The two 25-GHz clocks distilled from the data streams

are sent to a clock generator, which creates 2.5, 5, 10, and 12.5-GHz clocks for the subsequent deserializer. Here, we perform an additional 1:2 demuxing right after the CDR to relax the stringent speed requirement. The 1:5 demuxing can therefore be realized in a relaxed way, and finally five 4:1 MUXes are incorporated to produce five 10-Gb/s outputs. A complete 4×25-Gb/s receiver can easily be implemented by using two identical chipsets proposed here.

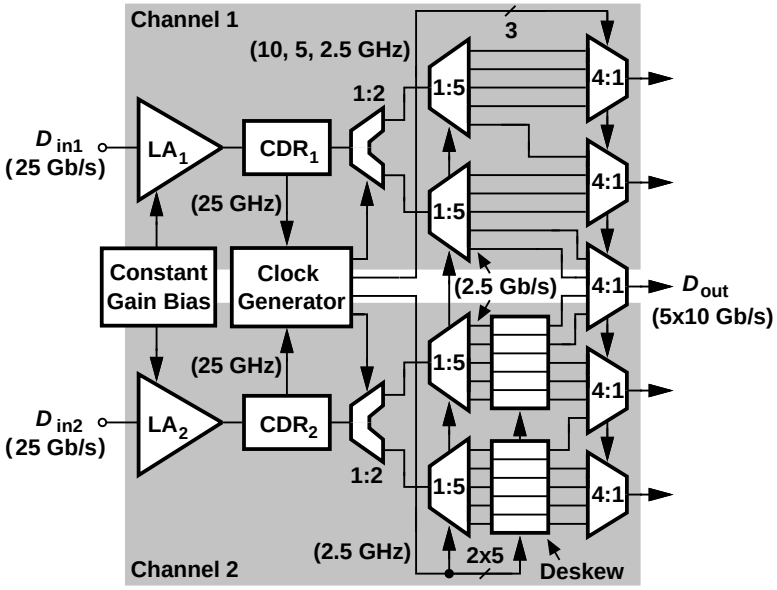
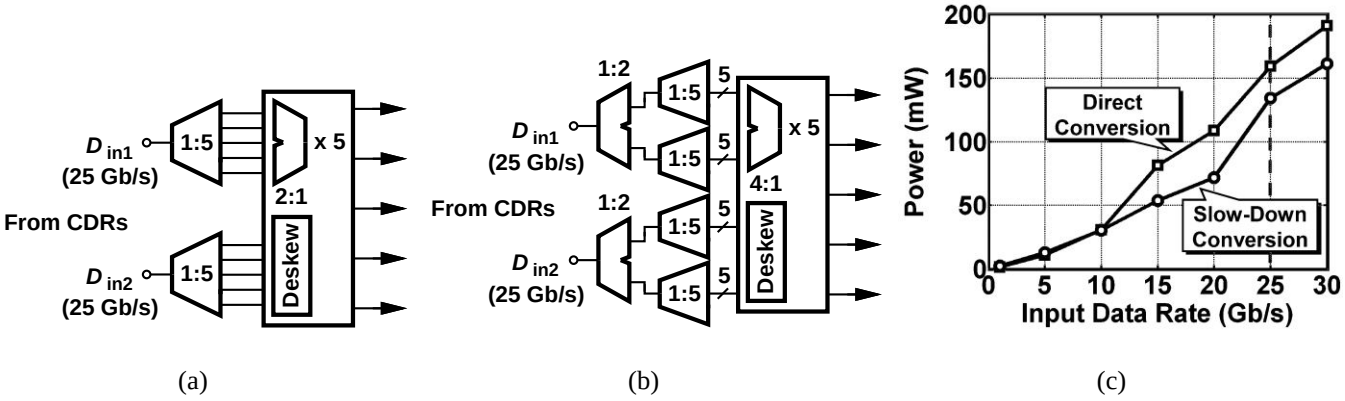


Fig. 4.24 100GbE gearbox with 2:5 demultiplexing.

The two channels may suffer from significant skew due to channel imbalance. The phase error can be removed by placing a deskew circuit in channel 2, which lines up the 10×2.5-Gb/s data streams. The adjustment is mandatory because the middle 4:1 MUX has to handle inputs from both channels. Without this realignment, wrong data be sampled. Note that skews larger than one bit can be removed by the bit alignment circuit, which consists of shift registers.

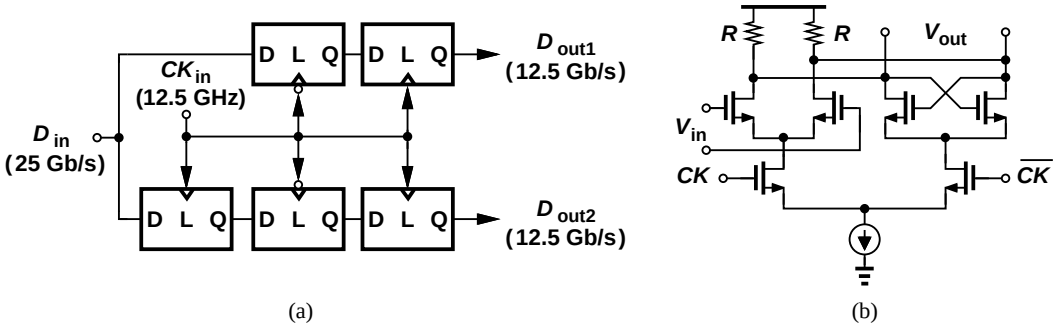
The outputs of the two CDRs are then deserialized into five subrate outputs. We have two possible solutions to do so. Shown in Fig. 4.25(a) is a straightforward approach, which uses two 1:5 DMUXes to parallelize the two 25-Gb/s data streams into 10×5-Gb/s lines, and conbines every two of them as 5×10-Gb/s outputs. Such a direct conversion suffers from a few difficulties.



(a) (b) (c)

Fig. 4.25 2:5 demultiplexing approaches. (a) Direct conversion. (b) Slow-down conversion. (c) Power efficiency comparison.

First it is quite stringent to design a 25-Gb/s 1:5 DMUX with reasonable power. Second, the two sets of lower-speed lines need to be aligned before final combination (2:1 MUXing), and the deskew circuit would consume significant power as well. Finally, the routing of high-speed lines makes the layout even more complicated.



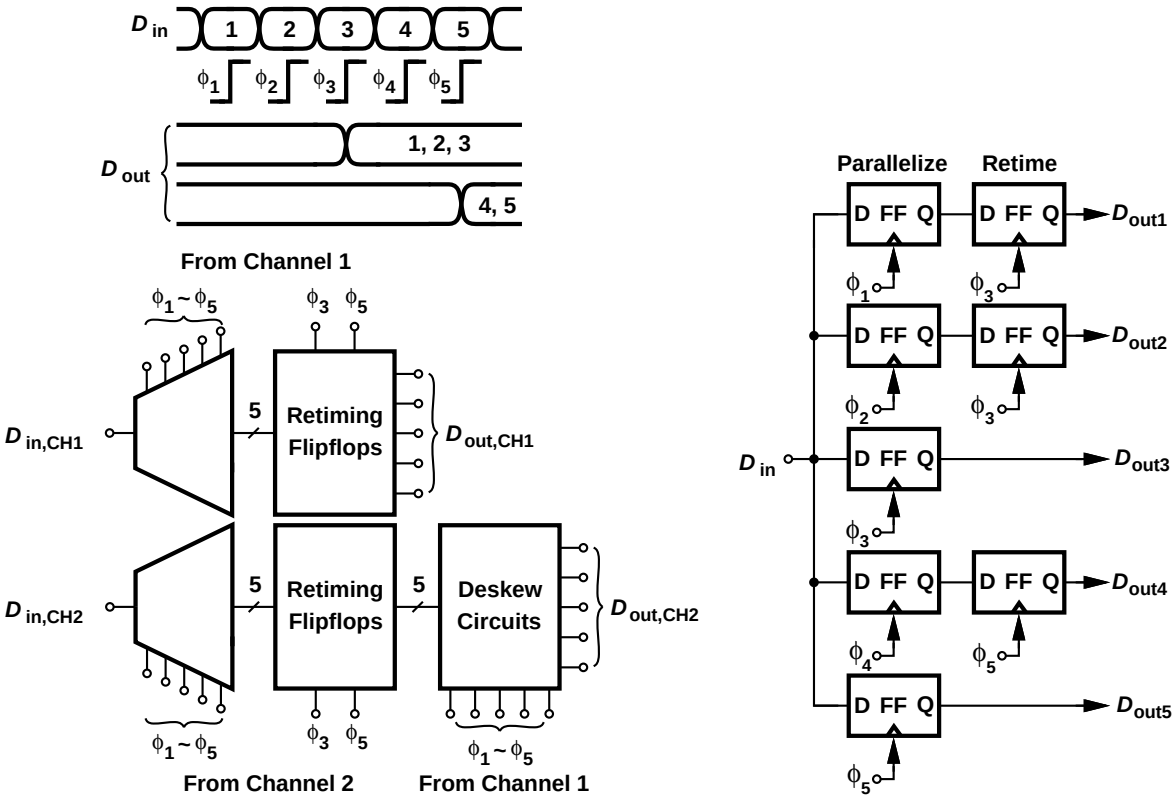
(a) (b)

Fig. 4.26 (a) 1:2 demultiplexer design. (2) CML design.

In this approach, we insert one more stage of DMUX in front of the 1:5 DMUXes to slow down the operation of subsequent circuits [Fig. 4.25(b)]. As a result, the 1:5 DMUXing and 4:1 MUXing can be realized in half-rate. Fig. 4.25 illustrates the power efficiency of the two structures. In 65-nm CMOS, for example, the slow-down conversion consumes less power than the direct conversion if the data rate is higher than 10 Gb/s. At D_{in} = 25 Gb/s, the overall power

of the former is less than that of the latter by 25 mW because most of the circuits are now in lower speed.

As shown in Fig. 4.26, the 25-Gb/s 1:2 DMUX is made of CML flipflops (FFs) with two outputs aligned in phase [9]. The alignment between the input data and clock is not an issue because both of them are to be aligned with the 25-GHz clock, i.e., retiming flipflops in CDR and the first $\div 2$ circuit are triggered with the same 25-GHz clock.



(a) (b)

Fig. 4.27 (a) 1:5 demultiplexing scheme. (b) DMUX with retiming sensing (to ϕ_3 and ϕ_5).

The 1:5 DMUX is much more complicated. It necessitates proper phase arrangement to produce the 20 $\times$ 2.5-Gb/s data.As shown in Fig. 4.27(a), a five-phase 2.5-GHz clock is used to sample the 12.5-Gb/s incoming data sequentially. Here, the outputs need to be separated by an angle as close as $180°$. Since the whole phase circle is divided into five pieces, we pick up two phases which are most apart from each other, say, ϕ_3 and ϕ_5, to do the retiming. In other words,

D_{out1}, D_{out2}, D_{out3} are launched simultaneously at the rising edge of ϕ_3 while D_{out4}, D_{out5} are initiated by the rising edge of ϕ_5. This operation is realized as the setup in Fig. 4.27(b), where the first, second, and fourth outputs are retimed by ϕ_3 and ϕ_5, respectively. The 1:5 DMUX in channel 2 basically follows the same operation except that a deskew circuit is added to ensure proper sampling. The deskew curcuit design can be found in [10].

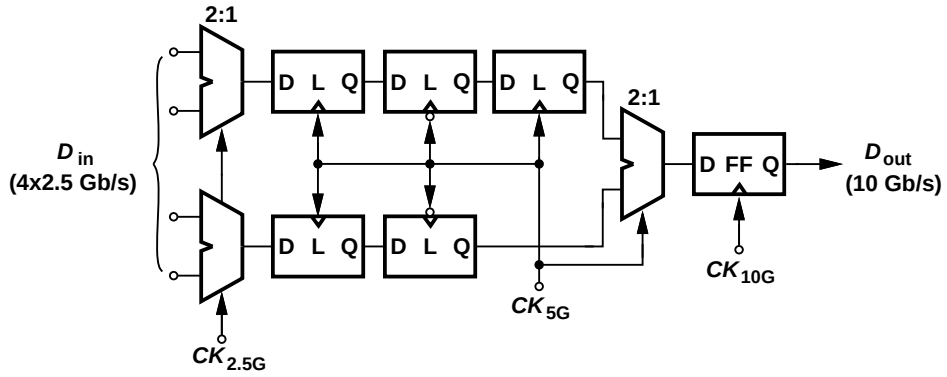


Fig. 4.28 4:1 multiplexer design.

The 4:1 multiplexer is depicted in Fig. 4.28. Since the four data inputs have been aligned in the preceding 1:5 DMUX stage, the circuit does not need 2.5-Gb/s shift latches as a conventional MUX does. A 10-Gb/s retimer is placed to clean up the final output data, eliminating possible imbalance caused by data duty cycle error.

4.5 HIGH-SPEED BUILDING BLOCK

In this section, we discuss the implementation of high-speed building blocks commonly used in wireline communication systems. Focusing on design of CMOS circuits, these blocks are realized in CML.

4.5.1 Logic Gates

All logic functions can be made in CML. The differential topology of a CML allows dual outputs (e.g., AND/NAND) depending on the definition of polarity. Unlike digital circuits, there is no need

to put inverters behind the logic gates to get the complementary results. The implementation of buffer/inverter, AND/NAND, OR/NOR, and XOR/NXOR gates in CML have been shown in Fig. 4.29. To obtain balanced rise/fall times, we need proper sizing for circuits with stacked devices. For example, the AND/NAND gate in Fig. 4.29(b) may have $(W/L)_1 = (W/L)_2 = 2(W/L)_3 = 2(W/L)_4$. Note that the inputs $A/\overline{A}$, $B/\overline{B}$ are of normal logic swing (i.e., 500 mV or longer in 1.2-V supply), and the switching devices are wide enough to accommodate the tail current. Peaking inductors can be added in series of loading resistors to accelerate the operation.

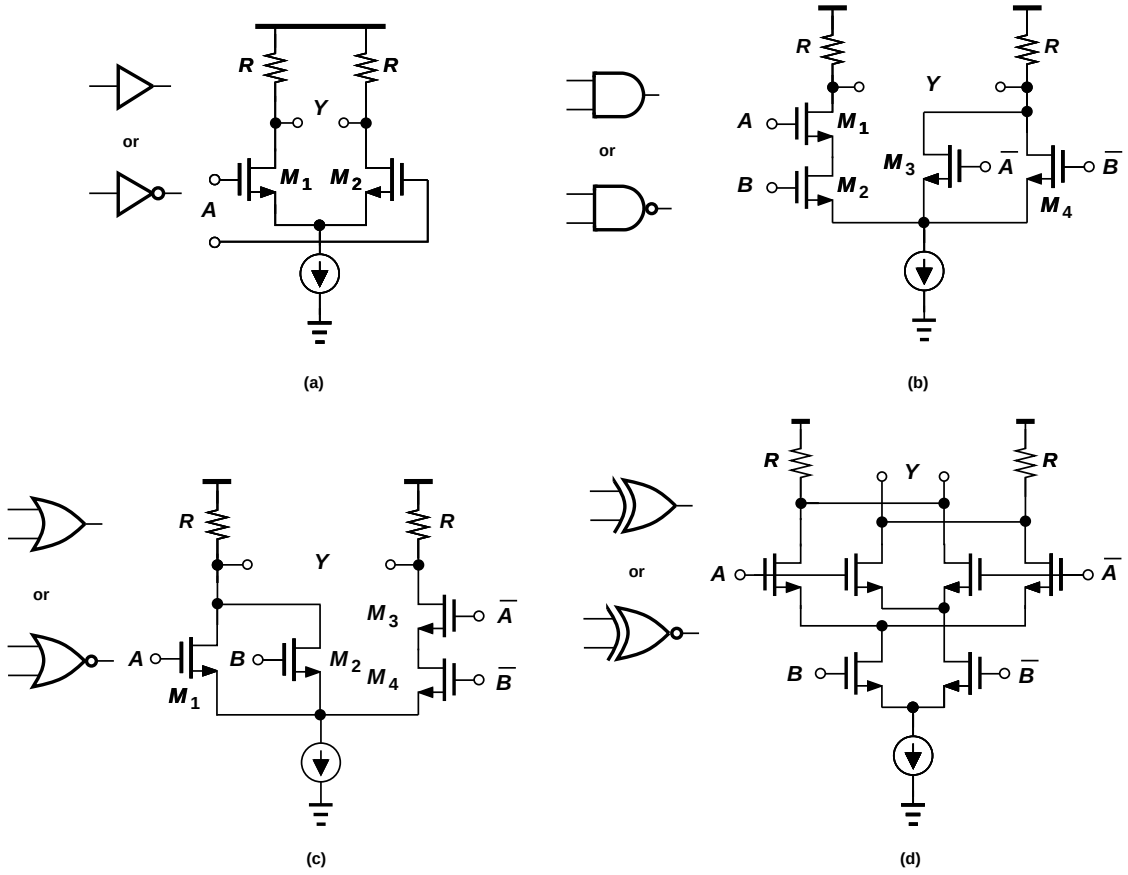


Fig. 4.29 Logic gates implement in CML: (a) buffer/inverter, (b) AND/NAND, (c) OR/NOR, (d) XOR/NXOR.

4.5.2 Analog Building Blocks

The XOR gate in Fig. 4.29(d) is actually a Gilbert cell, which has been extensively used as a mixer. A mixer usually deals with longer signals as comported with that in RF applications, leading to more relaxed tradeoffs in design. A typical mixer in 90 nm CMOS with 20 GHz bandwidth can be found in Fig. 4.30(a).

Another important blocks in analog signal processing is the delay cell. Figure 30(b) illustrates such a design, which incorporates cross-coupled pair $M_{3,4}$ and inductive peaking.

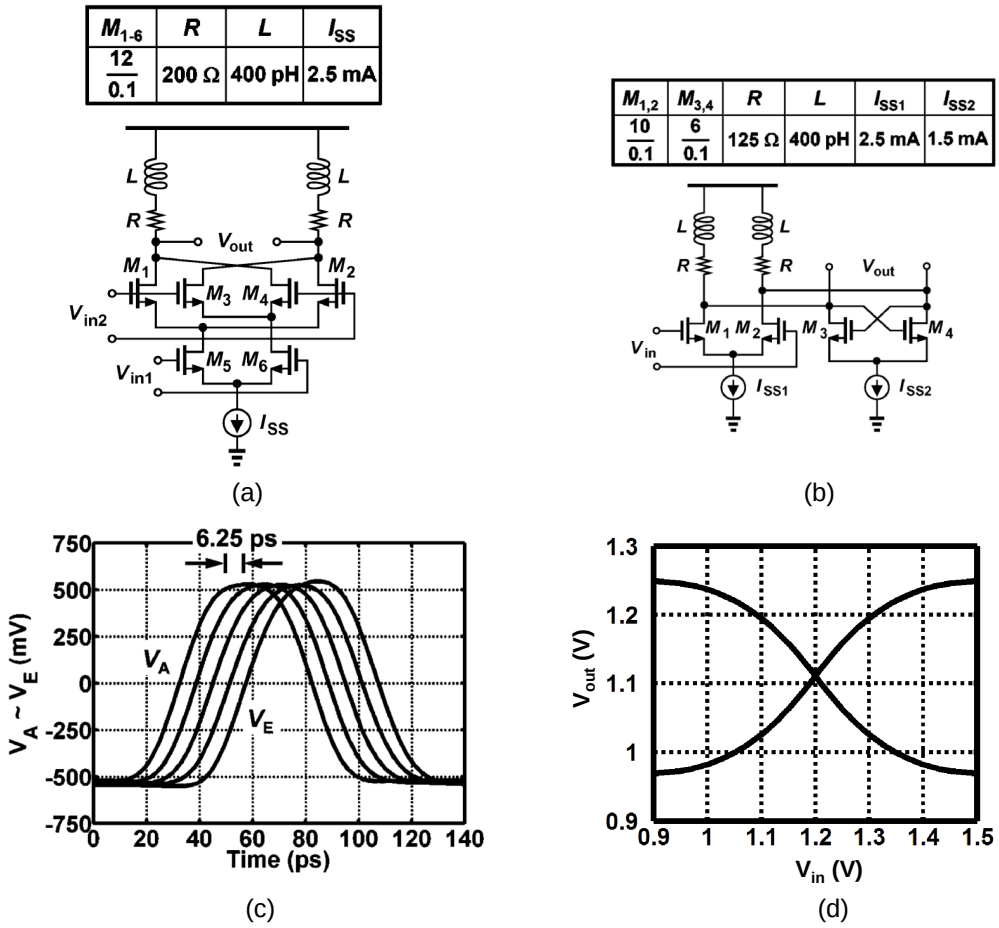


Fig. 4.30 Analog building blocks in CML: (a) mixer, (b) delay cell, transient waveform of a delay chain, (d) input-output characteristic of (b).

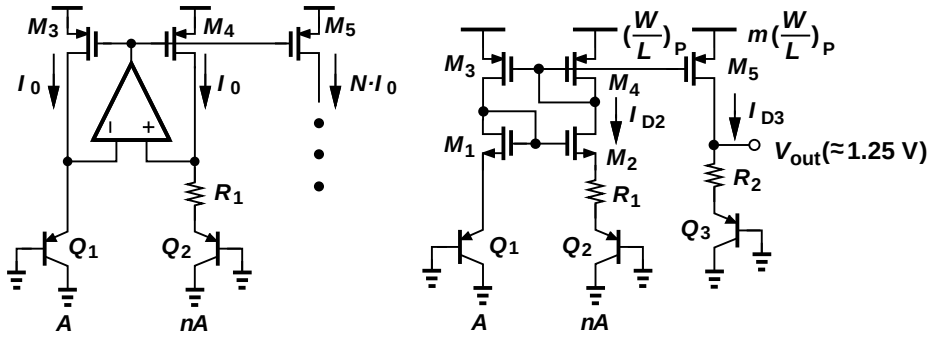
Under large-signal inputs, the cross-coupled pair $M_{3,4}$ provides hysteresis characteristic, creating

significant delay without degrading the bandwidth. Placing 4 identical cells realized in 90 nm CMOS in a row, we arrive at approximately 25-ps delay while consuming 24 mW (with a 1.0-V supply). Delay tuning can be accomplished by adjusting the two tail currents I_{SS1} and I_{SS2}. The power dissipation could be further reduced if more advanced technologies are used. Figure. 30(d) depicts the dc characteristic. Such a hysteresis buffer is actually quite useful in many situations. For example, it can sharpen a very slow sinusoidal wave to a square wave. More details can be found in [11].

4.6 CALIBRATION CIRCUITS

Typical calibration techniques such as bandgap references and low-dropout (LDO) regulators are popular in communication frontends. In general, we may need a constant value for voltage, current, IR drop, resistance, small-signal gain, and even frequency in our design. We summarize these techniques in this section.

4.6.1 PTAT Current



(a) (b)

Fig. 4.31 (a) PTAT current. (b) bandgap reference.

Creating a current which is linearly proportional to absolute temperature (PTAT) is essential to other calibration circuits. Shown in Fig. 4.31(a) is a standard structure, whose upper PMOS current sources are governed by the feedback Opamp to provide equal current I_0. Here, we have

$(W/L)_{M5}$ = N·$(W/L)_{M3}$ = N·$(W/L)_{M4}$. Owing to the longer size of Q_2, R_1 can be accommodated so that $V_{BE1} = V_{BE2} + I_0 R_1$. It follows

$$I_0 = \frac{\ln n}{R_1} \cdot V_T = \frac{\ln n}{R} \cdot \frac{kT}{q}, \tag{4.9}$$

where k is Boltzmann's constant (= 1.38 $\times$ 10^{-23} $m^2 kg S^{-2} K^{-1}$) and g the electron charge (= 1.6 $\times$ 10^{-19} coulombs). Since M_5 is mirrored from $M_{3,4}$, we create an PTAT output current $N \cdot I_0$.

It is instructive to check the feedback polarity. There are two feedback paths in Fig. 4.31(a). The factor β in such a voltage-voltage feedback is given by

$$\begin{aligned} \beta^+ &= g_{m3,4} \cdot \frac{1}{g_{m1}} \\ \beta^- &= g_{m3,4} \cdot (R_1 + \frac{1}{g_{m1}}) > \beta^+. \end{aligned} \tag{4.10}$$

Thus, the whole circuit forms a negative feedback, expected to be stable under proper design.

4.6.2 Bandgap Reference

A PTAT current can be used as a temperature sensor by placing an external, low temperature coefficient resistor as a loading to ground. More importantly, it can be used to form a bandgap reference circuit, which provides a constant voltage immune to PVT variations. Shown in Fig. 4.31(b) is an example, in which the feedback Opamp is replaced by the double mirros $M_{1,2}$ and $M_{3,4}$. Since $(W/L)_1$ = $(W/L)_2$ and $(W/L)_3$ = $(W/L)_4$, I_{D2} and I_{D3} are still PTAT currents. Also, $(W/L)_5$ = m·$(W/L)_4$, the output voltage is equal to

$$\begin{aligned} V_{out} &= V_{BE3} + I_{D3} R_2 \\ &= V_{BE3} + m \cdot \frac{R_2}{R_1} \cdot \ln n \cdot V_T. \end{aligned} \tag{4.11}$$

As we know, V_{BE} has a negative temperature coefficient (≈xx mV/K) whereas V_T has a positive on (xx mV/K). Thus, we arrive at a voltage with zero temperature coefficient in the vicinity of room temperature if

$$m \cdot \frac{R_2}{R_1} \cdot \ln n \cong 17. \tag{4.12}$$

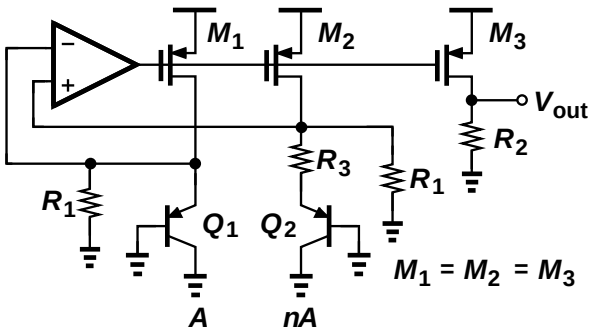


Fig. 4.32 Sub-1V bandgap reference.

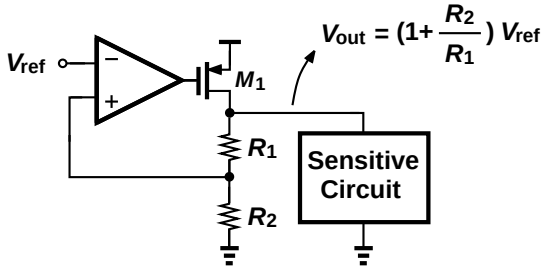


Fig. 4.33 Creating arbitrary supply voltage.

As a result, we obtain a bandgap reference voltage approximately equal to 1.25V.

In core devices of today's technologies , designers need an improved version of bandgap reference to create sub-1V reference. Figure 32 illustrates such a design, where two side resistors (R_1) have been added. Again, we assume $(W/L)_1$ = $(W/L)_2$ = $(W/L)_3$, the output voltage becomes

$$V_{out} = \frac{R_2}{R_1} V_{BE3} + \frac{R_2}{R_3} \cdot \ln n \cdot V_T. \tag{4.13}$$

As long as

$$\frac{R_2}{R_1} : \frac{R_2}{R_3} \ln n = 1 : 17, \tag{4.14}$$

a bandgap voltage can still be created. For instance, if $V_{BE} \approx$ 0.7 V, R_1 = 2 kΩ, R_3 = 270 Ω, and n = 10, we get $V_{out} \cong$ 563 mV. The reader can prove that the circuit in Fig. 4.32 is still stable.

A sub-1V bandgap voltage can be extended to generate reference voltages above 1V. Shown in Fig. 4.33 is an example, where the sensitive block necessitates a dedicated supply. A topology

similar to LDO is therefore employed. The feedback loop forces $V_{out} = (1 + R_1/R_2) \cdot V_{ref}$, where V_{ref} comes from bandgap reference. A mini LDO is there created locally. There are many way to implement the sub-1V Opamp in Fig. 4.32. Here, we introduce two approaches. Illustrated in Fig. 4.34(a) is a two-stage topology designed in 1.2-V supply, targeting high gain (50 dB) with large output dynamic range. The dc gain is given by

$$A_{V,dc} \cong g_{m1,2}(r_{o2}//r_{o4})g_{m6} \cdot (r_{o6}//r_{o7}). \tag{4.15}$$

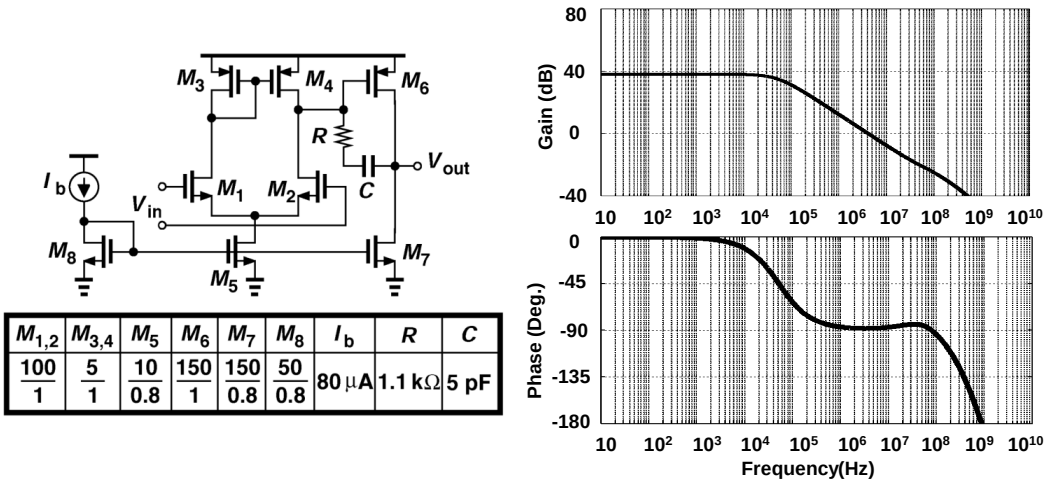
The loop stability must be handled with care because 1) the two-stage Opamp introduces two internal poles and 2) a third pole exists in the feedback path of bandgap circuits. To stabilize the loop, we have to push all the nondominant poles away from the origin. First, a compensation capacitor C (=5 pF) and a zero-shifting resistor R (= 1.1 KΩ) are placed between the two stages to achieve a large phase margin of xx°. Also, to minimize additional phase shift caused by the circuits in the feedback loop, the feedback path must have low gain and high bandwidth (i.e., much higher than the unity-gain frequency of Opamp, which is 10 MHz). Simulation shows that all loops maintain overall phase margins greater than xx°. Another possibility to realize a low-supply Opamp can be found in Fig. 4.34(b). The input difference is first translated into current lay the $M_{1,2}$ pair, and gets converted back to voltage by mirroring. Large output range is preserved, and the dc gain now becomes

$$A_{V,dc} \cong g_{m1,2}(r_{o6}//r_{o8}). \tag{4.16}$$

Pole-splitting is not needed here. Figure. 34(b) also shows the Bode Plot of it, suggesting a xx dB dc gain and xx MHz unity-gain bandwidth.

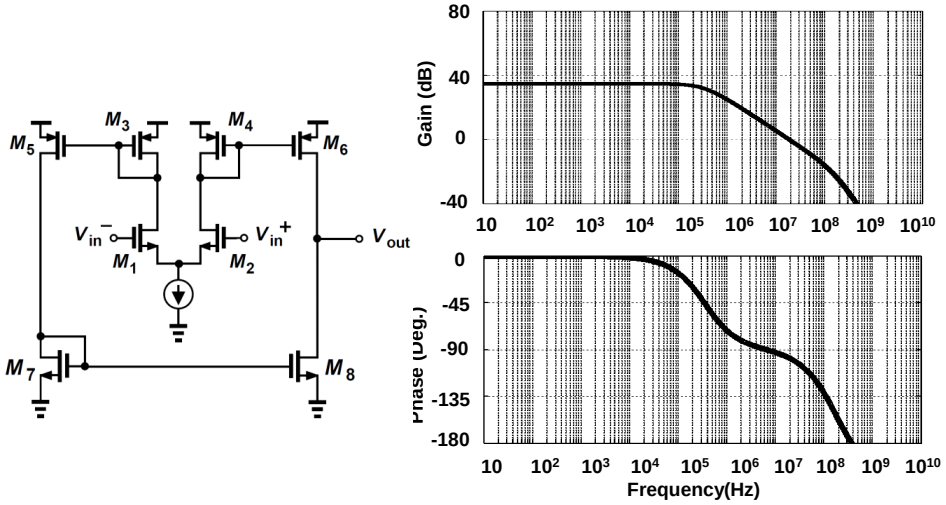
4.6.3 PTAT Constant IR Drop/Constant Circuits

One important application of bnadgap reference circuit in data link is to create a constant IR drop. Indeed, all data in CML must maintain a proper (and uniform) swing so as to ensure signal integnty. Such a circuit can be realized as depicted in Fig. 4.35. Here, the bandgap reference circuit generate a constant voltage of V_{BG} = 1.25 V, which is equal to I_6R_6 became of the negative feedback loop. Mirroring I_6 all the way from M_6 to M_9, we obtain the tail current of the CML



$M_{1,2}$	$M_{3,4}$	M_5	M_6	M_7	M_8	I_b	R	C
100/1	5/1	10/0.8	150/1	150/0.8	50/0.8	80 μA	1.1 kΩ	5 pF

(a)



(b)

Fig. 4.34 Low-supply Opamps.

buffer to be $n \cdot I_6$. Since R_6 and R_7 are realized on chip (with the same geometric outline) and $R_7 = m \cdot R_6$, the buffer's output swing is given by $\pm mnI_6R_6$. Here we assume the input signal is large enough to switch the tail current completely, which is the case for most CML buffer.

The circuit in Fig. 4.35 can also provide constant current if R_6 is placed externally. Here, an accurate and low temperature dependent resistor R_6 loyally translate the bandgap voltage to a

constant current I_6. Surface-mount devices (SMD) resistors with temperature coefficient on the order of 10 ppm/K is not difficult to find in the market. The reader can prove that a constant IR drop biasing can be created by means of the sub-1V bandgap circuit as well.

4.6.4 PTAT Constant Resistance

Unsilicided resistors may present generic inaccuracy as large as $\pm 15\%$ and high temperature coefficient of xx ppm/K. To achieve invariant resistance for loading, we need to introduce another device whose resistance is tunable. Putting a triode device in parallel with a real resistor is one way to do it. Figure 36 depicts a biasing circuit which provides both constant resistance and constant IR drop. On the right-hand side, the constant current I_{SS} and constant voltage (0.7 V) coming from a bandgap reference define the equivalent resistance of R and M_6 combination. That is,

$$I_{SS} \cdot (R//R_{eg,M6}) = 0.5, \tag{4.17}$$

where $R_{eg,M6}$ denotes the equivalent resistance of M_6. Same tail current and biasing voltage can be applied to a CML buffer on the left-hand side. Since $M_4 = M_5 = M_6$, the differential pain $M_{1,2}$ experiences constant output swing (i.e., IR drop), given that the input is large signal data. The capacitance introduced by the PMOS devices can be resonated out if inductive peaking is included.

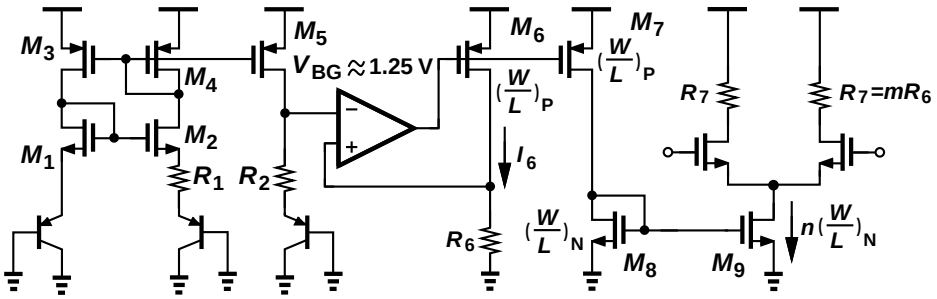


Fig. 4.35 Constant current and constant IR drop.

4.6.5 PTAT Constant Gain

A special circuit providing constant gain biasing for low-gain amplifiers is illustrated in Fig. 4.37. Here, $M_3 = M_4$, and M_2 is k-times larger than M_1 As known as supply insensitive biasing,

this circuit creates current independent of supply variation to the first order. It can be shown that

$$I_0 = \frac{2}{\mu_n C_{ox}(W/L)_1} \cdot \frac{1}{R^2} \cdot (1 - \frac{1}{\sqrt{k}})^2, \tag{4.18}$$

where $(W/L)_1$ denotes the dimension of M_1. Mirroring this current to M_5 (assume M_5 = M_1), the differential pair $M_{6,7}$ prevents a small-signal gain of

$$\begin{aligned} A_v &= g_{m6,7} \cdot R_D \\ &= \sqrt{\frac{2(W/L)_{6,7}}{(W/L)_1}} \cdot \frac{1}{R} \cdot (1 - \frac{1}{\sqrt{k}}) \cdot R_D. \end{aligned} \tag{4.19}$$

If $M_{6,7}$ and M_1 have the same tendency of deviation, the voltage gain here can be kept constant regardless of PVT variations.

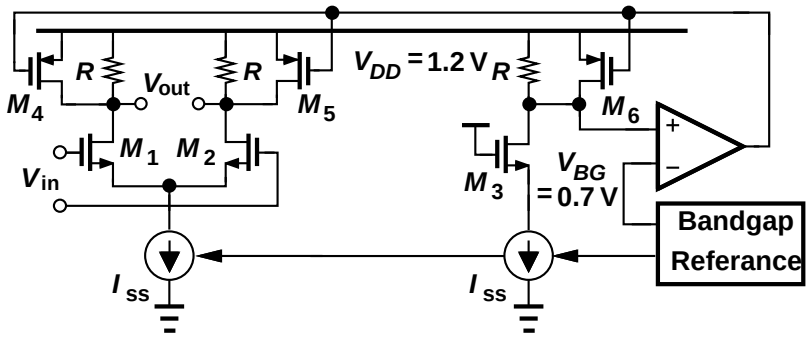
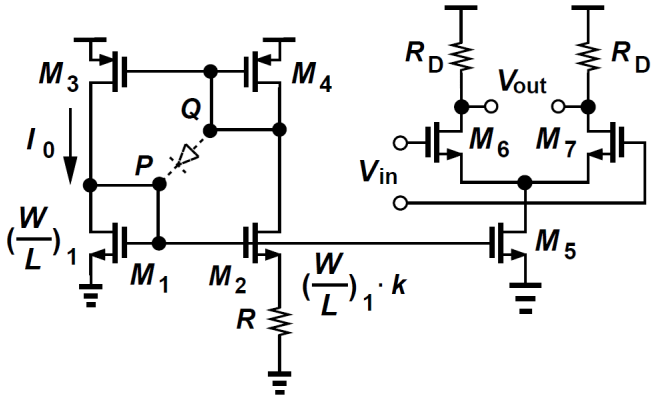


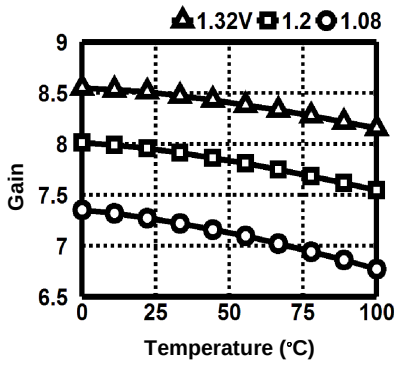
Fig. 4.36 Constant resistance (and also constant IR drop) circuit.

In reality, the biasing circuits M_1-M_4 are usually made pretty bulk (i.e., large L) so as to minimize the effect of channel-length modulation. In other words, the variation of $M_{6,7}$ and M_1 may not cancel out each completely. Other second-order effects would cause non idealities as well. Nonetheless, typical performance of this circuit with nominal gain of xx dB in 40 nm CMOS has been revealed in Fig. 4.37, suggesting a maximum gain deviation of $\pm xxdB$ across all variations. Note that a start-up diode (a real diode or diode connected MOS) between nodes P and Q is required to "wake up" the circuit at power up, forcing non-zero current flowing through it. The wake up diode must be turned off afterwards. Proper design regarding the supply and node voltages is mandatory.



Size

$M_{1,5}$	$M_{3,4}$	$M_{6,7}$	R_D	R	k
$\frac{1}{0.2}$	$\frac{1}{0.2}$	$\frac{4}{0.02}$	10kΩ	5k Ω	16



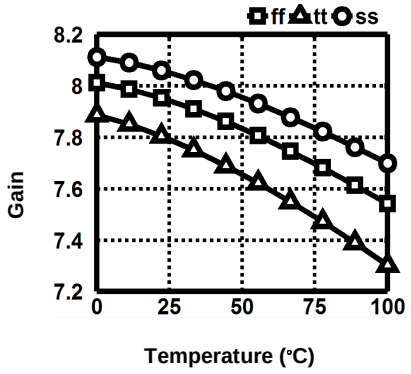


Fig. 4.37 Constant gain circuit.

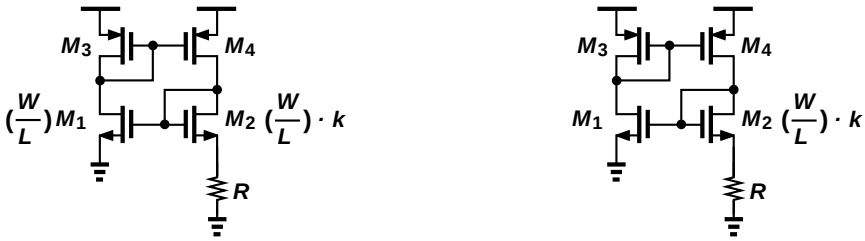


Fig. 4.38 Stability calculation.

REFERENCES

[1] B. Razavi, *Principles of Data Conversion System Design*, IEEE Press, 1995.

[2] S. C. Chan *et al*., "Distributed differential oscillators for global clock networks," *IEEE J. Solid-State Circuits*, vol. 41, no. 9, pp. 2083V2094, Sep. 2006.

[3] A. P. Jose and K. L. Shepard, "Distributed loss-compensation techniques for energy-efficient low-latency on-chip communication," *IEEE J. Solid-State Circuits*, vol. 42, no. 6, pp. 1415V1424, Jun. 2007.

[4] T. Suzuki *et al*., "A 90 Gb/s 2:1 multiplexer IC in InP-based HEMT technology," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2002, pp. 192V193.

[5] T. Yamamoto *et al*., "A 43 Gb/s 2:1 selector IC in 90 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2004, pp. 238V239.

[6] S. Galal and B. Razavi, "40 Gb/s amplifier and ESD protection circuit in 0.18-μm CMOS technology," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2004, pp. 480V481

[7] V. Balan *et al*., "A 4.8-6.4-Gb/s serial link for backplane applications using decision feedback equalization," *IEEE J. Solid-State Circuits*, vol. 40, pp. 1957V1967, Sep. 2005.

[8] K.-L. Wong and C.-K. Yang, "A serial-link transceiver with transition equalization," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2006, pp. 757V758.

[9] K. Kanda *et al*., "40 Gb/s 4:1 MUX/1:4 DEMUX in 90 nm standard CMOS," in *IEEE IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2005, pp. 152V153.

[10] K. Wu *et al*., "A 2 25xGb/s Receiver With 2:5 DMUX for 100-Gb/s Ethernet," *IEEE J. Solid-State Circuits*, Vol. 45, no. 11, pp. 2421-2432, NOV. 2010

[11] J. Lee *et al*., "A 75-GHz phase-locked loop in 90-nm CMOS technique," *IEEE J. Solid-State Circuits*, vol. 43, no. 6, pp. 1414V1426, Jun. 2008.

[12] Y. Amamiya *et al.*, “A 40 Gb/s multi-data-rate CMOS transceiver chipset with SFI-5 interface for optical transmission systems,” in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2009, pp. 358V359.

EQUALIZERS

5

5.1 CHANNEL IMPAIRMENTS

5.1.1 Line Loss

In electrical media, two main sources cause channel loss to degrade high frequency response: skin effect and dielectric loss. At high frequencies, current in a conductor tends to flow through the surface of it rather than the inner part. Known as skin effect, this phenomenon not only attenuates the magnitude but alters the phase. Dielectric loss, on the other hand, happens when the dielectric in the channel is not a perfect insulator. It only involves decay in magnitude.

In general, we model the loss of coaxial cables and backplane traces as a function of length and frequency. It is usually represented as

$$C(f) = \exp\Big[- k_s l(1+j)\sqrt{f} - k_d l f \Big], \tag{5.1}$$

where k_s and k_d are coefficients denoting skin effect and dielectric loss, respectively, and l the cable/trace length. At low frequencies, the substrate conductance contributes negligible loss compared with the skin effect, yielding a simplified transfer function of the channel

$$C(f) = \exp\Big[- k_s l(1+j)\sqrt{f} \Big]. \tag{5.2}$$

Here, the magnitude and phase are bound together: 10-dB and 20-dB loss correspond to phase shifts of 66° and 132°, respectively, regardless of the length and frequency. As frequency increases, the dielectric loss becomes significant, leading to a more rapid drop in magnitude. A typical

transfer function is depicted in Fig. 5.1. In order to specify the critical point where the skin effect and dielectric losses are equivalent in magnitude, we define a critical frequency $f_c = (k_s/k_d)^2$. Note that this critical frequency varies across a wide range for different media: typical cables (e.g., RG-58) and PCB traces (e.g., FR-4) would have f_c on the order of GHz, whereas some high-quality cables may have f_c as high as a few hundreds of GHz. Transmitting data at tens of gigabits per second across different channels results in different types of attenuation.

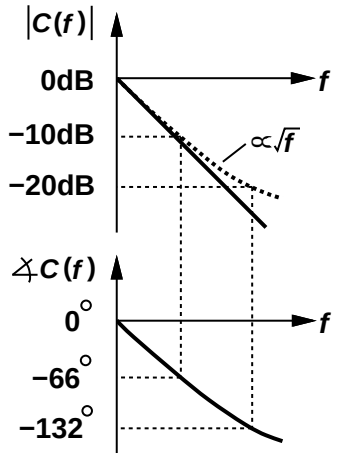


Fig. 5.1 Typical transfer function of coaxial cables and backplane traces.

5.1.2 Low f_c Channels

If f_c is much less than the data rate, dielectric loss dominates for most of the spectrum. We approximate the cable's characteristic as

$$C_1(f) = \exp(-k_d l f). \tag{5.3}$$

The impulse response is readily available by taking the inverse Fourier Transform of $C_1(f)$:

$$c_1(t) = \int_{-\infty}^{\infty} C_1(f) \exp(j2\pi f t)\ df \tag{5.4}$$

$$= \frac{2k_d l}{k_d^2 l^2 + 4\pi^2 t^2}, \tag{5.5}$$

as plotted in Fig. 5.2(a). Interestingly, $c_1(t)$ can be loosely considered as an impulse, and it really becomes $\delta(t)$ as l approaches zero. Now we apply a single bit $x(t)$ into the channel and see what happens. By convolution we have

$$y_1(t) = x(t) * c_1(t) \tag{5.6}$$

$$= \int_{-\infty}^{\infty} x(\tau) c_1(t-\tau)\, d\tau \tag{5.7}$$

$$= \int_{0}^{T_b} \mathrm{V}_0 \frac{2k_d l}{k_d^2 l^2 + 4\pi^2(\mathrm{t}-\tau)^2}\, d\tau \tag{5.8}$$

$$= \frac{\mathrm{V}_0}{\pi}\left[\tan^{-1}\frac{2\pi t}{k_d l} - \tan^{-1}\frac{2\pi(t-T_b)}{k_d l}\right], \tag{5.9}$$

where V_0 and T_b denote the swing and bit period of the input data, respectively. Equating the derivative of $y_1(t)$ to 0, we obtain the maximum value of $y_1(t)$ and realize that it occurs at $t = T_b/2$:

$$y_{1,\max} = y_1(\frac{T_b}{2}) = \frac{2\mathrm{V}_0}{\pi}\tan^{-1}\left(\frac{\pi T_b}{k_d l}\right). \tag{5.10}$$

It follows that

$$Eye\ Closure = \frac{2(\mathrm{V}_0 - y_{1,\max})}{\mathrm{V}_0} = 2 - \frac{4}{\pi}\tan^{-1}\left(\frac{\pi T_b}{k_d l}\right). \tag{5.11}$$

For example, if a cable presents 10-dB loss at 1/(2T_b) (i.e., Nyquist frequency), $y_{1,\max} = 0.597\mathrm{V}_0$ and ISI = 80.6%. Furthermore, an eye closure of 100% would happen if a cable shows more than 13.65-dB loss at 1/(2T_b).

5.1.3 High f_c Channels

If f_c is much greater than the data rate, only skin effect loss is significant at the frequencies of interest. Again we neglect phase shift and model the channel as

$$C_2(f) = \exp(-k_s l\sqrt{f}). \tag{5.12}$$

We then calculate the impulse response as

$$c_2(t) = \int_{-\infty}^{\infty} C_2(f)\exp(j2\pi f t)\, df \tag{5.13}$$

$$= 2\int_{0}^{\infty} \exp(-k_s l\sqrt{f})\cos(2\pi f t)\, df. \tag{5.14}$$

Unfortunately, Eq. (5.14) has no explicit solution. It can be proven that $c_2(t)$ peaks at $t = 0$ and decays to zero as t approaches infinity [Fig. 5.2(b)]. Similar to $c_1(t)$, $c_2(t)$ degenerates to $\delta(t)$ as l approaches zero. Now we can determine the single-pulse response $y_2(t)$. By convoluting $x(t)$ and $c_2(t)$, we get

$$y_2(t) = x(t) * c_2(t) \tag{5.15}$$

$$= \int_0^{T_b} \mathrm{V}_0 \cdot 2 \int_0^{\infty} \exp(-k_s l\sqrt{f}) \cos[2\pi f(t-\tau)]\, df\, d\tau \tag{5.16}$$

$$= 2\mathrm{V}_0 \int_0^{\infty} \exp(-k_s l\sqrt{f}) \int_0^{T_b} \cos[2\pi f(t-\tau)]\, d\tau\, df. \tag{5.17}$$

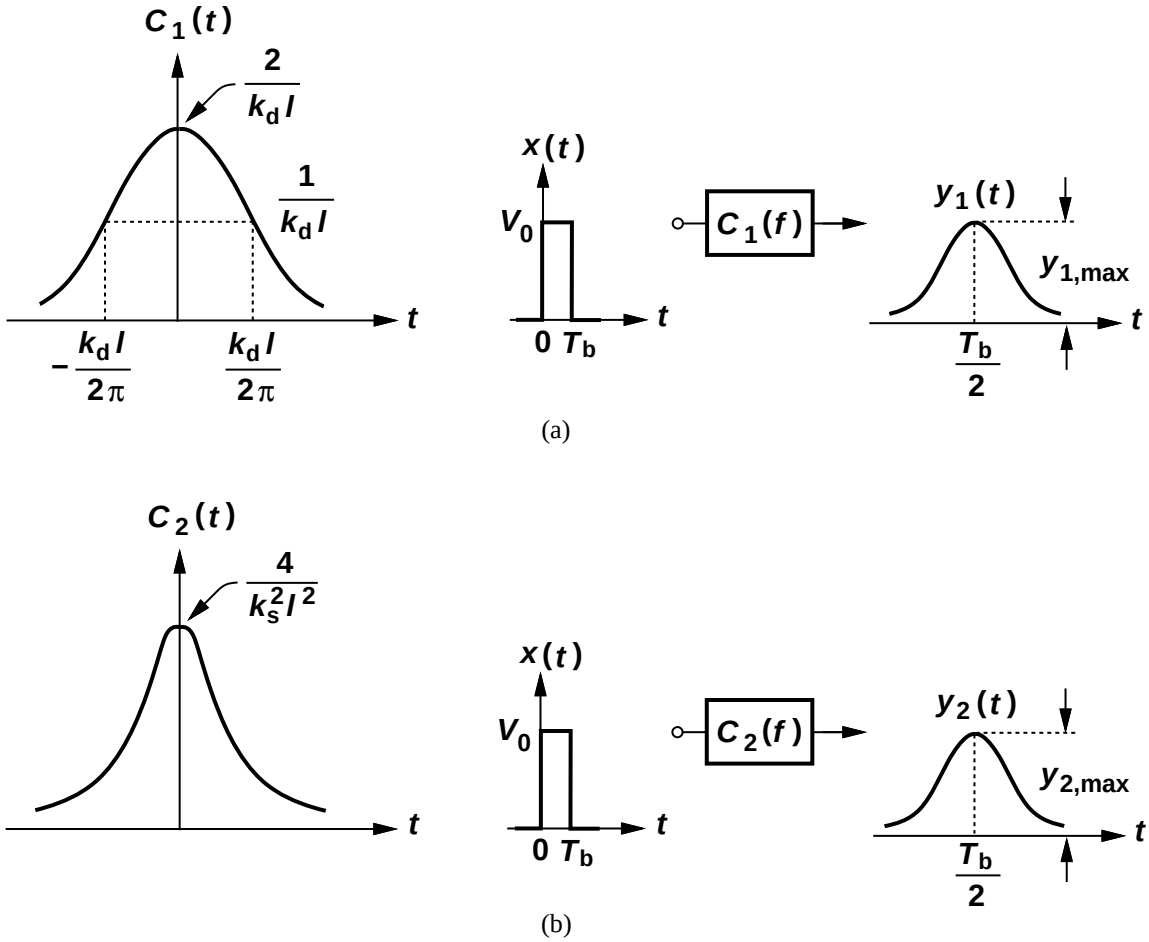


Fig. 5.2 Impulse response of (a) low-f_c, (b) high-f_c channels.

Owing to the symmetry, $y_2(t)$ still peaks at $t = T_b/2$. By the same token, we obtain

$$y_{2,\max} = y_2(\frac{T_b}{2}) = \frac{2\mathrm{V}_0}{\pi}\int_0^\infty \exp(-k_s l\sqrt{f})\frac{1}{f}\sin(\pi f T_b)\ df. \tag{5.18}$$

For a cable with transfer function of Eq. (5.12) and 10-dB attenuation at $1/(2T_b)$, $y_{2,\max} \approx 0.48V_0$ and no eye opening can be observed in the eye diagram. The 100% eye closure occurs if the cable exhibits 9.61-dB loss at half data rate.

The above analysis neglects phase shift effect. In reality, the phase discrepancy induced by channel loss also causes substantial jitter. To see this, we apply an ideal PRBS7 data into two transfer functions with and without phase shift effect [i.e., Eq. (5.2) and Eq. (5.12)]. Both channels experience 6-dB loss in magnitude at Nyquist frequency. As illustrated in Fig. 5.3, the actual jitter presented at crossover point should be xx UI rather than xx UI. Such an under-estimation could be avoided if accurate channel model (e.g., s-parameter) is included. A typical impulse response of a channel is illustrated in Fig. 5.4.

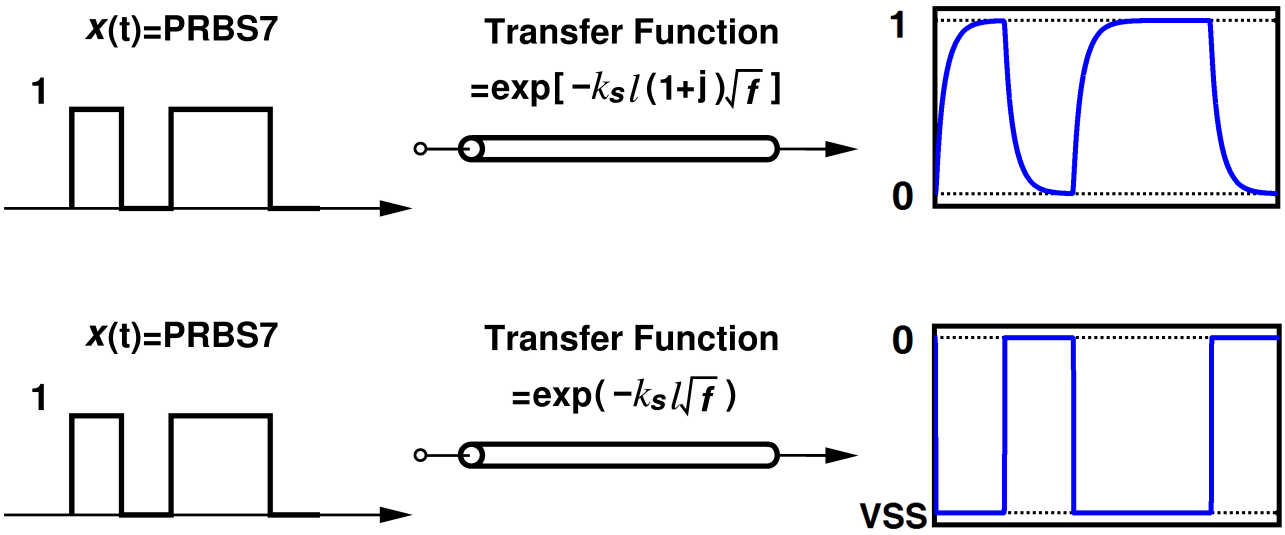


Fig. 5.3 Response of PRBS7 data stream going through high f_c channels with and without phase shift considered.

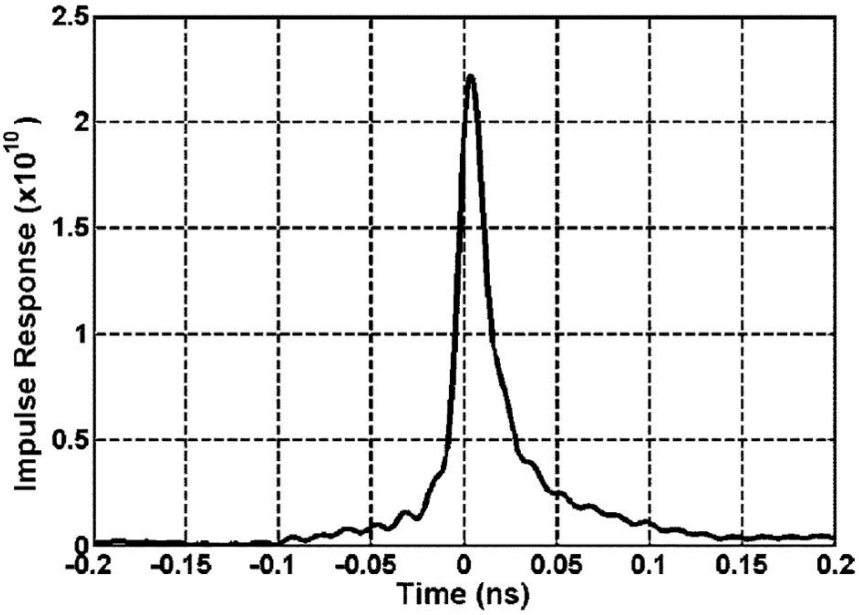


Fig. 5.4 Impulse response of a 5-meter AWG18 cable path.

5.1.4 Reflection

Reflection occurs when a channel presents impedance discontinuities, such as bondwires, vias, connectors, and terminators. Traditionally, reflections are classified into three categories: resistive, capacitive, and inductive [HHMOO]. As we know, the voltage reflection coefficient Γ in a transmission line with characteristic impedance Z_0 and termination Z_L is given by

$$\Gamma = \frac{Z_L - Z_0}{Z_L + Z_0}. \tag{5.19}$$

Lots of components along the data channel can cause reflection. The termination resistance on the TX and RX sides may deviate from the desired value. Parasitic capacitance and inductance in wirebonding and package induce capacitive and inductive reflections. Vias and connectors lead to transmission-line discontinuity as well. Reflections on a practical data link channel is actually a combination of these effects.

An example is made in Fig. 5.5 to illustrates this issue. If the I/O ports on both sides experience termination inaccuracy and parasitics, the output at RX side reveals residual pulses (as returned wave bouncing back and forth) with “strings” (capacitive and inductive reflections).

Obviously, these little pulses contaminate subsequent data bits. Since the perturbation happens after the main bit (i.e., they are post-cursors), a common approach to cancel out reflection is to use a decision feedback equalizer. Also, a proper way to evaluate the reflection effect is to check the s-parameter to evaluate the reflection effect of the channel. In system design, it is suggested to keep the end-to-end S_{11} and S_{22} less than -10 dB over the bandwidth (from dc to data rate or at least Nyquist frequency) in order to preserve signal integrity as much as possible.

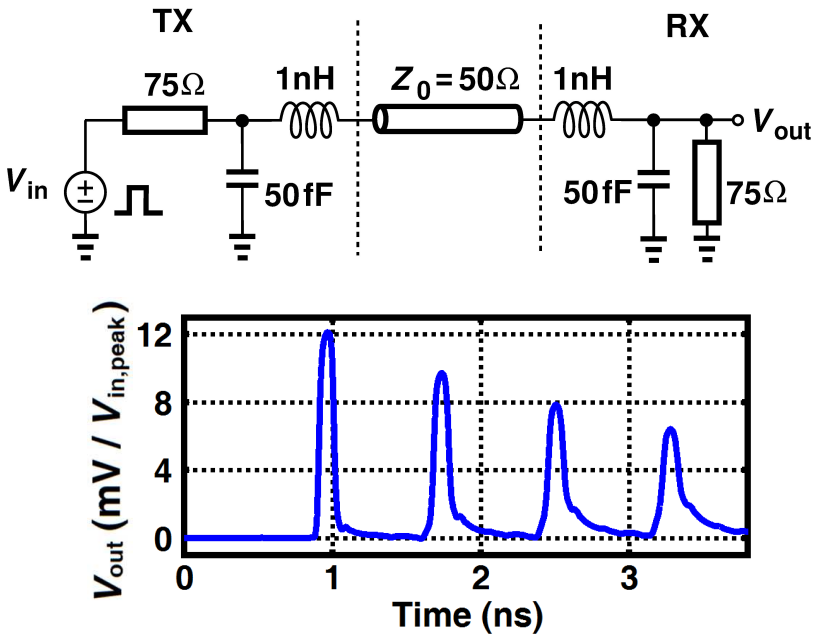


Fig. 5.5 Example of reflection effect on a practical channel.

5.2 CHANNEL CHARACTERISTICS

Now that we realize a typical channel suffers from insertion loss, return loss, crosstalk, and other nonidealities, we need to investigate the response of data stream flowing through it. A typical channel may present more irregular response than what is shown in Fig. 5.5. To further characterize the response, an ideal single pulse $\{\cdots, 0, 0, 1, 0, 0, \cdots\}$ (a ONE preceded and followed by runs of ZEROs with bit period T_b), is applied into the channel. For simplicity, the input's magnitude is

normalized to unity, e.g., 1V. Thus, the output at far-end side could be observed. Defining the peak value $x[0]$ as main cursor, we name the values at kT_b ($k = 1, 2, 3, \cdots$) ahead as "pre-cursors" $x[-1]$, $x[-2]$, $x[-3]$, $\cdots$, and the values at kT_b ($k = 1, 2, 3, \cdots$) behind as "post-cursors" $x[1]$, $x[2]$, $x[3]$, $\cdots$. These cursors are quite important in determining equalizer coefficients. Since they are sampled every T_b seconds, discrete mathematics and digital signal processing will be used in analysis.

It is instructive to see that, if the channel dissipates no dc energy, the total amount of all cursors is equal to unity:

$$\sum_{k=-\infty}^{\infty} x[k] = 1 \, . \tag{5.20}$$

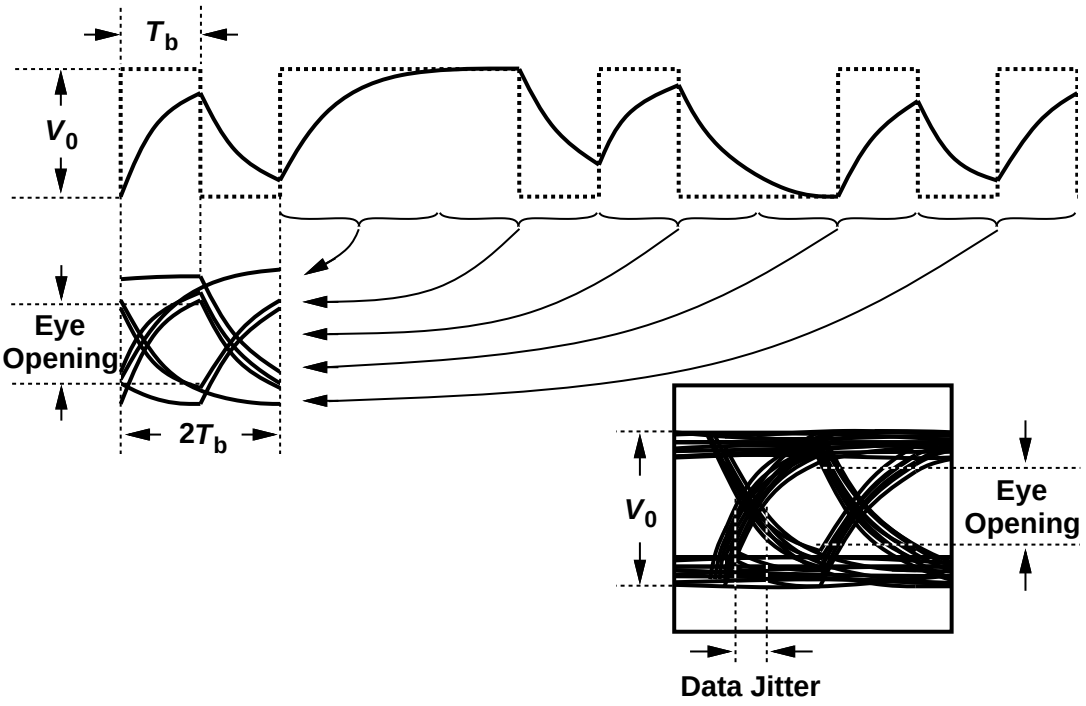


Fig. 5.6 Eye diagram.

A simple approach can demonstrate this property. Consider two worst cases $\{\cdots, 0, 0, 1, 0, 0, \cdots\}$ and $\{\cdots, 1, 1, 0, 1, 1, \cdots\}$ as illustrated in Fig. 5.7(a). If the channel is lossless in dc, it

presents no dc drop, i.e., the output magnitude returns back to 1 if consecutive ONEs are applied into the channel. In other words, these two cases are identical, and the peak in the former and the valley in the latter have the same height $x[0]$. Since the valley of the second case is a combination of all ONE bits, it is equivalent to the sum of all cursors without the main cursor ($= \sum_{k \neq 0} x[k]$). As a result,

$$x[0] + \sum_{k \neq 0} x[k] = 1. \tag{5.21}$$

What happens if a PRBS is applied into a channel with limited bandwidth? The waveform at far-end side would be a sequence of incomplete pulses, as illustrated in Fig. 5.6. By folding it every two bit periods ($2T_b$) and redrawing the waveform, we obtain an eye diagram. The widest opening of the "eye" is named eye opening, usually presented as a percentage as compared with the bit magnitude V_0. Similarly, eye closure is defined as $1-$eye-opening. Also known as inter symbol interference (ISI), the eye closure is indeed caused by incomplete pulses. To see why, let us feed a $\{\cdots, 0, 0, 1, 0, 0, \cdots\}$ sequence into the channel [Fig. 5.7(b)]. Since the output is the linear combination of two pulses, the middle bit is corrupted by both pre-cursor $x[-1]$ and post-cursor $x[1]$. An error occurs if the sum of $x[-1]$ and $x[1]$ is less than 0.5. The bottom line for a channel to deliver detectable data is that the main cursor must be greater than 0.5, if no equalization is employed. Situation becomes more stringent in the presence of additive noise, reflection, crosstalk, and other nonidealities.

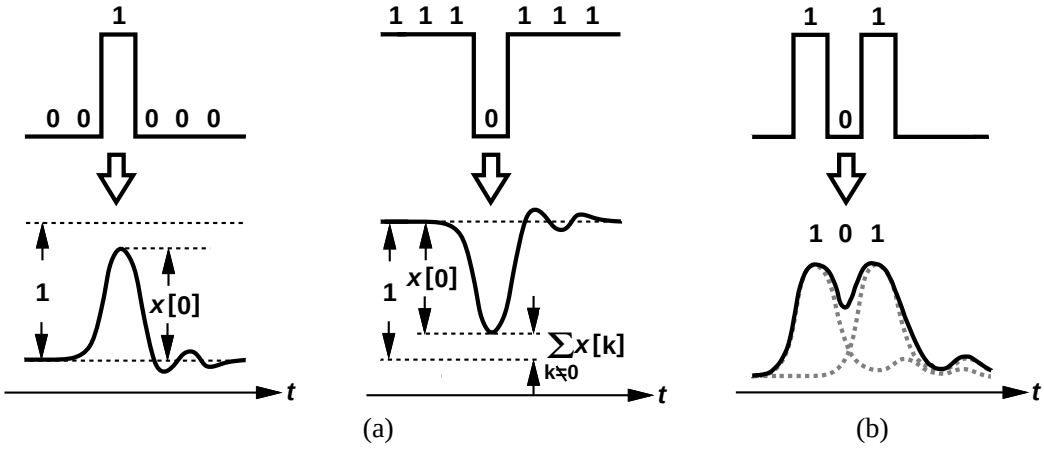


Fig. 5.7 (a) Worst-case scenario of eye closure, (b) ISI accumulation.

Example 5.1

Suppose a channel can be modeled by a RC network with time constant $\tau = \text{RC}$ (Fig. 5.8). (a) Determine the cursors. (b) How much loss at Nyquist frequency would lead to 100% eye closure?

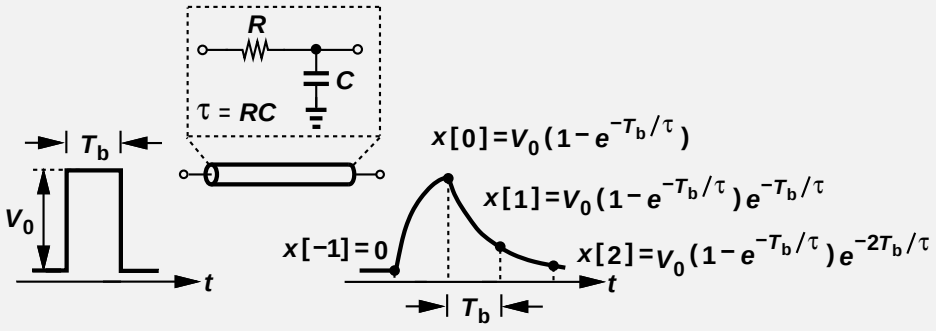


Fig. 5.8 Response to a RC network.

Solution:

(a) A single-pulse input results in an exponentially climbing up for T_b period and exponentially rolling down afterwards. As shown in Fig. 5.8, we have

$$x[0] = 1 - e^{-T_b/\tau} \tag{5.22}$$

$$x[1] = [1 - e^{-T_b/\tau}]e^{-T_b/\tau} \tag{5.23}$$

$$x[2] = [1 - e^{-T_b/\tau}]e^{-2T_b/\tau} \tag{5.24}$$

$$\vdots$$

The sum of all cursors is

$$\sum_{k=-\infty}^{\infty} x[k] = (1 - e^{-T_b/\tau}) \cdot [1 + e^{-T_b/\tau} + e^{-2T_b/\tau} + \ldots] \tag{5.25}$$

$$= 1, \tag{5.26}$$

verifying Eq. (5.20). It is because the output side of the cable is unloaded as no voltage division is presented for dc signal.

Example 5.1 ***(Continued)***

(b) For $\mathrm{x}[0] = 1/2$, we arrive at

$$1 - e^{-T_b/\tau} = \frac{1}{2}, \tag{5.27}$$

$$\frac{T_b}{\tau} = 0.693. \tag{5.28}$$

The RC-network transfer function is given by $H(s) = \dfrac{1}{1+s\tau}$. The magnitude of it at Nyquist frequency becomes

$$\left| H\left(j \cdot 2\pi \frac{1}{2T_b} \right) \right| = \frac{1}{\sqrt{1 + 4\pi^2 \cdot \dfrac{\tau^2}{2^2 T_b^2}}} \tag{5.29}$$

$$= 0.215. \tag{5.30}$$

That is, the channel loss at Nyquist frequency must be higher than 13.3 dB.

The above analysis manifests the importance of equalizers. Recall our discussion on bit error rate. Errors begin to occur if the eye opening is degraded to 14 times of rms noise or less, which may correspond to only $6 \sim 8$ dB loss at Nyquist frequency. In typical backplane applications, for example, users are looking at transceivers with $20 \sim 30$ dB loss tolerance at Nyquist frequency. Signal loss at high frequencies must be recovered as much as possible.

Three different kinds of equalizers are commonly used for high-speed data links: feedforward equalizers (FFEs), continuous-time linear equalizers (CTLEs), and decision-feedback equalizers (DFEs). While the FFEs are usually put at the TX side, the other two are placed in the RX side with adaptability in most cases. FFEs and CTLEs are linear equalizers, whereas DFEs are non-linear. Equalizers are more pronounced for signal attenuation and dispersion, and are somewhat useful for reflection. However, it may worsen the crosstalk. Equalizers sometimes have to be codesigned with other building blocks in order to work properly, e.g., FFE with output driver and DFE with CDR. Equalization is not a panacea. It rescues damaged signal and restores signal integrity to

some extent. The unsolved issues (e.g., jitter) would be taken care of by the subsequent blocks.

We look at the different types of equalizers in the following sections.

5.3 FEEDFORWARD EQUALIZERS

Feedforward equalizers come from the idea of finite-impulse response (FIR) filters. With proper parameter setting, we are capable of generating a high-pass response from dc to Nyquist frequency with desired shape to compensate for the channel loss. Consider the simplest FIR filter structure with two taps as shown in Fig. 5.9(a). The input signal $x(t)$ is delayed by T_b, and the output $y(t)$ is the sum of $x(t)$ and $x(t - T_b)$ with weighting 1 and α, respectively. The transfer function is given by

$$H(s) = 1 + \alpha e^{-sT_b}, \tag{5.31}$$

and its magnitude and phase are

$$|H(j\omega)| = \sqrt{(1 + \alpha^2) + 2\alpha \cos(\omega T_b)} \tag{5.32}$$

$$\measuredangle H(j\omega) = \tan^{-1}\left[\frac{-\alpha \sin(\omega T_b)}{1 + \alpha \cos(\omega T_b)}\right]. \tag{5.33}$$

Indeed, as $-1 < \alpha < 0$ we arrive at a transfer function whose magnitude raised up monotonically from dc to $1/(2T_b)$. It is intuitively obvious to see that if more taps are used, a more delicate shape can be generated. The channel loss would be compensated more completely. An FFE with $2M + 1$ taps are shown in Fig. 5.9(b). In discrete-time systems, it is represented as

$$y[n] = \alpha_{-M} x[n] + \alpha_{-M+1} x[n + 1] + \ldots + \alpha_M x[n - 2M]. \tag{5.34}$$

The transfer function in Z-domain becomes

$$H[z] = \alpha_{-M} + \alpha_{-M+1} \cdot z^{-1} + \cdots + \alpha_M z^{-2M} \tag{5.35}$$

$$= \sum_{k=0}^{2M} \alpha_{-M+k} z^{-k}. \tag{5.36}$$

Note that an FFE can have any number of taps (which need not be odd number). Readers can easily link the S- and Z-domain by $z = e^{sT_b} = e^{j\omega T_b}$. Since an FFE is an FIR filter, it is unconditionally stable.

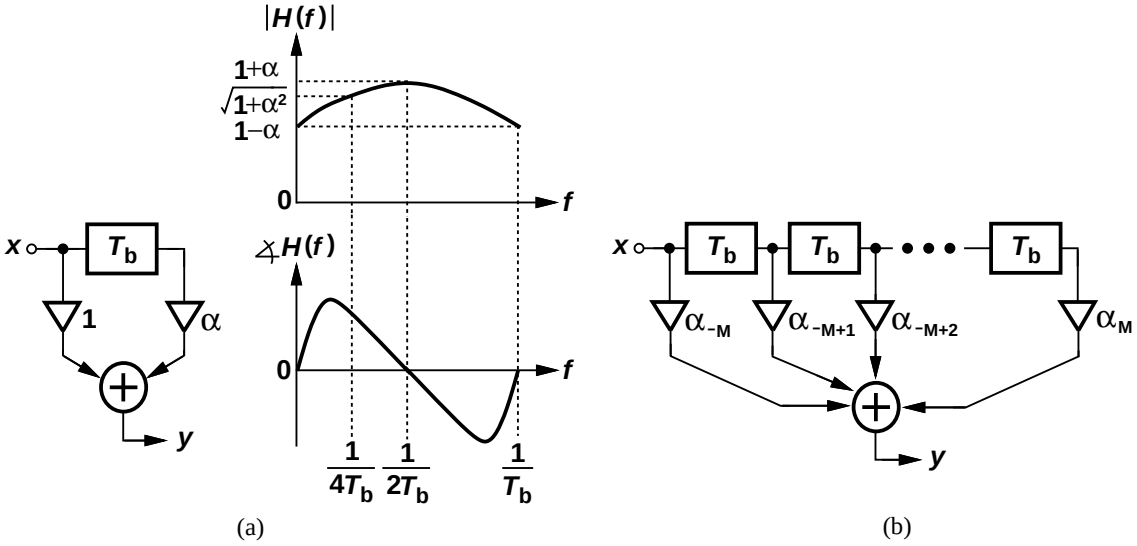


Fig. 5.9 (a) Two-tap FFE, (b) (2M+1)-tap FFE.

Example 5.2

If a channel can be modeled as a simple RC network with time-constant $\tau = \text{RC}$. Use a two-tap FFE to optimize the output y.

Solution:

The waveform at different points are redrawn in Fig. 5.10. If we choose

$$\alpha = -\exp\left(\frac{-T_p}{2}\right), \tag{5.37}$$

then $y(t)$ will exponentially climb up from 0 to $1+\alpha$ in a period of T_b. In that sense all transitions are degenerated to two traces, and no ISI is observed. In other words, we have 100% eye opening and no jitter in the eye diagram. It can be proven that even with a distributed RC model, an FFE can still perfectly restore the waveform by two taps.

Example 5.2 ***(Continued)***

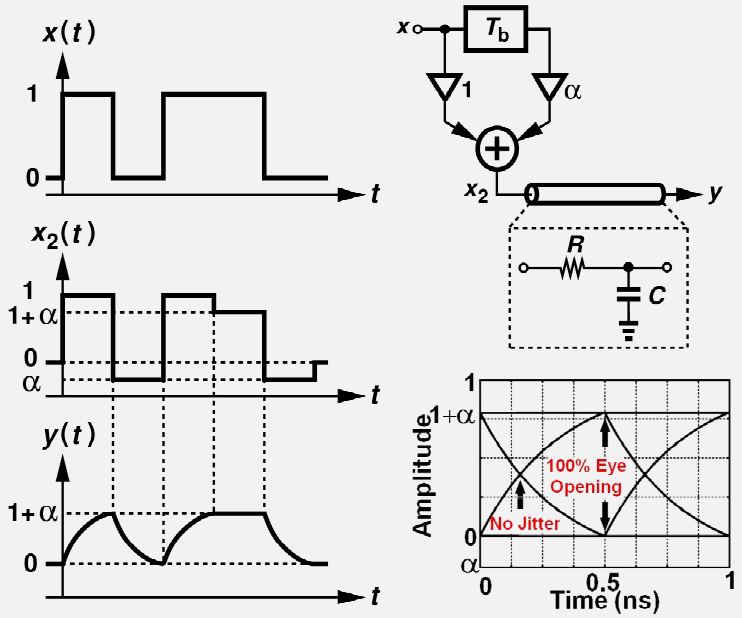


Fig. 5.10 Using 2-tap FFE to equalize a channel.

In reality, it is impossible to clean up all cursors by using two taps. The RC model is too simple to represent the actual response of a channel.

But how to setup coefficients for the taps? To answer this question, we resort to time-domain analysis. Let us start with a 3-tap FFE as shown in Fig. 5.11. The single-pulse response of the channel is also characterized as $\{\cdots, 0, 0, 0.2, 0.85, -0.2, 0.1, 0.05, 0, 0, \cdots\}$. Since any pulse would form such a response at the far end and y is a linear combination of all responses, we could adjust weightings (coefficients) to eliminate pre-cursors and post-cursors as much as possible. For this 3-tap FFE, we have

$$\begin{bmatrix} 0.85 & 0.2 & 0 \\ -0.2 & 0.85 & 0.2 \\ 0.1 & -0.2 & 0.85 \end{bmatrix} \begin{bmatrix} \alpha_{-1} \\ \alpha_0 \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}. \tag{5.38}$$

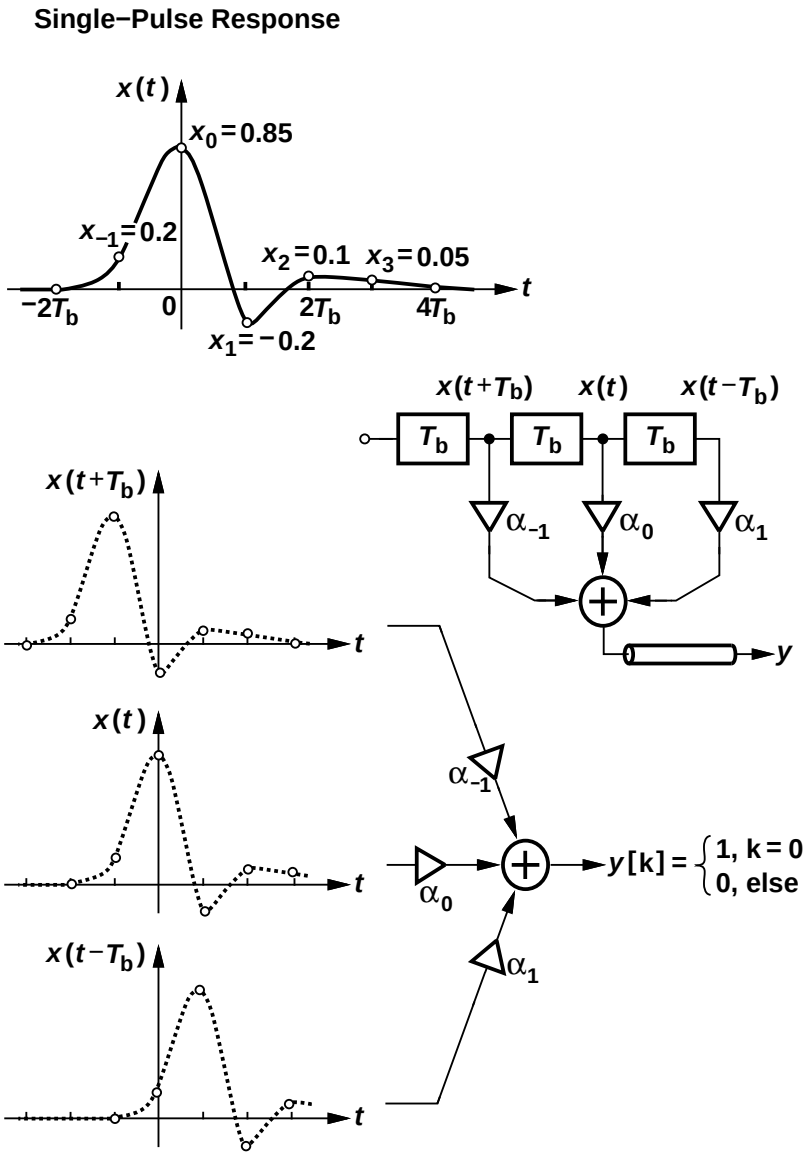


Fig. 5.11 Calculating 3-tap FFE coefficients.

It follows that

$$\begin{bmatrix} \alpha_{-1} \\ \alpha_0 \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} -0.25 \\ 1.05 \\ 0.28 \end{bmatrix}. \tag{5.39}$$

Here all cursors but main cursor are set to 0, so ISI is minimized. Inevitably, some cursors can not be cleaned up due to the limited number of taps. The more taps an FFE has, the wider range of cursors it correct. Since the same rules apply to whole data stream, a clean eye diagram is expected.

Also known as "zero-forcing" technique, this method can be extended to

$$\begin{bmatrix} x_0 & x_{-1} & \dots & x_{-M} & \dots & x_{-2M+1} & x_{-2M} \\ x_1 & x_0 & \dots & x_{-M+1} & \dots & x_{-2M+2} & x_{-2M+1} \\ \vdots & \vdots & & & & & \\ x_M & x_{M-1} & \dots & x_0 & \dots & x_{-M+1} & x_{-M} \\ \vdots & \vdots & & & & & \\ x_{2M-1} & x_{2M-2} & \dots & x_{M-1} & \dots & x_0 & x_{-1} \\ x_{2M} & x_{2M-1} & \dots & x_M & \dots & x_1 & x_0 \end{bmatrix} \begin{bmatrix} \alpha_{-M} \\ \alpha_{-M+1} \\ \vdots \\ \alpha_0 \\ \vdots \\ \alpha_{M-1} \\ \alpha_M \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \tag{5.40}$$

which nulls M pre-cursors and M post-cursors. Be aware that we only force the sampling points and pay no attention to transitions. Therefore, jitter is not considered.

The output combiners of FFEs are usually realized in current mode because (1) coefficient can be assigned and summed up easily; (2) the CML structure provides great bandwidth and impedance matching. Figure 5.12 illustrates a typical realization. Here, three taps are combined in current, which is converted to voltage by the loading resistors. In order to minimize the parasitic capacitance associated with the output port, each differential pair is sized to accommodate its potential maximum current. Similarity, switches are placed in front of differential pairs so that the signs of coefficients could be changed.

At the speed beyond 20 Gb/s, a full-rate FFE inevitably dissipates significant power, because every single block in it has to be made in CML. Half-rate architecture, however, can leverage against the stringent speed requirement and save considerable power. It is primarily because the half-rate data and clock can usually be handled in digital domain. Inverter-based buffers and TSPC

flipflops save significant powers. A typical half-rate FFE is depicted in Fig. 5.13. Here, two sub-rate inputs are fed into their delay paths directly, alternately coming out for data combination. Note that clock is driven in half-rate as well, so the delay elements are latches (i.e., half a flipflop) instead of a full flipflop.

To fairly the power of full-rate and half-rate structures, we implement both FFEs in 65-nm CMOS. Their power consumptions for different portion are listed in detail. In half-rate FFE, all signal (including data and clock) are rail-to-rail except the output driver. TSPC latches, high-speed clock drivers, and 2-to-1 selectors are discussed in Chaper xx.

Choosing a proper number of FFE taps at 20^+ Gb/s involves the tradeoff between bandwidth and signal integrity. For a CML combiner, adding more taps implies an almost linear increase of parasitic capacitance at the output node. If we denote C_0 and C_1 as the capacitance caused by main tap and pre/post taps, the total parasitic capacitance of an N-tap combiner is estimated as

$$C = C_0 + (N-1)C_1. \tag{5.41}$$

Since bandwidth is inversely proportional to C, the maximum data rate would roll off as the number of taps increases. On the other hand, for large-signal operation, the output eye opening is more important. Even with identical boosting at Nyquist frequency, an FFE with less taps suffers from larger ISI. Fig. 5.14 illustrates the bandwidth and ISI effects for a typical FFE designed in 65-nm CMOS technology. Here, we set the total tail currents of all taps to 12 mA (which corresponds to a maximum swing of 300 mV). Transistor-level simulation suggests that to keep sufficient bandwidth, the tap number N must be less than or equal to 4. Meanwhile, it requires at least 3 taps so as to maintain an eye opening larger than 75%.

The half-rate FFE architecture inevitably suffers from pulse-width distortion if the clock presents duty cycle other than 50%. It is obvious that multiple traces would appear in the output data and cause jitter. Duty-cycle correction circuit and careful layout can minimize this issue.

How to design an FFE at even higher data rate, say, 40 Gb/s? At such a high speed, the half-rate structure has marginal advantage, since 20 GHz/20 Gb/s is still too fast for CMOS logics. Figure 5.15 depicts one design example, where a 4-tap full-rate FFE is demonstrated. Quite a few difficulties may arise. First, for a flipflop to operate at 40 Gb/s, output buffers must be added

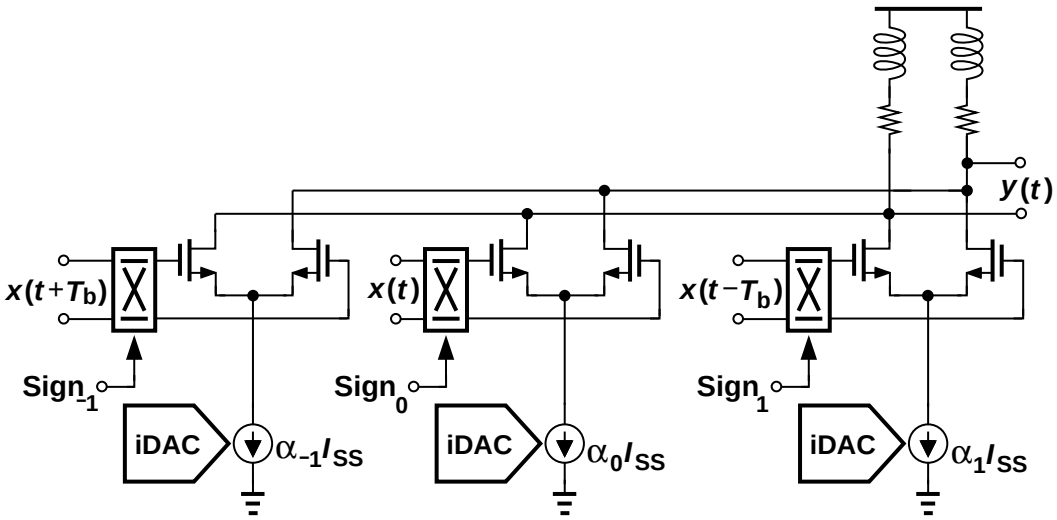


Fig. 5.12 Typical combiner.

in order to drive the combiner and the next flipflop. Even with a CML structure, the parasitic still causes serious problems. It creates a clock-to-Q delay as large as 15 ps in 65-nm CMOS, which is very significant to one bit period (25 ps). As a result, the next flipflop suffers from misalignment [Fig. 5.15(b)], i.e., the data output will be shifted to the right and thus the clock edge no longer falls in the center of the data eye. The flipflop would have insufficient time for data regeneration, resulting in inter-symbol interference (ISI). In addition, we need a large clock buffer tree to drive the loading, which not only consumes significant power but increases layout difficulties. Experimental results suggest that this approach achieves maximum data rate around 30$\sim$35 Gb/s. Even though we put delays in the data and clock paths to cancel out the clock-to-Q delay, the overall performance is still limited by the bandwidth of the flipflops. A much better solution in 40-Gb/s FFE is to use passive elements as delay unit. We discuss the details in Case Study.

We study an important property to finish the discussion on FFEs. Let us consider a general FFE with N taps as shown in Fig. 5.16(a). In real circuit implementation such as a CML summer, we usually require a constant tail current in the output driver so as to keep a fixed common-mode level. That means the sum of all coefficients (the absolute values) is a constant, which can be

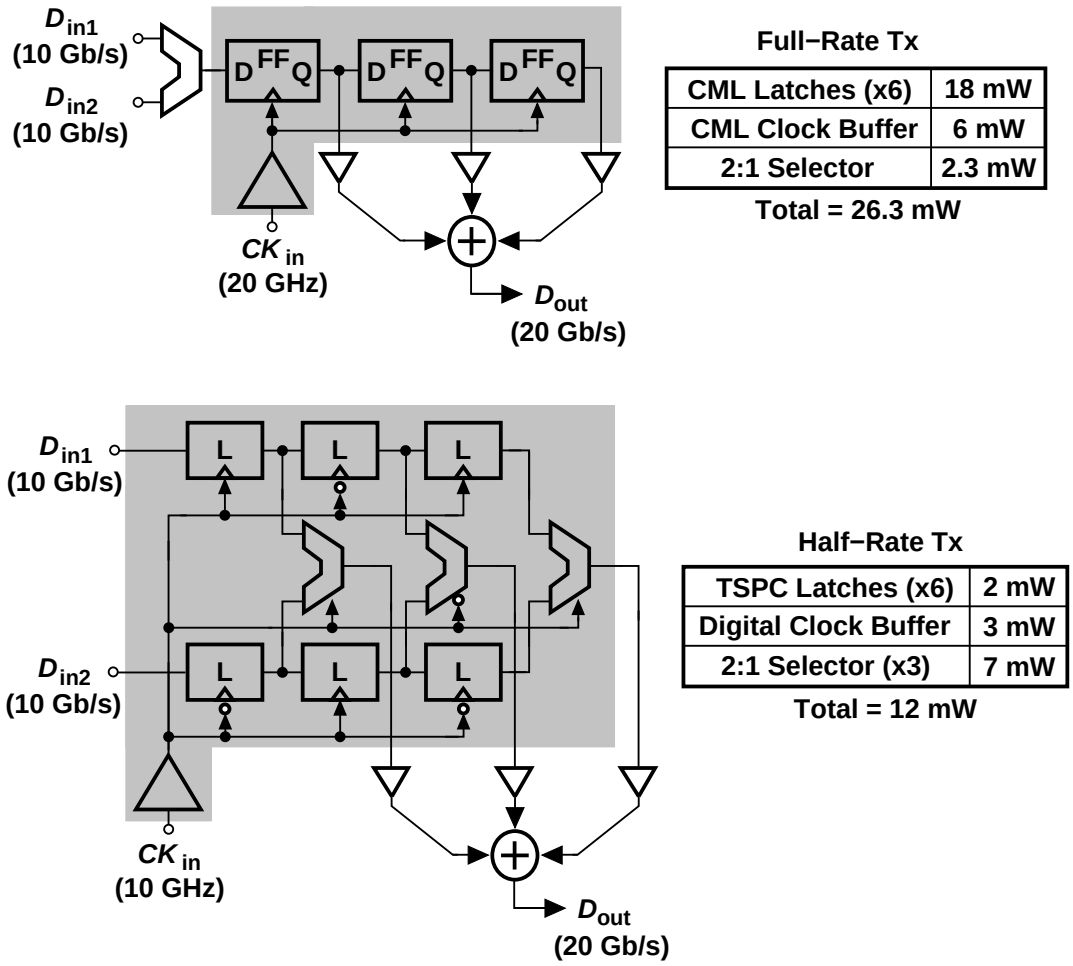


Fig. 5.13 Comparison between full-rate and half-rate FFEs.

normalized to unity. We thus denote the first tap as $1 - \sum_{k=1}^{N-1} |\alpha_k|$, while the second to the Nth taps as $+\alpha_1, +\alpha_2 \cdots + \alpha_{N-1}$, respectively. We arrive at the transfer function from x to y:

$$H(z) = 1 - \sum_{k=1}^{N-1} |\alpha_k| + \sum_{k=1}^{N-1} \alpha_k z^{-k}. \tag{5.42}$$

To gain more insight, we convert the above discrete analysis to continuous domain. That is,

$$H(j\omega) = 1 - \sum_{k=1}^{N-1} |\alpha_k| + \sum_{k=1}^{N-1} \alpha_k \exp(-jk\omega T_b), \tag{5.43}$$

where T_b denotes the bit period. Equation (5.43) implies important properties for the maximum boost that an FFE can provide. Now, if we keep the total amount of all coefficients other than the

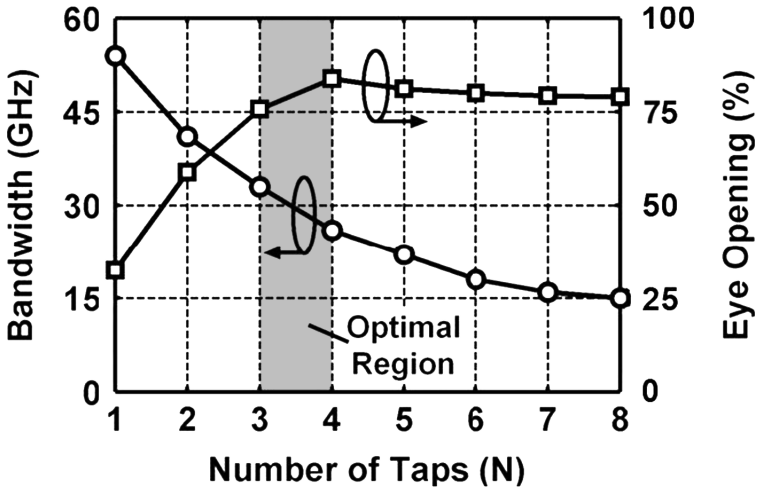


Fig. 5.14 Bandwidth and eye opening of FFE with different tap number.

first one as K (i.e., $K = \sum_{k=1}^{N-1} |\alpha_k|$), the maximum boost at Nyquist frequency $1/(2T_b)$ becomes

$$\frac{H(j\frac{\pi}{T_b})}{H(j0)} = \frac{1 - K + \sum_{k=1}^{N-1} \alpha_k (-1)^k}{1 - K + \sum_{k=1}^{N-1} \alpha_k} \tag{5.44}$$

$$\leq \frac{1}{1 - 2K}\,, \quad \left(0 < K < \frac{1}{2}\right) \tag{5.45}$$

The equation holds when $\alpha_1 < 0$, $\alpha_3 < 0 \cdots$ and $\alpha_2 = \alpha_4 = \cdots = 0$. In other words, if the current ratio between the first tap and the other taps is a constant, the maximum boosting is also a constant, regardless of the number of taps. Note that K must locate between 0 and 1/2 in order to perform high-frequency boosting. For $K > 1/2$, the response actually presents an attenuation rather than a boost for high frequencies. Depending on the dc loss that a system can tolerate, we can select an optimal K. For example, if a minimum amplitude as large as 1/3 of the original full swing is acceptable for the receiver, we have $K = 1/3$ and the FFE creates 9.5 dB boost at $1/(2T_b)$. Using more taps would help reshape the response and better fit the inverse of the channel loss, but give no additional compensation. Actually as expected, the more taps we use, the better the response fits into the desired response. With $K = 0.18$ and 0.3, we plot the response of different FFEs having 2, 3, and 4 taps [Fig. 5.16(b)]. Here, the dc gain is normalized to unity for a fair comparison. The

FFE with 3 or more taps reveals sufficient fitting quality to the compensation curve, whereas the one with 2 taps provides only limited accuracy. Note that the desired response here is obtained by transforming the pulse response of a 20-cm FR4 channel into frequency domain.

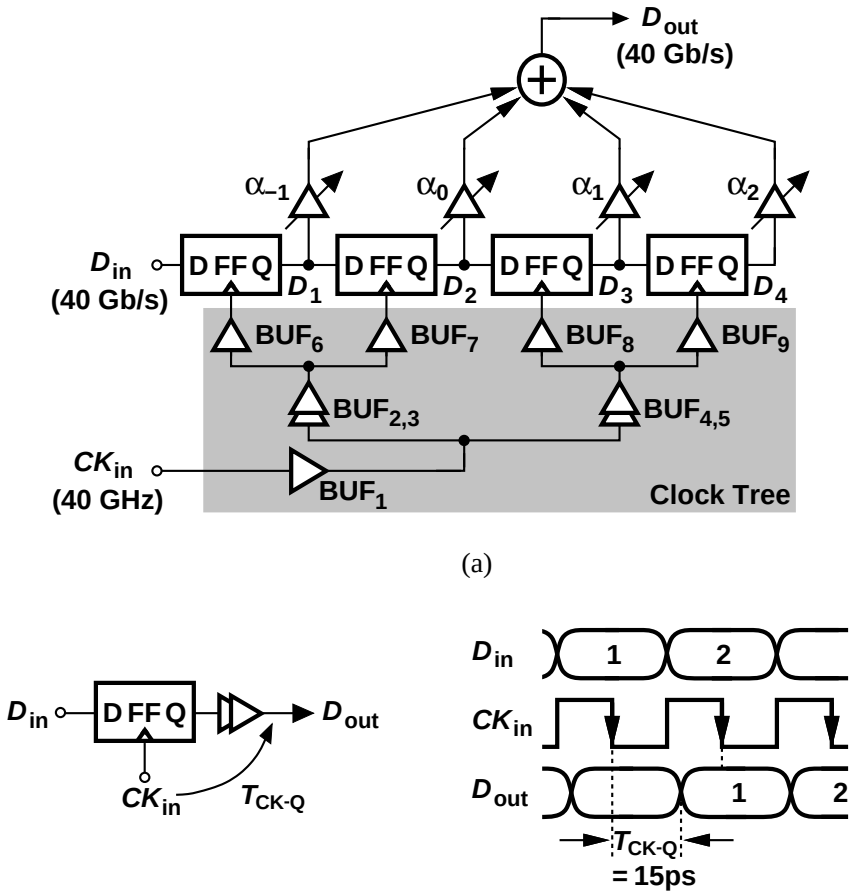


Fig. 5.15 40-Gb/s FFE design: (a) architecture, (b) phase misalignment issue due to clock-to-Q delay.

5.4 CONTINUOUS-TIME LINEAR EQUALIZERS

5.4.1 Boosting Filters

Perhaps the most intuitive and straight-forward approach to compensate for high-frequency loss is a continuous-time linear equalizers (CTLEs). It is quite obvious that we need a high-pass

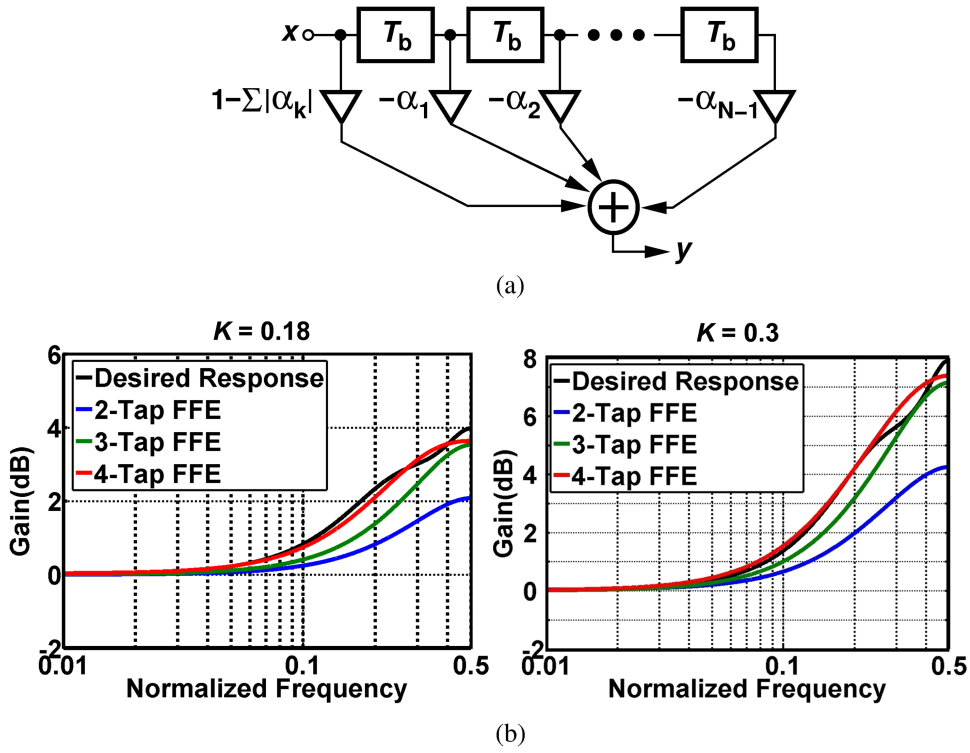


Fig. 5.16 (a) N-tap FFE with constant coefficient amount, (b) frequency response.

filter (from low frequency to Nyquist) M order to boost up the high frequency response. A simple RC filter [Fig. 5.17(a)] provides such a characteristic, but it can not be used in data path simply because the dc gain is 0. To tolerate long runs a dc path must be created between input and output. Fig. 5.17(b) illustrates an example, where two resistors and two capacitors (C_1, C_2) form a transfer function as:

$$H(s) = \frac{R_2}{R_1 + R_2} \cdot \frac{1 + R_1 C_1 s}{1 + \dfrac{R_1 R_2}{R_1 + R_2}(C_1 + C_2)s}. \tag{5.46}$$

One zero and one pole are created such that the voltage gain raises from $R_2/(R_1+R_2)$ to $C_1/(C_1+C_2)$. While providing decent linearity, this passive implementation reveals no signal gain but loss. As a result, the SNR would degenerate significantly.

Can we utilize the peaking technique introduced in Chapter xx to create boosting at high frequencies? Indeed, an underdamped inductive peaking presents a peaking response, as illustrated

in Fig. 5.17(c). Denoting the transconductance of $M_{1,2}$, peaking inductor, the loading resistor, and the parasitic capacitance as $g_{m1,2}$, L, R_S, and C, respectively, we obtain the transfer function

$$\frac{V_{\text{out}}}{V_{\text{in}}} = g_{m1,2}\, R_S \cdot \frac{s + 2\zeta\omega_n}{s^2 + 2\zeta\omega_n s + \omega_n^2} \cdot \frac{\omega_n}{2\zeta}, \tag{5.47}$$

where $\omega_n^2 = (LC)^{-1}$ and $\zeta = (R_S/2)\sqrt{C/L}$. That is, the voltage gain goes up from $g_{m1,2}\, R_S$ to $g_{m1,2}\, R_S/(4\zeta^2)$, forming a ramp of approximately 40 dB/dec.If $\zeta = 0.2$, the total boosting is equal to 16.5 dB. However, a fatal drawback prevents it from being widely used in equalizers. The complex conjugate poles in Eq. (5.47) results in a ringing phenomenon. No matter what kind of input is applied, the output always contains the term

$$y_1(t) = e^{-\zeta\omega_n t}\cos(\sqrt{1-\zeta^2}\,\omega_n t). \tag{5.48}$$

Since ζ is pretty low, the cosine wave decays slowly and creates significant ISI. By the same token, time-domain jitter in the crossover region is severely large. Figure 5.18 shows data eye diagrams before and after an underdamped peaking circuit with 10-dB boosting at Nyquist frequency. A much better way to implement a boosting filter is capacitive degeneration. Illustrative in Fig. 5.19, it has resistors and capacitors (varactors) inserted into the common-source node of M_1-M_2 pair. Denoting the loading and degeneration resistors and capacitors as R_D, R_S, C_L, and C_S, respectively, we obtain the transfer function

$$\frac{V_{\text{out}}}{V_{\text{in}}}(s) = \frac{g_{m1}R_D}{1 + \dfrac{g_{m1}R_S}{2}} \cdot \frac{1 + \dfrac{s}{\omega_{z1}}}{\left(1 + \dfrac{s}{\omega_{p1}}\right)\left(1 + \dfrac{s}{\omega_{p2}}\right)}, \tag{5.49}$$

where $\omega_{z1} = 1/(R_S C_S)$, $\omega_{p1} = (1+g_{m1}R_S/2)/R_S C_S$, $\omega_{p2} = 1/(R_D C_L)$, and g_{m1} the transconductance of M_1. A typical plot of the transfer function is also depicted in Fig. 5.19(c). To continuously tune the boosting, a control voltage (V_{ctrl}) is applied to the MOS resistor and varactors. As V_{ctrl} goes up, both R_S and C_S go down, leading to a milder boost and vice versa [Fig. 5.19(d)]. Note that in tuning, ω_{p1} shifts in the same direction as ω_{z1} does, but with minor movement. Meanwhile, readers shall be aware that the boosting at high frequencies is actually accomplished by suppressing the low-frequency port. Additional amplifier stages need to be added so as to maintain data swing. Since the two poles of Eq. (5.49) are real, ringing issue is minimized.

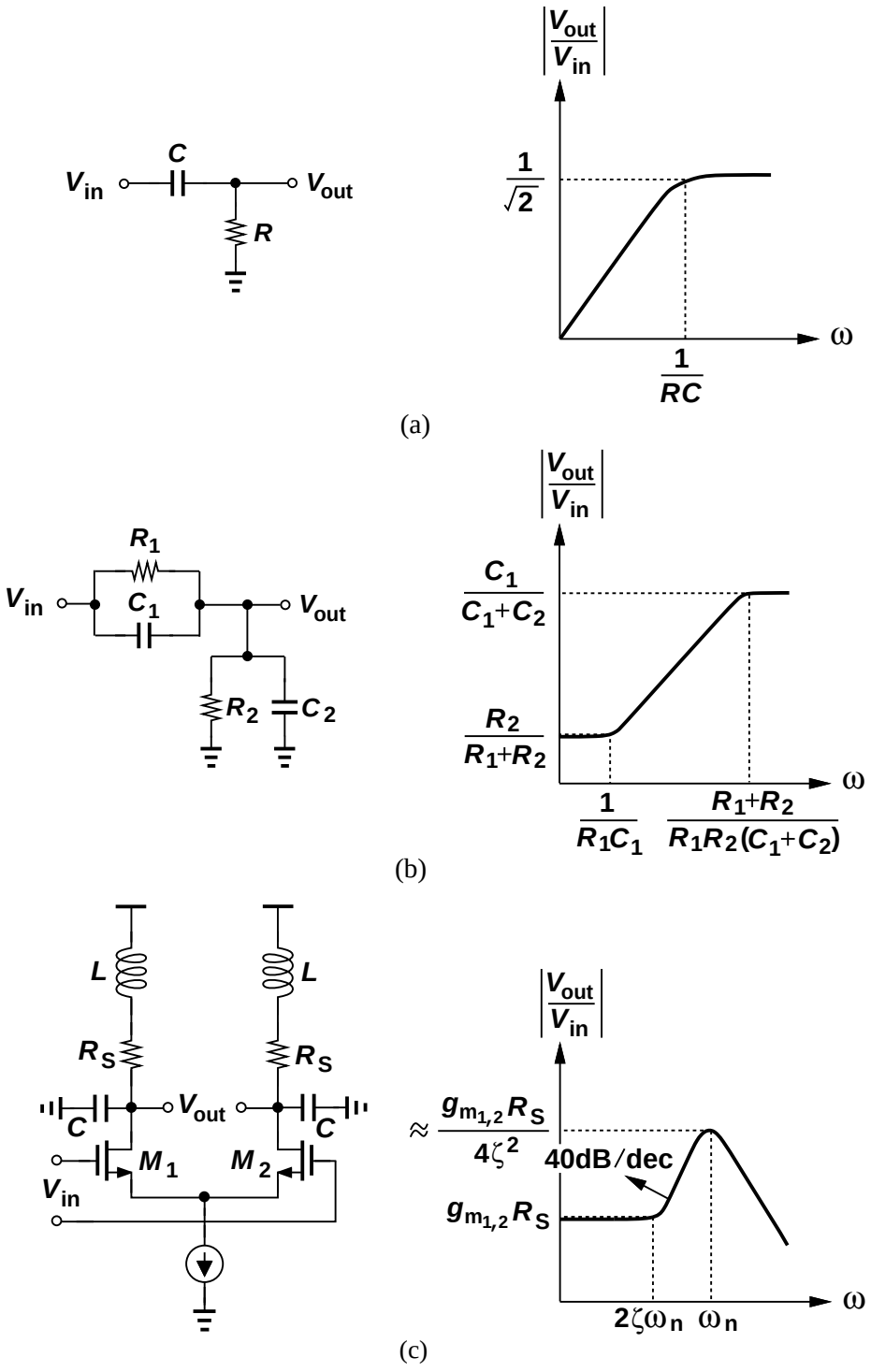


Fig. 5.17 Implementing high-pass filters (a) simple RC, (b) double RC network, (c) underdamped peaking.

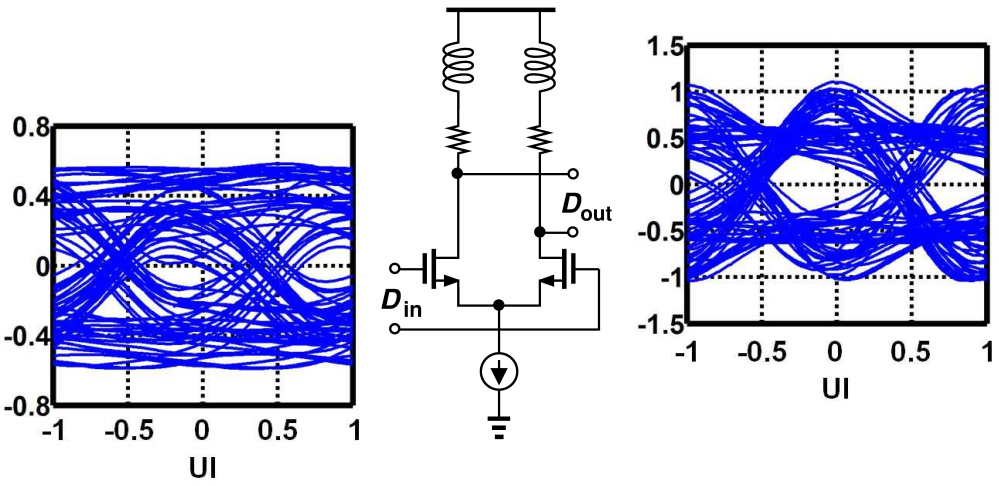


Fig. 5.18 Data eye diagrams before and after an underdamped peaking circuit.

The above topology, however, still suffers from limited bandwidth and insufficient compensation at high frequencies. It is because ω_{p1} exceeds ω_{z1} by a factor of $1 + g_{m1}R_S/2$, and the dc gain drops by the same amount of factor. In other words, $g_{m1}R_S$ must stay low so as to avoid large dc loss. This issue limits the maximum achievable boost in magnitude and phase. For example, if $g_{m1}R_D = g_{m1}R_S = 2$ and $\omega_{p2} = 4\omega_{z1}$, the maximum magnitude is only 3.3 dB. Such a filter fails to provide reasonable performance at high data rate even with multiple stages in cascade. A modification can be found when we introduce inductive peaking in the output of the filter [Fig. 5.20(a)]. We arrive at a new transfer function:

$$\frac{V_{\text{out}}}{V_{\text{in}}}(s) = \frac{g_{m1}R_D}{1+\dfrac{g_{m1}R_S}{2}} \cdot \frac{1+\dfrac{s}{\omega_{z1}}}{1+\dfrac{s}{\omega_{p1}}} \cdot \frac{1+\dfrac{s}{\omega_{z2}}}{1+\dfrac{2\zeta}{\omega_n}s+\dfrac{s^2}{\omega_n^2}}, \tag{5.50}$$

where $\omega_{z2} = 2\zeta\omega_n$, $\zeta = (R_D/2)\sqrt{C_L/L_P}$, $\omega_n = 1/\sqrt{L_P C_L}$, and ω_{z1} and ω_{p1} remain unchanged. This configuration creates a second zero, ω_{z2}, extending the gain boosting and phase compensation at high frequencies by canceling the first pole ω_{p1} [Fig. 5.20(b)]. It can be shown that for $g_{m1}R_D = g_{m1}R_S = 2$, $\omega_{z2} = \omega_{p1} = 2\omega_{z1}$, and $\omega_n = 4\omega_{z1}$, the maximum magnitude compensation (for one stage) are equal to 18 dB. Note that the peaking inductors have only perform critical damping in

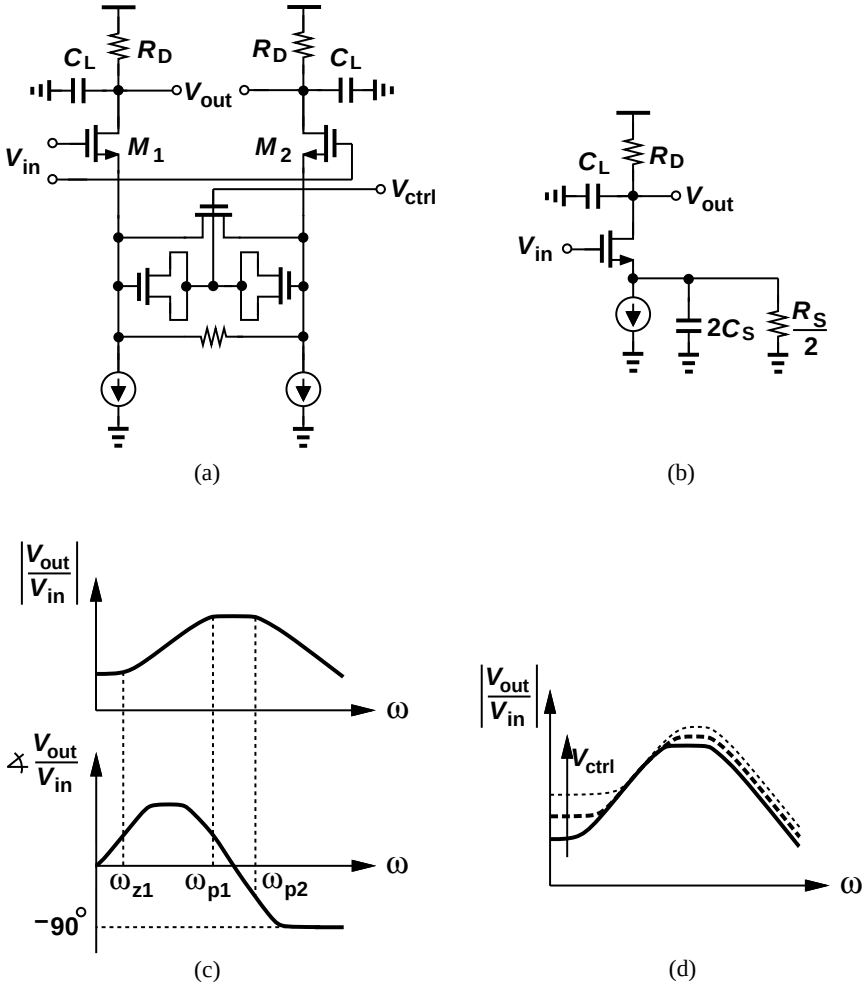


Fig. 5.19 (a) Filter stage with capacitance degeneration, (b) single-ended model, (c) it's response, (d) tuning.

Eq. (5.50). Therefore, it contributes little ringing. It is in contrast to the filter in Fig. 5.17(c), whose boosting entirely relies on underdamped peaking.

5.4.2 Architecture and Adaptation

A typical CTLE architecture is depicted in Fig. 5.21. Usually we require two or more stages of boosting filters and gain stages in order to accommodate high loss cases. To keep reasonable swing along the data path, it is recommendable to place these stages alternately. Since a CTLE is

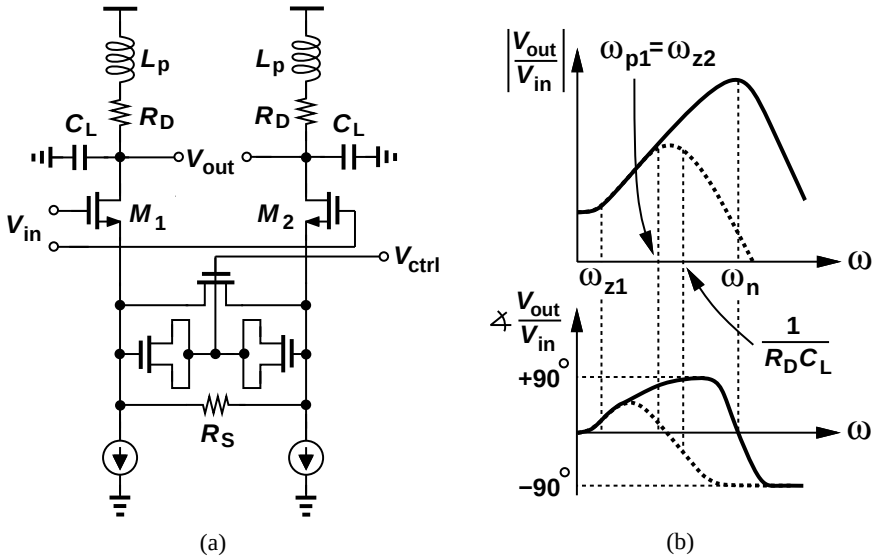


Fig. 5.20 (a) Filter stage with inductive peaking, (b) its transfer function.

usually adapted in RX side, it is preferable to include adaptability. Conventional designs incorporate a slicer and a power detector to detect whether the boosting is optimal. Other structures are introduced in section 5.4.xx.

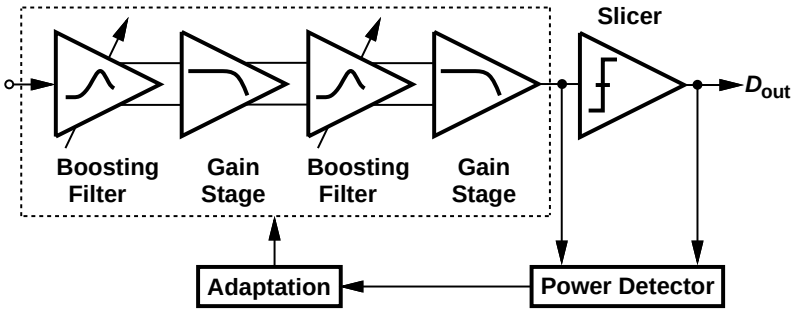
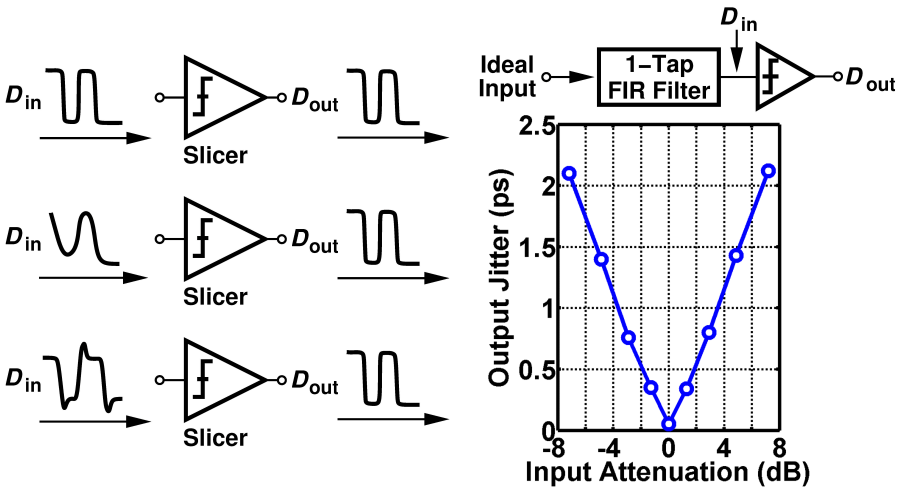


Fig. 5.21 Typical CTLE architecture.

How do we recognize the compensation of an CTLE is optimized? To answer this question, we need to understand what a slicer is. A slicer is defined as a buffer which has (1) high gain, (2) large bandwidth, and (3) capability to clean up ISI. As illustrated in Fig. 5.22, a slicer "restores" the input data (no matter it is under- or over-compensated) to an ideal pulse sequence with minimal ISI.[1] A

[1]Assumed the input at least has eye opening.

slicer is nothing more than a digitizer, although we are dealing with CML data in most cases. A differential pair with inductive peaking may serve as a slicer, which saturates the output to a full CML swing. The key point is that even though a slicer can fix an incomplete or overshooting waveform, large jitter still remains. A minimum jitter appears at the output of the slicer only if the input data is critically (perfectly) compensated. To see this effect, let us consider the setup shown in Fig. 5.22(b). An ideal data is over- or under-compensated by 2-tap FIR filter, which presents -7 to $+7$ dB boosting at Nyquist frequency. Applying this result in Fig. 5.22(b). The jitter reaches a minimum as the slicer's input ($D_{\rm in}$) is very close to an ideal pulse. In other words, we can optimize the boosting by checking the similarity of the waveforms before and after a slicer. Once the slicer's input is as good as the output, we conclude that the filter provides a optimal compensation and the final output jitter is minimized. The adaptation criteria is that the pulse sequence before and after the slicers present similar power spectral density. The more alike, the better.



(a) (b)

Fig. 5.22 Slicer's response in large signal, (b) output jitter as a function of input data integrity.

With the above analysis, we introduce dual-loop adaptation method here. To measure likeness between $D_{\rm in}$ and $D_{\rm out}$ in Fig. 5.22(b), we resort to the comparison of their spectra. As shown in Fig. 5.23, the two sinc functions are first lined up at dc power, i.e., $\mathrm{A} = \mathrm{A}'$. Once it is achieved, we compare the high-frequency power (e.g., the power beyond certain point f_c) and adjust the

boosting accordingly. We optimize the boosting steady and minimize the output data jitter after the loop converges to a state. Since the adjustment is taken place all the time, an adaptive equalization is achieved.

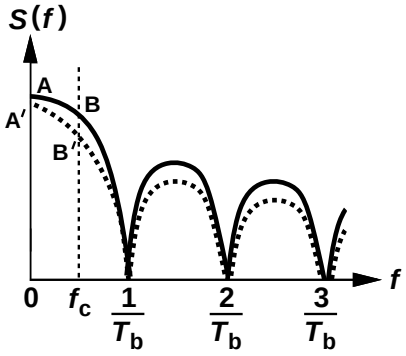


Fig. 5.23 Spectrum for equalizers adaptation.

Figure 5.24 illustrates two examples of adaptive equalizers utilizing this method. In Fig. 5.24(a), the input data goes through two amplifiers before entering the slicer: a broadband amplifier (upper) and a boosting amplifier (lower). Since the slicer's output is fixed, the upper loop tunes the dc gain to make $A = A'$. The lower loop, on the other hand, adjusts the boosting to make $B = B'$. Note that the broadband amplifier must be real broadband with respect to the data rate. Figure 5.24(b) presents another example. Here, the dc points are equalized by tuning the slicer's tail current. High-frequency power are compared to optimize the boosting in the second loop.

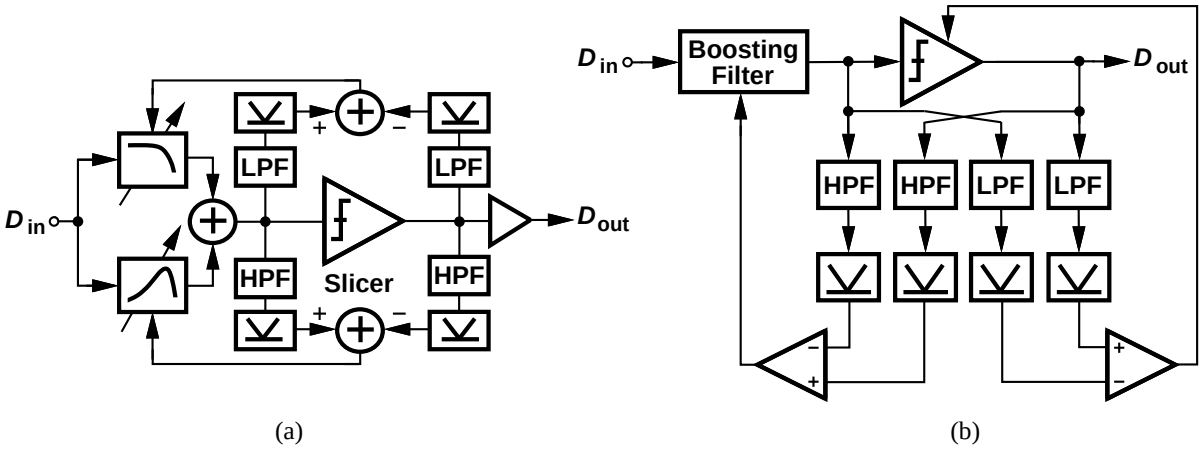


Fig. 5.24 CTLE adaptation examples.

Example 5.3

A popular way to do power detection is to take the common source of a differential pair as the output (Fig. 5.25). Derive the average V_{out} with the assumption that M_1 and M_2 are completely switched.

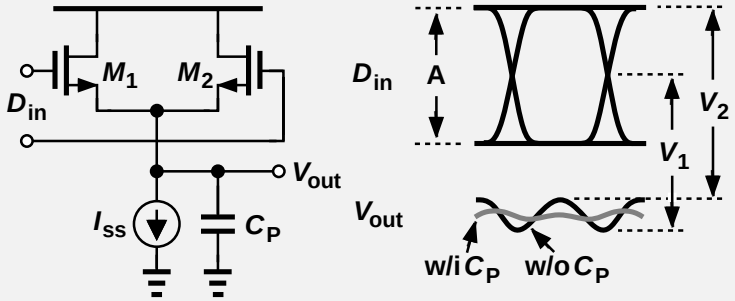


Fig. 5.25 Power detector.

Solution:

From basic eletronics we have

$$V_1 = \sqrt{\frac{I_{ss}}{\mu_n C_{ox}(W/L)_{1,2}}} \tag{5.51}$$

$$V_2 = \sqrt{\frac{2I_{ss}}{\mu_n C_{ox}(W/L)_{1,2}}}, \tag{5.52}$$

Example 5.3 ***(Continued)***

as both transistors carry currents during data transition. Neglect $C_{\rm p}$ for the time being. Suppose the common source voltage $V_{\rm out}$ varies like a sinusoidal, we arrive at the swing of $V_{\rm out}$ as

$$V_{\rm out}\ \text{Swing} \approx \frac{\frac{\rm A}{2} + V_1 - V_2}{2}, \tag{5.53}$$

where A denotes the input data swing. Note that it represents 100% data transition rate. For a purely random data stream, probability to have transition between two adjacent bits is 50%. Thus, the actual $V_{\rm out}$ swing would be further divided by 2. With averaging capacitor $C_{\rm p}$ included, we obtain the average $V_{\rm out}$ level

$$\overline{V_{\rm out}} = V_{\rm DD} - V_2 - \frac{\frac{\rm A}{2} + V_1 - V_2}{4} \tag{5.54}$$

$$= V_{\rm DD} - \frac{\rm A}{8} - \frac{V_1}{4} - \frac{3V_2}{4}. \tag{5.55}$$

That is, $\overline{V_{\rm out}}$ is in proportion to A with offsets.

The foregoing examples work nicely in the vicinity of 10 Gb/s. At higher data rate, the use of slicer itself causes a series of problems. A slicer is to generate a clean, unaffected waveform for comparison. However, it is quite difficult to keep high gain and large bandwidth simultaneously at high speed. Ringing or other unwanted coupling may go through a slicer and present itself in the output. An adaptive CTLE without using a slicer is illustrated in Fig. 5.26. A novel approach is illustrated here to alleviate the above difficulties. Consider an ideal random binary data. The normalized spectrum can be expressed as

$$S_x(f) = T_b\left[\frac{\sin(\pi f T_b)}{\pi f T_b}\right]^2, \tag{5.56}$$

where T_b denotes the bit period of the data stream, and

$$\int_0^\infty S_x(f)df = \frac{1}{2}. \tag{5.57}$$

To restore the waveform properly, an equalizer must present an output spectrum as close as an ideal one. In other words, we can examine the equalizer's output, determining whether the high-frequency part is under or over compensated, and adjusting the boost accordingly. Note that the slicer is no longer needed here and issues such as imbalanced swings are fully eliminated.

To decompose the spectrum, we recognize a frequency f_m that splits the spectrum into two parts with equal power. That is,

$$\int_0^{f_m} S_x(f)df = \int_{f_m}^{\infty} S_x(f)df = \frac{1}{4} \tag{5.58}$$

and

$$f_m \approx \frac{0.28}{T_b}. \tag{5.59}$$

To be more specific, the high and low frequency power (above and below f_m) are denoted as P_H and P_L, respectively. Fig. 5.**??**(a) depicts the spectra of three different conditions, namely, overcompensated ($P_H > P_L$), critical-compensated ($P_H = P_L$), and under-compensated ($P_H < P_L$). Note that for a dc-balanced data pattern such as 8B/10B coding, the dc power vanishes, resulting in a slightly higher f_m.

Based on the foregoing observation, the equalizer can be realized as shown in Fig. 5.26(b). Here, two voltage-controlled boosting stages interspersed with gain buffers are cascaded to provide large boosting at high frequencies, and the output is directly fed into the power detector. The equalizing filter is designed to achieve a maximum peaking of 15$\sim$20 dB at 10 GHz. A compact design of power detector compares the average power of low and high frequencies (P_L and P_H) by means of the (first-order) low- and high-pass filters and a high-gain rectifier. Rather than an integrator in conventional designs, a V/I converter along with a capacitor C_p follow the power detector, generating appropriate control voltage for the equalizing filter. Such a configuration obviates the need for high-gain error amplifier and preserves flexibility for offset cancellation.

It is worth noting that the setup of $f_m = 0.28/T_b$ is valid for purely random or at least pseudo-random data sequence. The splitting frequency is subject to change if specific patterns/codings are used.

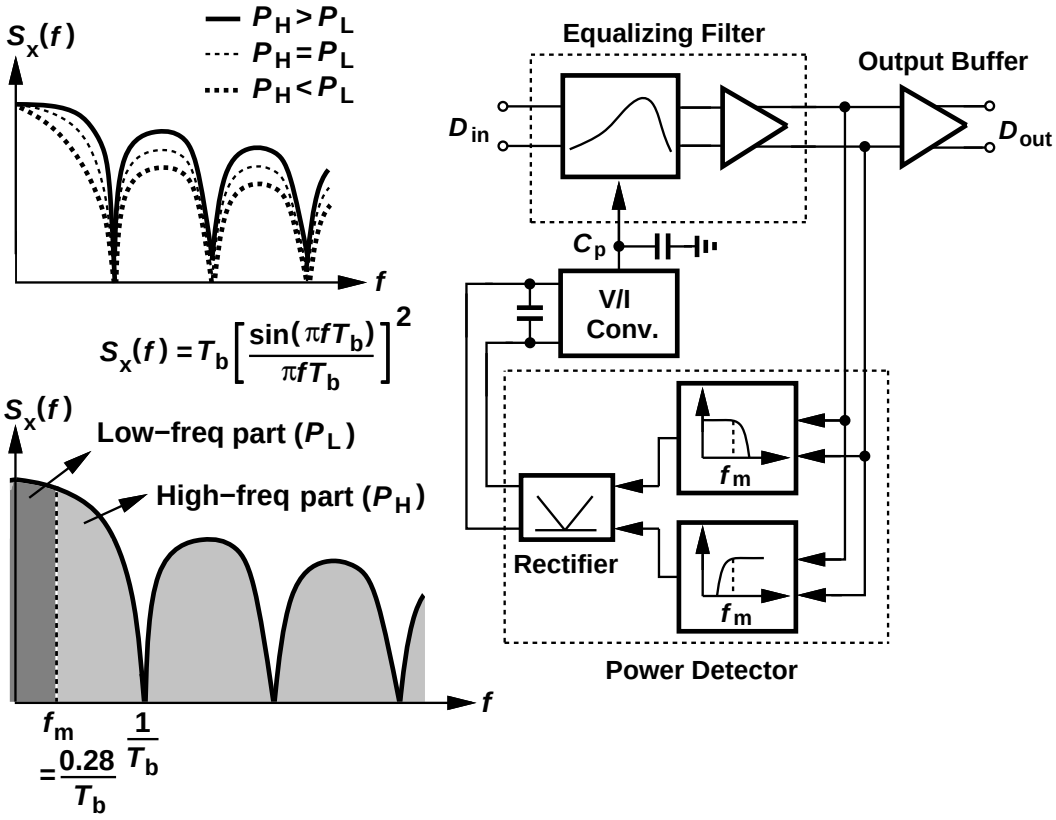


Fig. 5.26 Adaptive CTLE without slicer.

5.5 DECISION-FEEDBACK EQUALIZERS

The decision-feedback equalizers (DFEs) come from the idea of infinite impulse response (IIR) filters. Again we start our discussion from a first-order IIR filter (Fig. 5.27), the transformer function is given by

$$H(s) = \frac{1}{1 + \alpha_1 e^{-sT_b}} \tag{5.60}$$

or equivalently,

$$|H(j\omega)| = \frac{1}{\sqrt{(1+\alpha_1^2) + 2\alpha_1 \cos(\omega T_b)}} \tag{5.61}$$

$$\measuredangle H(j\omega) = \tan^{-1}\left[\frac{\alpha_1 \sin(\omega T_b)}{1 + \alpha_1 \cos(\omega T_b)}\right]. \tag{5.62}$$

In discrete system, it becomes

$$H(z) = \frac{1}{1 + \alpha_1 z^{-1}}. \tag{5.63}$$

The readers can prove Eq. (5.60) and Eq. (5.63) are two different expressions with identical

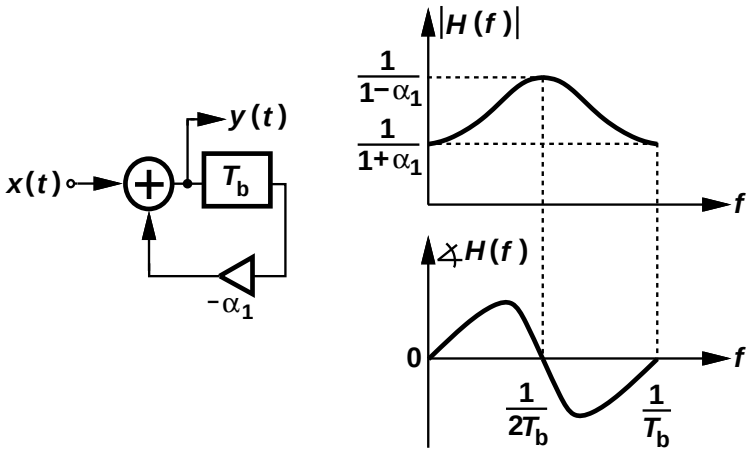


Fig. 5.27 First-order IIR filter and its response.

meaning. The transfer function reveals a monotonic boosting in magnitude from dc to Nyquist frequency, a typical character of equalizers.

The only difference between a DFE and a IIR filter is that the former digitizes the summation result before feeding it back to the delay chain. Figure 5.28 depicts a typical realization. Since the output $y(t)$ and all its delayed versions are rounded to either 0 or 1, a DFE amplifies no noise. It is in contrast to a CTLE, which amplifies high-frequency noise while boosting the signal. A DFE can be made adaptive easily, as it is meant to be dealing with incomplete data in the receiver side. We demonstrate how to set the coefficients α_1, α_2, $\cdots \alpha_N$.

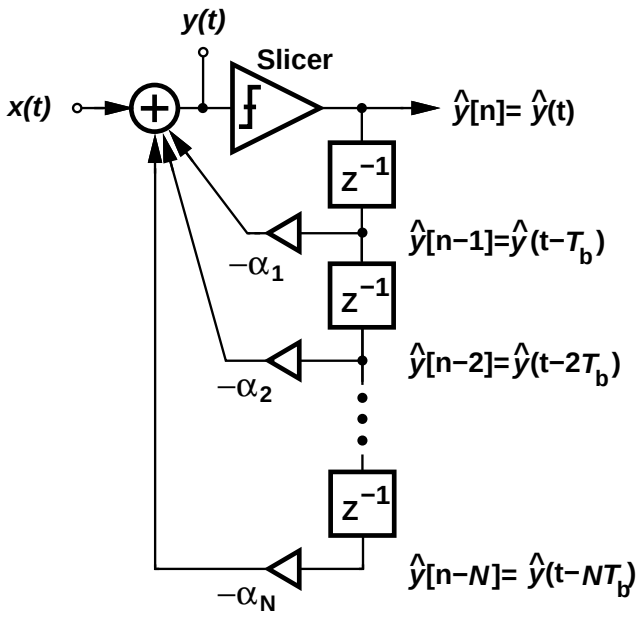


Fig. 5.28 N-tap DFE.

Example 5.4

Use the single-pulse response of Fig. 5.11 as the input, determine the coefficients of a 3-tap DFE that minimize the post-cursors.

Solution:

We use discrete expression for implicity. The goal is to have $\hat{y}[n] = \{\cdots, 0, 0, 1, 0, 0, \cdots\}$, as we come up with the following equations:

$$\hat{y}[0] = 1 \tag{5.64}$$

$$\hat{y}[1] = 0 = -\alpha_1 + x_1 \tag{5.65}$$

$$\hat{y}[2] = 0 = -\alpha_2 + x_2 \tag{5.66}$$

$$\hat{y}[3] = 0 = -\alpha_3 + x_3 \tag{5.67}$$

As a result, we have $[\alpha_1, \alpha_2, \alpha_3] = [-0.2, 0.1, 0.05]$.

Example 5.4 reveals a fact that the optimal DFE coefficients to equalize a single pulse are the post-cursors. Indeed, a DFE can only handle post-cursors, since it needs a “1” (after rounding) to

trigger the feedback compensation. A standalone DFE fails to work if the incoming data has no data transmit at all. In other words, a DFE must cooperate with a FFE for most cases. Of course, an adaptive DFE must optimize its coefficients without knowing the post-cursors. We introduce adaptation algorithm later.

As DFE is one kind of transformation from IIR filters, it is prone to instability if the coefficients are not properly assigned. Neglecting the digitization process, a DFE degenerates to a regular IIR filter

$$H(z) = \frac{1}{1 + \alpha_1 z^{-1} + \alpha_2 z^{-2} + \ldots + \alpha_N z^{-N}}. \tag{5.68}$$

The system is boundary input boundary output (BIBO) stable if and only if the unit circle is contained in the region of convergence (ROC) of $H(z)$. Since all coefficients are real, $H(z)$ can be expressed as a partial-fraction expansion containing real poles and/or complex conjugate poles. They appear as

$$\text{real pole term} : \frac{1}{1 - az^{-1}} \tag{5.69}$$

$$\text{complex-conjugate term} : \frac{1 - a\cos\omega_0 z^{-1}}{1 - 2a\cos\omega_0 z^{-1} + a^{-2}z^{-2}} \text{ or} \tag{5.70}$$

$$\frac{a\sin\omega_0 z^{-1}}{1 - 2a\cos\omega_0 z^{-1} + a^{-2}z^{-2}}. \tag{5.71}$$

Meanwhile, our system is always causal, resulting in a ROC for all these terms

$$|z| > |a|. \tag{5.72}$$

In other words, for a system to be BIBO stable, we require $|a| < 1$ for all partial-fraction terms of $H(z)$. A DFE design sure needs to obey this rule at least. However, due to the nonlinearity (i.e., digitization), conditions for a DFE to be stable are more restrictive.

Example 5.5

Consider the stability of the 1-tap DFE shown in Fig. 5.29.

Example 5.5 ***(Continued)***

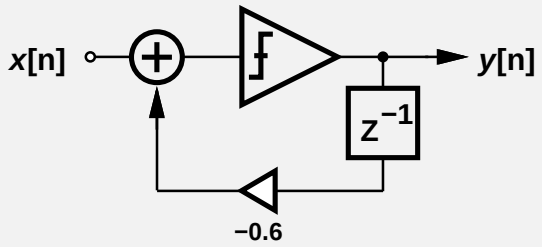


Fig. 5.29 A 1-tap DFE.

Solution:

Disregard the slicer, the 1-tap DFE becomes a first-order IIR filter with transfer function

$$H(z) = \frac{1}{1 - 0.6z^{-1}}, \tag{5.73}$$

which is BIBO stable.

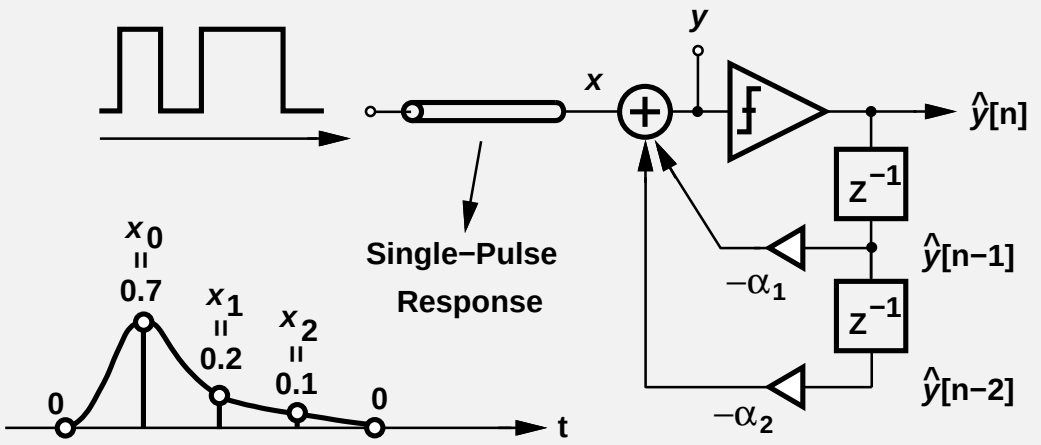
However, as a DFE, it is unstable. For example, applying $\delta[n] = \{\ldots, 0, 0, 1, 0, 0, \ldots\} = x[n]$, we have $y[n] = \{\ldots, 0, 0, 1, 1, 1, 1, \ldots\}$ because of rounding.

The coefficient setting in Example 5.xx is good for a single pulse. How do we choose them for a random data sequence? Recall from our discussion of CTLE, we realize that the optimal setting is to make the pre-slicer waveform resemble the post-slicer one as much as possible. Same rules can be applied to a DFE. That is, the adaptation criteria is to make the summation result $y(t)$ in Fig. 5.28 a critical-compensated data sequence. Suppose $y(t)$ and $\hat{y}[n]$ are first adjusted to have equal swing magnitude (i.e., dc power $A = A'$ in Fig. xx) and the real-time digitization error e is defined as $y(t) - \hat{y}[n]$, we surmise that the optimal coefficients are obtained as $e^2 = |y - \hat{y}|^2$ reaches a minimum. We learn how to calculate the optimal coefficients in the following examples.

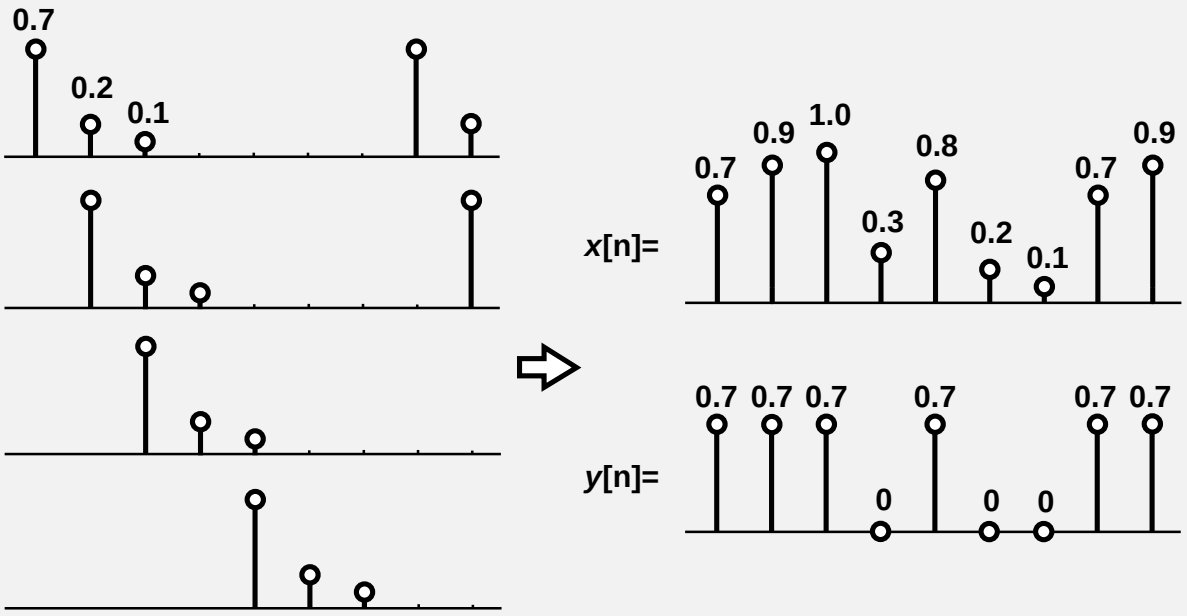
Example 5.6

Consider a 2-tap DFE shown in Fig. 5.30. It is to equalize the loss of a channel, whose single-pulse response is also shown. (a) Determine α_1 and α_2. (b) Plot y in discrete format if a PRBS of length $2^3 - 1$ is sent into the channel.

Example 5.6 ***(Continued)***



(a)



(b)

Fig. 5.30 (a) Calculate the coefficient of 2-tap DFE, (b) case for PRBS3.

Solution:

(a) The response has post-cursor only. For a bit ZERO, it could locate at different positions depending on its preceding bits. To be more specific, we have{1, 1, $\underline{0}$ }, {0, 1, $\underline{0}$ }, {1, 0, $\underline{0}$ }, and

Example 5.6 ***(Continued)***

{0, 0, $\underline{0}$ } 4 conditions. Their values as a logic "0" at x are

$$\{1, 1, \underline{0}\} \rightarrow \quad 0.1 + 0.2 - \alpha_1 - \alpha_2 \tag{5.74}$$

$$\{0, 1, \underline{0}\} \rightarrow \quad 0.2 - \alpha_1 \tag{5.75}$$

$$\{1, 0, \underline{0}\} \rightarrow \quad 0.1 - \alpha_2 \tag{5.76}$$

$$\{0, 0, \underline{0}\} \rightarrow \quad 0. \tag{5.77}$$

Each condition has equal probability of 1/4. Note that the first three cases have feedback components. The quantization error's power is given by

$$e^2 = (0.3 - \alpha_1 - \alpha_2)^2 + (0.2 - \alpha_1)^2 + (0.1 - \alpha_2)^2, \tag{5.78}$$

which needs to be minimized:

$$\frac{\partial\, e^2}{\partial\, \alpha_1} = 0 \tag{5.79}$$

$$\frac{\partial\, e^2}{\partial\, \alpha_2} = 0. \tag{5.80}$$

As a result, $\alpha_1 = 0.2$, $\alpha_2 = 0.1$. For a bit ONE, same procedure applies. The 4 possible values as a logic "1" at x become

$$\{1, 1, 1\} \rightarrow \quad 0.7 + 0.2 + 0.1 - \alpha_1 - \alpha_2 \tag{5.81}$$

$$\{0, 1, 1\} \rightarrow \quad 0.7 + 0.2 - \alpha_1 \tag{5.82}$$

$$\{1, 0, 1\} \rightarrow \quad 0.7 + 0.1 - \alpha_2 \tag{5.83}$$

$$\{0, 0, 1\} \rightarrow \quad 0.7. \tag{5.84}$$

Subtracting 0.7 in each case and squaring them individually, we arrive at the same error power. Thus, same results are obtained.

(b) A sequence of ideal PRBS pulses with length $2^3 - 1$ is {1, 1, 1, 0, 1, 0, 0}. When appearing at the far-end side x, it becomes

$$x[n] = \{0.7, 0.9, 1.0, 0.3, 0.8, 0.2, 0.1, 0.7, \ldots\}\ . \tag{5.85}$$

Example 5.6 ***(Continued)***

After equalization, we have

$$x[n] = \{0.7, 0.7, 0.7, 0, 0.7, 0, 0, 0.7, \ldots\}, \tag{5.86}$$

as depicted in Fig. 5.30(b).

Example 5.7

Repeat Example 5.6 for a 1-tap DFE.

Solution:

(a) For a 1-tap DFE, the 4 posible values as logic "0" are

$$\{1, 1, \underline{0}\} \rightarrow \quad 0.1 + 0.2 - \alpha_1 \tag{5.87}$$

$$\{0, 1, \underline{0}\} \rightarrow \quad 0.2 - \alpha_1 \tag{5.88}$$

$$\{1, 0, \underline{0}\} \rightarrow \quad 0.1 \tag{5.89}$$

$$\{0, 0, \underline{0}\} \rightarrow \quad 0. \tag{5.90}$$

The error's power becomes

$$e^2 = (0.3 - \alpha_1)^2 + (0.2 - \alpha_1)^2 + 0.1^2. \tag{5.91}$$

As a result, $\alpha_1 = 0.25$.

(b) The summation result $y[n]$ becomes

$$y[n] = \{0.7, 0.65, 0.75, 0.05, 0.8, -0.05, 0.1, 0.7, \ldots\}. \tag{5.92}$$

Figure 5.31 shows the result.

Example 5.7 ***(Continued)***

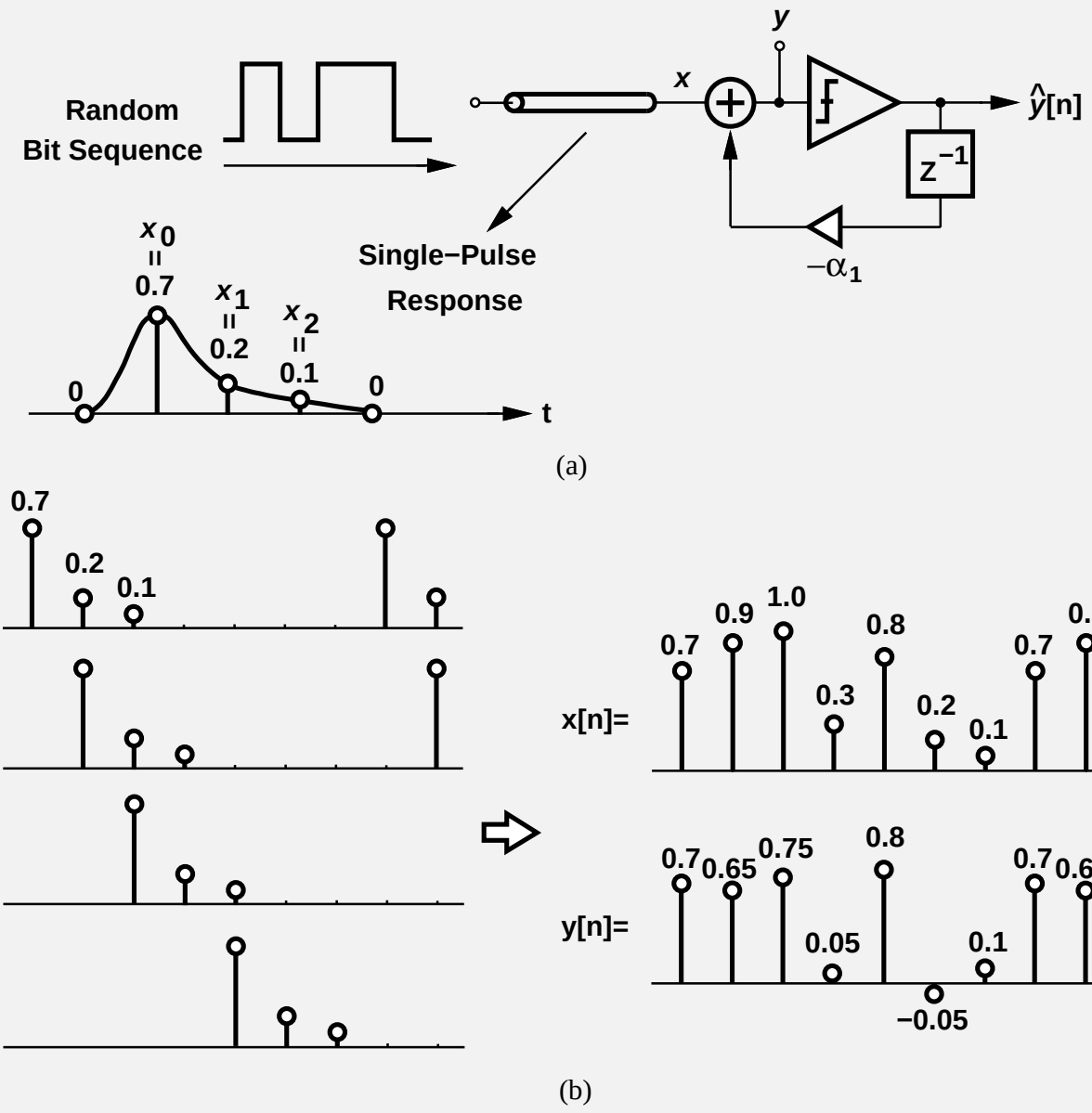


Fig. 5.31 (a) Calculate the coefficient of 1-tap DFE, (b) case for PRBS3.

The above examples suggest the following facts. (1) If a DFE has long enough taps to cancel out all post-cursors, the optimal coefficients are the post-cursors themselves. (2) Otherwise, a set of optimal coefficients would be obtained by minimum error method. They will be slightly different from the post-cursors, and trivial errors would remain after equalization. Similar to a CTLE, the

magnitude of DFE's output (i.e, the summation result y) gets shrunk from 1 to 0.7 in these two cases. It is a typical phenomenon for all kinds of equalizers that the high-frequency boosting is accomplished by suppressing low-frequency power (or equivalently, dc swing). If necessary, amplification must be imposed on the signal path to restore it.

Another point of view to understand a DFE is that, it dynamically adjust the threshold level based on the previous results to make the present transition easier to happen. Consider a 1-tap DFE again with $\alpha = 0.2$. Now we set the threshold of the slicer to be 0.4 instead of 0.5. Suppose the previous data is logic 1. A value of -0.2 will be added up to the present input x. That is, if the present input is less than 0.6, it would be considered a logic 0. By the same token, if the previous data is logic 0, the present input would be considered logic 1 if it is greater than 0.4. In other words, the threshold level of the whole DFE is actually either 0.4 or 0.6, depending on the previous state. In such a way, transitions become easier and high-frequency port gets boosted. Figure 5.32 illustrates a 1-tap DFE design. To accelerate the feedback, we merge the adder and the slicer into the flipflop. Now, the output directly feeds back to the input with a coefficient $-\alpha$, which is implemented in current mode. The pair $M_{11} - M_{12}$ carries the feedback signal. It is equivalent to dynamically adjust the threshold level of the sampler based on the previous result. That is, if the previous bit is "0", the current bit will be considered "1" if the output crosses $V_{TH,L}$, and vice versa. Note that the total tail current of the adder and the master latch remains constant in order to keep a fixed data swing. The current of adder pair $M_{11} - M_{12}$ is also steered by $M_{13} - M_{14}$ synchronously with the master latch, resetting the feedback when the comparison (or "slicing") is accomplished. As a result, the master latch maintains a constant output swing in locking state, where the regeneration pair $M_3 - M_4$ carries all the tail current (I_{SS}). Note that the shorter feedback path in the DFE not only increases the operation speed but provides a larger margin of phase for sampling.

How many taps in a DFE do we need for a given power budget? Let us neglect the effect of slicer again. A simple model of it can be found in Fig. 5.33(a), where N delayed outputs are fed back to the input with corresponding coefficients $-\alpha_1, -\alpha_2 \ldots - \alpha_N$. The input x is applied to the summer directly. Similar to the case of FFE, the maximum achievable boosting is determined by

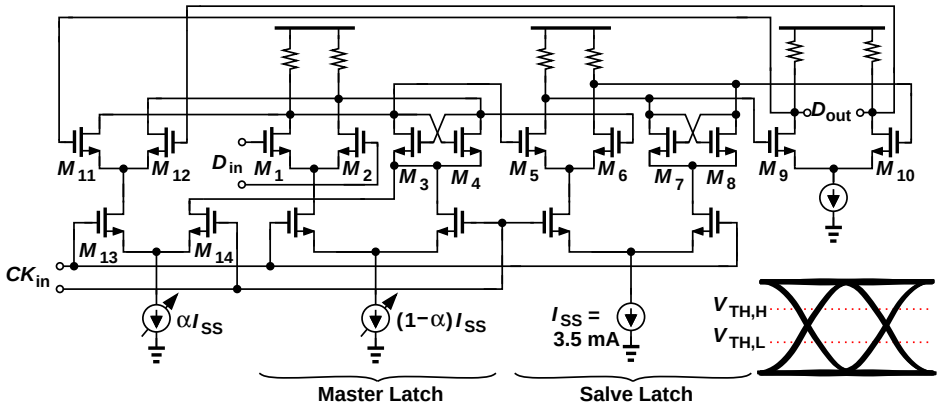


Fig. 5.32 High-speed 1-tap DFE.

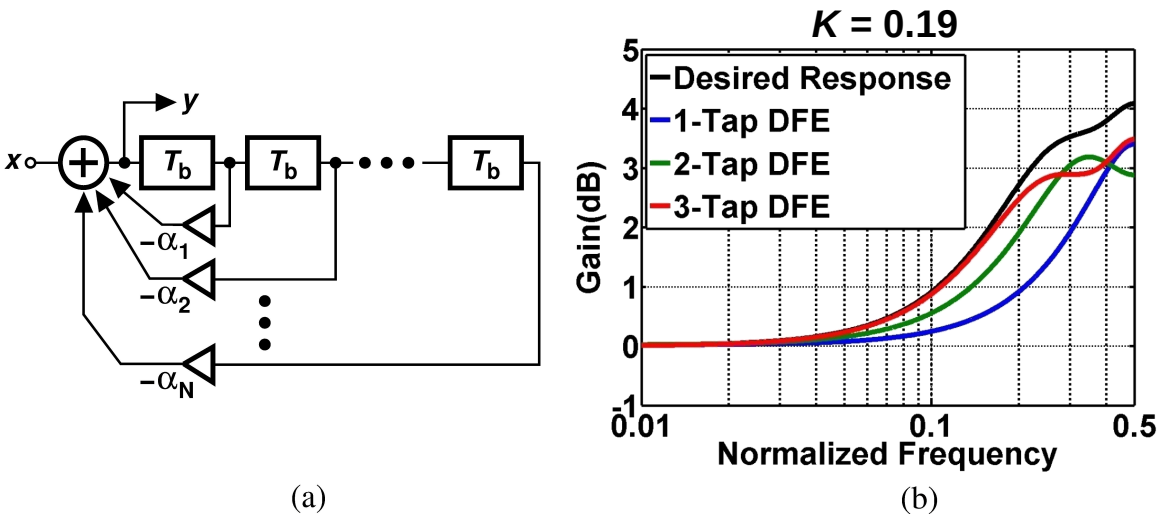
the coefficient amount rather than the number of taps. If we fix the total amount of all coefficients as K (i.e., $\sum_{k=1}^{N}|\alpha_k| = K$), we obtain the maximum Nyquist boosting as

$$\frac{H(j\frac{\pi}{T_b})}{H(j0)} = \frac{1+\sum_{k=1}^{N}\alpha_k}{1+\sum_{k=1}^{N}\alpha_k(-1)^k} \tag{5.93}$$

$$\leq \frac{1+K}{1-K}\,, \quad 0<K<1. \tag{5.94}$$

The equation holds when $\alpha_1 > 0$, $\alpha_3 > 0 \ldots$ and $\alpha_2 = \alpha_4 = \ldots = 0$. In other words, if we fix the total amount of the feedback coefficients, the maximum boost at Nyquist frequency is also fixed regardless of the tap number N. Again, using more taps only improve the equalization quality but not the amount of boosting. In Fig. 5.33(b), we plot the DFE responses with different taps and have them compared with the desired response. A DFE with three or more taps provides better fitting. For high-speed operation, however, we may use fewer taps due to the excessive parasitic capacitance and circuit complexity. In real circuits, the slicer in a DFE not only digitizes the summation result but help boosting. Taking the saturation effect into consideration, we realize that a DFE with a slicer actually generates larger compensation at high frequencies. Fig. 5.34 reveals the simulated response of a 20-Gb/s, 1-tap DFE with and without slicer in 65-nm CMOS technology. Using slicer improves the dc gain and boosting by 3 and 5 dB. Time-domain waveforms are also shown in Fig. 5.34 with the same setup. The slicer increases eye opening by 200 mV. DFE usually

suffers from very stringent timing requirement. It has to accommodate clock-to-Q delay, coefficient multiplication, summation, digitization, and setup time around the feedback loop within one clock cycle (T_b). Several structures have been developed to overcome their issue. Figure 5.35(a) illustrates the idea. Here, no analog summation is taken place in feedback. Rather, we place two pre-set slicers ahead. Two possible conditions $+\alpha$ and $-\alpha$ are loaded to the slicers, and the selector picks the right one based on the previous bit. Since the feedback loop has been unrolled to some extent, the timing requirement becomes much more relaxed. Furthermore, we can do one more step to unroll two taps and make it a 2-bit feedback. As shown in Fig. 5.35(b), 4 pre-sets must be ready for the slicers. Obviously, the circuit complexity and power consumption grow up exponentially. The slicers themselves need to maintain low offset in order to minimize BER.



(a) (b)

Fig. 5.33 (a) N-tap DFE without a slicer, (b) response for a given K.

Shown in Fig. 5.36 is another way to relax the timing issue. Called half-rate structure, it splits the input data into two paths alternately producing outputs. Since all components are operated in half rate, the 1:2 demultiplexing function is naturally included and feedback timing is extended. Note that although the half-rate structure increases routing complexity slightly, it saves power in high-speed applications. Figure 5.37 illustrates 20-Gb/s, 1-tap, full-rate, and half-rate DFEs designed and optimized in 65 nm CMOS. The half-rate structure incorporates 4 latches (i.e., two flipflops in 10 Gb/s), a 10 GHz clock buffer, and two adders, with a total power consumption of 16

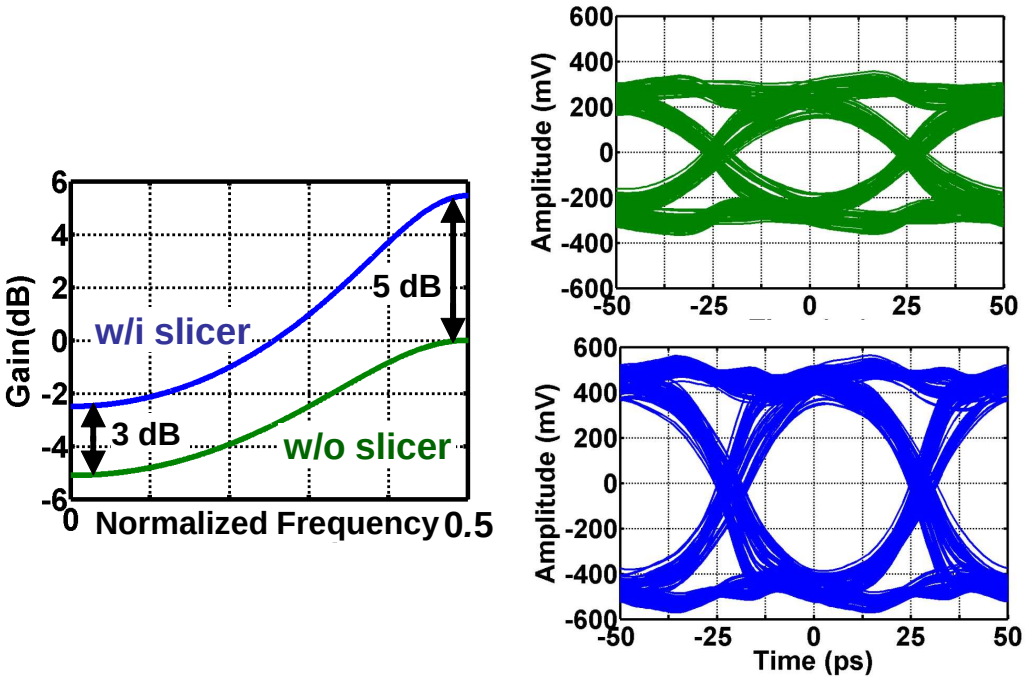
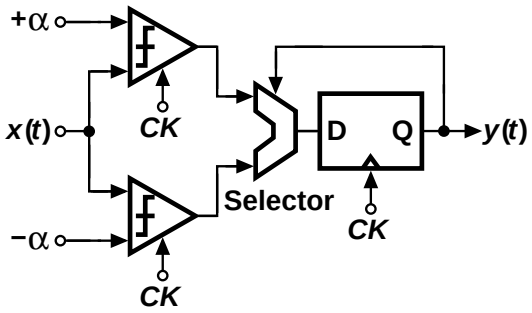


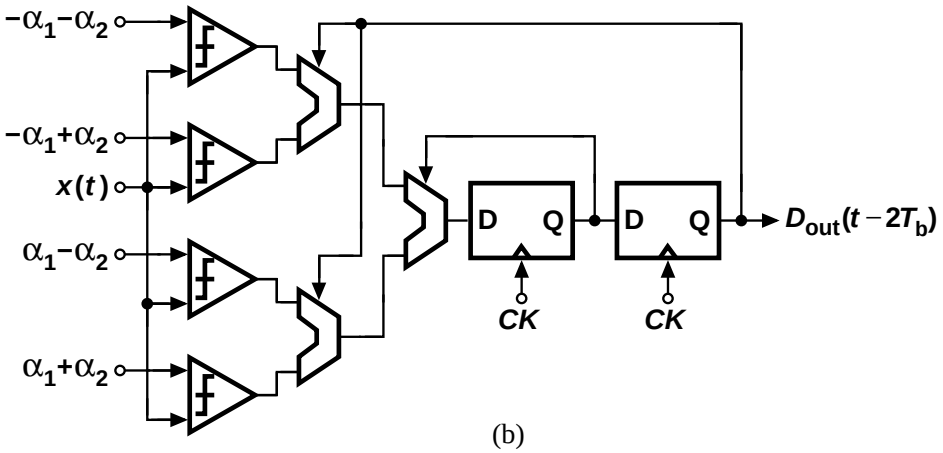
Fig. 5.34 Simulation results for a 20-Gb/s 1-tap DFE with and without a slicer.

mW. The slicers are merged to the flipflops. The full-rate DFE, on the other hand, necessitates a 20-Gb/s flipflop, 20 GHz clock buffer, an adder, a divided-by-2 circuit, and a 2-to-1 selector. The overall power dissipation would be as large as xx mW.

Finally, we discuss the adaptation method. As we mentioned earlier, a DFE's coefficients is optimized if and only if the data before slicer (i.e., $y[n]$) resembles the data after slicer (i.e., $\hat{y}[n]$). To make a DFE adaptive, we need an algorithm that dynamically adjusts the coefficients in the optimal positions. Many algorithms have been developed to do this task, but they are usually too complicate to be implemented in silicon. For example, Newton's method has been used to find numerical solutions for polynomial roots, but the hardware for doing this would be too costly. The most commonly used algorithm here is called "Sign-Sign LMS", which belongs to least-mean-square (LMS) algorithm family. Generally speaking, optimal coefficients can be found by N-dimensional iterative searching that starts at arbitrary point in the vector space, and progressively moves towards the destination. Sign-Sign LMS method is believed to be the simplest realization.



(a)



(b)

Fig. 5.35 Loop unrolling DFEs: (a) 1-tap, (b) 2-tap speculation.

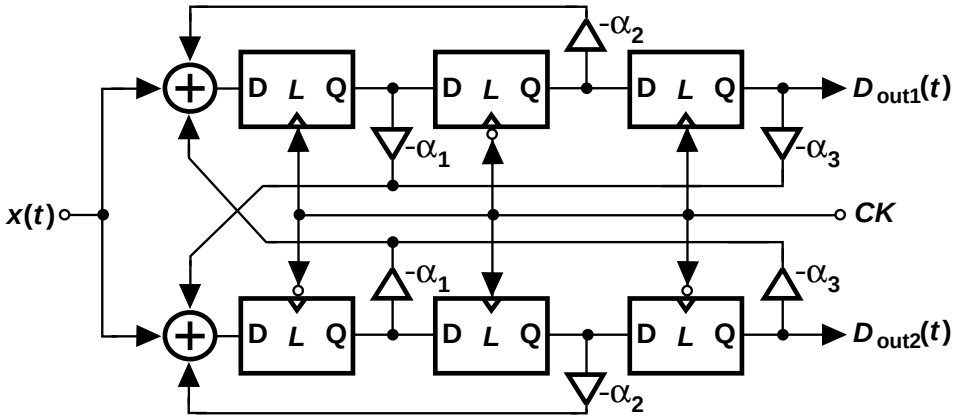
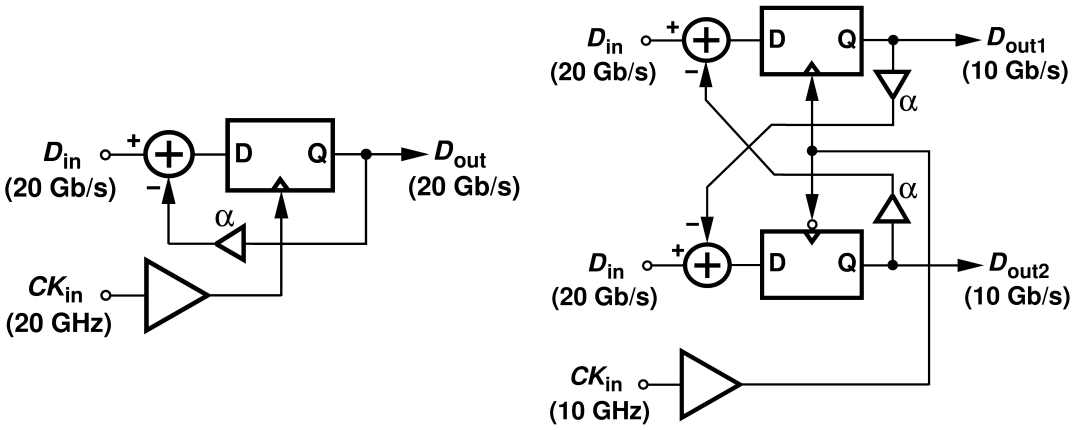


Fig. 5.36 Half-rate DFE.



2 Latches @ 20Gb/s	6 mW
Clock Buffer @ 20GHz	3 mW
1 Adder	3 mW

(a)

4 Latches @ 10Gb/s	8 mW
Clock Buffer @ 10GHz	3 mW
2 Adders	5 mW

(b)

Fig. 5.37 (a) One-tap fu1l-rate DFE and (b) one-tap Half-rate DFE.

Let us reconsider a N-tap DFE as shown in Fig. 5.27. The analog sum $y[n]$ is equal to

$$y[n] = x[n] - \alpha_1\hat{y}[n-1] - \alpha_2\hat{y}[n-2] - \ldots - \alpha_N\hat{y}[n-N], \tag{5.95}$$

where the digital data $\hat{y}$ is either 0 or 1. $y[n]$ is actually a function of multiple variables α_1, α_2, $\ldots \alpha_N$. Since $e = y(t) - \hat{y}[n]$, we have

$$\frac{\partial\, e^2}{\partial\, \alpha_k} = 2 \cdot e \cdot \frac{\partial\, e}{\partial\, \alpha_k} = 2 \cdot e \cdot \frac{\partial\, y}{\partial\, \alpha_k} = -2 \cdot e \cdot \hat{y}[n-k]. \tag{5.96}$$

Note that all $\hat{y}$ are constants. To find optimal coefficients (variables) α_1, α_2, $\ldots \alpha_N$, we adjust them in small step $\triangle$ ($\triangle$ is positive) toward the correct direction. That is, for α_k,

$$\text{if} \frac{\partial\, e^2}{\partial\, \alpha_k} > 0 \quad \Rightarrow \alpha_k[n+1] = \alpha_k[n] - \triangle \tag{5.97}$$

$$\text{if} \frac{\partial\, e^2}{\partial\, \alpha_k} < 0 \quad \Rightarrow \alpha_k[n+1] = \alpha_k[n] + \triangle \tag{5.98}$$

Equivalently, we have

$$\alpha_k[n+1] = \alpha_k[n] - \triangle \cdot \text{sign}\{\frac{\partial\, e^2}{\partial\, \alpha_k}\} \tag{5.99}$$

$$= \alpha_k[n] + \triangle \cdot \text{sign}\{e[n] \cdot \hat{y}[n-k]\}. \tag{5.100}$$

Figure 5.38 depicts the algorithm. Since the sign of $\{e[n] \cdot \hat{y}[n-k]\}$ can be easily obtained, it can be easily implemented.[2] Coefficients will keep tracking until the optimal points (e.g., $\alpha_{k,opt}$) which minimizes e^2. Since all coefficients are independent, the algorithm is executed for all coefficients simultaneously. It can be shown that all α_1, α_2, $\ldots \alpha_N$ will converge to a certain point, gives that their ranges are properly assigned. Tricky conditions such as saddle point do not happen here. Like adaptation in CTLE, we need data transitions to make this algorithm work.

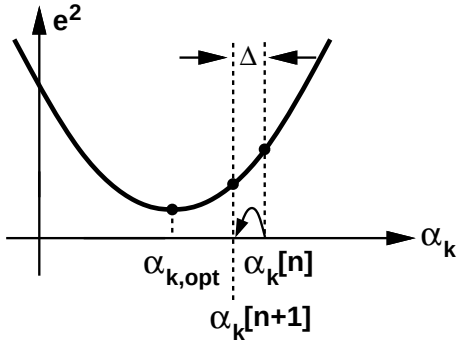


Fig. 5.38 Sign-Sign LMS algorithm.

It is worth noting that, the adaptation procedure of DFE is still based on the fact that the dc power levels before and after the slicer are equal. Otherwise, it won't be able to conduct a fair comparison between $y(t)$ and $\hat{y}[n]$. The power detector introduced in Example 5.xx could be used, but it only works for signals with swing from $V_{\rm DD}$ to $V_{\rm DD}$-IR drop. The feedback paths in DFEs actually lead to inconstant swings and common-mode levels. Therefore, a more sophisticated control system should be developed for DFEs.

An alternative approach to realize adaptive DFEs is to build up a dynamic level tracking loop. The idea is to set up the common-mode level as well as the upper (logic 1) and lower (logic 0) levels based on the present condition of $y(t)$. As shown in Fig.5.39, an unequalized $y(t)$is jittery and full of ISI. Suppose we create these reference levels, $V_{\rm ref}^{+}$, $V_{\rm cm}$, and $V_{\rm ref}^{-}$ for signal processing. $V_{\rm ref}^{+}$ and $V_{\rm ref}^{-}$ are located above and below $V_{\rm cm}$ and they can be adjusted symmetrically with respect to $V_{\rm cm}$ (i.e., $V_{\rm ref}^{+} - V_{\rm cm} = V_{\rm cm} - V_{\rm ref}^{-}$). The adaptive operation can be performed as

[2]The readers shall not be confused by the notation. For example, in our previous discussion, logic "1" means "positive", and logic "0" means "negative".

follows. The first step is to line up $V_{\rm cm}$ with the actual common-mode level of $y(t)$. Next, move $V_{\rm ref}^{+}(V_{\rm ref}^{-})$ to the nominal (average) logic level 1 (logic 0). At this moment, $y(t)$ has a relatively fair reference to optimize DFE coefficients, as $V_{\rm ref}^{+}$ and $V_{\rm ref}^{-}$ dynamically track the optimal logic levels for comparison. Finally, we conduct sign-sign LMS algorithm and optimize each coefficient of DFE. Once all procedures are converged, the waveform of $y(t)$ would be optimized with minimum jitter and ISI.

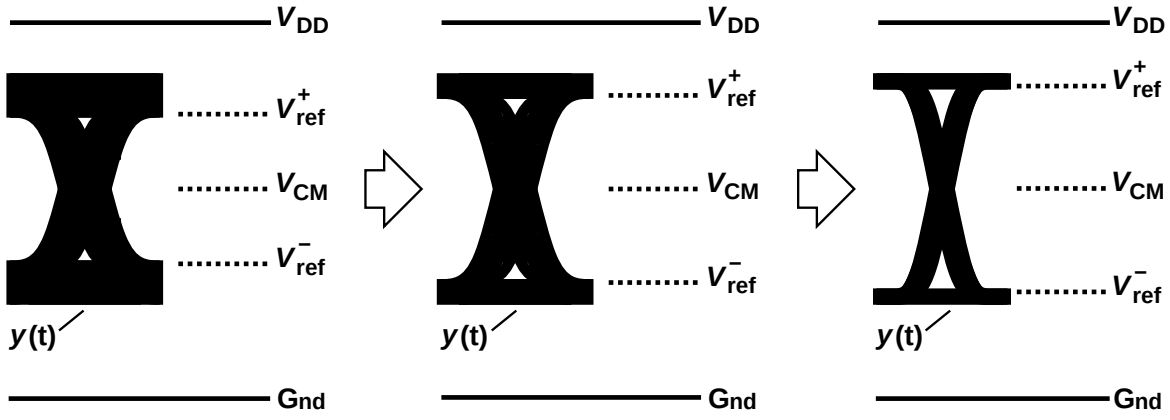


Fig. 5.39 Reference generator.

The above approach relies on an exquiste reference generator. An simple yet powerful design can be found in Fig.5.40. Here, a tilted differential pair $M_{1,2}$, two current sources (I_1), and an Opamp form a servo loop. The negative feedback along the loop forces $V_{\rm cm}$ to be equal to the common-mode level of $y(t)$. The two references are

$$V_{\rm ref}^{+} = V_{\rm DD} - I_1 R \tag{5.101}$$

$$V_{\rm ref}^{-} = V_{\rm DD} - (I_{\rm DAC} + I_{\rm ss} + I_1)R, \tag{5.102}$$

as M_1 carries all the tail currents of $I_{\rm DAC}$ and $I_{\rm ss}$. By tuning $I_{\rm DAC}$, we have I_1 changed accordingly. That results in symmetric adjustment on $V_{\rm ref}^{+}$ and $V_{\rm ref}^{-}$ with respect to $V_{\rm CM}$, which is fixed to the common-mode level of $y(t)$.

The complete DFE which such a level tracking algorithm is depicted in Fig.5.41. Here, the reference adjuster cooperates with the reference generator to conduct the governs the timing and

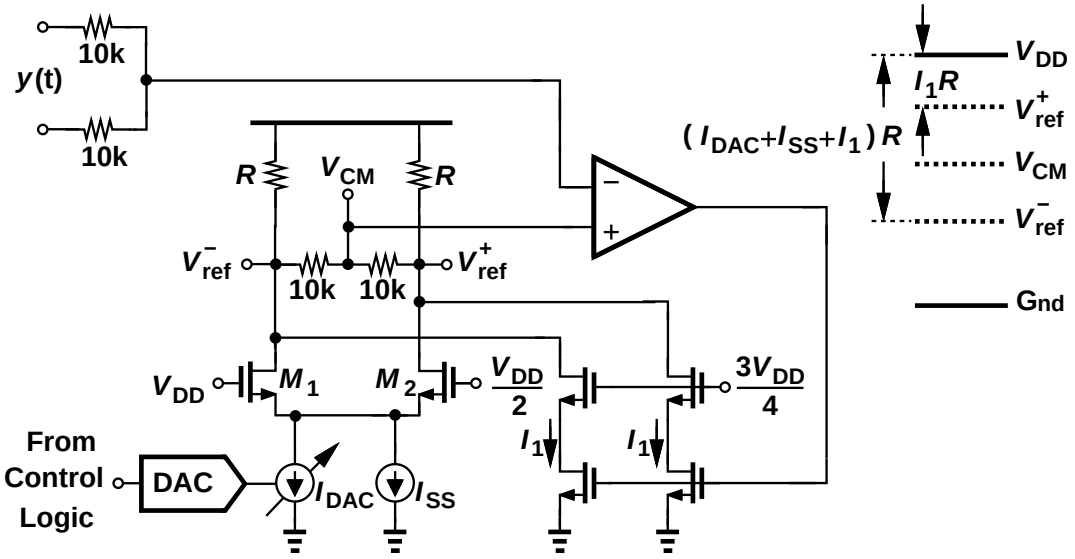


Fig. 5.40 Reference generator.

convergence sequence so that the whole system operates smoothly. The reference adjuster can be realized as shown in Fig.5.42. Two additional slicers (comparators) are employed to examine the status of sampled data. That is, this arrangement checks the sampled point. If it is above $V_{\rm ref}^{+}$ or below $V_{\rm ref}^{-}$, the reference levels should be pushed away from $V_{\rm CM}$. Otherwise, they ought to be moved toward $V_{\rm CM}$. The reader can prove three XOR gates can provide the necessary logic here.

5.6 CASE STUDY

In this section, we present two works designed for

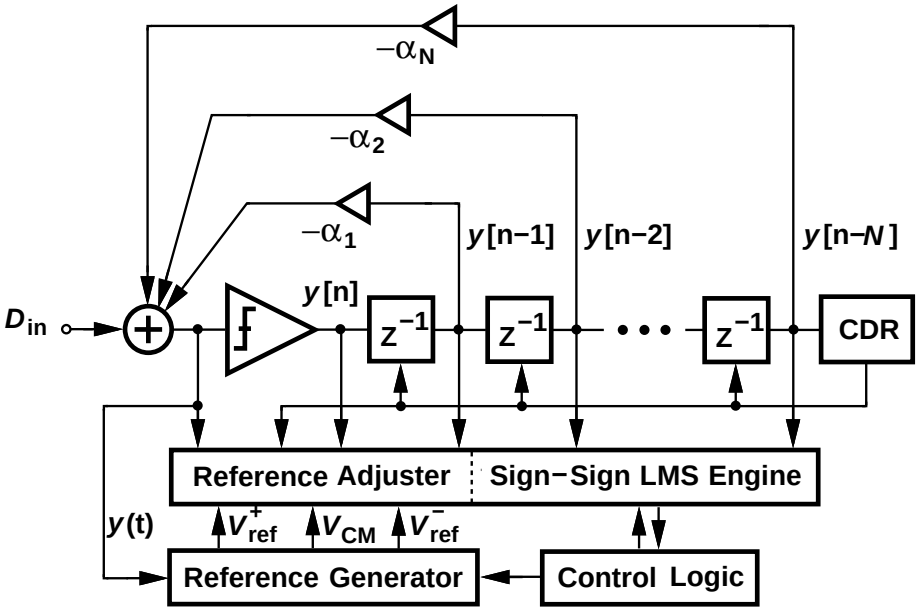


Fig. 5.41 Adaptive DFE with dynamic level tracking.

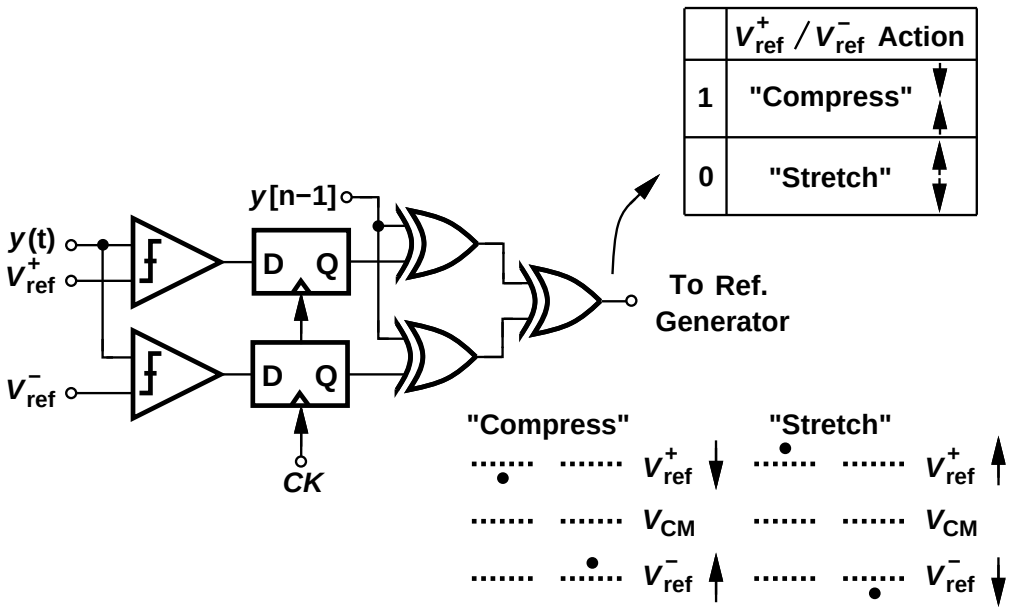


Fig. 5.42 Reference adjuster.

REFERENCES

[1] J.S. Choi, M.S. Hwang ,and D.K. Jeong, "A 0.18-μm CMOS 3.5-Gb/s Continuous-Time Adaptive Cable Equalizer Using Enhanced Low-Frequency Gain Control Metho," *IEEE J. Solid-State Circuits*, vol. 39, pp. 419-425, Mar. 2004.

[2] S. Gondi, J. Lee, D. Takeuchi and B. Razavi, "A 10Gb/s CMOS Adaptive Equalizer for Backplane Applications," *ISSCC Dig. Tech. Papers*, pp. 328-329, Feb. 2005.

[3] Simon Haykin, "Adaptive Filter Theory," *Prentice Hall*, 2001.

[4] H. Wang, C. Lee, A. Lee, and Jri Lee, "A 21-Gb/s 87-mW Transceiver with FFE/DFE/Linear Equalizer in 65-nm CMOS Technology," *Digest of Symposium on VLSI Circuits*, pp. 50-51, Jun. 2009.

OSCILLATORS

Oscillators have been playing critical roles in communication systems for decades. Tunable oscillators such as voltage-controlled oscillators (VCOs) have tremendous influence on the overall performance of the system. Today's CMOS technologies allow us to develop VCOs over 100 GHz with adequately large output power and low phase noise. On the other hand, new techniques continue to emerge, achieving even better performance for next generations's applications.

6.1 REVIEW OF OSCILLATION THEORY

We begin our discussion on fundamental theories of oscillation. It is well-known that a negative feedback system with transfer function of A(s)=A(s)/[1+A(s)] becomes oscillating at certain frequency ω_{osc} if Barkhausen criteria are satisfied:

$$|A(j\omega_{osc})| \geq 1 \tag{6.1}$$

$$\angle A(j\omega_{osc}) = 180^{\circ}. \tag{6.2}$$

Figure 1 illustrates the idea. The overall phase shift along the loop [i,e., $\angle A(j\omega_{osc})$] must be 2π or its multiple if a positive feedback is presented. The noise component at ω_{osc} circulates and survives around the loop and gets larger and larger. In the end, if forms a steady oscillation and the average loop gain becomes exactly 1. Same phenomenon can be explained by observing the poles of the closed-loop gain H(s) [Fig. 6.1(b)]. There must be a conjugate pair on the right hand side as oscillation begins, and gradually shifted to the imaginary axis as the oscillation becomes steady. The distance between the origin and these poles is equal to the oscillation frequency ω_{osc}. Similar

approach can be found by using Nyquist plot [Fig. 6.1(c)]. For a steady oscillation, the open loop gain A(s) pases through the point $(-1,0)$.

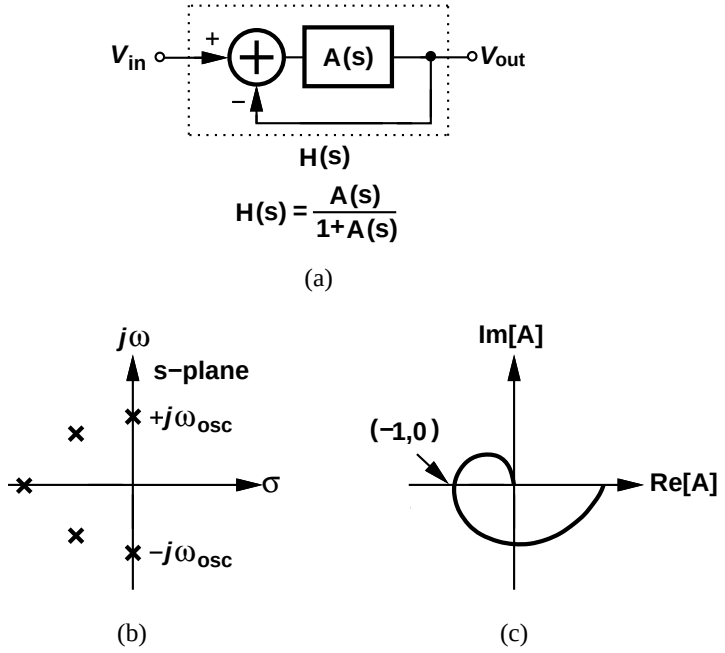


Fig. 6.1 Oscillation condition.

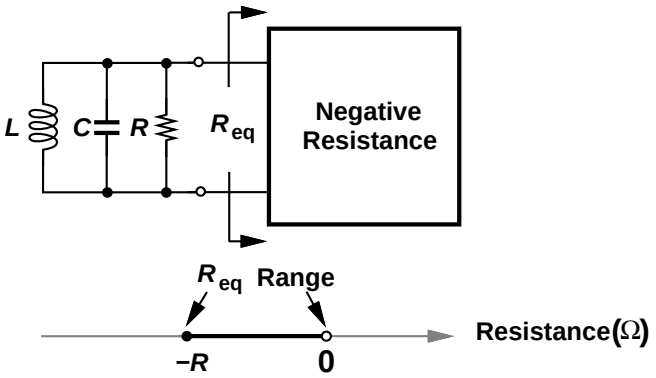


Fig. 6.2 One port oscillation theory.

For oscillators using resonant elements, e.g., inductors, the one-port theory can be applied to examine the occurrence of oscillation. Shown in Fig. 6.2 is an RLC network in parallel with a circuit generating (equivalent) negative resistance $R_{\rm eq}$. If the magnitude of R_{eq} (a negative value)

is smaller than or equal to the positive resistor R, oscillation world occur. In regular cases, the magnitude of R_{eq} would be smaller than R in the beginning and eventually become equal to R in steady oscillation. It can be thought as the negative resistance supplies energy to the RLC network, compensating the loss due to R. In steady oscillation, R and $-R$ cancel each other, leading to oscillation frequency $\omega_{osc} = 1/\sqrt{LC}$.

Example 6.1

Sketch the waveform of the follwing in Fig3 , where a RLC network experiences an input of impulse $V_0\delta(t)$.

Solution:

V_{out}(s) can be expressed as

$$V_{out}(s) = V_0 \cdot \frac{sL||\dfrac{1}{sC}}{sL||\dfrac{1}{sC} + R} = V_0 \frac{\left(\dfrac{\omega_0}{Q}\right)s}{s^2 + \left(\dfrac{\omega_0}{Q}\right)s + \omega_0^2}, \tag{6.3}$$

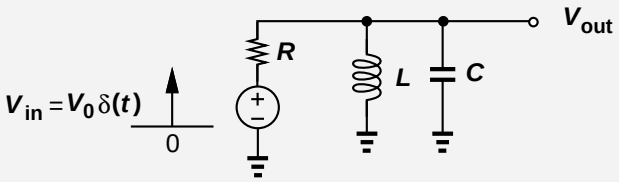


Fig. 6.3 RLC network.

where

$$\omega_0 = \frac{1}{\sqrt{LC}} \tag{6.4}$$

$$Q = R\sqrt{\frac{C}{L}}. \tag{6.5}$$

Example 6.1 ***(Continued)***

For $Q > \frac{1}{2}$, V_{out}(s) presents complex conjugate poles, leading to ringing in it's time domain waveform. That is,

$$V_{out}(s) = \left(\frac{\omega_0}{Q}\right) \cdot V_0 \cdot \frac{s}{\left(s + \frac{\omega_0}{2Q}\right)^2 + \left(1 - \frac{1}{4Q^2}\right)\omega_0^2}. \tag{6.6}$$

Taking inverse Laplace transform, we obtain

$$V_{out} \propto e^{\frac{-\omega_0 t}{2Q}} \cdot cos\left(\sqrt{1 - \frac{1}{4Q^2}} + \omega_0 t + \phi\right) \tag{6.7}$$

where

$$\phi = tan^{-1}\left[\frac{1}{\sqrt{4Q^2 - 1}}\right]. \tag{6.8}$$

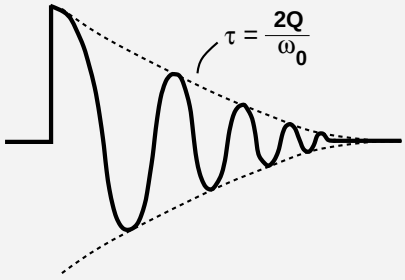


Fig. 6.4 $V_{out}(t)$.

As expected, V_{out} decays exponentially with a time constant $\tau = 2Q/\omega_0$ (Fig. 6.4). Note that for $Q < 1/2$, no ringing occurs. In other words, a resonating oscillators such as an LC tank fails to oscillate if the inductor's quality factor Q is less than 1/2, no matter how much power is burned.

Note that most oscillators are made to be tunable in frequency. That could be done by adjusting the associated capacitors (for LC tank oscillators), changing the equivalent resistance (for ring oscillators), or other methods. Now we are ready to investigate different oscillator topologies, starting from LC tank oscillators.

6.2 LC-TANK OSCILLATORS

6.2.1 Spiral Inductors

On-chip inductors have been applied in wireless and wireline communications over two decades. A commonly used model for an on-chip inductor is illustrated in Fig. 6.5, where L, C, and R_p denotes the inductance, parasitic capacitance, and equivalent loss, respectively. The impedance $Z(s)$ looking into the RLC tank is plotted as well, where $\triangle\omega$ defines the -3-dB bandwidth of it. The phase migrates from $+90^\circ$ to -90° as the peak of magnitude coincides with 0° of phase. Overall speaking, we have

$$Z(s) = \frac{R(\frac{\omega_0}{Q})s}{s^2 + (\frac{\omega_0}{Q}) \cdot s + \omega_0^2} \tag{6.9}$$

$$\measuredangle Z = 90^\circ - tan^{-1}[\frac{\omega_0\omega}{Q(\omega_0^2 - \omega^2)}]. \tag{6.10}$$

Usually, the quality factor Q of an inductor can be defined in different ways:

(i) $Q \triangleq \dfrac{R_p}{\omega_0 L} = R_p\sqrt{\dfrac{C}{L}}$

(ii) $Q \triangleq \dfrac{\omega_0}{\Delta\omega}$

(iii) $Q \triangleq 2\pi \cdot \dfrac{Energy\ Stored}{Energy\ Dissipated\ in\ One\ Cycle}$

(iv) $Q \triangleq \dfrac{\omega_0}{2} \cdot \left|\dfrac{d\phi}{d\omega}\right|$

For $Q \gg 1$, the following example proves the 4 definitions of Q are identical. We see how it works in the following Example.

Example 6.2

Demonstrates the above definitions are equivalent.

Solution:

Example 6.2 ***(Continued)***

using the same notation as Example 6.1 now $R = R_P$ and adopting the 2nd-order filter theory, we realize that the -3-dB frequencies $\omega_{1,2}$ are given by

$$\omega_{1,2} = \omega_0 \sqrt{1 + \frac{1}{4Q^2}} \pm \frac{\omega_0}{2Q}. \tag{6.11}$$

Thus, $\Delta\omega_0 = \omega_2 - \omega_1 = \omega_0/2Q$. Meanwhile, by observing the waveform in Fig. 6.4, we calculate magnitude of the sinusoidal-like gets attenuated by a factor of $e^{-\pi/Q}$ in each cycle. The power of it therefore shrinks by $e^{-2\pi/Q}$ in one cycle. If $Q \supset 2\pi$, we have $e^{-2\pi/Q} \approx 1 - 2\pi/Q$. In other words,

$$2\pi \cdot \frac{1}{1 - (1 - 2\pi/Q)} = Q. \tag{6.12}$$

The reader can easily prove that (iv) is also an equivalent statement.

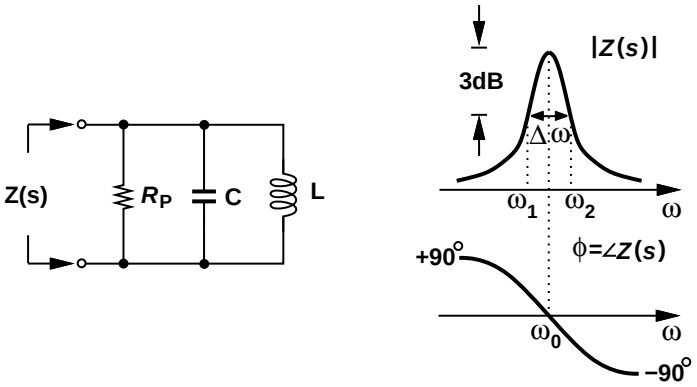
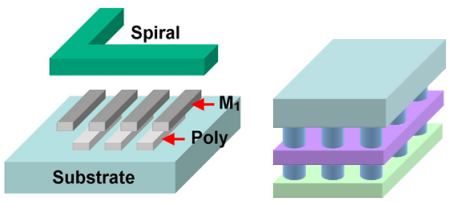


Fig. 6.5 Definition of inductor Q.



(a) (b)

Fig. 6.6 (a) Physical and (b) geometric improvements on inductor Q.

A spiral inductor is made of the top-most layer(s) of metal to reduce parasitic capacitance. It suffers from 3 types of loss : Ohm loss, Eddy current and Skin effect. We can do little about the Ohm loss, as copper is already the metal with second-best conductivity. Eddy current can be prevented or minimized by placing "crossties" underneath the spiral and perpendicular to the current flowing direction. Here, we usually put poly sticks with minimum width and space to shield the coupling to the substract. By doing so, most lines of electric force would be terminated on the crossties rather than the substract, inducing much less Eddy current. Metal 1 sticks can be also used to help fill up the gaps and further improve the Q. Note that all shielding sticks must be connected to ground or other dc level. Floating shield still couples energy from spiral to substract. The skin effect can be alleviated by increasing the surface area of a conductor. Unfortunately, with a fixed metal thickness, the only thing a designer can do here is to shunt multiple layers in parallel. The Q is expected to be improved at a coot of higher parasitic capacitance. The lower metals however only mirror help as their thickness is less than that of the top metal.

It is worth noting that the self-resonance frequency of an inductor L is given by

$$\omega_{SR} = 2\pi f_{SR} = \frac{1}{\sqrt{LC}}, \tag{6.13}$$

where C denotes the equivalent capacitance lumped in parallel with L. This is the physical upper bound of oscillation frequency for an oscillator made of such an inductor.

Different geometric structures have been developed to achieve better inductor design. Figure 7(a) illustrates the fundamental spiral with square shape. Taking 40 nm CMOS process with 9 metal layers as an example. If we design an 0.5nH inductor, the area would be around 3355 um^2 and the Q reaches a peak of 12.87 at 30 GHz. The right angle layout may imitate designers with good PCB layout experience. An octangle spiral can be found in Fig. 6.7(b). With the same desired inductance (0.5nH), the area is about 5616 um^2 and the peak Q becomes 14.6 at 26 GHz. Yet another topology dedicated to differential circuits is to wrap the spiral symmetrically [Fig. 6.7(c)]. Due to the differential operation, the effective substrate loss is reduced by a factor of 2, leading to a higher Q. The only side effect is that the spacing between turns needs to be wider so as to minimize the interwinding capacitance. A 0.5-nH inductor of this structure occupies 5148 um^2 while the peak Q is 16.73. Vertical stacking is another useful technique to shrink the occupied

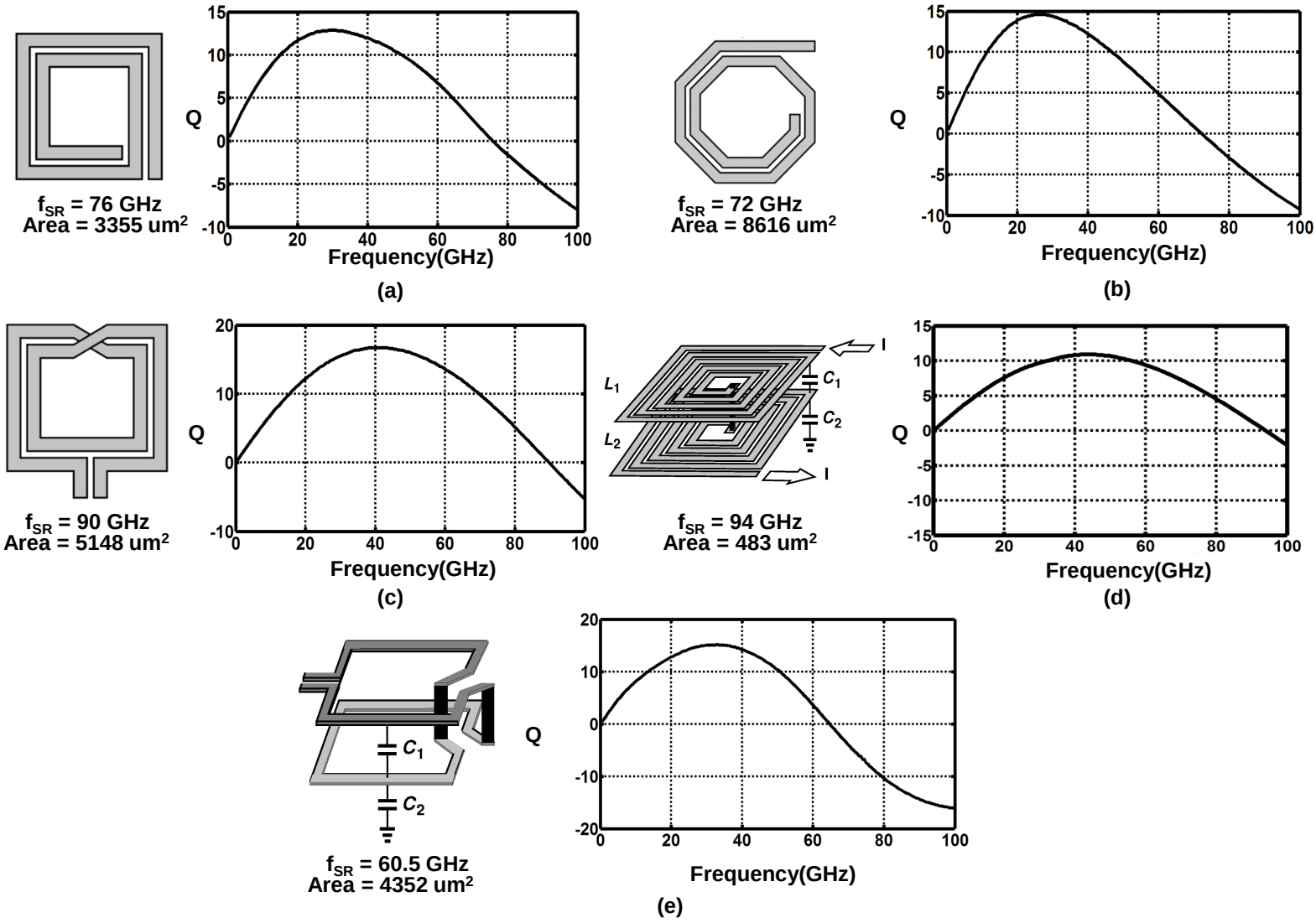


Fig. 6.7 Different inductor topologies with corresponding Q and f_{SR} for L = 0.5 nH.

area for a given inductance [Fig.7(d)]. Depending on the mutual coupling factor, the inductance of a two-layer structure is around 3.5 $\sim$ 4 times larger than that of a single layer one with the same area. Note that the two layers should be kept as far as possible in order to maximize the self-resonance frequency. For a 0.5-nH design, using two layers (M_9-M_{10}) takes only 483 um^2, arriving at an area saving solution. The inductor Q us inevitably lower become of the use of lower layers. With 3 layers in series (M_8-M_9-M_{10}), the area further reduces to 4352 um^2. A method combining these two techniques is depicted in Fig. 6.7(c). Recognized as a differentially-stacked inductor, it preserves the benefits from both structures. More details can be found in [1], [2], [3].

6.2.2 Output Swing

A typical realization of cross-coupled VCOs can be found in Fig. 6.8(a), where the pair M_1-M_2 provides negative resistance $-2/g_{m1,2}$ (differentially) to compensate for the inductor loss R_P. At resonance, these two resistances cancel each other and the oscillation frequency is given by

$$\omega_{osc} = \frac{1}{\sqrt{LC_P}}, \tag{6.14}$$

where L and C_P denote the loading inductor and parasitic capacitance at output nodes, respectively, and M_3-M_4 the MOS varactors. Barkhausen criteria imply that we must have $g_{m1,2} \geq (1/R_P)$ to make the circuit oscillate, while practical design would choose a higher value (≈ 3) to ensure oscillation over PVT variations.

It is instructive to derive an alternative expression for ω_{osc} with simplified conditions to examine what factors actually limit the operation frequency. Modeling the VCO as Fig. 6.8(b), we obtain R_P and ω_{osc} in stable oscillations as

$$R_P = Q \cdot \omega_{osc} L = \frac{1}{g_{m1,2}} \tag{6.15}$$

$$\omega_{osc} = \frac{1}{\sqrt{2L(\frac{C_P}{2} + \frac{C_{GS}}{2})}}, \tag{6.16}$$

where Q represents the quality factor of the tank, and C_{GS} the average gate-source capacitance contributed by $M_{1,2}$. Here, we take off the varactors for simplicity. If C_P is negligible as compared

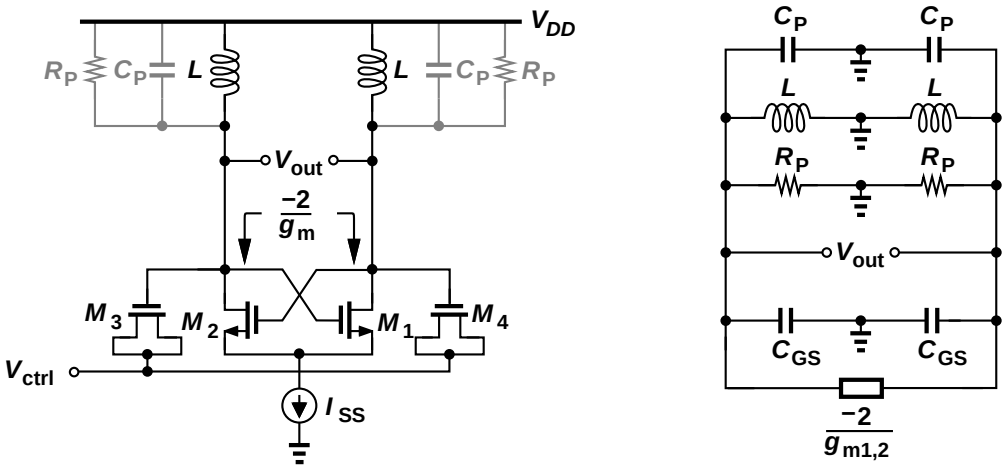


Fig. 6.8 Calculating oscillation frequency of LC tank oscillator.

with C_{GS} (which is basically true at high frequencies), we arrive at

$$\begin{aligned}\omega_{osc} &\approx \frac{1}{\sqrt{LC_{GS}}} \\ &= \frac{1}{\sqrt{\frac{1}{g_m Q\omega_{osc}}C_{GS}}} \\ &= \sqrt{Q\omega_T\omega_{osc}},\end{aligned} \tag{6.17}$$

where ω_T denotes the transit frequency of $M_{1,2}$. It follows that

$$\omega_{osc} = Q \cdot \omega_T. \tag{6.18}$$

In other words, a cross-coupled oscillator can possibly operate at very high frequencies, given that the inductors provide a sufficiently high Q. In reality, however, several issues discourage ultra high-speed oscillation: (1) the on-chip inductors usually have a self-resonance frequency (f_{SR}) of only a few hundred GHz; (2) the varactors present significant loss at high frequencies, and it could eventually dominate the Q of the tank; (3) even (2) is not a concern, the on-chip inductors can never reach a very high Q due to the physical limitations; (4) C_P may not be negligible in comparison with other parasitics. Nonetheless, cross-coupled VCOs are still expected to operate

at frequencies close to device f_T . For example, 50- and 96- and 140-GHz realizations have been reported in 0.25-μm, 0.13-μm, and 90-nm CMOS technologies [4], [5], [6]. The final example illustrates oscillation above f_T.

It is important to know that, if the varactor's capacitance is much greater than other parasitics, the tuning range of an LC VCO approaches a constant and has nothing to do with the inductance. Figure. 9 illustrates such an effect. On the other hand, lowering the inductance leads to smaller swing and puts the oscillator in danger of failing unless the current is increased (assuming that R_P = Qω_0L decreases). Consequently, it is always desirable to use inductors as large as possible.

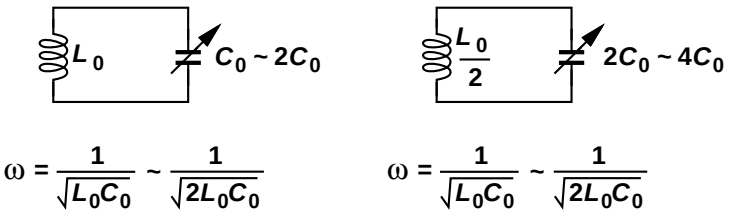
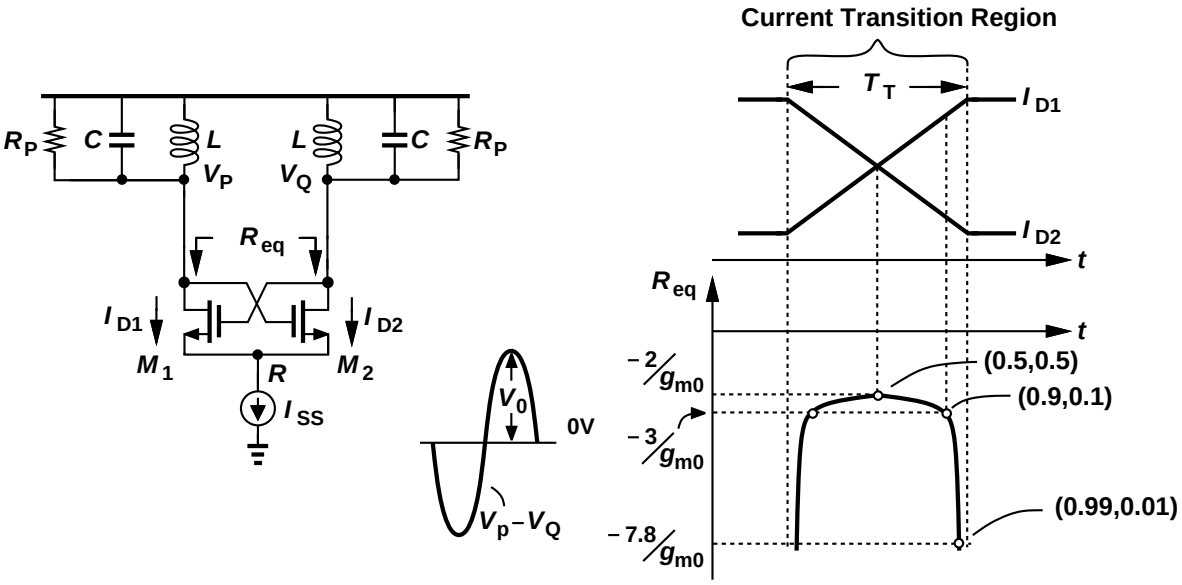


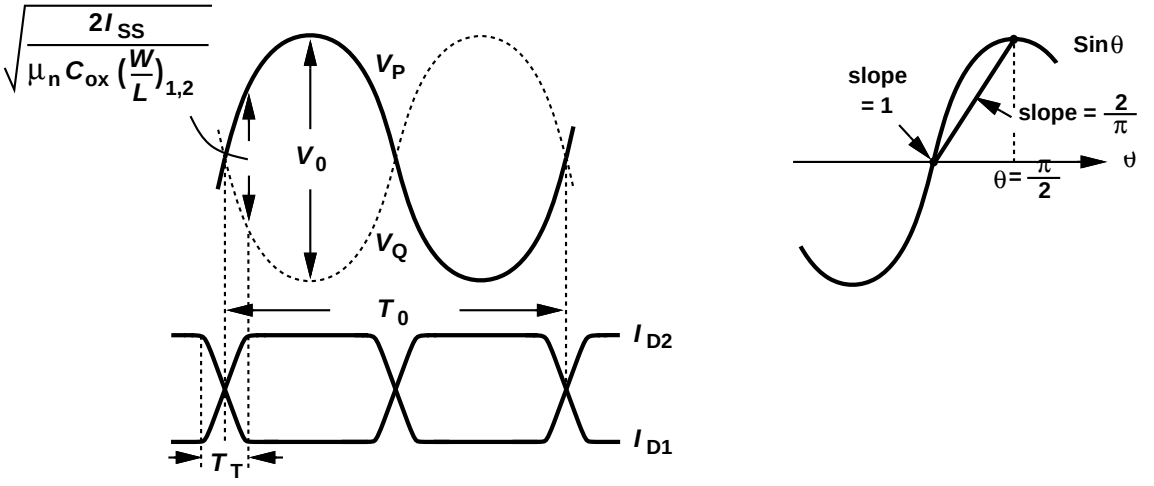
Fig. 6.9 LC networks with equal resonance frequency and tuning range.

What is the output swing of an LC-tank oscillator? Well, the above small-signal analysis only reveals parts of the fact. In real operation, LC-tank oscillators are operated in large signal: the swing is so large that the tail current I_{SS} in Fig. 6.8 is completely switched by M_1 and M_2 most of the time. To determine the output swing, we need to derive large-signal analysis.

Let us redraw an LC-tank oscillator in Fig. 6.10(a). Due to the abrupt and violent switching, M_1 carries full I_{SS} for half cycle and stays off for the other half cycle (and so does M_2). The negative resistance is generated only during current transitions, as both M_1 and M_2 are carrying currents. Denoting g_{m0} as the transconductance of M_1 and M_2 when it carries I_{SS}/2, we obtain the equivalent resistance $R_{eq} = -2/g_{m0}$ at the point $I_{D1} = I_{D2} = I_{SS}/2$. Assuming currents are swept linearly across the transition region, we further observe that R_{eq} stays quite close to $-2/g_{m0}$ mostly during current transition, and drops very abruptly to minus infinity at the edges of transition region. For example, for $I_{D1} : I_{D2} = 9 : 1$, $R_{eq} = -3/g_{m0}$ and for $I_{D1} : I_{D2} = 99 : 1$. $R_{eq} = -7.8/g_{m0}$.We simplify the large-signal model as $R_{eq} = -2/g_{m0}$ during transition and $R_{eq} = \infty$ outside the transition region.



(a)



(b)

Fig. 6.10 Calculation swing of LC tank oscillators.

Owing to the resonance between L and C, the output waveforms V_P and V_Q are close to sinusoids. Recognizing that the M_1-M_2 pair switches all tail current to one side if

$$|V_P - V_Q| \geq \sqrt{\frac{2I_{SS}}{\mu_n C_{ox}(\frac{W}{L})_{1,2}}}, \tag{6.19}$$

we calculate the current transition region based on the ratio of magnitude. Assuming peak value of $V_P - V_Q$ is $\pm V_0$ and denoting transition time as T_T, we arrive at

$$\frac{\sqrt{\frac{2I_{SS}}{\mu_n C_{ox}(\frac{W}{L})_{1,2}}}}{V_0} = \frac{\pi}{2} \cdot \frac{\frac{T_T}{2}}{\frac{T_0}{2}}, \tag{6.20}$$

where T_0 denotes one clock period. The modification factor $\pi/2$ comes from the slope difference between tangent at origin and the straight line connecting peak and origin [Fig. 6.10(b)]. The key point here is that the average"transconductance" $(1/R_{eg})$ is exactly equal to $(2R_p)^{-1}$:

$$\frac{g_{m0}}{2} \times \frac{2T_T}{T_0} + 0 \cdot \frac{(T_0 - 2T_T)}{T_0} = \frac{1}{2R_P}. \tag{6.21}$$

Combining these two equations, we obtain

$$V_0 \cong 0.9 I_{SS} R_P. \tag{6.22}$$

That is, the differential output ($V_P - V_Q$) of an LC-tank oscillator presents swing of $\pm 0.9 I_{SS} R_P$. This is an important result as we will need it in phase noise calculation. Figure 11 illustrations the predicted and simulated swings of it.

6.2.3 Phase Noise

Phase noise only considers the time-domain wandering (i.e., jitter) around the crossover points. In calculating phase noise we focus on the center of current transition region, where $g_{m1} = g_{m2} = g_{m0}$. Both the thermal noise and the flicker noise of M_1 and M_2 are shaped by the response of the tank. Since R_P is cancelled out by the negative resistance R_{eg}, the tank's impedance Z becomes

$$Z = sL // \frac{1}{sC}. \tag{6.23}$$

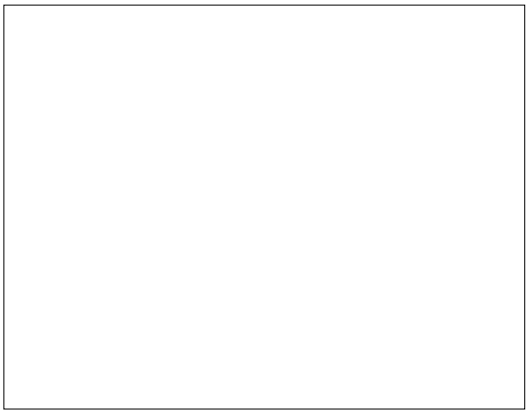


Fig. 6.11 Simulated and calculated swings.

It goes to infinity at $\omega_0 = (LC)^{-1/2}$. The square of magnitude of Z at $\triangle\omega$ deviating from the resonance frequency is therefore obtained as

$$\begin{aligned} |Z|^2 &= \left| \frac{j(\omega_0 + \Delta\omega)C}{-(\omega_0 + \Delta\omega)^2 + \omega_0^2} \right|^2 \\ &\cong \frac{1}{4C^2 \triangle \omega^2}, \end{aligned} \tag{6.24}$$

which is inversely proportional to $\Delta\omega^2$. Multiplying the current noise spectrums of thermal and flicker noise by $|Z|^2$. we arrive at

$$\begin{aligned} S_{n,out} &= (\overline{I_{n,M1,T}^2} + \overline{I_{n,M2,T}^2}) \cdot |Z|^2 + (\overline{I_{n,M1,1/f}^2} + \overline{I_{n,M2,1/f}^2}) \cdot |Z|^2 \\ &= 4kT\gamma \cdot 2g_{m0} \cdot \frac{1}{4C^2\Delta\omega^2} + \frac{K}{C_{ox}(WL)_{1,2}} \cdot 2g_{m0}^2 \cdot \frac{1}{4C^2\Delta^3}. \end{aligned} \tag{6.25}$$

As shown in Fig. 6.12, the output noise spectrum has a steeper slope for in-band (i.e., closer to ω_0) noise. Namely, $S_{n,out}$ falls at a rate of $\Delta\omega^{-3}$ in low offset region and a rate of $\Delta\omega^{-2}$ in medium offset part. The intersection is approximately equal to

$$\Delta\omega^* = \frac{K \cdot g_{m0}}{4C_{ox}(WL)_{1,2}kT\gamma}. \tag{6.26}$$

To obtain phase noise, $S_{n,out}$ must be divided by the signal power. Expressing it in decibels, we obtain

$$L(\Delta\omega) = 10\log_{10}\frac{S_{n,out}}{\frac{(0.9I_{SS}R_P)^2}{2}} \cong 10\log_{10}(\frac{2.5S_{n,out}}{I_{SS}^2R_P^2}) \quad (dBc/Hz). \tag{6.27}$$

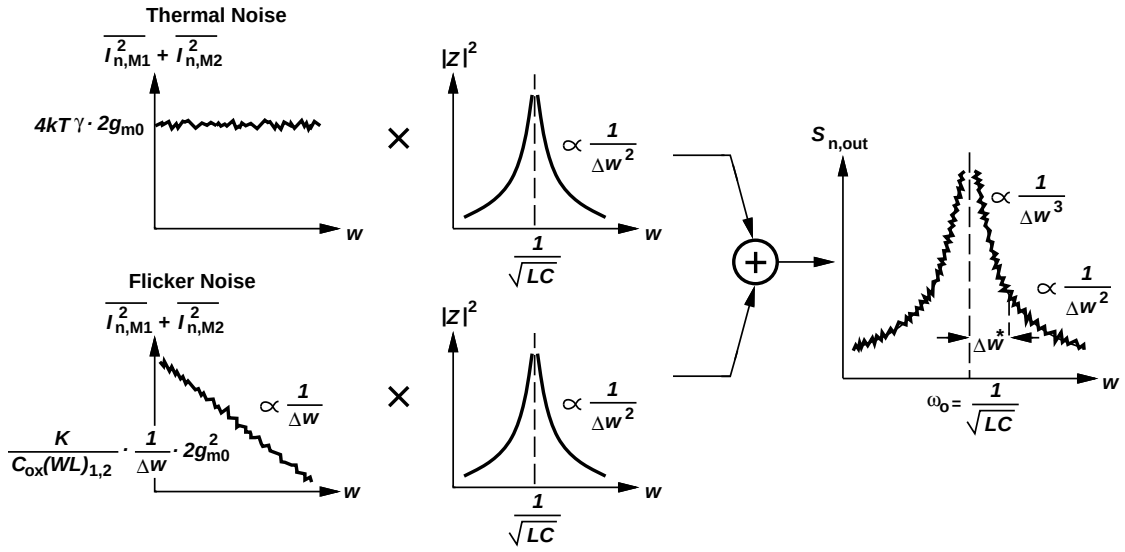


Fig. 6.12 Noise calculation.

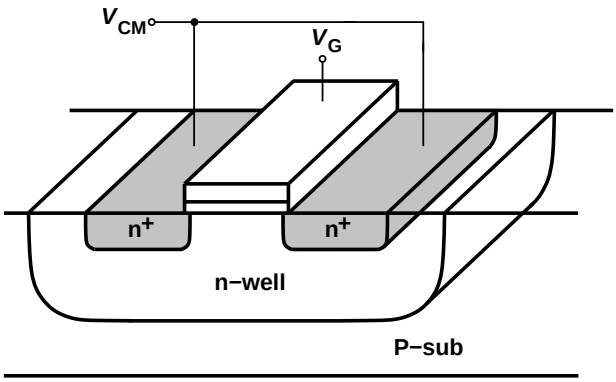


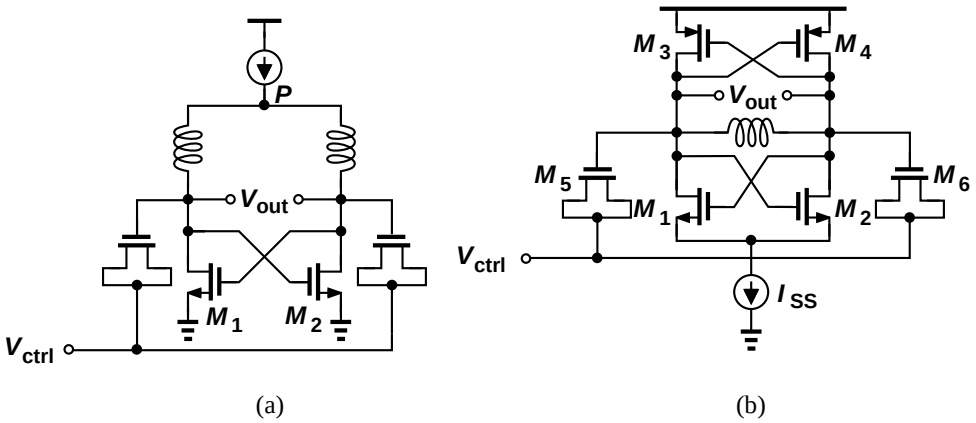
Fig. 6.13 Varactor in CMOS process.

Not only the cross-coupled pair, the tail current presents noise to the output, too. Since the common-source point R in Fig. 10(a) experiences 2nd-order harmonic swing, the tail current noise around $2\omega_0$ down-converts to ω_0 by the mixing behavior of M_1-M_2 pair and adds itself to the output. More specifically, the output noise contributed by the tail current is given by

$$S_{n,out,M3} = \overline{I_{n,M3}^2} \cdot (\frac{2}{\pi})^2 \cdot \frac{1}{4C^2\Delta\omega^2} \quad (V^2/Hz), \tag{6.28}$$

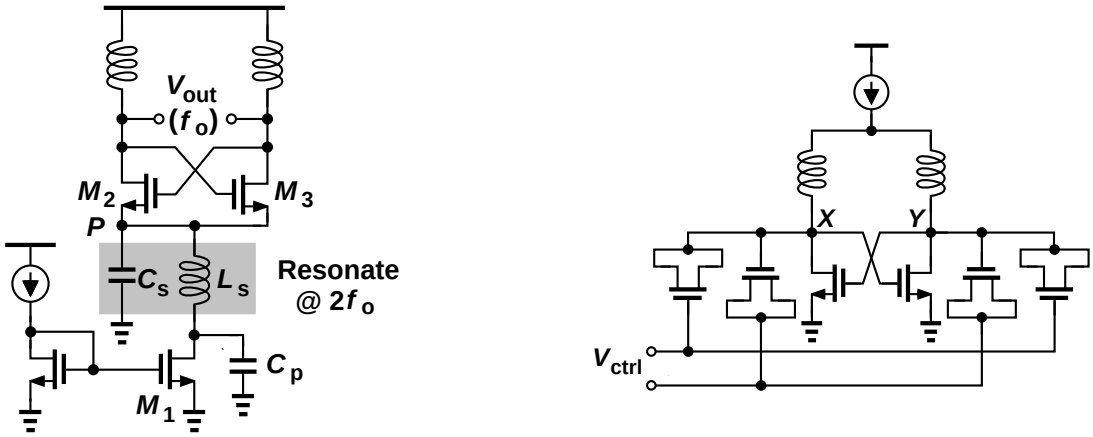
where $2/\pi$ denotes the mixing gain and $\overline{I_{n,M3}^2}$ the noise current of tail current. The tail current contributes roughly commensurate amount of noise as the cross-couple pair does.

LC-tank VCOs utilize varactors to tune the frequency. The varactors are actually realized as a NMOS device in a n-well. Depending on the relative potential between gate (V_G) and source -drain combination (V_{CM}), the device channel forms a variable capacitor. Figure 13 illustrates the structure. The capacitance of such a varactor is tuned monotonically as a function of $V_G - V_{CM}$. The maximum value could be more than twice as much as the minimum.



(a) (b)

Fig. 6.14 LC oscillators with (a) top biasing, (b) dual pairs.



(a) (b)

Fig. 6.15 (a) Tail current noise rejection (b) differential control.

Several techniques have been developed to improve the performance of LC oscillators. Figure 14(a) shows a popular topology moving the tail current to the top to setup the output common-mode around V_{DD}/2. The noise of the top current source may disturb the voltage at node P and hence modulate the frequency, resulting in higher phase noise. Figure 14(b) incorporates PMOS devices, but the oscillation frequency (or tuning range) degrades due to the extra parasitic capacitance. Note that the tail current plays important role here, because it defines the bias current (and hence the output amplitude if the inductor Q is known) while giving a high impedance to ground so as to maintain a more constant quality factor for oscillation. The tail current in Fig. 6.14(a) and (b) can be removed to accommodate low-supply operation but at a cost of higher supply sensitivity. The noise of tail current could be blocked to improve the phase noise performance. As illustrated in Fig. 6.15(a), a large bypass capacitor C_p absorbs the noise of the current source M_1. With the $L_s - C_s$ network resonating at twice the output frequency, the common-source node P still experiences a high impedance to ground. Differential voltage control is also achievable by adding two sets of varactors with opposite direction, as illustrated in Fig. 6.15(b). The differential operation improves the common-mode rejection by 10-20dB.

A few techniques are commonly-used to extend the tuning range. A straightforward method is to employ a capacitor array (preferably binary-weighted for better efficiency) to tune the VCO coarsely [Fig. 6.16(a)] [7]. Such a band selection mechanism sometimes benefits the PLL design because the VCO gain becomes smaller.

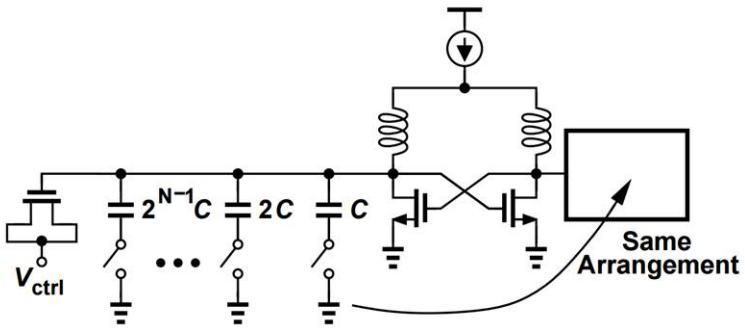
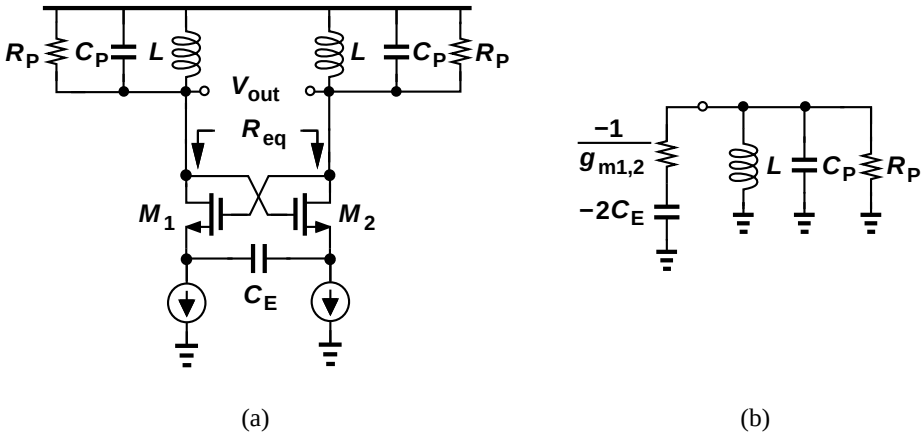


Fig. 6.16 LC-tank VCO with switched capacitance.

It is noteworthy that the basic cross-coupled oscillators are very efficient, and careless modification of the structure can lead to unpredictable results. One example using capacitive degeneration is illustrated in Fig. 6.17(a). Like a relaxation oscillator, the impedance seen looking into the cross-coupled pair is given by

$$R_{eq} = -\frac{2}{g_{m1,2}} - \frac{1}{sC_E}, \tag{6.29}$$



(a) (b)

Fig. 6.17 Cross-coupled VCO with capacitive degeneration.

and the equivalent small-signal model is shown in Fig. 6.17(b). Intuitively, such a degeneration provides a negative capacitor to cancel out part of the positive capacitor C_P, raising the oscillation frequency. In reality, however, this frequency boosting is accomplished at a cost of weakening the negative resistance, making the circuit harder to oscillate. To see why, let us first consider a general transformation between series and parallel networks. As shown in Fig. 6.18(a), a series circuit containing R_1 and C_1 can be converted to a parallel one by equating the impedance. Defining Q_d as $1/(R_1C_1\omega)$, we arrive at

$$\frac{R_2}{R_1} = 1 + Q_d^2 \tag{6.30}$$

$$\frac{C_2}{C_1} = \frac{Q_d^2}{1 + Q_d^2} \tag{6.31}$$

where R_2 and C_2 are of the equivalent parallel combination. Figure 18(b) plots R_2 and C_2 as a function of Q_d. Obviously, depending on Q_d, the transformed network behaves differently. For

$Q_d \gg 1, C_2 \approx C_1$ and $R_2 \approx Q_d^2 R_1$; whereas for lower Q_d, both R_2 and C_2 degrade. Applying this result into fig. 6.17(b), we arrive at the small-signal model in Fig. 6.19. The resonance frequency now becomes

$$\omega_{osc} = \frac{1}{\sqrt{L(C_P - 2C_E \cdot \frac{Q_d^2}{1+Q_d^2})}}. \tag{6.32}$$

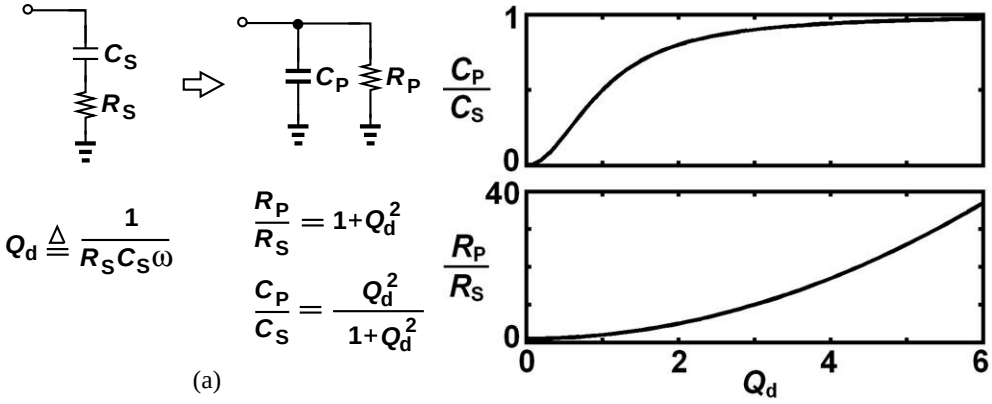


Fig. 6.18 Conversion between series and parallel RC network.

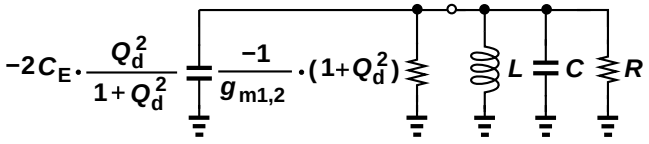


Fig. 6.19 Modification of Fig 17. (b)

Although degraded, it is indeed a boost in the frequency. However, a more difficult condition is imposed on the start-up oscillation:

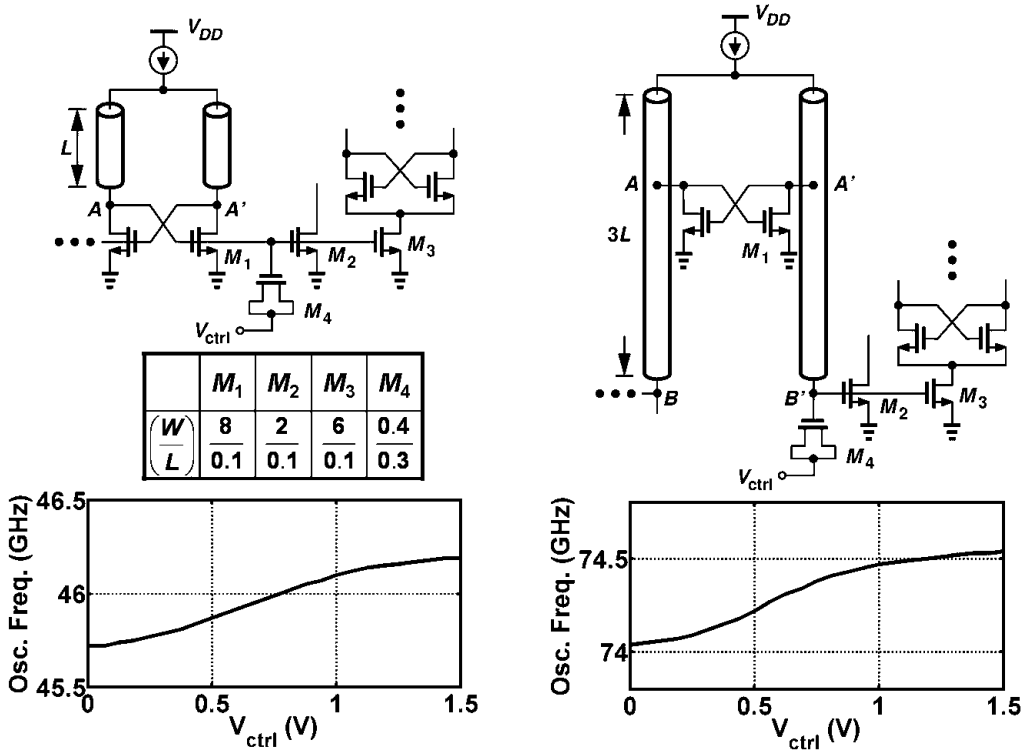
$$g_{m1,2} \geq \frac{1+Q_d^2}{R_P}. \tag{6.33}$$

For a Q_d of 3, this circuit needs a transconductance 10 times larger in order to ignite (and maintain) the oscillation. As a result, wider devices may be required to implement $M_{1,2}$, leading to less improvement or even deterioration in oscillation frequency. The circuit may consume more power as well.

6.3 ADVANCED LC-TANK OSCILLATORS

The fundamental LC-tank architecture can be further modified or transformed to create oscillators with more functions or better performance. We study a few representative techniques in this section.

6.3.1 $\lambda/4$ Oscillators



(a) (b)

Fig. 6.20 (a) Conventional LC tank VCO, (b) 3λ/4 transmission-line VCO.

The cross-coupled LC-tank VCO introduced in 6.2 can also be modeled as a short-circuited quarter-wavelength ($\lambda/4$) resonator. Figure 20(a) illustrates such a design based on transmission lines. It consists of a simple buffer (M_3) , an injection locked divider (M_4 and M_6-M_7), and a MOS varactor (M_5) . The circuit oscillates at a frequency such that the corresponding wavelength is 4 times as large as the equivalent length L , leaving the ends (node A and A') as maximum

swings. However, as the resonance frequency increases, the loading of the varactors, the buffers, and the dividers becomes significant as compared with that of the cross-coupled pair itself. These indispensable capacitances burden the VCO substantially. Note that none of these devices can be made arbitrarily small: M_1-M_2 pair must provide sufficient negative resistance, transistor M_4 needs to inject large signal current, and M_5 has to provide enough frequency tuning. With the device dimension listed in Fig. 6.20(a), the circuit oscillates at only 46 GHz. Note that the device sizes have approached the required minimum and further shrinking may cause significant swing degradation.

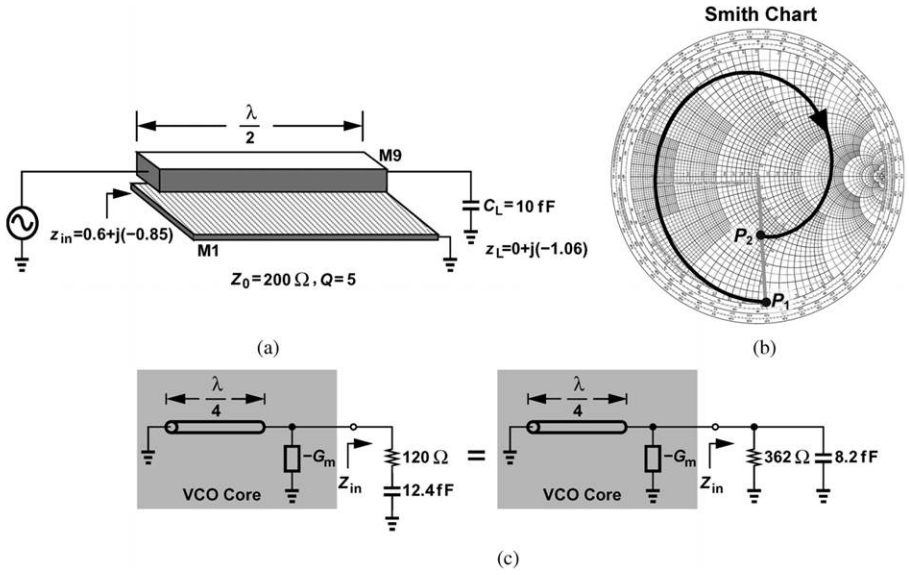


Fig. 6.21 Impedance transformation: (a) half-wavelength microstrip line, (b) rotation on Smith chart, (c) series-to-parallel conversion.

To overcome the above difficulty, we introduce transmission lines equivalent to three-quarter wavelength ($3\lambda/4$) of a 75-GHz clock to distribute the loading and boost the oscillation frequency. As can be shown in Fig. 6.20(b), these lines have one end short-circuited and the other open-circuited, resonating differentially with the cross-coupled pair M_1-M_2 providing negative resistance. Connecting to the one-third points of the lines (nodes A and A'), this pair forces the transmission lines to create peak swings at these nodes. The waves thus propagate and reflect along the lines, forming the second maximum swings with opposite polarities at nodes B and B'. That is,

node A(A') and node B(B') are 180° out of phase. As a result, the buffers, dividers, and varactors can be moved to these ends to relax the loading at nodes A and A', making the two zenith positions bear approximately equal capacitance. With the same device dimension [M_{1-5} in Fig. 6.20(a)], the oscillation frequency raises up to around 75 GHz, which is a 60% improvement without any extra power dissipation.

The reader may wonder why the loading capacitance at node B(B') would look differently at node A(A'). Indeed, the loading at nodes A(A') and B(B') will appear identical if the transmission line is lossless, since the $\lambda/2$ line rotates the loading impedance by exactly 360° along the outmost circle of the Smith chart. However, in a lossy line the equivalent capacitance seen from node A(A') toward the load does become lower. The magnitude attenuation translates the purely capacitive loading into a lossy but smaller capacitor. Consider a typical microstrip line with $\lambda/2$ length as shown in Fig. 6.21(a). Made of 1-μm wide M_9 on top of M_1 ground plane, this transmission line would present a characteristic impedance (Z_0) of about 200 Ω and a quality factor (Q) of 5. Denoting the real and imaginary parts of the propagation constant as α and β, we have $Q = \beta/(2\alpha)$ and therefore $\alpha = \pi/(Q\lambda)$. The 10-fF loading capacitor (representing the capacitance of the buffer, the divider, and the varactor) locates at P_1 with a normalized impedance $z_L = 0 + j(-1.06)$. To calculate the input impedance, we rotate clockwise z_L by 360° with the radius decreasing by a factor of exp$[-2\alpha \cdot (\lambda/2)]$:

$$exp\left(-2\alpha \cdot \frac{\lambda}{2}\right) = exp\left(-2 \cdot \frac{\pi}{Q\lambda} \cdot \frac{\lambda}{2}\right) = exp\left(-\frac{\pi}{Q}\right) = 0.53. \tag{6.34}$$

As depicted in Fig. 6.21(b), the new location P_2 represents the normalized impedance z_{in} which is 0.6+$j(-0.85)$. It corresponds to a 12.4 fF capacitor in series with a 120-Ω resistor, which can be further translated to a parallel network (8.2 fF and 362 Ω) at 75 GHz [Fig. 6.21(c)]. In other words, the three-quarter wavelength VCO in Fig. 6.20(b) experiences 18% less capacitance from M_3-M_5 at a cost of higher loss, which can be compensated by the negative resistance from the cross-coupled pair. Note that the capacitance reduction becomes higher if Z_0 goes higher.

The transmission lines could be replaced by spiral inductors to increase Q and save area. With the use of spiral inductors, an alternative way to explain the frequency boosting can be found by using the lumped model in Fig. 6.22. Here, we assume M_1 and M_3-M_5 in Fig. 6.20 present

equivalent capacitance of $C/2$ (which is true in our design), and each inductor is denoted as L. Since nodes A and B oscillated at the same frequency, there must exist a virtual ground point x located at somewhere along the third inductor such that the Network I and Network II have the same resonance frequency ω_0:

$$\omega_0 = \frac{1}{\sqrt{[L||(2-x)L)] \cdot \frac{C}{2}}} = \frac{1}{\sqrt{xL \cdot \frac{C}{2}}}. \tag{6.35}$$

It follows that

$$x = 0.59 \tag{6.36}$$

and

$$\omega_0 = \frac{1.84}{\sqrt{LC}}. \tag{6.37}$$

Such a first-order model implies a frequency improvement of 84%.

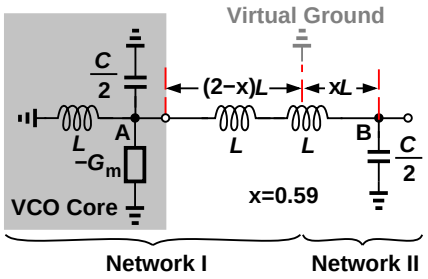


Fig. 6.22 Frequency estimation of quarter wavelength VCO with lumped model.

The above analyses also imply that, although the varactors hang on nodes B and B', the cross-coupled pair still "sees" the loading variation at these far ends through the two-third segments of the lines. Since the resonance frequency is determined by the inductance of the first one-third segment and the overall equivalent capacitance associated with node $A(A')$, the tuning of the VCO presents monotonic increasing, similar to that of a regular LC tank VCO. Fig. 6.23 shows the simulated waveforms on different nodes of the VCO.

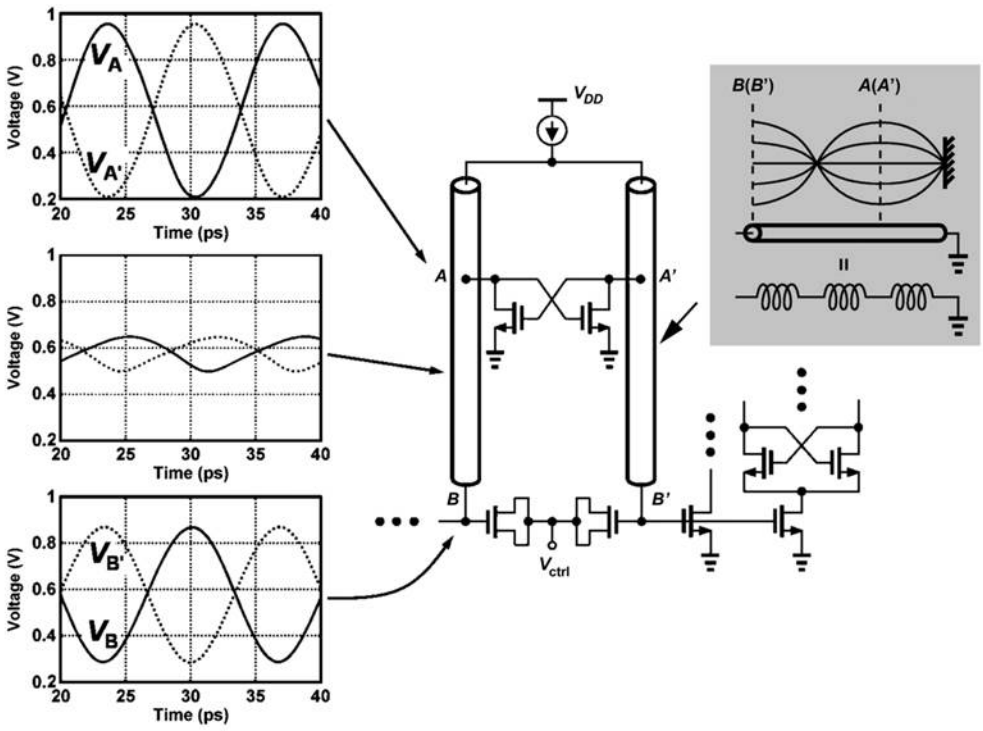


Fig. 6.23 Simulated waveforms of $3\lambda/4$ transmission line VCO.

6.3.2 Temperature Compensation

LC-tank VCOs usually have limited tuning range, and it is especially true for those operating at high frequencies. We might lose the full coverage of the bands of interest as temperature varies. Temperature compensation becomes important if we need to keep a roughly constant oscillation frequency over wide temperature range.

As temperature goes up, threshold voltage of a device also goes up. That is, for an LC-tank VCO with constant tail current(from bandgap), the cross-couple pair $M_{1,2}$ must raise up their V_{GS} (i.e., output common-mode level) so as to accommodate the constant current. As a result, the oscillation frequency decreases as temperature increases. A simple modification can efficiently suppress the deviation. As shown in Fig. 6.24, a PTAT current pulls out part of current from I_{SS}, keeping the output common-mode level unchanged. In other words, the oscillation frequency maintains constant as temperature varies.

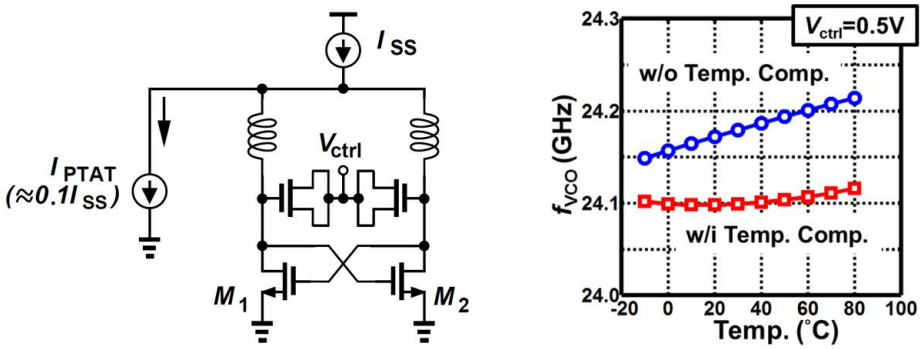


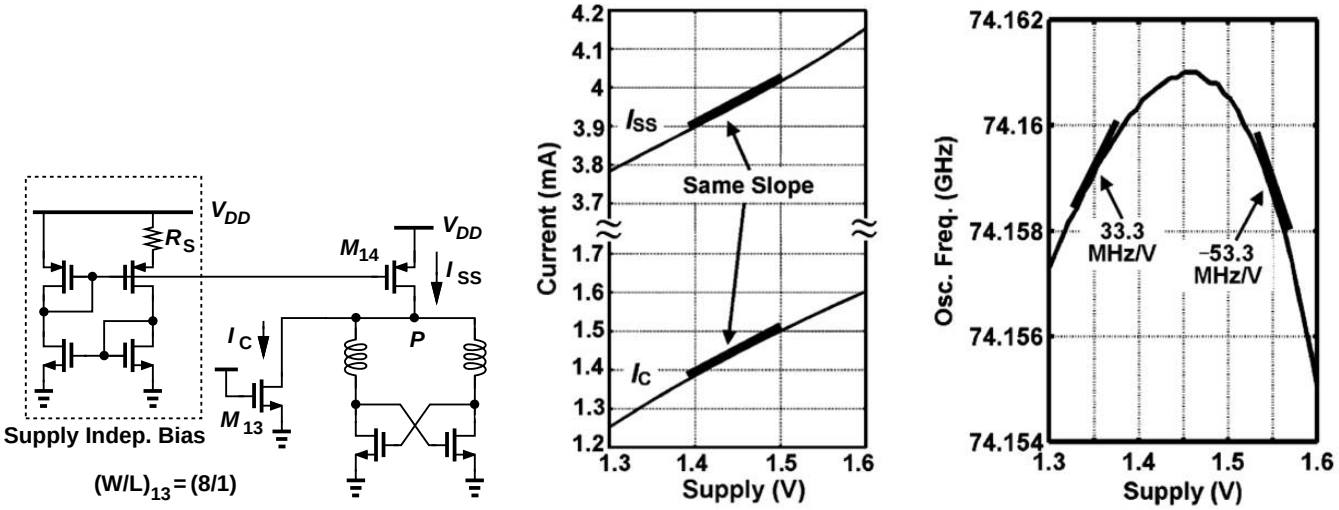
Fig. 6.24 Temperature compensation technique.

6.3.3 Supply Insensitive Biasing

Supply noise could also be applied to VCOs, disturbing the oscillation frequency. Again we take the design in 6.3.1 as example. To suppress the coupling from power lines, the VCO can be biased with a supply-independent circuit (M_9-M_{12} and R_S), as illustrated in Fig. 6.25(a). Here, we introduce M_{13} to absorb extra current variation caused by channel-length modulation to further reject the supply noise. That is, by proper sizing we set

$$\left|\frac{\partial I_{SS}}{\partial V_{DD}}\right| = \left|\frac{\partial I_C}{\partial V_{DD}}\right|, \tag{6.38}$$

letting the current flowing into M_1-M_2 pair remain constant [8]. By the same token we used in temperature compensation, we can minimize the frequency deviation here. Fig. 6.25(b) shows the currents through $M_{13}(I_C)$ and $M_{14}(I_{SS})$ as functions of supply voltage, suggesting an equal slope in the vicinity of 1.45 V. In other words, the voltage at node P is fixed, leaving the resonance frequency insensitive to supply perturbation [Fig. 6.25(c)]. The power penalty of M_{13} can be restrained to as low as 20 - 30% with proper design. The performance of this open loop compensation would slightly degrade if PVT variations occur. For example, the supply sensitivity becomes 33.3 MHz/V and -53.3 MHz/V at 1.35-V and 1.55-V supplies, respectively [Fig. 6.25(c)]. Nonetheless, these results are still much better than that of a conventional design without M_{13}.



(a) (b) (c)

Fig. 6.25 (a) Supply independent biasing, (b) current variations, (c) oscillation frequency as a function of supply voltage.

6.3.4 Wideband $LC-$tank VCOs

Some applications may need to cover a very wide frequency range. The $10 \sim 15\%$ typical range for LC-tank VCOs can hardly be enough, let along the extra margin required by PVT variations. An area efficient way to implement multi-band, wide-range LC-tank VCO without using switched capacitors is depicted in Fig. 6.26. Here, severed negative resistors (created by cross-couple pair) are deployed along the resonating elements (i.e., transmission lines or spiral inductors). Only one of the N negative resistors is turned on. Since the part of resonating elements between node P and the active pair is considered $\lambda/4$ of the oscillation wave, we arrive at a very wide range. Varactors are used to tune the frequency. A one-to-N selector is placed subsequently to take out the output accordingly. Note that unlike in Fig. 6.16, the equivalent inductor here can be chosen as large as possible for a given tuning range, which theoretically leads to a better phase noise performance.

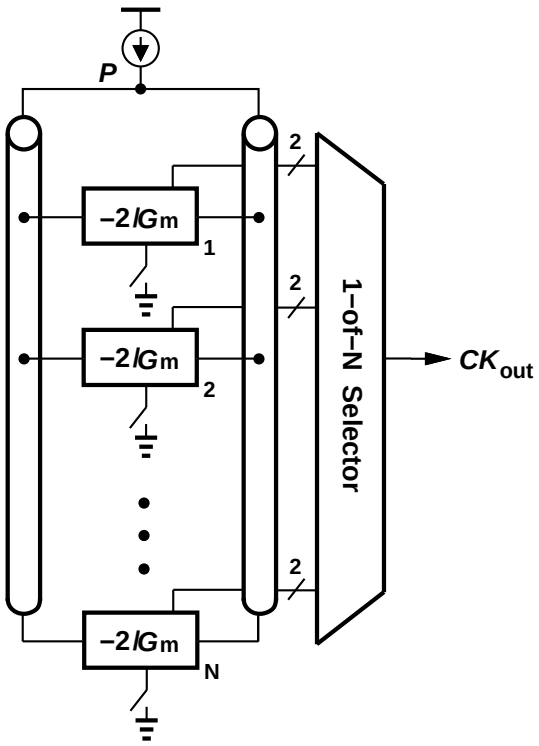


Fig. 6.26 Wide-band LC-tank VCO based on $\lambda/4$ oscillation.

6.3.5 Multiphase VCOs

Many systems require quadrature or even semi-quadrature VCOs to provide clocks with multiple phase. The most famous quadrature VCO structure is the so-called coupling quadrature VCO (QVCO). As illustrated in Fig. 6.27, it combines two basic LC-tank VCOs by direct coupling though the two M_3-M_4 pairs. Be aware of the coupling polarity: one is direct whereas the other is inversed (crossover in the enter). That gives us the model shown on the right. Two outputs V_I and V_Q are coupled through the M_3-M_4 pairs, arriving at

$$V_I \; G_{m3,4}(Z//\frac{-1}{G_{m1,2}}) = V_G \tag{6.39}$$

$$-V_Q \; G_{m3,4}(Z//\frac{-1}{G_{m1,2}}) = V_I, \tag{6.40}$$

where Z denotes the impedance of LC-tank, and $G_{m1,2}(G_{m3,4})$ the average transconductances of $M_{1,2}(M_{3,4})$, respectively. It follows that

$$V_I = \pm jV_Q, \tag{6.41}$$

indicating they are indeed in quadrature. The two possible oscillation frequencies are

$$\omega_{1,2} = \frac{\omega_0}{2Q}\left[\sqrt{4Q^2 + \frac{G_{m3,4}^2}{G_{m1,2}^2}} \pm \frac{G_{m3,4}}{G_{m1,2}}\right]. \tag{6.42}$$

Again, $\omega_o = 1/\sqrt{LC}$. Note that $\omega_1 \cdot \omega_2 = \omega_o^2$. If $(W/L)_{1,2} = 2(W/L)_{3,4}$, $Q = 10$, and $I_{SS1} = 2I_{SS2}$, we have $G_{m1,2} \cong 2G_{m3,4}$ and $\omega_{1,2} = \omega_0(1 \pm 0.025)$. Frequency tuning can be done by either adjusting the ratio of I_{SS1}/I_{SS2}, or placing varactors as we did for normal LC-tank oscillators. The phase noise performance here is usually worse than a typical LC-tank VCO simply because the oscillation frequency deviates from ω_o (where $|d\phi/d\omega|$ reaches a maximum).

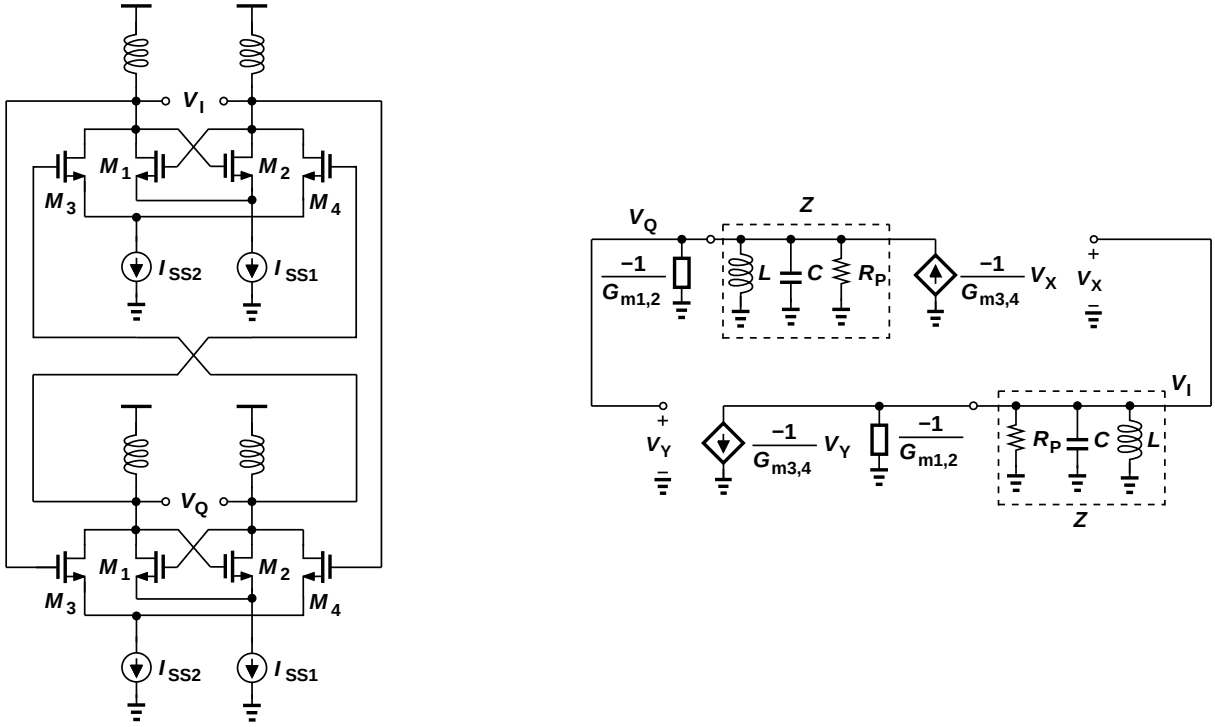


Fig. 6.27 Quadrature VCO with coupling.

The structure using 4 tail currents in Fig. 6.27 can be somewhat modified by serial pairs. To satisfy the coupling concept of Fig. 6.28(a) with less power consume, we reuses the tail currents.

Figure 28 (b)∼(d) present different coupling methods whose phase noise is independent of coupling strength. The phase noise is expected to be better than that of Fig. 6.27. Certainly, the tail currents of Fig, 27 could be merged as that in Fig. 6.28(e). It is not difficult to create clock phases finer than $90°$. Depicted in Fig. 6.29 is an example, which 4 tuned amplifiers (in differential mode) are placed in cascade with negative feedback around them. As a result, each stage is responsible for $\pm 45°$ phase shift, revealing a semi-quadrature VCO. Again, the VCO does not operate at ω_o, LC resonance frequency suggesting higher phase noise. The multiphase VCOs in Fig. 6.27∼6.29 share a serious issue: there are two possible oscillation frequencies. <<Analysis>>

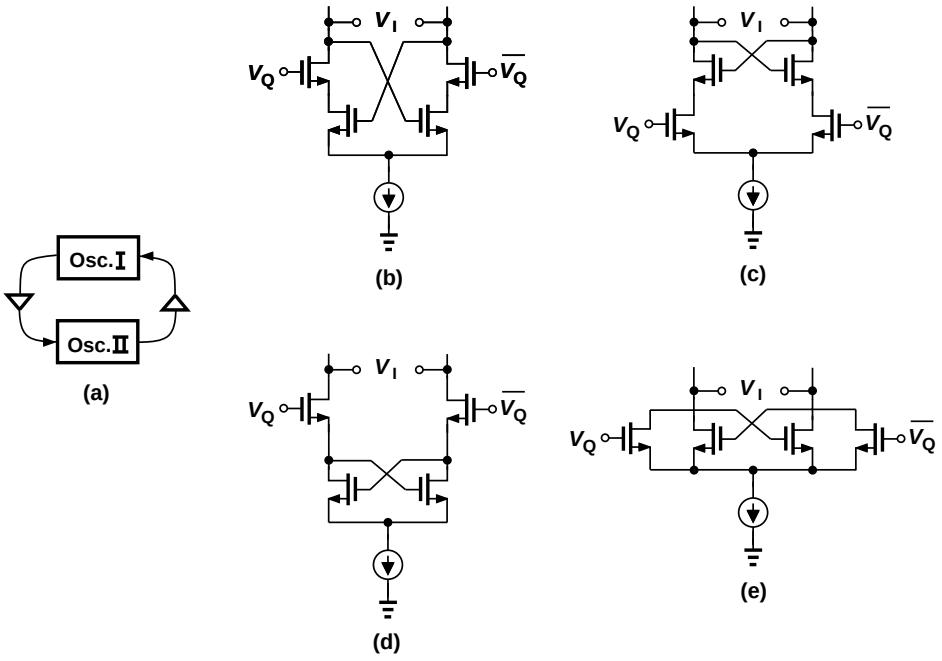


Fig. 6.28 Quadrature VCO with different coupling methods.

Nonetheless, to ensure clock phase sequence, we can use the polarity check as shown in Fig. 6.30. Taking quadrature clocks as an example, the sequence can be determined by sampling one phase with the other. Two different outputs would be obtained depending on the polarity of phase difference. Same technique can be applied to frequency detection as well. We look at it in chapter chapter 8.

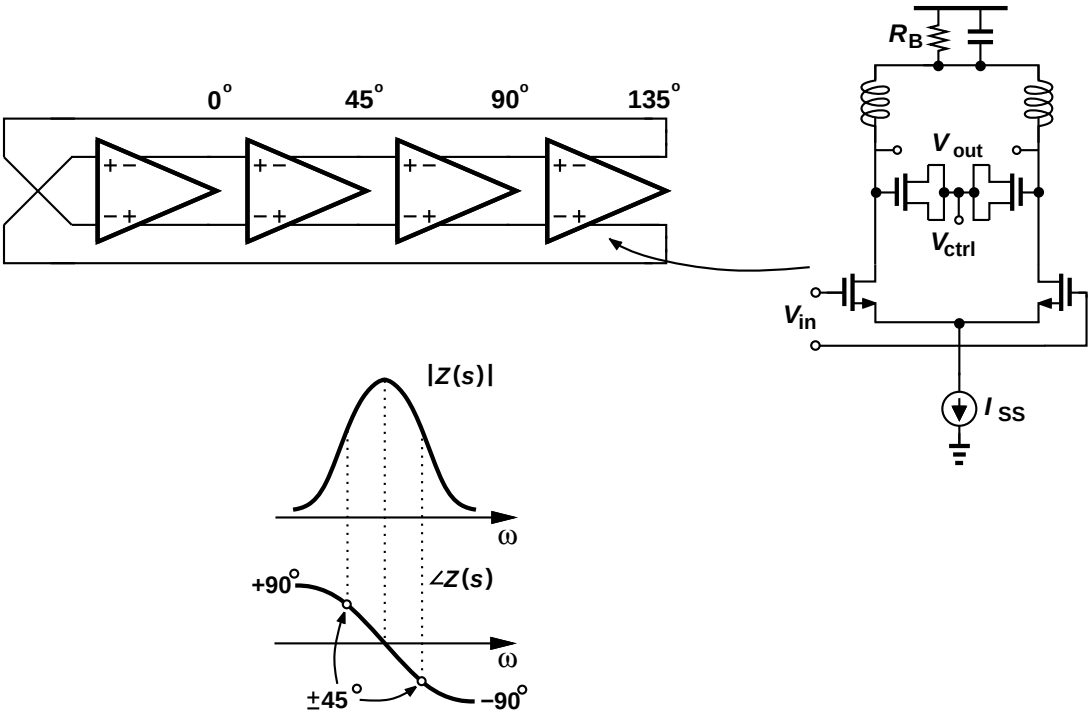


Fig. 6.29 Semi-quadrature VCO with different ring structure.

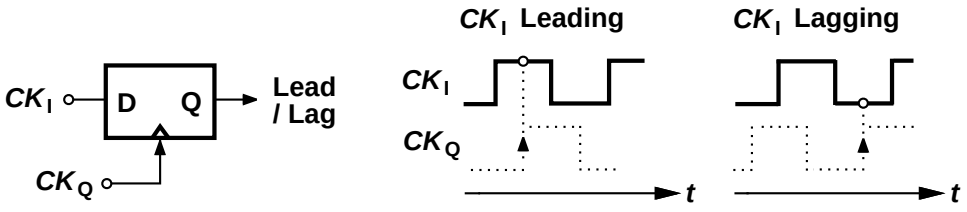


Fig. 6.30 Polarity checker for quadrature clocks.

6.4 COPITTS OSCILLATORS

Another important VCO topology that has been widely used in high-speed systems is Colpitts oscillator. First proposed in 1920's [9], this type of oscillator could be operated with only one transistor. In modern times, the abundance of transistors and the desire for differential circuits favors a symmetric Colpitts oscillators.

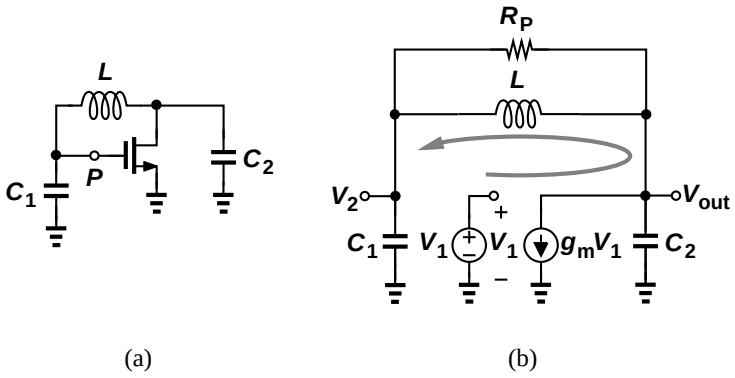


Fig. 6.31 (a) Colpitts oscillator, (b) its linear model with feedback at node P.

A Colpitts VCO can be easily understood by examining a resonating circuit shown in Fig. 6.31(a), where an inductor is sitting across the drain and gate of a MOS with two capacitors C_1 and C_2 connected to these nodes. In order to oscillate, the signal in the feedback path through the C-L-C network must satisfy Barkhausen criteria. Breaking the loop and exciting it with an input V_1 [Fig. 6.31(b)], we obtain the loop gain as

$$\frac{V_2}{V_1}(s) = -g_m \cdot \frac{R_P + sL}{s^3 LC_1C_2R_P + s^2L(C_1+C_2) + sR_P(C_1+C_2)}. \tag{6.43}$$

To make the oscillation happen at a frequency ω_{osc}, we have $|V_2/V_1| \geq 1$ and $\angle(V_2/V_1) = 0°$. In other words, at $\omega = \omega_{osc}$, the ratio of the real and imaginary parts of the numerator must be equal to that of the denominator:

$$\frac{R_P}{\omega_{osc}L} = \frac{-\omega_{osc}^2 L(C_1+C_2)}{\omega_{osc}R_P(C_1+C_2) - \omega_{osc}^3 LR_PC_1C_2}. \tag{6.44}$$

It follows that

$$\omega_{osc} = \frac{1}{\sqrt{\frac{LC_1C_2}{C_1+C_2}}} \cdot \sqrt{1+\frac{1}{Q^2}} \approx \sqrt{\frac{1}{L}\left(\frac{1}{C_1}+\frac{1}{C_2}\right)}. \tag{6.45}$$

Here, Q denotes the quality factor of the inductor and $R_P = \omega_0 LQ$. The loop gain requirement yields

$$\left|\frac{V_2}{V_1}\right|(j\omega_{osc}) = \frac{g_m R_P}{\omega_{osc}^2 L(C_1+C_2)} \geq 1. \tag{6.46}$$

As a result, we have the following condition for oscillation:

$$g_m R_P \geq \frac{(C_1+C_2)^2}{C_1C_2} \geq 4. \tag{6.47}$$

An alternative explanation of a Colpitts oscillator is to investigate the impedance seen looking into the gate-drain port of such a circuit [Fig. 6.32(a)]. It can be shown that R_{eq} is given by

$$R_{eq} = \frac{g_m}{C_1C_2s^2} + \frac{1}{C_2s} + \frac{1}{C_1s}, \tag{6.48}$$

which is equivalent to a negative resistance $-g_m(C_1C_2\omega^2)$ in series with a capacitor $C_1C_2/(C_1+C_2)$. If the quality factor of this RC network is high, we can approximate it as a parallel combination as shown in Fig. 6.32(b). Obviously the circuit may oscillate if the negative resistance is strong enough to cancel out the inductor loss R_P. As expected, the oscillation frequency is equal to

$$\omega_{osc} = \sqrt{\frac{1}{L}\left(\frac{1}{C_1}+\frac{1}{C_2}\right)}, \tag{6.49}$$

same result as Eq. (5.22). Equation. (5.24) can be obtained with a similar approach.

Depending on the bias, the prototype in Fig. 6.32(a) provides three topologies of Colpitts oscillators [10]. Among them, Fig. 6.33(a) reveals the greatest potential for high-speed operation, since C_1 can be realized by the intrinsic capacitance C_{GS} of M_1. The capacitor C_2 is replaced by a varactor M_2 to accomplish the frequency tuning. At resonance, all the components oscillate at the same frequency ω_{osc}, including the drain current of M_1. That allows us to place a loading R_D at drain and take the voltage output from this node. Inductive peaking could be an option here if the output needs to drive large capacitance.

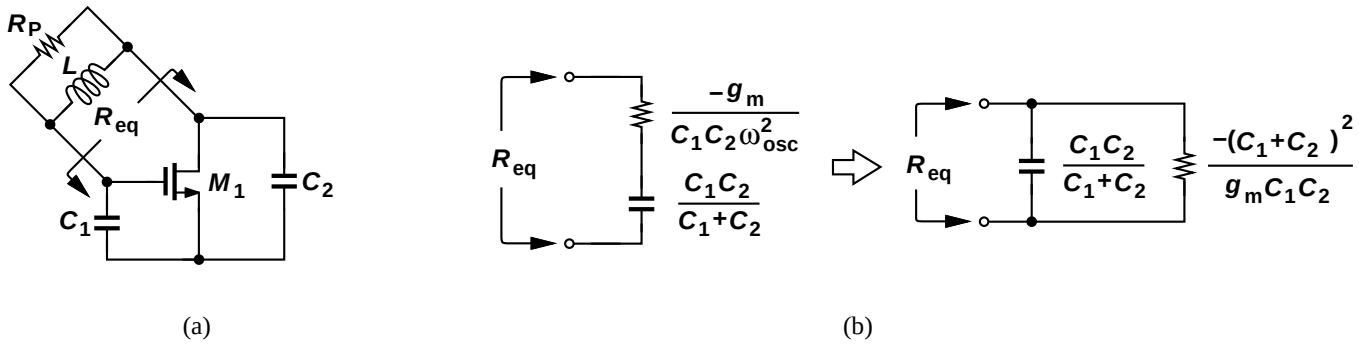
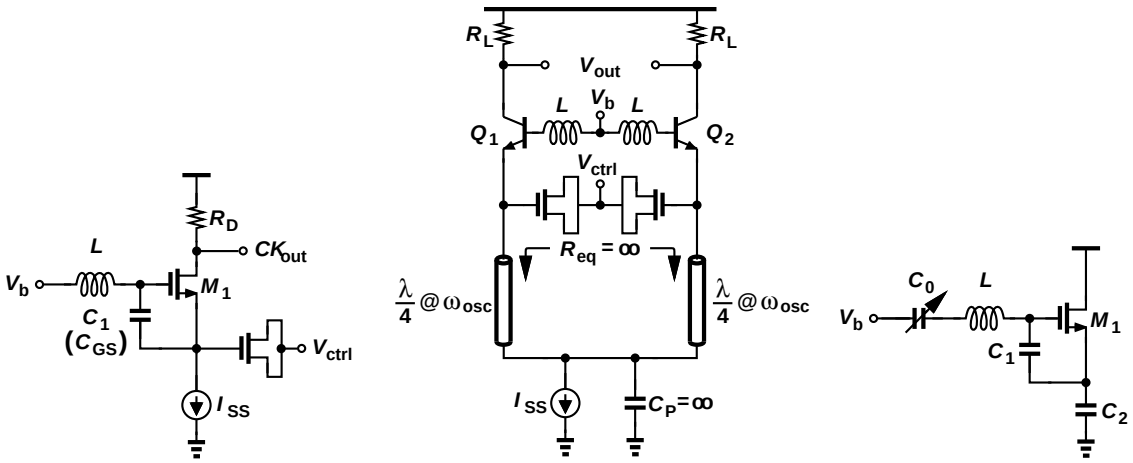


Fig. 6.32 Alternative approach to analyze oscillators by examining the equivalent impedance.

Despite many advantages, the circuit in Fig. 6.33(a) still suffers from two drawbacks: the single-ended operation makes the oscillator vulnerable to supply noise, and the capacitance contributed by the tail current source degrades the oscillation frequency. To remedy these issues, we usually implement the Colpitts oscillator as a differential configuration with $\lambda/4$-lines between $Q_{1,2}$ and I_{SS}. Figure 33(b) illustrates such a realization. The combined bias points of the symmetric circuit facilitate differential operation, and $\lambda/4$-transmission lines make the equivalent impedance looking down (R_{eq}) become infinity. Colpitts VCOs operating at 60 GHz and beyond with silicon compound technologies have been reported extensively [11], [12], [13]. In fact, the current source can be replaced with a "choke" inductor, or a sufficiently large inductor such that the impedance to ground is dominated by the capacitance for proper feedback. A Colpitts oscillator taking this approach is presented in [14], which demonstrates 104 GHz operation in a 90-nm CMOS technology.

The circuit in Fig. 6.33(a) tunes the frequency at the risk of losing stability or failing the oscillation. According to Eq. (5.24), $g_m R_P$ must be greater than $(C_1 + C_2)^2/(C_1 C_2)$, which varies as the control voltage changes. To guarantee safe margin for oscillation, one can introduce another capacitor C_0 (which is variable) in series with L and level C_1 and C_2 fixed as depicted in Fig. 6.33(b). The oscillation frequency therefore becomes

$$\omega_{osc} = \sqrt{\frac{1}{L}\left(\frac{1}{C_0} + \frac{1}{C_1} + \frac{1}{C_2}\right)}. \tag{6.50}$$



(a) (b) (c)

Fig. 6.33 (a) Common-drain Colpitts oscillator in CMOS, (b) differential realization in bipolar, (c) Clapp oscillator.

Also known as “Clapp oscillator” , this circuit inevitably suffers from less tuning range.

One important application of Colpitts oscillators is the so-called “Pierce oscillator”. As shown in Fig. 6.34, it incorporates a piezoelectric crystal (serving as an inductor) and two capacitors C_1 and C_2 to form a Colpitts oscillator. Here, the crystal can be modeled as a series RLC network (i.e., L, C_S and R_S) in parallel with another capacitor C_P and $C_P \gg C_S$. Similar to M_1 in Fig. 6.32(a), the inverter-like amplifier M_1 and M_2 provides negative resistance to compensate for the loss. Note that the circuit is self-biased through R_1 such that both M_1 and M_2 are in saturation. The reader can easily prove that the oscillation frequency is equal to

$$\omega_{osc} \approx \frac{1}{\sqrt{LCs}}, \tag{6.51}$$

which is an unchangeable value for a given crystal. To increase oscillation stability, R_2 can be added in the loop to dampen the higher order harmonics. Such a crystal-based oscillator achieves marvellous frequency stability in the presence of temperature variation, and is extensively used as a reference clock in various applications.

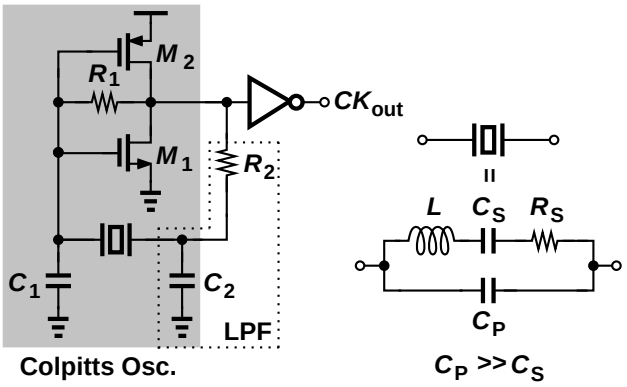


Fig. 6.34 Example of Pierce oscillator.

6.6 PUSH-PUSH OSCILLATORS

One important application of $\lambda/4$ transmission line technique is the push-push oscillators. As suggested by its name, this type of oscillator takes the 2nd-order harmonic from the common-mode node, and amplifies it properly as an output. Note that second-order harmonic is generated by the nonlinearity of the circuit, which manifests itself in large-signal operation. Figure 35 reveals an example, where V_P needs to swing up and down at twice the fundamental frequency so as to maintain a constant I_{SS}. Similar to a frequency doubler, the desired harmonic can be extracted while the others are suppressed.

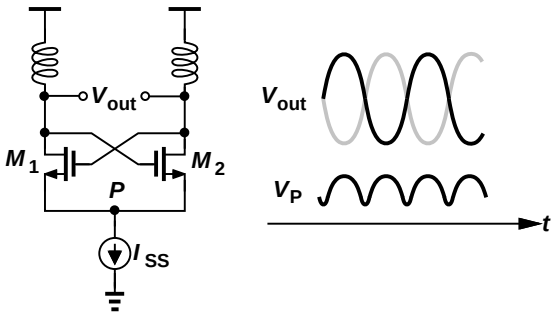


Fig. 6.35 Generation of 2nd-order harmonic.

Since node P suffers from large parasitic capacitance, we usually resort to other common-mode points to obtain the output. Two examples of circuit-level realization based on cross-coupled

and Colpitts structures are illustrated in Fig. 6.36. The $\lambda/4$ lines in both cases reinforce the $2\omega_{osc}$ signal by providing an equivalent open at node P when looking into it, and the output power could be quite large if proper matching is achieved. Compared with typical frequency doublers, this topology consumes less power and area, resulting in a more efficient approach. More details are described in [15], [16].

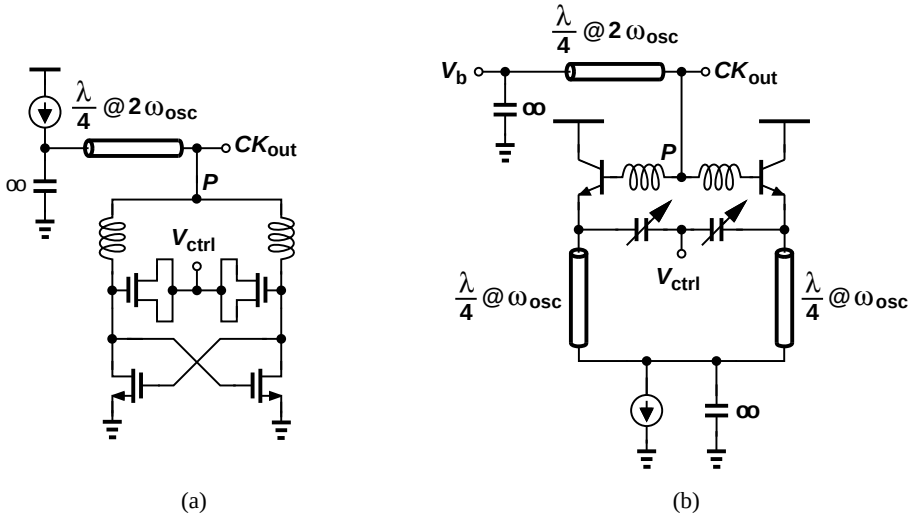


Fig. 6.36 Push-push VCOs based on (a) cross-coupled, (b) Colpitts topologies.

The push-push oscillator can only provide a single-ended output. In addition, tuning the fundamental frequency could result in a mismatch in the $\lambda/4$ lines, potentially leading to lower output power.

6.6 DISTRIBUTED OSCILLATORS

Another distinctive VCO topology shooting for high-speed operation is the distributed oscillator. As shown in Fig. 6.37, the output of a distributed amplifier is returned back to the input, yielding wave circulation along the loop. Oscillation is therefore obtained at any point along the transmission line. Here, the transmission line loss is overcome by the gain generated along the line. To be more specific, we assume the two propagation lines in Fig. 6.37 to be identical, i.e., the characteristic impedances, group velocities, and physical lengths are the same. The oscillation

period under such circumstances is nothing more than twice the propagation time along the length l:

$$f_{osc} = \frac{1}{2l\sqrt{L_0 C_0}} \tag{6.52}$$

where L_0, C_0 denote the equivalent inductance and capacitance (with the MOS capacitance included) per unit length. It can be shown that the oscillation frequency is commensurate with the device f_T [7].

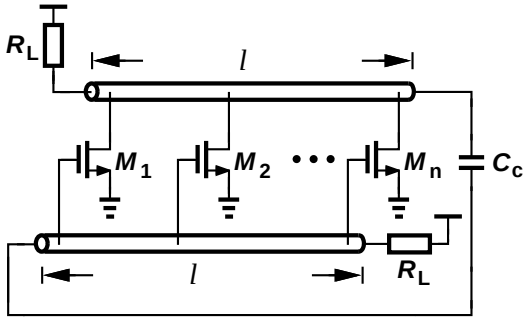


Fig. 6.37 Distributed oscillator.

While looking attractive, the distributed oscillator suffers from a number of drawbacks: (1) the group velocities along the two lines may deviate from each other due to the difference between the gate and drain capacitance; (2) the circuit would need larger area and higher power dissipation, (3) the frequency tuning could be difficult. The third point becomes clear if we realize that adding any varactors to the lines can cause significant degradation on the oscillation frequency and the quality factor Q. Varying the bias voltage of the transistors may change the intrinsic parasitics (and therefore the oscillation frequency) to some extent, but the imbalanced swing and the mismatch between the lines could make things worse. The circuit may even stop oscillating in case of serious deviation. Note that placing a "short-cut" on the lines by steering the current of two adjacent transistors is plausible as well [17]: it is hard to guarantee that the wave still propagates appropriately along the lines while both devices are partially on.

A modification of distributed oscillators can be found if we terminate a transmission line by itself. The circuit is based on the concept of the differential stimulus of a closed-loop transmission line at evenly-spaced points, as illustrated conceptually in Fig. 6.38(a). In contrast to regular

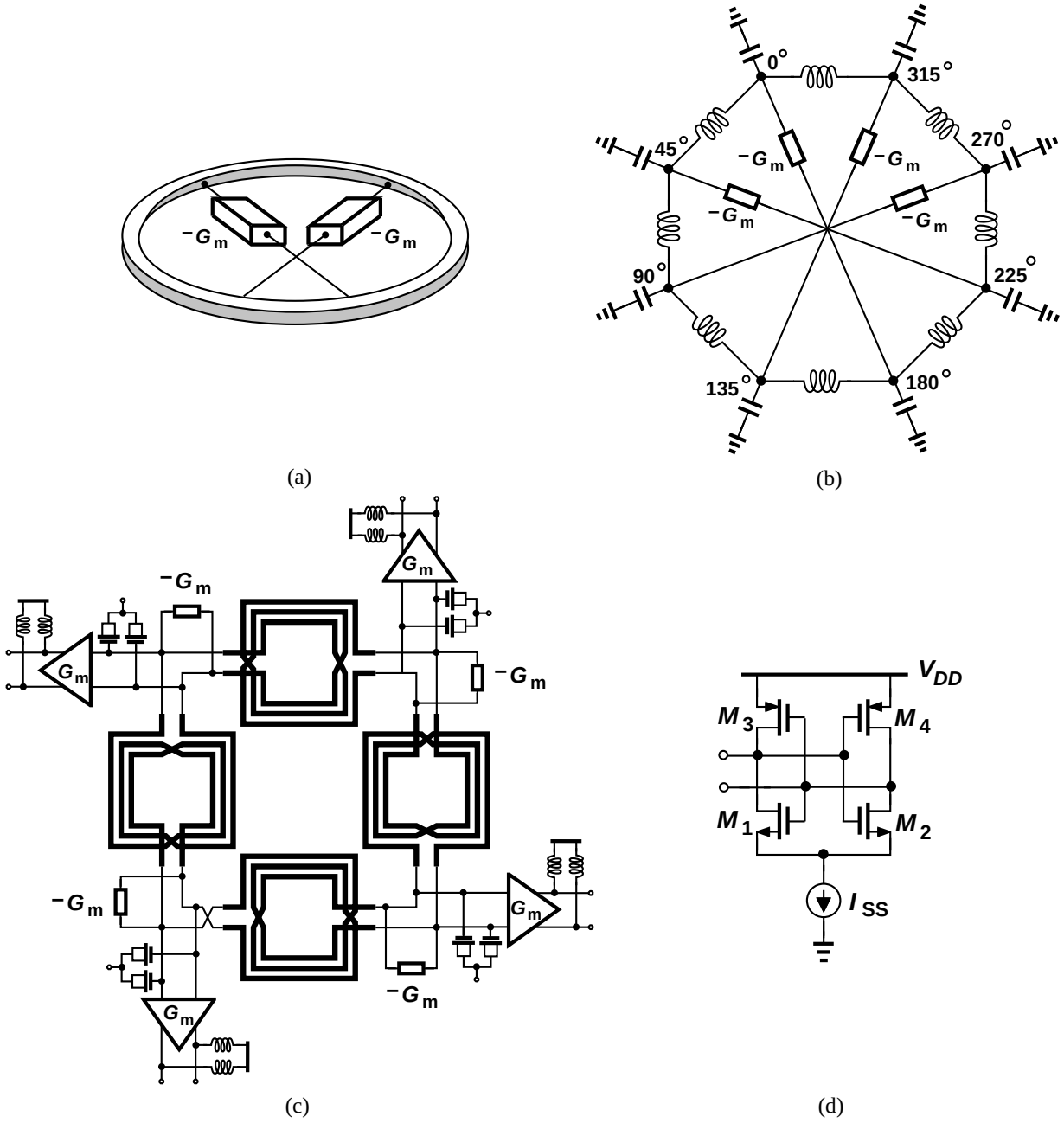


Fig. 6.38 (a) Oscillator based on closed-loop transmission line, (b) half-quadrature realization, (c) modification of (b), (d) implementation of $-G_m$ cell.

distributed oscillators, the transmission line requires no termination resistors, lowering phase noise and enlarging voltage swings. The circuit can be approximated by lumped inductors and capacitors, and one example is shown in Fig. 6.38(b). Here, eight inductors form a loop with four differential negative $-G_m$ cells driving diagonally opposite nodes. In steady state, the eight nodes are equally separated by 45°, providing multiphase output if necessary.

The oscillation frequency of the circuit is uniquely given by the travel time of the wave around the loop. We write the oscillation frequency of this topology as

$$f = \frac{1}{8\sqrt{LC}} \tag{6.53}$$

where L and C, respectively, denote the lumped inductance and capacitance of each of the eight sections. The circuit can be further modified as shown in Fig. 6.38(c) to avoid long routing, and the negative $-G_m$ cell can be simply implemented as Fig. 6.38(d). The PMOS transistors help to shape the rising and falling edges while providing lower $1/f$ noise.

One interesting issue in such a VCO is that, due to symmetry, the wave may propagate clockwise rather than counterclockwise. To achieve a more robust design, a means of detecting the wave direction is necessary. Since nodes that are 90° apart in one case exhibit a phase difference of -90° in the other case, a flipflop sensing such nodes generates a constant high or low level, thereby providing a dc quantity indicating the wave direction. Other approach to avoid direction ambiguity can be found in [18].

REFERENCES

[1] M. Danesh *et al*., "A Q-factor ehancement technique for MMIC inductors," *IEEE Radio Frequency Integrated Circuits (RFIC) Symposium Dig. Pape,* pp. 217-220, June. 1998.

[2] A. Zolfaghari *et al*., "Stacked inductors and transformers in CMOS technology," *IEEE J. Solid-State Circuits,* vol. 36, no. 4, pp. 620-628, Apr. 2001.

[3] J. Lee, "High-speed circuit designs for transmitters in broadband data links," *IEEE J. Solid-State Circuits,* vol. 41, no. 5, pp. 1004-1015, May 2006.

[4] H. Wang *et al*., "A 50 GHz VCO in 0,25gm CMOS," *IEEE ISSCC Dig. of Tech. Papers*, pp. 372-373, Feb. 2001.

[5] C. Cao *et al*., "192 GHz push-push VCO in 0.13 gm CMOS," in *Electron. Lett.*, vol. 42, pp. 208-210, Feb. 2006.

[6] C. Cao *et al*., "A 140-GHz Fundamental Mode Voltage-Controlled Oscillator in 90-nm CMOS Technology," in *Microwave and Wireless Components. Lett.*, vol. 16, pp.555-557, Oct. 2006.

[7] E. Hegavi *et al*., "A Filtering Technique to Lower Oscillator Phase Noise," in *ISSCC Dig. of Tech. Papers*, pp. 364-365, Feb. 2001.

[8] M. Mansuri and C. K. K. Yang, "A low-power adaptive bandwidth PLL and clock buffer with supply-noise compensation," in *IEEE J. Solid-State Circuits*, vol. 38, no. 11, pp. 1804V1812, Nov. 2003.

[9] E. H. Colpitts *et al*., "Carrier current telephony and telegraphy," in *Journal AIEE,*, vol. 40, no.4, pp. 301-305, Apr. 1921.

[10] B. Razavi, *Design of Integrated Circuits for Optical Communications*, New York: McGrawHill, 2002.

[11] W. Winkler *et al*., "60 GHz transceiver circuits in SiGe:C BiCMOS technology," in *Proc. of European Solid-State Circuits Conf*, pp. 83-86, Sep. 2004.

[12] B.A. Floyd *et al*., "SiGe Bipolar transceiver circuits operating at 60 GHz," in *Proc. IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 156-167,Jan. 2005.

[13] S.T. Nicolson *et al*., "Design and scaling of SiGe BiCMOS VCOs above 100GHz," in *Proc. Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 1-4, Oct. 2006.

[14] B. Heydari, "Low-Power mm-Wave Components up to 104GHz in 90nm CMOS," *ISSCC Dig. of Tech. Papers*, pp. 200-201, Feb. 2007.

[15] P. Huang *et al*., "A low-power 114-GHz push-push CMOS VCO using LC source degeneration," in *IEEE J. Solid-State Circuits*, vol. 42, no. 6, pp. 1230-1239, June 2007.

[16] R. Wanner *et al*., "SiGe integrated mm-wave push-push VCOs with reduced power consumption," in *IEEE Radio Frequency Integrated Circuits (RFIC) Symposium Dig. Paper*, pp. 483-486, June. 2006.

[17] H. Wu and A. Hajimiri, *et al*., "Silicon-based distributed voltage-controlled oscillators," in *IEEE J. Solid-State Circuits*, vol. 36, no. 3, pp. 493-502 , Mar. 2001.

[18] N. Tzartzanis, *et al.*, "A reversible poly-phase distributed VCO," in *ISSCC Dig. of Tech. Papers*, pp. 2452- 2461, Feb. 2006.

FREQUENCY DIVIDERS AND MULTIPLIERS 7

Frequency division and multiplication are essential to the whole SerDes system. Dividers can be made in various topologies so as to accommodate different frequencies of operation. Three main divider structures, namely, static, Miller, and injection-locked, account for 99% of the dividers used in high-speed data links. Other than those, dividers with programmable modulus are needed as well in wireline transceivers if variable frequency operation such as spread spectrum are adopted. In special applications, frequency doublers or triplers are required in order to relax the core VCO design. We study frequency dividers and multipliers in this chapter.

7.1 STATIC DIVIDERS

7.1.1 Divided-by-2 Circuits

We begin our discussion with $\div 2$ circuits. One of the simplest $\div 2$ realizations is to place an edge-triggered flipflop (composed of two latches) in a negative feedback loop, as illustrated in Fig. 7.1(a). Differentially driven by the input clock, the two latches provide quadrature outputs running at half the input frequency[1] [Figure 7.1(b)]. Since the stored information can be held in the latches forever, the static frequency dividers can theoretically operate at arbitrarily low frequencies. Such a simple yet robust configuration manifests itself in low to moderate speed.

Almost any type of latch can serve as a vehicle for static $\div 2$ circuit. Figure 7.2 illustrates a few commonly seen latch topologies. We studied CML and TSPC flipflops in chapter 4, and C^2MOS

[1] $CK_{out,I}$ and $CK_{out,Q}$ are separated by exactly 90° if the two latches experience the same loading.

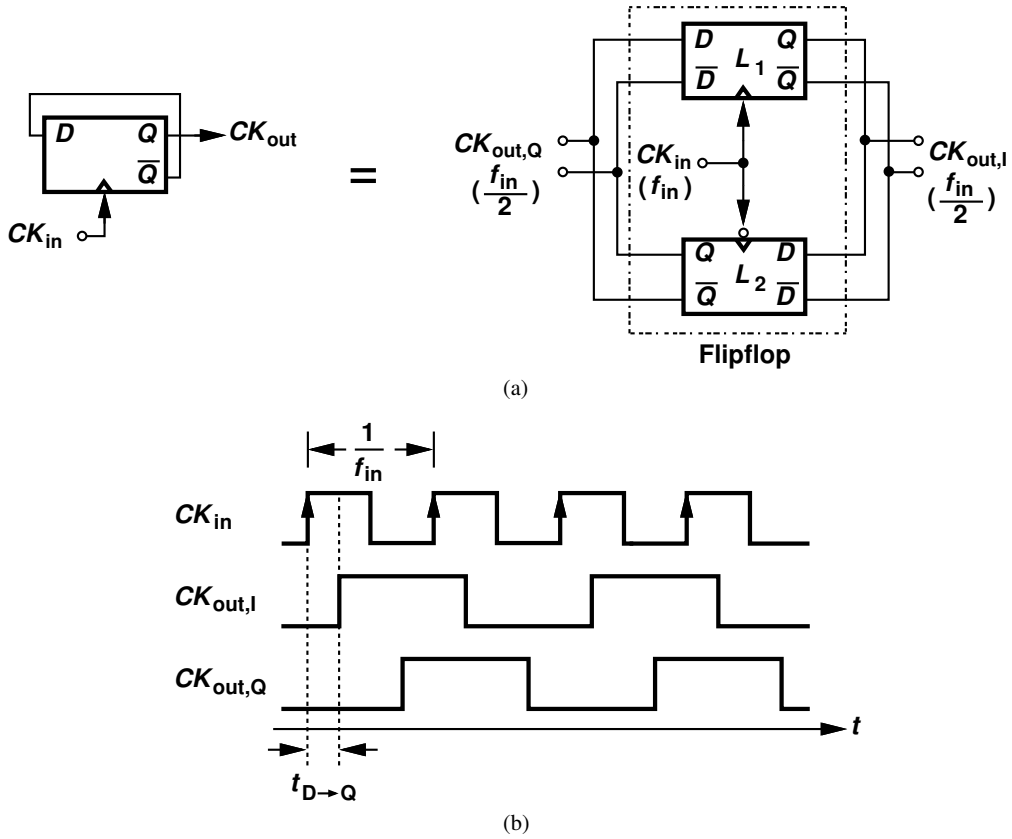


Fig. 7.1 Typical static divider (a) topology, (b) waveforms.

can be easily understood as well. Like other building blocks, tradeoffs exist among bandwidth, power, robustness and signal integrity. For pure digital implementations such as C^2MOS or TSPC latches, the stacking of devices and rail-to-rail operation lead to long rising and falling times. The single-ended structure also suffers from supply noise coupling, potentially introducing jitter in the output. Meanwhile, the abrupt switching of the circuit would pull significant current from V_{DD} momentarily during transitions, which induces voltage bounce and perturbs the quiet analog region nearby. Even if supplies are separated, unwanted coupling can occur through the substrate or package. A better choice is to use current-mode logic (CML), if power consumption is not an issue. Shown in Fig. 7.2(a) and (b) are the corresponding latches for CMOS and bipolar implementations. Controlled by CK_{in} and $\overline{CK_{in}}$, it samples (amplifies) the input while $M_{1,2}$ ($Q_{1,2}$) pair is activated, and holds (regenerates) the data by means of the cross-coupled pair $M_{3,4}$ (Q_3-Q_6). The emitter

followers (Q_5 and Q_6) in Fig. 7.2(b) serve as level shifters for Q_3 and Q_4. The constant tail currents and differential operation alleviate a number of design issues.

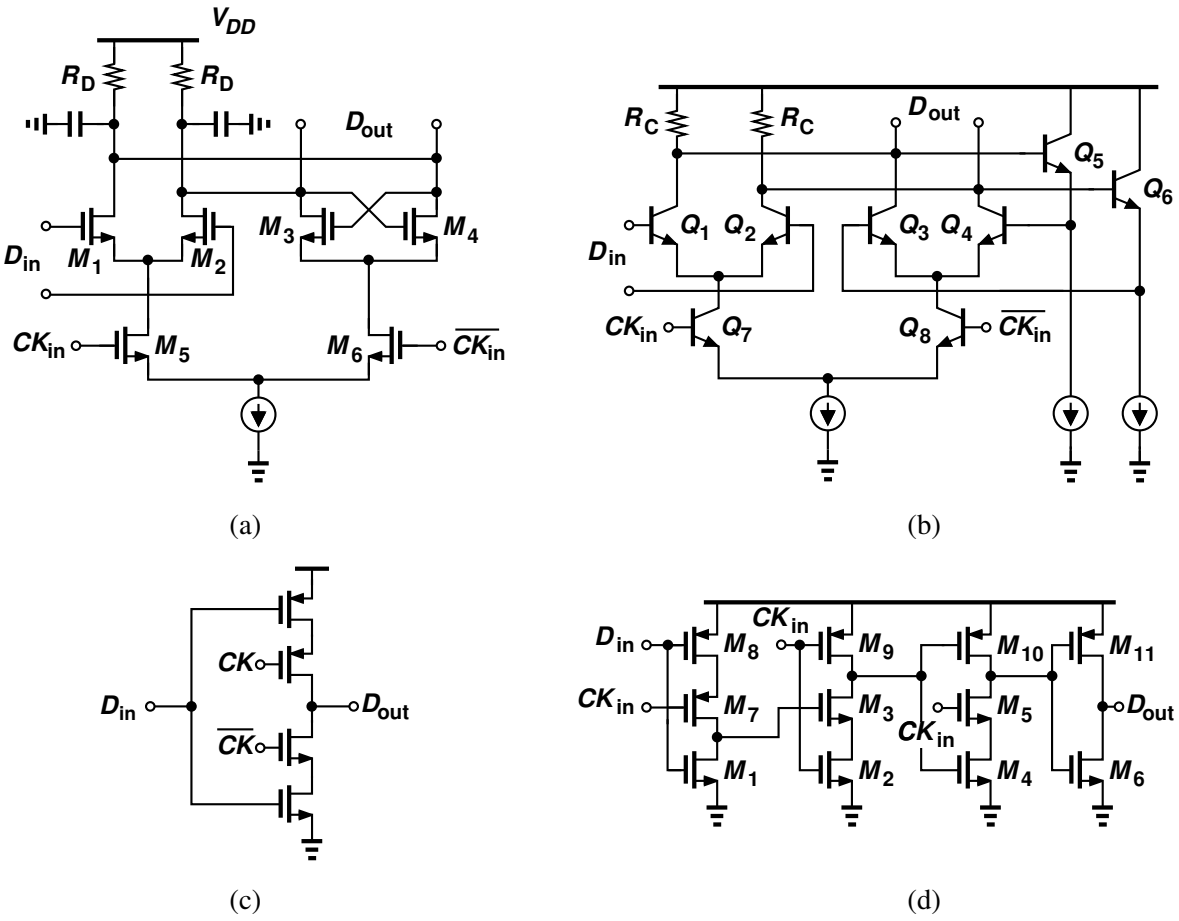


Fig. 7.2 Latches in static (a) CMOS CML, (b) bipolar CML, (c) C^2MOS, (d) TSPC.

Now let us consider the operation of a static divider with CML latches. At low frequencies, the latches lock the sampled data and wait until the next clock phase comes in. Apparently, the loop gain of the positive feedback (e.g., M_3-M_4 pair and two resistors of R_D) must exceed unity, and the output looks like square wave under such a condition. As frequency goes up, the idle time decreases, and the divider would work properly as long as the input pair M_5-M_6 switches the current completely. Afterwards, the divider encounters a self-resonance frequency, where the divider operates as a two-stage ring oscillator. At this moment, the regenerative pairs provide sufficient hysteresis such that each latch contributes 90° of phase shift, and no input power is

required. In other words, the static divider oscillates at such a frequency as it satisfies Barkhausen criteria. Beyond this frequency, the dividers acts as a driven circuit again. It hits a limit as the frequency reaches the bandwidth of the circuit. That is, the D-to-Q delay of the latches ($t_{D\rightarrow Q}$) approaches half input cycle $[1/(2f_{in})]$. As can be clearly explained in Fig. 7.1(b), the timing sequence becomes out of order in such a circumstance, failing the division no matter how large the input power is. Figure 7.3 reveals the simulated input sensitivity (i.e., minimum required power) as a function of input frequency of a typical static divider in 90-nm CMOS technology. A notch can be found around 15 GHz as it oscillates here.

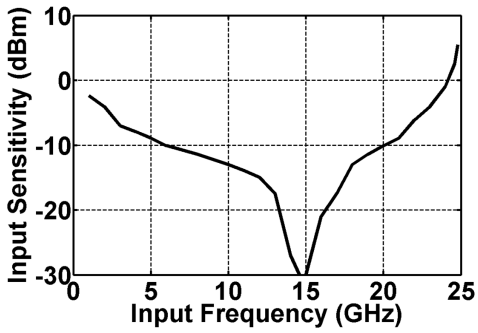


Fig. 7.3 Typical input sensitivity of CML static ÷2 circuit in 90-nm CMOS.

To make a fair comparison, we design divided-by-2 circuits made of different CMOS latch topologies and plot the power efficiency in Fig. 7.4. Here, the dividers are designed to have a fan-out-of-4 loading. It is clearly shown that for a 10 GHz input, TSPC reaches only 1/4 of power as compared with CML. TSPC dividers presents a maximum operation frequency of around 15 GHz. C^2MOS, on the other hand, could barely operation beyond 10 GHz, even though it reaches the lowest power consumption. Nonetheless, CML structure is still the best choice for high-speed operation.

Quite a few techniques have been developed to enhance the performance of CML static dividers. Figure 7.5(a) reveals a modified version with inductive peaking and class-AB biasing. It is obvious that the operation frequency would be pushed up with the help of peaking. As we increase L (and decrease R_D to a roughly constant loading impedance), the divider approaches a

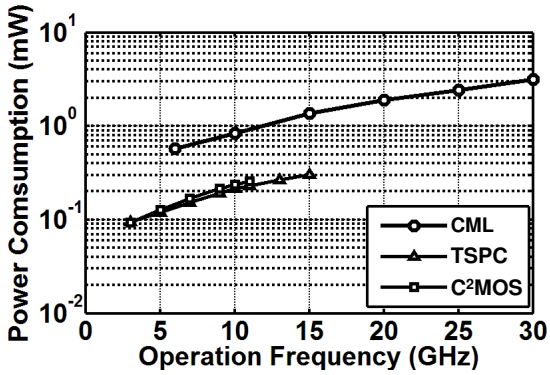


Fig. 7.4 Power and operation range for different CMOS latch implementations. (40-nm COMS).

higher operation frequency. However, the R-L-C combination no longer sustains proper phase relationship at very low frequencies. As a result, the bandwidth enhancement is achieved at a cost of sacrificing low frequency band. Figure 7.5(b) illustrates the phenomenon. Here, we design a 40-GHz static $\div 2$ circuit with inductive peaking in 40-nm CMOS. Given a 0-dBm clock as input, the divider achieves operation range from 14 to 51 GHz. If we optimize the static $\div 2$ circuit for different targeting frequencies and record the corresponding range of operation, we arrive at a plot as shown in Fig. 7.5(c). Roughly speaking, the operation range is inversely proportional to the center frequency. The class-AB biasing helps to create larger peak currents of M_5 and M_6, leading to higher gain for amplification at M_1-M_2 and regeneration at M_3-M_4. The operation bandwidth is improved roughly by 10%. Note that class-AB biasing improves power efficiency simply due to the lower dc currents consumed in the circuit.

Another technique to improve the operation range is to insert resonating inductors in the internal nodes. Figure 7.6 depicts such an idea. Since CK_{in} is differential, it is preferable to put a combined inductor L_P between nodes P and Q rather than using two inductors separately. Assuming both P and Q have associated parasitic capacitance of C_P, we must have

$$2\pi f_{in} = \frac{1}{\sqrt{L_P C_P/2}}, \tag{7.1}$$

to resonate out the parasitics. Thus, more signal currents can be applied into the latch, increasing the operation range.

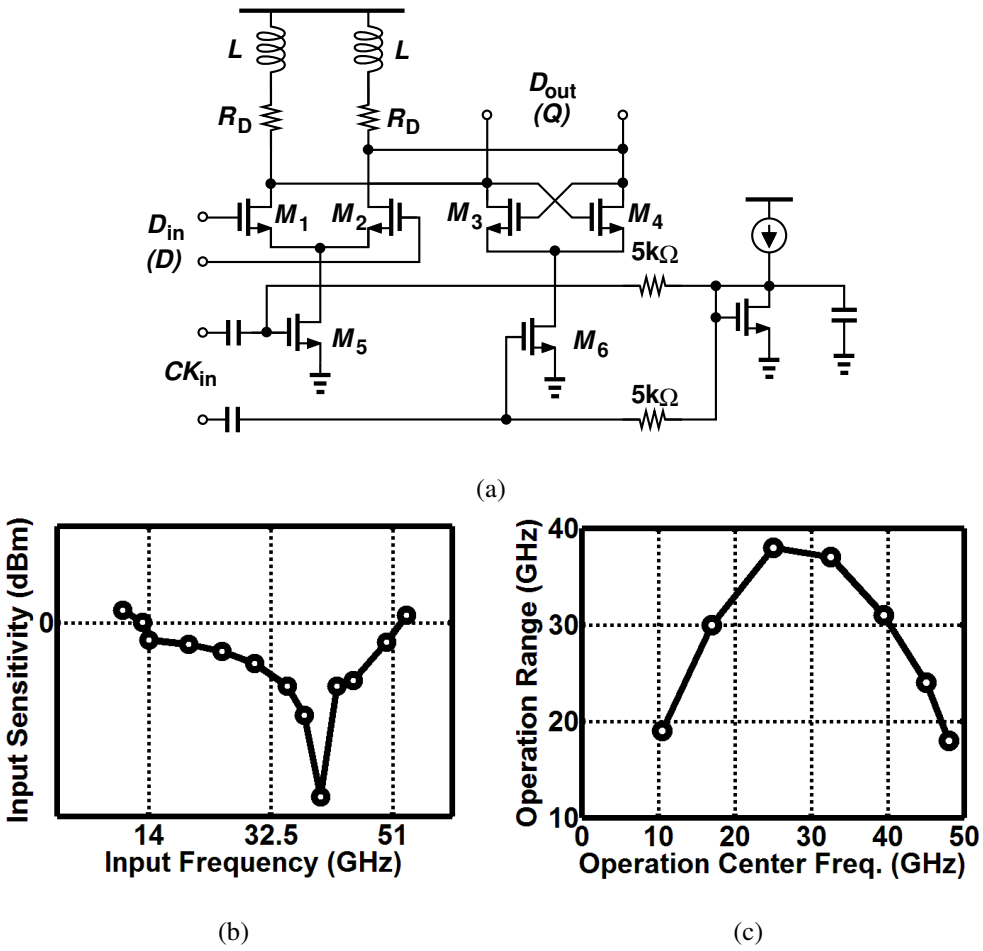


Fig. 7.5 (a) CML latch with inductive peaking and class-AB biasing, (b) tradeoff between operation frequency and range.

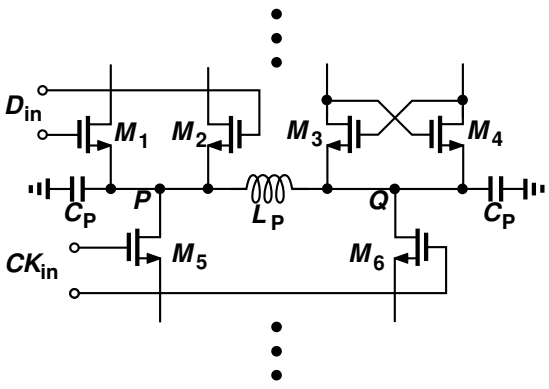


Fig. 7.6 Peaking technique applied to internal nodes.

Example 7.1

Consider an alternative static divider shown in Fig. 7.7, where the latches are implemented as a digital differential pair M_1-M_2 with positive feedback loading M_3-M_4 and a switch (M_5) on the bottom. CK_{IN} is rail-to-rail as well. Determine the device ratio of $(W/L)_{1,2}$ and $(W/L)_{3,4}$ such that the divider could operate properly. Assume $V_{THN} = |V_{THP}| \triangleq V_{TH}$.

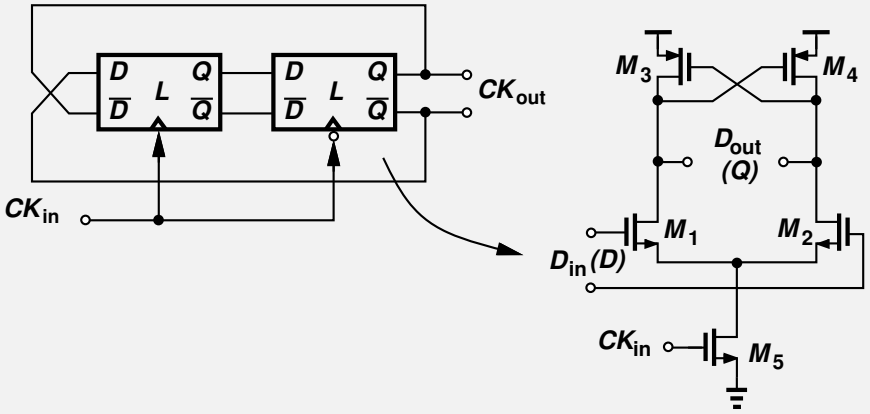


Fig. 7.7 Divided-by-2 circuit made of alternative rail-to-rail latches.

Solution: We determine the minimum requirement for a latch to flip the data. Figure 7.8(a) reveal the case, where $V_P = V_{DD}$ and $V_Q = 0$ in the beginning. As data input comes in, M_1 turns on and M_2 turns off. Since M_3 is also on, we need V_P to be lower than $V_{DD} - V_{TH}$ so as to turn on M_4. Since M_2 is off, V_Q rises up and weakens M_3. The positive feedback continues until $V_P \triangleq 0$ and $V_Q = V_{DD}$. Here, we neglect the effect of bottom switch M_5.

The critical condition is thus $V_P = V_{DD} - V_{TH}$ in the beginning of regeneration

$$\frac{1}{2}\mu_n C_{ox}(\frac{W}{L})_{1,2}[2(V_{DD} - V_{TH})^2] = \frac{1}{2}\mu_p C_{ox}(\frac{W}{L})_{3,4}[2(V_{DD} - V_{TH})V_{TH} - V_{TH}^2]. \tag{7.2}$$

Note that M_3 stays in triode region if $V_{DD} > 2V_{TH}$. Defining k as the size ratio of $M_{1,2}$ and $M_{3,4}$, we obtain the condition for the current data regeneration to occur

$$k \triangleq \frac{(W/L)_{1,2}}{(W/L)_{3,4}} > \frac{\mu_p}{\mu_n} \cdot \frac{(V_{DD} - V_{TH})V_{TH} - V_{TH}^2/2}{(V_{DD} - V_{TH})^2}. \tag{7.3}$$

For example, if $V_{DD} = 3V_{TH}$, $(W/L)_{1,2} > (1/8) \cdot (W/L)_{3,4}$. This is not a tough condition to meet. However, if the M_1-M_2 pair is small, it takes a longer time to complete data sampling, leading to

Example 7.1(*Continued*)

slower operation. Figure 7.8(b) plots the maximum operation speed of the divided-by-2 circuit as a function of k. Here, we use 40-nm process as a testing vehicle with $V_{DD} = 1$ V, $V_{THN} = 468.1$ mV, $V_{THP} = 455.2$ mV, and an inverter-buffered clock as input.

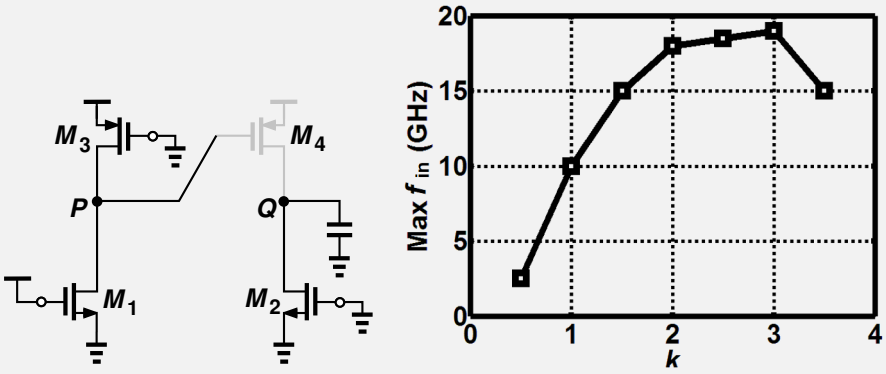
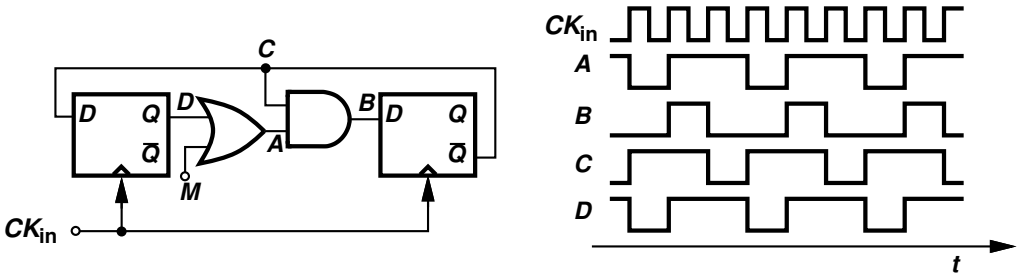


Fig. 7.8 (a) Calculation proper device ratio for Fig. 7. (b) Operation frequency as a function of size ratio $k[= (W/L)_{1,2}/(W/L)_{3,4}]$.

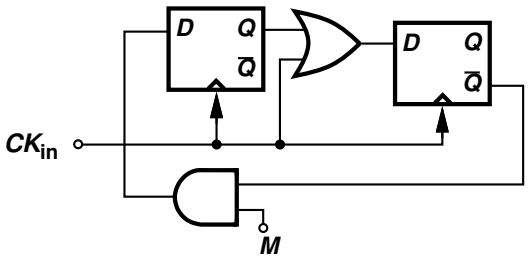
7.1.2 Dividers with Other Moduli

The versatility of static dividers manifests in realization of dividers with other moduli. For example, a divided-by-3 circuit could be achieved by using 2 flipflops with a logic gate [1]. A more useful implementation is to combine the ÷2 and ÷3 circuits. Shown in Fig. 7.9 is a commonly-used structure, where modulus control bit M defines its mode. As $M = 1$, we have $A = 1$ and $B = C$. The circuit degenerates to a sample ÷2 circuit. As $M = 0$, the OR gates becomes transparent, arriving at a ÷3 circuit with the waveforms of all nodes shown on the right. Note that the output duty cycles of this ÷2/3 are either 1/3 or 2/3, not 50%. Programmable dividers with higher moduli (e.g., ÷3/4, ÷4/5, ÷8/9, etc) can be found in the literature. The programmable dividers are widely used in fractional-N frequency synthesizers, which would be discussed in chapter 8.

Perhaps the most powerful divider with programmable modulus is the so-called multi-modulus dividers. Imagine a divider chain which is composed of ÷3 is set, it would be executed only once.



(a)



(b)

Fig. 7.9 (a) Classical ÷2/3 circuit and its waveforms of ÷3 mode, (b) alternative approach.

The divide modulus returns back to 2 afterwards. Such an arrange allows us to program the desired modulus over a wide range. Figure 7.10 illustrates a typical realization, where N stages of ÷2/3 circuits are placed in cascade with modulus control bits $P_{N-1}, \cdots, P_1, P_1$. In addition, each stage has a mod_{out} bit feeding back to its preceding one as mod_{in}, performing the extinguishment of special ÷3 mode. As depicted in the timing diagram, 2^k cycles are inserted in one complete period if stage k is set to special ÷3 mode. Thus,

$$\text{Number of cycles in One Output Period} = 2^N + P_0 \cdot 2^0 + P_1 \cdot 2^1 + \cdots + P_{N-1} \cdot 2^{N-1}. \quad (7.4)$$

For example, if $N = 3$, we can create modulus from 8 to 15. The realization of ÷2/3 cell is also depicted in Fig. 7.10. The reader can easily show how it works.

The imbalanced duty cycle in ÷3 circuits can be corrected by introducing the third flipflop in it (Figure 7.11). By creating a delayed version of the original output, a 50% duty cycle output can be realized with the help of an OR gate.

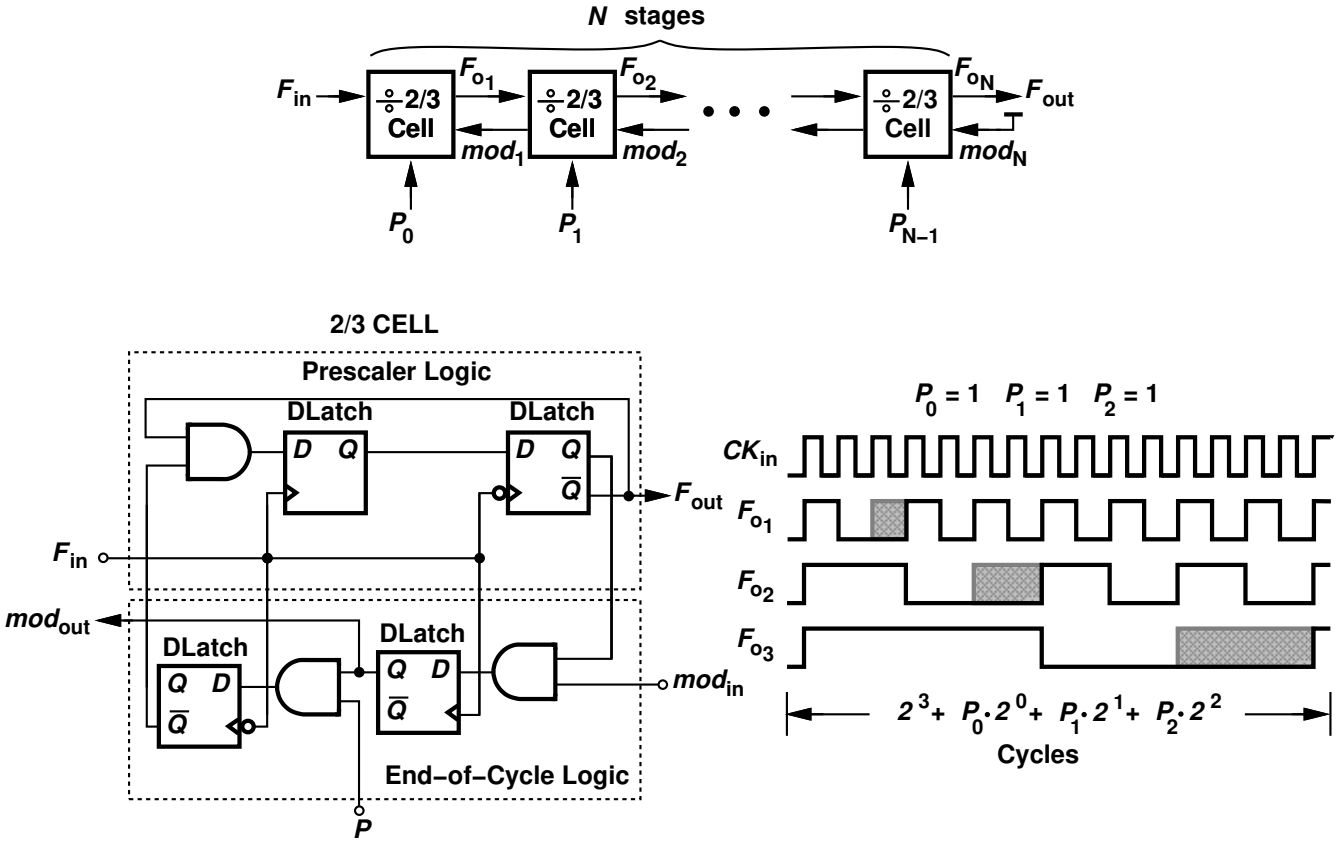


Fig. 7.10 Multi-modulus dividers.

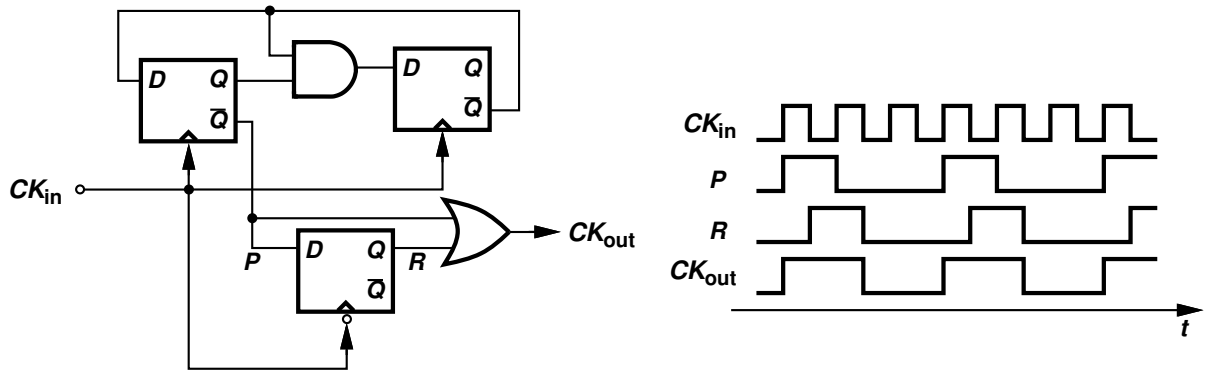


Fig. 7.11 ÷3 Circuit with 50% duty cycle.

7.2 MILLER DIVIDERS

Static dividers can safely work up to 20-30 GHz in today's CMOS technologies. We resort to other divider topologies. We resort to other divider topologies if frequency goes higher. The Miller divider (also known as regenerative divider or dynamic divider) provides purely analog operation with much higher bandwidth. Originally proposed by Miller in 1939 and becoming popular for bipolar devices in 1980's, the Miller divider is based on mixing the output with the input and applying the result to a low-pass filter (LPF), as shown in Fig. 7.12(a). Under proper phase and gain conditions, the component at ω_{in} survives and circulates around the loop. Since the device capacitances are absorbed in the low-pass filter, this topologies achieves a high speed and is widely adopted in the design of bipolar and GaAs dividers.

While providing an intuitive understanding of the circuit's operation, Fig. 7.12(a) fails to stipulate the conditions for proper division. For example, the low-pass filter may be realized as a first-order RC network [Fig. 7.12(b)], a reasonable model of the load seen at the output node of typical mixers. Neglecting nonlinearities in the mixer, we have

$$R_1C_1\frac{\mathrm{d}y}{\mathrm{d}t} + y = \beta y A\cos\omega_{in}t, \tag{7.5}$$

where β denotes the mixer conversion factor. Thus

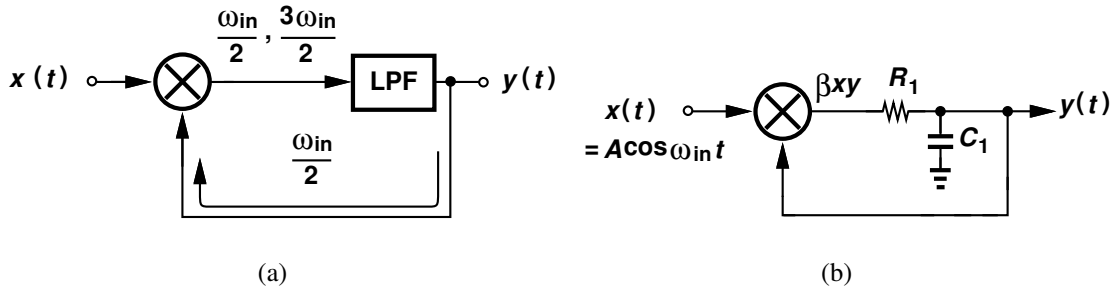
$$R_1C_1\frac{\mathrm{d}y}{\mathrm{d}t} = y(\beta A\cos\omega_{in}t - 1), \tag{7.6}$$

and hence

$$y(t) = y(0)exp(-\frac{t}{R_1C_1} + \frac{\beta A}{R_1C_1\omega_{in}}\sin\omega_{in}t). \tag{7.7}$$

Interestingly, $y(t)$ decays to zero with a time constant of R_1C_1, i.e., the circuit fails to divide regardless of the value of ω_{in} with respect to the LPF corner frequency, $(R_1C_1)^{-1}$. In other words, $\omega_{in}/2$ is not regenerated even though R_1C_1 is chosen to attenuate the third harmonic, $3\omega_{in}/2$ (and even if a noise current at $\omega_{in}/2$ is injected into the loop).

Let us now consider an extreme case where all time constants in the loop are negligible, all waveforms are rectangular, and the circuit operates correctly. As illustrated in Fig. 7.13(a), the

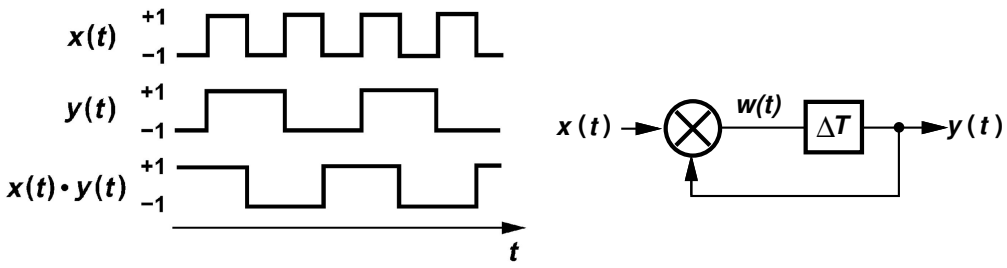


(a) (b)

Fig. 7.12 Dynamic (Miller) divider. (a) Generic topology. (b) Realized with an RC filter.

mixer output resembles $y(t)$ but shifted by a quarter period, suggesting that inserting a broadband delay $\Delta T = \pi/\omega_{in}$ in the loop permits correct division [Fig. 7.13(b)].

It is important to note that the RC network of Fig. 7.12(b) does not satisfy the condition required in Fig. 7.13(b). For example, the network cannot provide a phase shift of 90° at $\omega_{in}/2$ and 270° at $3\omega_{in}/2$. Furthermore, it attenuates the third harmonic considerably, failing to generate the idealized waveforms shown in Fig. 7.13(a).



(a) (b)

Fig. 7.13 90° phase shift operation. (a) Waveforms. (b) Model.

A typical bipolar implementation is shown on Fig. 7.14, which includes both low-pass filtering and delay. The loading resistor R and the parasitic capacitance associated with nodes X and Y from the low-pass filtering, and the emitter followers create the proper delay. We use simulations to plot the requisite delay as a function of RC [Fig. 7.14(c)], arriving at the solution space (on or above the line) for the choice of these two parameters.

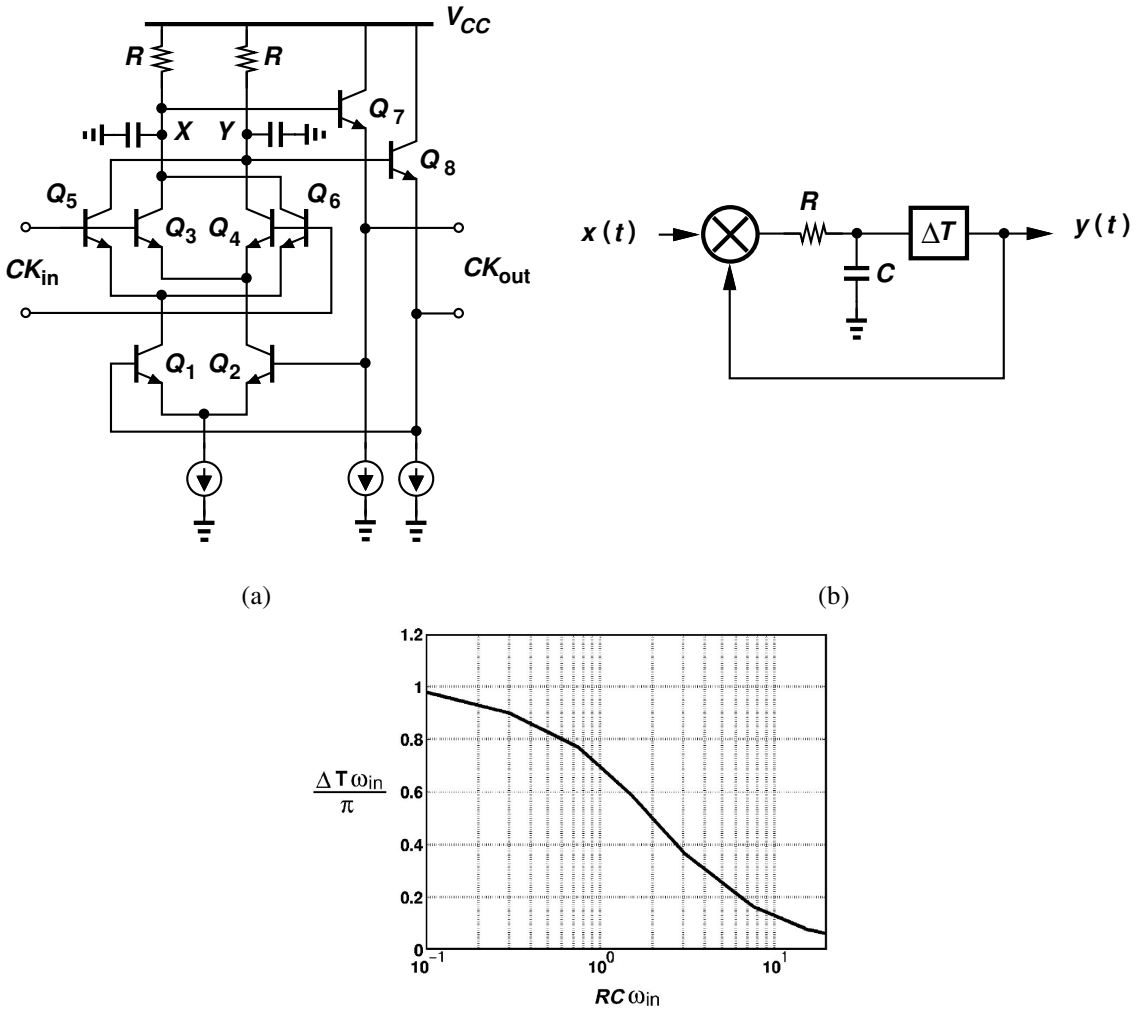


Fig. 7.14 (a) Bipolar Miller divider. (b) Simplified model. (c) Requisite delay as a function of RC.

Realizing that the LPF is to filter out the component at $3\omega_{in}/2$ and preserve that at $\omega_{in}/2$, we examine two cases to determine the operation range. As illustrated in Fig. 7.15, the rule of thumb is to keep $\omega_{in}/2$ inside the passband while rejecting $3\omega_{in}/2$ and other harmonics. In other words, we can roughly estimate the operation range as

$$\frac{\omega_{in,max}}{2} \leq \omega_c \;\; and \;\; \frac{3\omega_{in,min}}{2} \geq \omega_c, \tag{7.8}$$

and hence

$$\frac{2\omega_c}{3} \leq \omega_{in} \leq 2\omega_c. \tag{7.9}$$

Fig. 7.15 illustrates the sensitivity of a typical regenerative divider. Note that it has no notch as there is no condition to from a self-resonating loop.

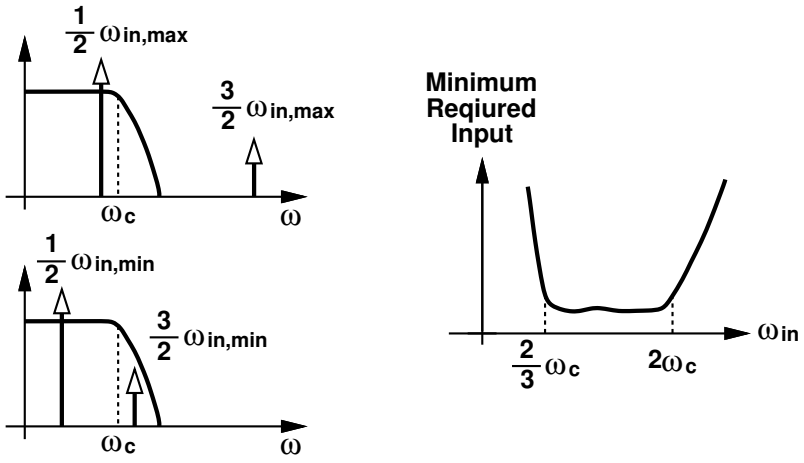


Fig. 7.15 Operation range determination.

However, the configuration of Fig. 7.14(a) is difficult to realize in CMOS technologies because the relatively low transconductance of CMOS devices arrives at a source follower with poor performance. It may consume substantial voltage headroom while attenuating the signal and discouraging the divider from high-speed operation. We introduce CMOS version Miller divider in the following section.

7.3 MODIFIED MILLER DIVIDERS

Now that the LPF-based structure is not suitable for CMOS implementation, we study another extreme case where the loop exhibits no delay at but enough selectivity to attenuate the third harmonic. Fig. 7.16(a) exemplifies this case, with the mixer injecting a current into the parallel tank and $\sqrt{LC} = 2/\omega_{in}$. We assume that the peaks of $x_1(t)$ and $x_2(t)$ are aligned and examine $x_1(t)x_2(t)$ and $y(t)$. As depicted in Fig. 7.16(b), the product waveform displays multiple zero crossings in each period due to the third harmonic, revealing that such a loop fails to divide

if this harmonic is not suppressed sufficiently, i.e., if $y(t)$ does not monotonically rise and fall. Fig. 7.16(c) illustrates the resulting waveforms for different values of the attenuation factor, α, experienced by the third harmonic with respect to the fundamental.To eliminate the extraneous zero crossings, we require that the slope of $y(t)$ not change sign between a positive peak and the next negative peak. Since

$$y(t) \propto \cos\frac{\omega_{in}t}{2} + \alpha\cos\frac{3\omega_{in}t}{2}. \tag{7.10}$$

We have

$$\frac{\mathrm{d}y}{\mathrm{d}t} \propto -\sin\frac{\omega_{in}t}{2} - 3\alpha\sin\frac{3\omega_{in}t}{2} < 0 \quad for\ 0 < t < \frac{2\pi}{\omega_{in}}. \tag{7.11}$$

Illustrated in Fig. 7.17(a), the terms $\sin(\omega_{in}t/2)$ and $3\alpha\sin(3\omega_{in}t/2)$ yield a positive sum if $0 < 3\alpha < 1$. Thus, the attenuation factor must satisfy

$$0 < \alpha < \frac{1}{3}. \tag{7.12}$$

The foregoing derivation assumes the third harmonic experiences no phase shift, contradicting the actual behavior of the RLC tank. Since the tank impresses a phase shift of approximately 90° upon this harmonic, Eq (7.10) must be rewritten as

$$y(t) \approx \cos\frac{\omega_{in}t}{2} + \alpha\sin\frac{3\omega_{in}t}{2}, \tag{7.13}$$

and $\mathrm{d}y/\mathrm{d}t$ must remain negative in a proper interval. Plotting the two components of $\mathrm{d}y/\mathrm{d}t$ in Fig. 7.17(b), we note that a positive sum results between t_1 and t_3 if $\sin(\omega_{in}t_2/2) - 3\alpha\cos(3\omega_{in}t_2/2) > 0$. Since the phase $\omega_{in}t/2$ reaches 60° at t_2, we have

$$0 < \alpha < \frac{1}{2\sqrt{3}} \tag{7.14}$$

which is a slightly more stringent condition than that in Eq (7.12). We now determine the selectivity required of the tank to guarantee Eq (7.14):

$$\frac{L^2\omega^2}{R^2(1-LC\omega^2)^2 + L^2\omega^2} = (\frac{1}{2\sqrt{3}})^2, \tag{7.15}$$

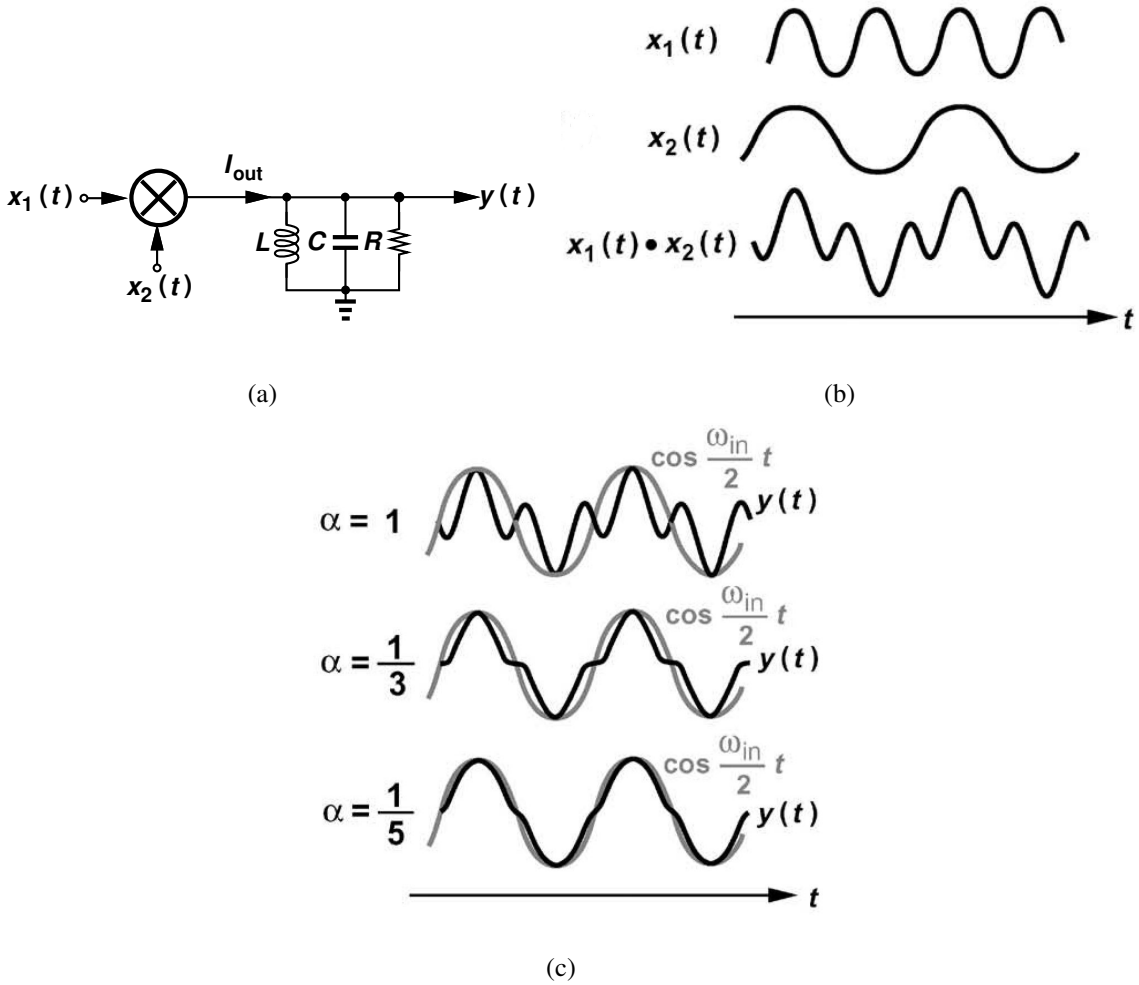
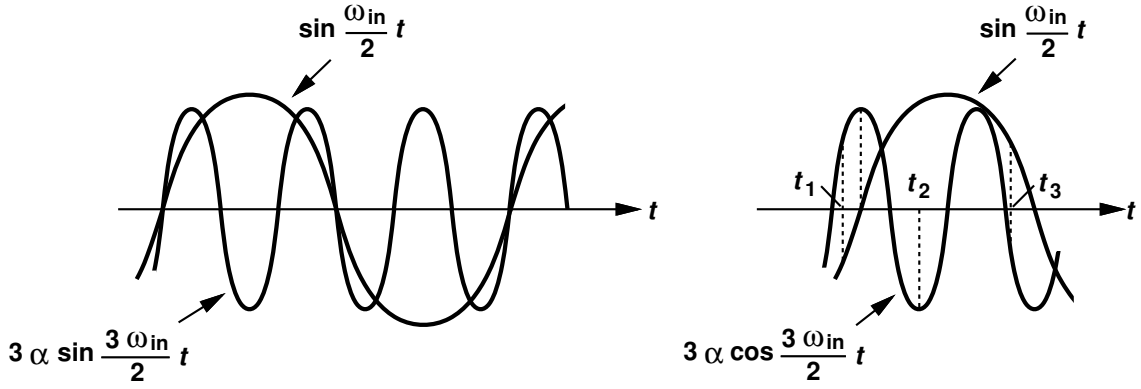


Fig. 7.16 (a) Mixer with selective network. (b) Input waveforms and (c) output waveforms for different values of α.

where $LC = (\omega_{in}/2)^{-2}$ and $\omega = 3\omega_{in}/2$. It follows that

$$\frac{R}{L\frac{\omega_{in}}{2}} = \frac{3\sqrt{11}}{8} \approx 1.24. \tag{7.16}$$

In other words, a tank Q of 1.24 at $\omega_{in}/2$ ensures enough attenuation of the third harmonic. Of course, it is assumed that the loop gain at $\omega_{in}/2$ is sufficient to sustain this component. We check this issue in the following discussion. With the fundamental knowledge established, we can employ an LC tank as the load in the Miller divider, as shown in Fig. 7.18. For this circuit to divide



(a) (b)

Fig. 7.17 Components of the slopes of output waveforms. (a) Simplified case. (b) Actual case.

properly, the loop gain at $\omega_{in}/2$ must be at least unity. Modeling the mixer as an ideal multiplier and assuming the following transfer function for the RLC tank:

$$H(s) = \frac{2\zeta\omega_n s}{s^2 + 2\zeta\omega_n s + \omega_n^2}, \tag{7.17}$$

where $2\zeta\omega_n = (RC)^{-1}$ and $\omega_n^2 = (LC)^{-1}$, we require that

$$\frac{\beta A}{2}\left|H(j\frac{\omega_{in}}{2})\right| \geq 1. \tag{7.18}$$

(The factor 1/2 arises from the product-to-sum conversion of sinusoids after multiplication.) That is,

$$\frac{\beta A}{2}\left|\frac{2\zeta\omega_n\frac{\omega_{in}}{2}}{\sqrt{(\omega_n^2 - \frac{\omega_{in}^2}{4})^2 + 4\zeta^2\omega_n^2\frac{\omega_{in}^2}{4}}}\right| \geq 1. \tag{7.19}$$

Thus, the minimum input amplitude necessary for correct division is given by

$$A \geq \frac{2}{\beta}\sqrt{1 + \frac{(1 - \frac{\omega_{in}^2}{4\omega_n^2})^2}{\zeta^2\frac{\omega_{in}^2}{\omega_n^2}}}. \tag{7.20}$$

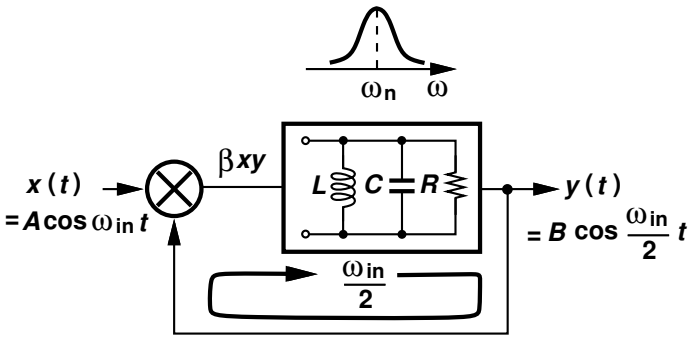


Fig. 7.18 Miller divider with bandpass filter.

As expected, the right-hand side falls to a minimum of $2/\beta$ for $\omega_{in} = 2\omega_n = 2/\sqrt{LC}$. For $\Delta\omega = |\omega_{in} - 2\omega_n| \ll 2\omega_n$, we have

$$\begin{aligned} 1 - \frac{\omega_{in}^2}{4\omega_n^2} &= \frac{(2\omega_n + \omega_{in})(2\omega_n - \omega_{in})}{4\omega_n^2} \\ &\approx \frac{4\omega_n(2\omega_n - \omega_{in})}{4\omega_n^2} \\ &\approx \frac{\Delta\omega}{\omega_n}. \end{aligned} \tag{7.21}$$

Consequently, since$\zeta = (2Q)^{-1}$, the fraction under the square root in Eq (7.20) can be reduced to $(Q\Delta\omega/\omega_n)^2$,yielding

$$A \geq \frac{2}{\beta}\sqrt{1 + (\frac{Q\Delta\omega}{\omega_n})^2}. \tag{7.22}$$

Fig. 7.19 plots the input sensitivity as a function of ω_{in}. For example, if we restrict the maximum input amplitude to $4/\beta$, then

$$\Delta\omega = \frac{\sqrt{3}}{Q}\omega_n. \tag{7.23}$$

As the input amplitude increases, the switching quad of the mixer eventually experiences complete switching, yielding a conversion factor of $2/\pi$ in the ideal case. The loop gain is then equal to $(2/\pi)g_m$ times the magnitude of the tank impedance, where g_m denotes the transconductance of the bottom differential pair of the mixer. Consequently, Eq (7.19) is modified to

$$\frac{2}{\pi}g_m \left| \frac{2\zeta\omega_n sR}{s^2 + 2\zeta\omega_n s + \omega_n^2} \right| \geq 1, \tag{7.24}$$

and (7.24) to

$$\frac{2}{\pi}g_m R \geq \sqrt{1+(\frac{Q\Delta\omega}{\omega_n})^2}. \tag{7.25}$$

That is,

$$\begin{aligned}\Delta\omega &= \frac{\omega_n}{Q}[(\frac{2}{\pi}g_m R)^2 - 1] \\ &\approx \frac{\omega_n}{Q}(\frac{2}{\pi}g_m R)^2. \end{aligned} \tag{7.26}$$

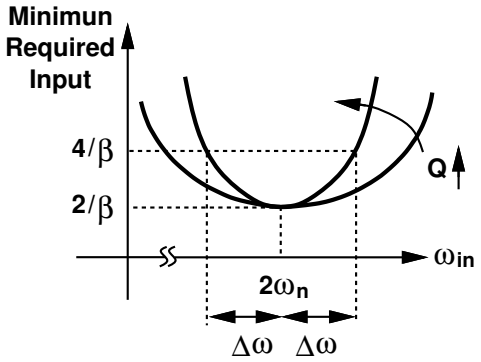


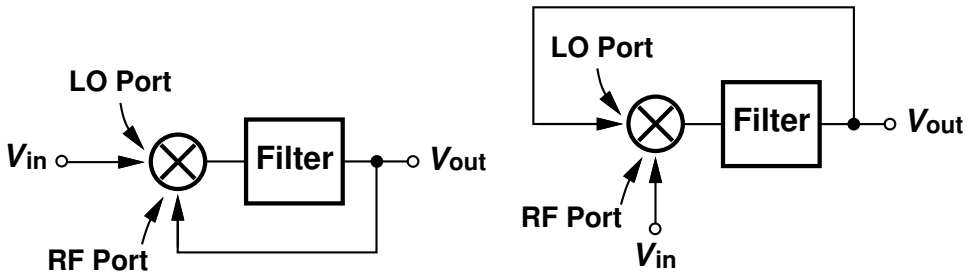
Fig. 7.19 Minimum input amplitude for correct division versus input frequency.

Now we are ready to build up Miller dividers based on BPF in CMOS.

It is interesting to note that a mixer has two input ports, that leads to two possible configurations of Miller dividers. As illustrated in Fig. 7.20, the output could either return to the RF port (type I) or the LO port (type II) of the mixer. Although conceptually indistinguishable, these two approaches still make difference in circuit implementation.

Figure 7.21(a) shows the type I Miller divider. Here, loading inductors L_1 and L_2 resonate with the parasitic capacitances at node X and Y and the input capacitance of M_1 and M_2, providing a few hundred Ω equivalent resistance at $\omega_{in}/2$ with negligible voltage headroom consumption.

The device dimensions and component values in this circuit must be chosen so as to provide both sufficient loop gain—to guarantee correct division—and large enough output swings necessary for the subsequent stage. Assuming abrupt, complete switching of M_3 - M_6, neglecting the effect of L_3 and parasitic capacitances, and simplifying the circuit to that shown in Fig. 7.21(b), we express



(a) (b)

Fig. 7.20 Regenerative divider with the output fed back to (a) RF port, (b) LO port.

the voltage conversion gain of the mixer (= loop gain) as $(2/\pi)g_{m1,2}R_p$, where $R_p = QL_{1,2}\omega$ denotes the equivalent parallel resistance of each tank. Since $g_m \approx 2\pi f_T C_{GS}$ and since the loop gain must exceed unity

$$\frac{2}{\pi}2\pi f_T C_{GS} Q L_{1,2}\frac{\omega_{in}}{2} \geq 1. \tag{7.27}$$

With all of the parasitics neglected, $\omega_{in}/2 \approx 1/\sqrt{C_{GS}L_{1,2}}$ and hence

$$Q \geq \frac{\pi}{4}\frac{f_{in}}{f_T}, \tag{7.28}$$

where f_{in} is the input frequency.[2] This result implies that, even for input frequencies as high as f_T, a Q of about unity suffices. However, the following effects necessitate a much higher Q.

1) The total capacitance at nodes A and B; even if the source/drain junction capacitances are neglected, M_3-M_6 create a pole around f_T at these nodes, wasting about half of the small-signal drain currents of M_1 and M_2.

2) The gradual switching of M_3-M_6 with a nearly sinusoidal drive converts part of the differential currents produced by M_1 and M_2 to a common-mode component.

3) The parasitic capacitances of the load inductors and the coupling capacitors lead to $\omega_n < 1/\sqrt{C_{GS}L_{1,2}}$. Simulations reveal that the Q must exceed 4.5 for correct division.

[2]Equation (28) holds for the center of the input frequency range, i.e., if the tank can be reduced to a single resistor R_P.

In summary, the required Q of the tank is determined by the following requirements: attenuation of the third harmonic, sufficient loop gain in the ideal case, and sufficient loop gain in the presence of parasiticsXwith the last dominating in this design.

Since all of the six transistors in this circuit are relatively wide, the total capacitance at the drains of M_1 and M_2 shunts a considerable portion of their small-signal drain current to ground. Inductor L_3 is, therefore, added to resonate with this capacitance. Since the feedback signal is applied to the RF port, the circuit produces a zero output when the LO input is zero. In contrast to the injection-locked oscillator, this topology is not prone to oscillation.

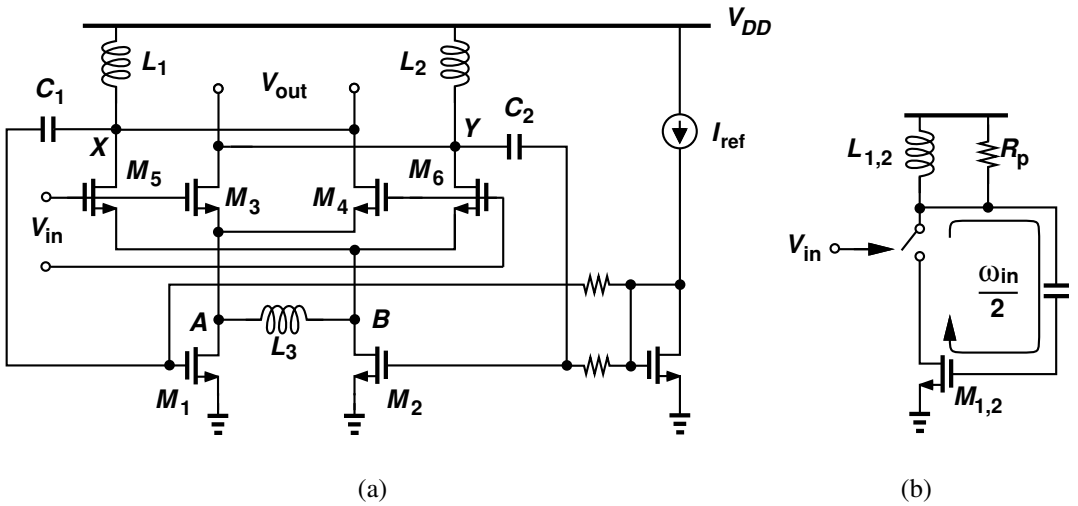


Fig. 7.21 (a) Type I Miller divider. (b) Simplification of (a).

What happens if the output is fed back to the LO port? Figure 7.22 depicts such a realization. In this case, the output is returned to the switching quad rather than to the bottom pair so as to present less capacitance to the first divider. This circuit in fact operates as an injection-locked oscillator if $(W/L)_{3,4} \neq (W/L)_{5,6}$: M_3 and M_4 form a cross-coupled pair, and M_5 and M_6 appear as diode-connected transistors, lowering the Q of the tank and, hence, increasing the lock range.[3] Inductor L_3 resonates with the capacitances at nodes A and B, widening the lock range to some extent [2]. In contrast to injection-locked dividers with a single-ended input [3], [2], this topology injects the differential phases of the 20-GHz signal into the tail nodes and the output nodes. Simulations

[3]In this design $(W/L)_{3,4} = (W/L)_{5,6}$ so that the circuit has no tendency to oscillate.

indicate that differential injection in this manner increases the lock range by 20%.It is possible to find a self-resonance frequency of the circuit if $(W/L)_{3,4} > (W/L)_{5,6}$ [4].

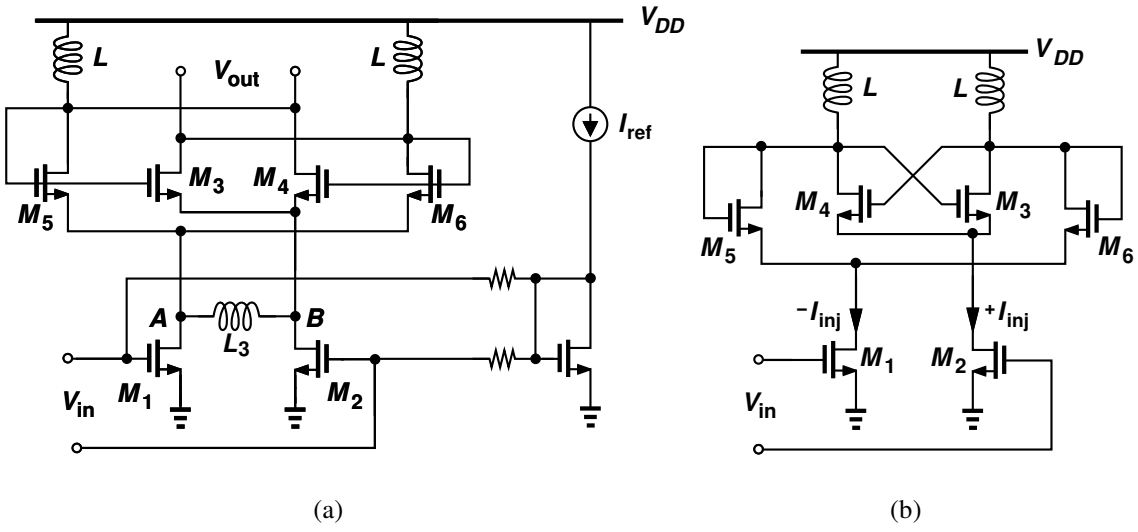


Fig. 7.22 (a) Type II Miller divider. (b) Redrawn to show injection locking.

Example 7.2

CMOS Miller dividers could be modified to implement moudli other than 2. Design divided-by-$(N+1)$ Miller dividers by inserting a $\times N$ and $\div N$ block in the loop. Compare their performance.

Solution: The simplest realizations can be found in Figure 7.23. Both circuits can achieve $\div(N+1)$ function. Obviously, Fig. 7.23(a) is superior in terms of BPF's selectivity, but frequency multiplier is somewhat more challenging to design. We look at multipliers in section 7.5.

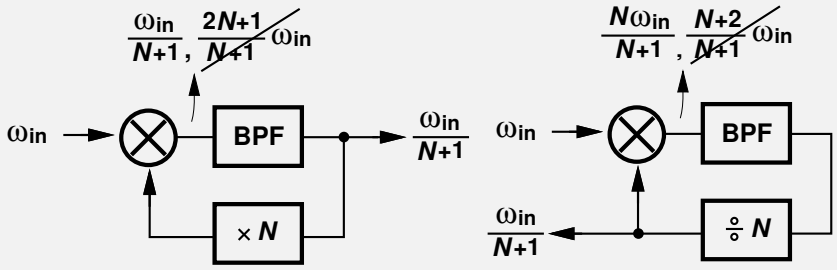


Fig. 7.23 Implementing $\div(N+1)$ Miller divider with (a) multiplier (b) divider in the loop.

7.4 INJECTION-LOCKED DIVIDERS

To achieve even higher frequencies, designers usually resort to injection-locking techniques. It can be early observed that, if an LC-tank oscillator experiences a 2nd-order harmonic input at any of its "common-mode" nodes (i.e., central line of a symmetric circuit), the fundamental output would be "locked" to exactly half of the input frequency. Recognized as an injection-locked divider, this approach is indeed an inverse operation of push-push oscillators. Among the existing divider topologies, it basically reaches the highest speed.

Many theories have been proposed over the past decades to analyze the injection locking phenomenon. From circuit's points of view, it could be best explained by the model shown in Fig. 7.24(a). If a resonant network (e.g., LC-tank) with nature frequency ω_0 undergoes an external injection I_{inj} (whose frequency ω_{inj} is slightly away from ω_0), the network would no longer oscillate at ω_0 but rather ω_{inj}. However, in order to accommodate the excess phase shift (i.e., $-\phi_0$), the overall current flowing into the network I_T must bear an opposite phase shift ϕ_0. After all, output voltage V_{out} and device current I_{osc} are in phase. That forms am angle ϕ_0 between I_T and I_{osc}. Note that I_T is composed of two phasors I_{osc} and I_{inj} Fig. 7.24(b), and all components are at frequency of ω_{inj}. By law of cosines, we have

$$I_{inj}^2 = I_{osc}^2 + I_T^2 - 2 \cdot I_{osc} \cdot I_T \cdot \cos\phi_0. \tag{7.29}$$

For given I_{osc} and I_{inj}, ϕ_0 reaches a maximum as I_T is perpendicular to I_{inj}, which also stands for the maximum tolerable range or lock range [Fig. 7.24(c)].

Example 7.3

Prove $\phi_{0,max}$ occurs as $I_T \perp I_{inj}$.

Solution: Consider $\cos\phi_0$ as a function of I_T and equal its derivative to 0, we arrive at

$$\frac{\mathrm{d}\cos\phi_0}{\mathrm{d}t} = 0 \;\Rightarrow\; I_T^2 = I_{osc}^2 - I_{inj}^2. \tag{7.30}$$

If the resonant network is made of R, L, and C in parallel, we obtain the phase shift in the vicinity of resonance.

$$\tan\phi_0 = \frac{I_{inj}}{I_T} \approx \frac{2Q}{\omega_0}(\omega_0 - \omega_{inj}). \tag{7.31}$$

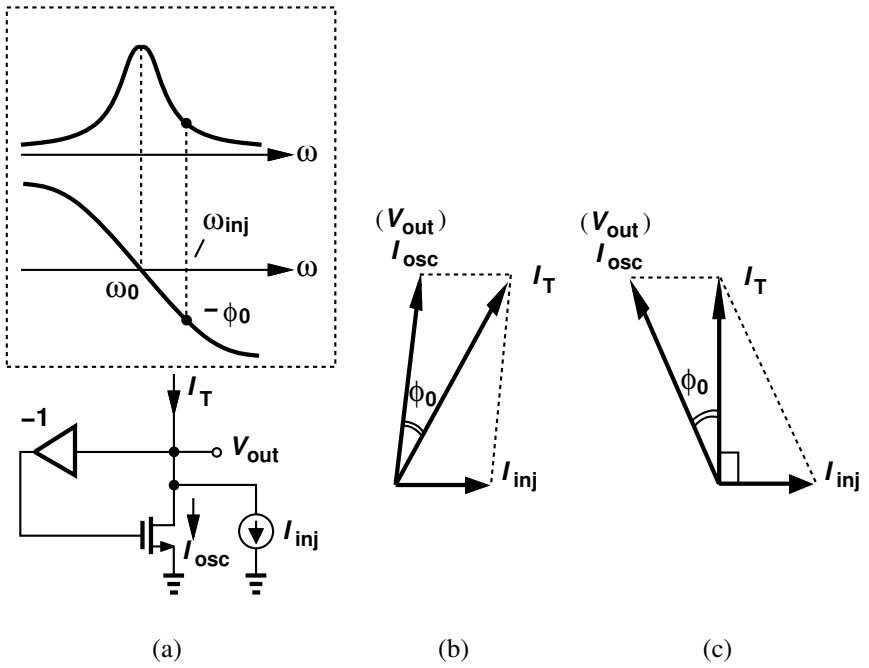


Fig. 7.24 Analysis of LC-tank oscillator under injection locking: (a) modeling, (b) typical phase relationship, (c) maximum tolerable ϕ_0 (lock range).

That is, the maximum lock range ω_L is given by

$$\omega_L = \omega_0 - \omega_{inj} \approx \frac{2Q}{\omega_0} \cdot \frac{I_{inj}}{I_{osc}} \cdot \frac{1}{\sqrt{1 - \dfrac{I_{inj}^2}{I_{osc}^2}}}. \tag{7.32}$$

The overall lock range actually counts on both sides, i.e., $\pm\omega_L$.

The analysis of injection locking derived above can be further extended to injection-locked dividers. As illustrated in Fig. 7.25, an injection-locked divider based on tank oscillator topology can be achieved by applying the input to the common source node P. Here, I_{inj} still denotes the injection current and I_B the bias tail current of the tank. At large signals, the cross-coupled pair serves as a mixer with conversion gain of $2/\pi$. The circuit is nothing more than an injection current I_{inj} ($\approx 2\omega_0$) gets down-converted by the output itself ($\approx \omega_0$). Using half-circuit equivalent circuit, we can approximate the division as a $(I_{inj}/2) \cdot (2/\pi)$ injection current applying to the oscillator with current $I_B/2$. Assuming the injection current is much less than the original current, we modify

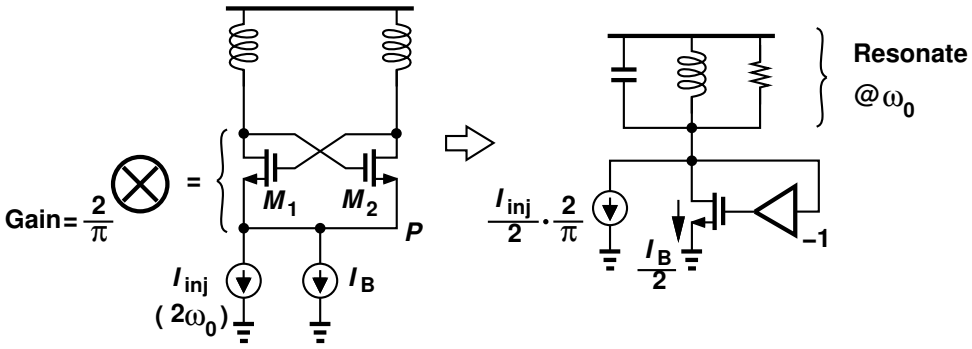


Fig. 7.25 Injection-locked divider and its model.

the lock range (of output, $\approx \omega_0$) as

$$\omega_{L,output} \approx \frac{\omega_0 \cdot I_{inj}}{Q \cdot \pi \cdot I_B}. \tag{7.33}$$

Referring it to the input frequency, we arrive at

$$\omega_L = \frac{2\omega_0 I_{inj}}{Q \cdot \pi \cdot I_B}. \tag{7.34}$$

It can be normalized as percentage

$$\text{Lock Range} = \frac{I_{inj}}{Q \cdot \pi \cdot I_B}. \tag{7.35}$$

As excepted, the lock range of an injection-locked divider is typically quite limited. For example, if tank $Q = 10$, $I_{inj}/I_B = 1/4$, the lock range is roughly equal to $\pm 0.8\%$. In reality, the lock range of injection-locked dividers could be even smaller, as the linear model is overoptimistic. As can be observed in Chapter 6, large-signed operation would turn off the transistor for significant amount of time, making the circuit less confined by the injected signal. Nonetheless, a plot of simulated and predicted lock range for a 40-GHz divider has been shown in Fig. 7.26.

A few modifications can be made to improve the performance of the divider. One issue of the circuit in Fig. 7.26(b) stems from the parasitic capacitance associated with node P. At high speed, it creates a path to ground, robbing significant portion of I_{inj} and undermining the injection. To modify it, an inductor L can be added to resonate out the capacitance C_P [Fig. 7.27], enlarging lock range without extra power consumption [5]. Other than the parasitic, the circuit in Fig. 7.26(b) is

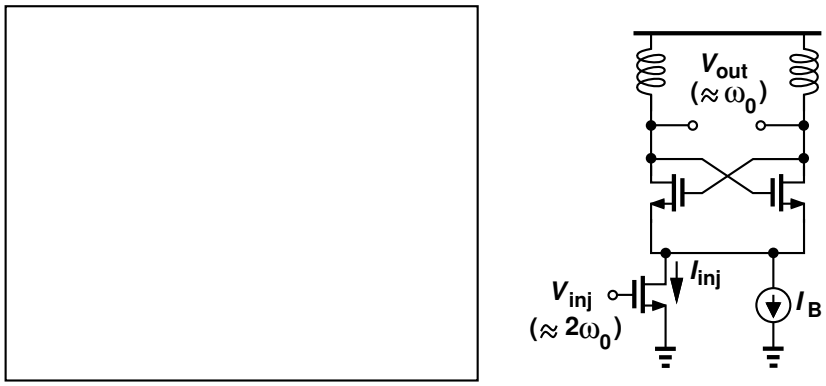


Fig. 7.26 Simulated and theoretical lock range.

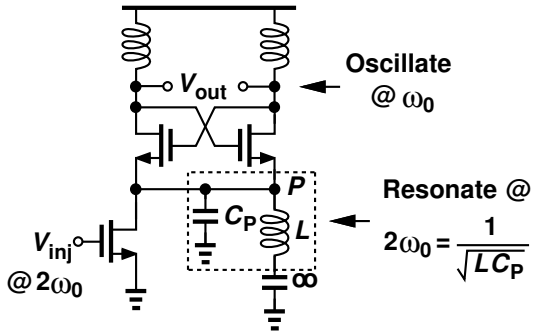


Fig. 7.27 Shunt peaking technique to improve locking range.

driven single-endedly, wasting 50% of the injection power. Another topology called direct injection is shown in Fig. 7.28 [6]. Here, the signal injection is accomplished by driving the two switches M_5 and M_6 differentially, which are sitting across the two outputs of the oscillator made of M_1-M_4 and L. Note that M_5 and M_6 are turned on and off almost simultaneously. Here, the input signals still drive the common-mode points, i.e., gates of M_5 and M_6. With proper design and biasing, the quasi-differential operation is expected to achieve a wider locking range.

The injection locking technique can be also utilized to implement dividers with modulus other than 2. Figure 7.29(a) reveals a possible realization of $\div 3$ circuit [7]. Here, transistors M_1-M_3 form a ring oscillator, and the input signal (approximately 3 times of the ring oscillation frequency) is injected into the common-mode point by means of M_4. Again with proper design, the ring would

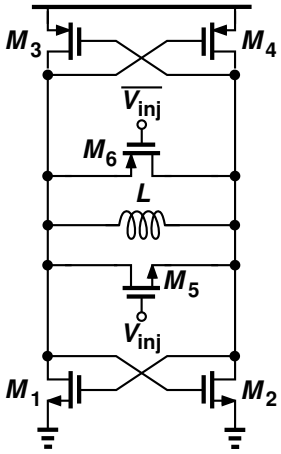
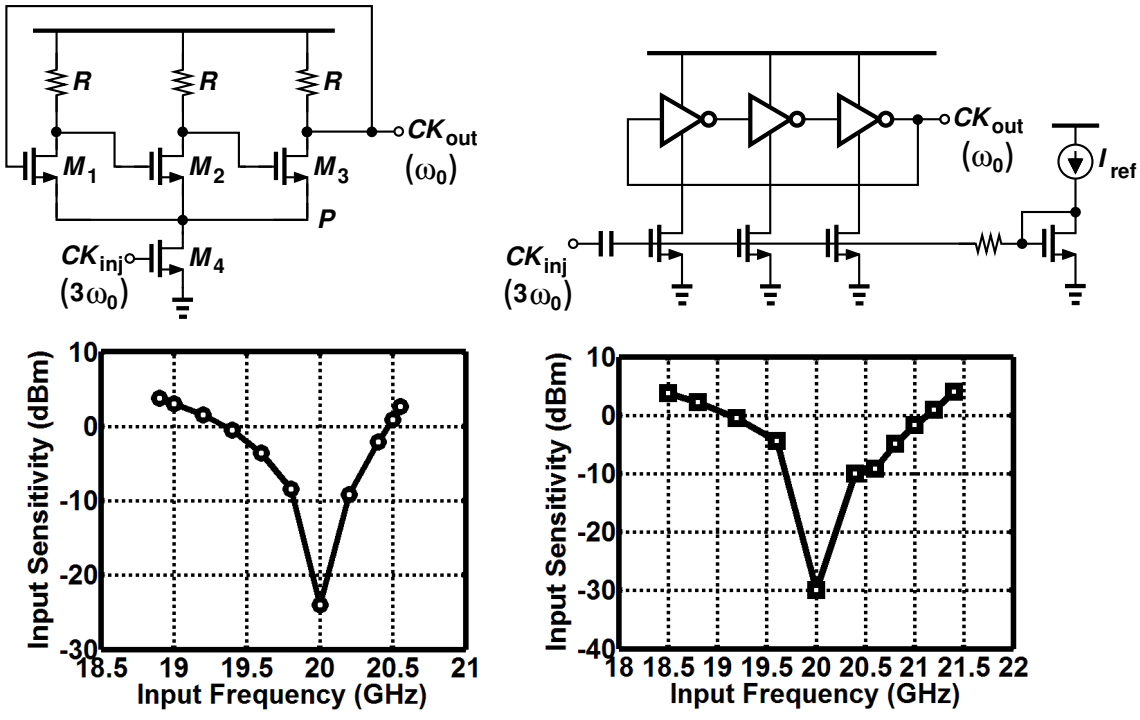


Fig. 7.28 Direct injection locking divider.



(a) (b)

Fig. 7.29 Divided-by-3 circuit with injection-locking technique: (a) RC ring, (b) inverter ring.

lock to one-third of the input frequency. Yet another ÷3 circuit with ring structure can be found in Fig. 7.29(b),where inverts are need with tail-currents governed by bias current I_B. Third-order

harmonic input is ac-coupled to tail currents, which in turn injection-locks to the fundamental ring. Again, I_B can be adjustable in order to overcome PVT variations. Simulated input sensitivity for 20 GHz $\div 3$ circuits are also plotted in Fig. 7.29, in which 40-nm CMOS technology is used.

The narrow locking range of injection-locked dividers usually necessitates careful design, skillful layout, as well as meticulous EM simulations. It is especially true at high speed since the deviation of natural frequency caused by PVT variations may destroy the locking.

7.5 FREQUENCY MULTIPLIERS

REFERENCES

[1] B. Razavi, *RF Microelectronics*, Second Ed., New Jersey: Prentice Hall, 2011.

[2] H. Wu and A. Hajimiri, A 19-GHz 0.5-mW 0.35-μmm CMOS frequency divider with shunt-peaking locking range enhancement, in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2001, pp. 412V413.

[3] H. R. Rategh and T. H. Lee, Superharmonic injection-locked frequency dividers, *IEEE J. Solid-State Circuits*, vol. 34, pp. 813V821, June 1999.

[4] J. Lee and B. Razavi, A 40-GHz frequency divider in 0.18-μm CMOS technology, *IEEE J. Solid-State Circuits*, vol. 39, no. 4, pp. 594-601, Apr. 2004.

[5] H. Wu and A. Hajimiri, A 19GHz 0.5μm CMOS frequency divider with shunt-peaking locking-range enhancement, *ISSCC Dig. of Tech. Papers*, pp. 412-413, Feb. 2001.

[6] M. Tiebout, A CMOS direct injection-locked oscillator topology as high-frequency low-power frequency divider, *IEEE J. Solid-State Circuits*, vol. 39, no. 7, pp. 1170-1174, Jul. 2004

[7] S. Verma et al., A multiply-by-3 coupled-ring oscillator for low-power frequency synthesis, *IEEE J. Solid-State Circuits*, vol. 39, no. 4, pp. 709-713, Apr. 2004.

CLOCK GENERATION

8.1 INTRODUCTION

Except for some asynchronous circuits, almost all electronic systems need a clock to launch data with the same tempo. In communication, phase-locked loops (PLLs) are extensively needed in wireless, RF, mm-wave, and wireline systems, providing ubiquitous solutions in our daily life. For example, a frequency synthesizer creates equally-spaced carrier frequencies for wireless channels [Fig. 8.1(a)]. These channels are separated by, says, 1 MHz, around the vicinity of 2.4 GHz. Setting $f_{ref} = 1$ MHz and programmable divide ratio M (= 2400 $\sim$ 2527), we can easily obtain 128 channels exactly located at desired positions. With proper design, one can create the carrier with the residual power at adjacent channels is at least 60 dB lower. If we were to use a passive filter to create the same carrier, the quality factor Q needs to be as high as 1.2×10^6! Let alone the frequency accuracy and other issues.

In data link, designers also employ PLLs to create clocks for the transmitters. Blocks including FFE, dividers and MUXes in different stages are need to be synchronized. In the receive side, clock and data recovery (CDR) circuits[1] extract clock from input data stream and retime it for the subsequent DMUXes, which need sub-rate clocks for synchronization as well. In some applications such as spread spectrum clocking (SSC) require the carrier to be modulated (in frequency) to minimize electro-magnetic interference (EMI). Actually radar systems adopt the same approach in so called frequency-modulated continuous-wave (FMCW) radars. Other applications of PLLs can be found in various fields of science and engineering, such as disk drive and motor control.

[1] As a special transformation of PLL, the CDR circuit is worth discussion in an independent chapter (Chapter 9).

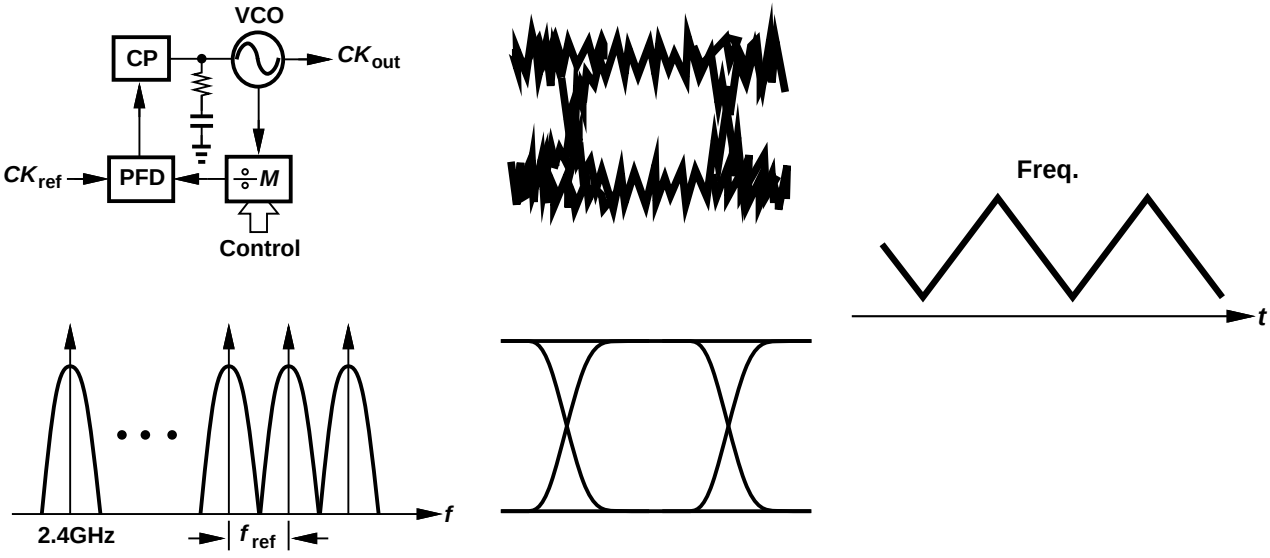


Fig. 8.1 PLL applications:(a) frequency divider, (b) clock and data recovery, (c) spread spectrum clocking.

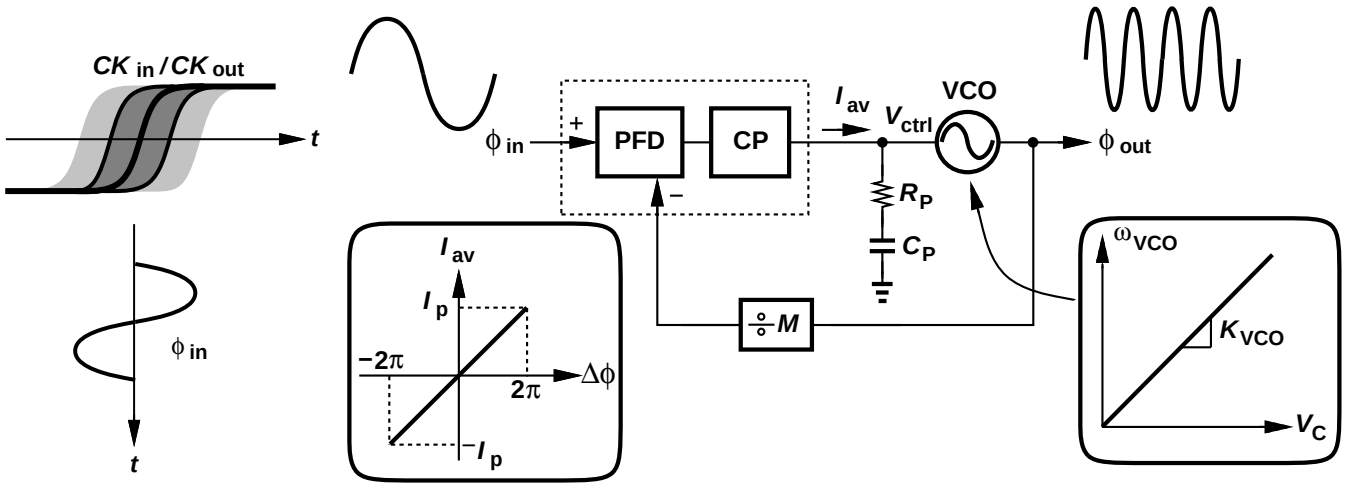
We focus our discussion on different types of PLLs and their associated circuits. Highly related blocks will also be included in this chapter.

8.2 INTEGER-N PLLS

8.2.1 Fundamental Model

Charge-pump PLLs (as called "type-II PLLs") are extensively need in modern communication systems due to their overwhelmingly advantages over simple ("type-I") PLLs, such as zero phase error, design flexibility, and infinite acquisition range.

We focus our discussion on type-II PLLs, beginning with a well-known linear model as illustrated in Fig. 8.2. Here the phase (Φ_{in} and Φ_{out}) can be considered the movement of zero-crossing point of a clock. Like other signals, it can be a function of time. It is easy to image if a clock frequency is N times higher than the other, then its phase is also N times larger (i.e., $\phi = \int \omega dt$). In other words, it is a phase domain model with input Φ_{in} and output Φ_{out}, but is identical if we put



(a) (b)

Fig. 8.2 (a) Definition of ϕ_{out} and ϕ_{in}, (b) Standard linear PLL model.

everything in frequency domain (f_{in} and f_{out}). The phase and frequency detector (PFD) together with the charge pump (CP) presents a linear characteristic. In the range of $\pm 2\pi$, it generates an average output current I_{av} proportional to the input phase difference. Based on the present phase error, a positive or negative pumping current I_{av} would be injected into the loop filter (R_p and C_p), which in turn changes the control voltage V_c and the VCO frequency ω_{osc}. The negative feedback loop eventually neutralizes the phase error of the PFD ($\Delta\phi = 0$) and makes $\omega_{out} = M \cdot \omega_{in}$. Here we emphasize "average" since many PFDs (e.g., type-IV PFD) are actually operated in pulses. In normal conditions, they can be considered continuous and s-domain analysis fits. We address this issue in the following paragraphes. The voltage-controlled oscillator (VCO) is modeled as a linear tuning characteristic with gain K_{VCO}. Since phase is the integration of frequency, we have

$$\phi_{out}(t) = \int K_{VCO} V_c(t) dt \tag{8.1}$$

$$\Phi_{out}(s) = \frac{K_{VCO}}{s} V_c(s). \tag{8.2}$$

Note that $s = j\omega$ here means the phase (or frequency) changing rate. In steady-state, we arrive at the loop gain as

$$H(s) \mid_{open} = \frac{K_{VCO} I_p (1 + sR_p C_p)}{2\pi M C_p s^2}, \tag{8.3}$$

and its Bode plot is depicted in Fig. 8.2(a). As ($K_{VCO}I_p/M$) increase, the magnitude moves up, improving the phase margin and stability. The closed-loop transfer function is given by

$$\frac{\Phi_{out}}{\Phi_{in}}(s) \triangleq H(s) = \frac{M(2\zeta\omega_n s + \omega_n^2)}{s^2 + 2\zeta\omega_n + \omega_n^2}, \tag{8.4}$$

where nature frequency ω_n and damping factor ζ are

$$\omega_n = \sqrt{\frac{I_p K_{VCO}}{2\pi C_p M}} \tag{8.5}$$

$$\zeta = \frac{R_p}{2}\sqrt{\frac{I_p C_p K_{VCO}}{2\pi M}}. \tag{8.6}$$

The poles of the closed-loop transfer function $H(s)$ is equal to

$$s_{1,2} = -\zeta\omega_n \pm \omega_n\sqrt{\zeta^2 - 1}. \tag{8.7}$$

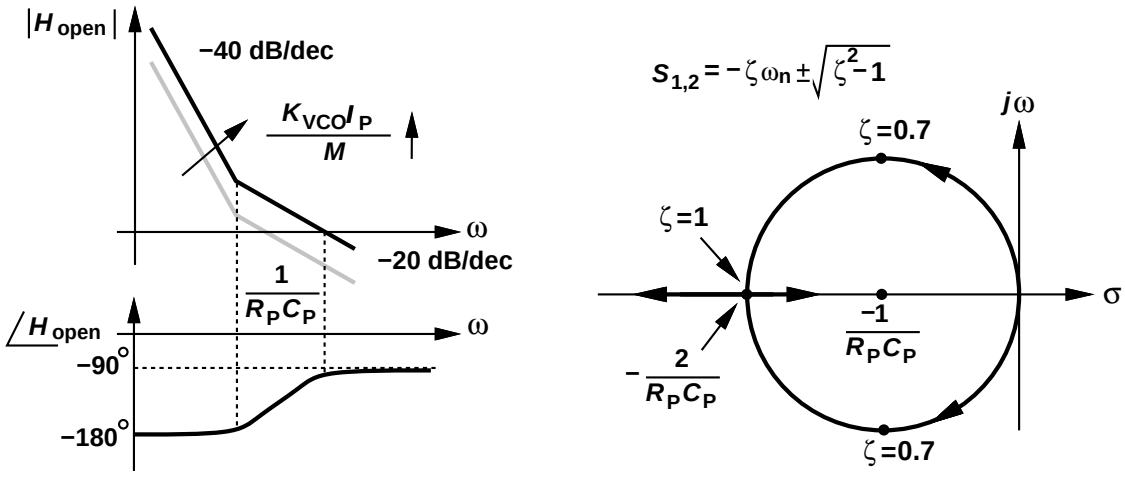


Fig. 8.3 (a) Bode plot of loop gain $H(s)\ |_{open}$, (b) root locus on closed-loop transfer function $H(s)$

Figure 8.3(b) depicts the possible solution of $s_{1,2}$ as a function of ζ (i.e., root locus). For $0 < \zeta < 1$ (under-damped), $s_{1,2}$ are complex conjugates, moving around the circle and merging at $\left(-\frac{2}{R_p C_p}, 0\right)$ as $\zeta = 1$. For $\zeta > 1$ (over-damped), $s_{1,2}$ split out and move toward opposite directions:

one goes to the right and eventually stop at $\left(-\frac{1}{R_pC_p}, 0\right)$, where as the other goes to the left (until infinity). This simple yet classic model is the fundamental of all PLLs, and lots of properties can be derived from it. For example, in most wireless systems, carrier needs to jump from one channel to another. To make it agile, ζ is chosen to be less than 1. Applying a step function in frequency into the PLL, the output frequency would be M-times of it with some ringing. The ringing decays exponentially with a time constant $\tau = (\zeta\omega_n)^{-1}$. That is, for a given ζ, we need a larger ω_n to speed up settling. However, as will be shown latter, increasing ω_n also introduces more noise from input. It forms a tradeoff between settling time and input noise. Sometimes other performance would be compromised as well. In practice, the settling time is roughly given by $10/(\zeta\omega_n)$. $M = 2400$, $K_{VCO} = 2\pi \cdot 200$ MHz/V, $I_p = 1$ mA, and $R_p = 1$ KΩ, we obtain a channel-jumping settling time of about 240 μs.

Example 8.1

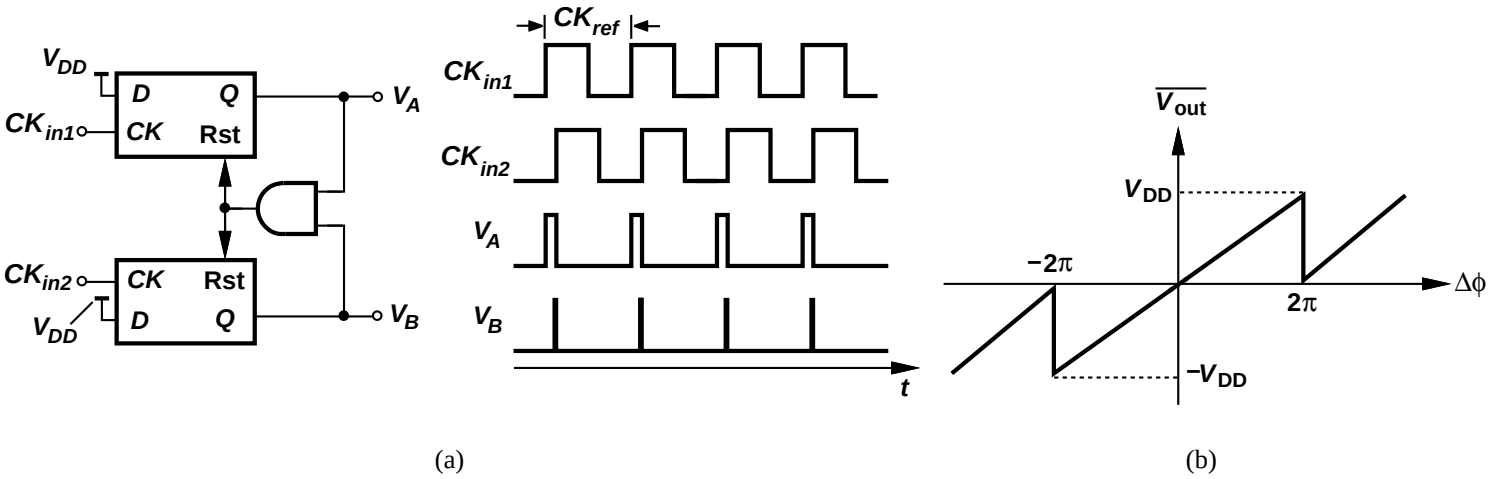
Figure 8.3(a) implies $|H_{open}|$ is greater than unity and $\angle H_{open}$ approaches -180° as $w \to 0$. Explain why a PLL does not become unstable.

Solution:

The barkausen criterion for oscillation are satisfied as $w = 0$. Indeed, a PLL "oscillates" at $w = 0$, meaning the phase(or the frequency) is stuck to a constant value. It can be imagined as the crossover point in Fig. 8.2(a) never changes from cycle to cycle. Do not confuse phase (or frequency) changing rate w with physical clock frequency.

8.2.2 PFD

Perhaps the most-commonly used PFD is the "type-IV" PFD as shown in Fig. 8.4(a). It consists of two resettable flip-flops and an AND gate, and two clock input CK_{in1} and CK_{in2} are compared in phase. The leading input raises its output until the lagging one arrives, and at that moment the reset signal generates. As a result, we obtain an input $(\Delta\phi)$ and output characteristic as shown in Fig. 8.4(b).



(a) (b)

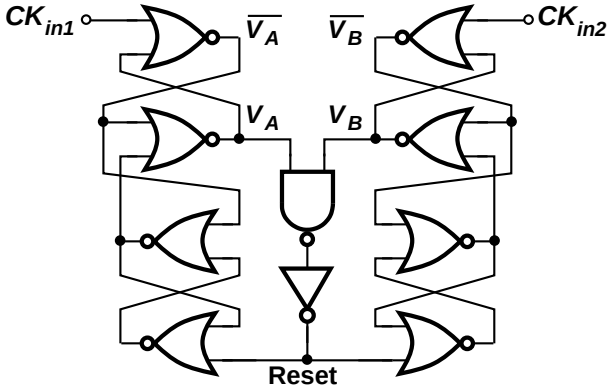
Fig. 8.4 Type-IV PFD.

As the loop gradually corrects the phase error, the two clock inputs eventually line up (i.e., $\Delta\phi = 0$). The whole type-IV PFD can be made in pure digital logic, as it would most likely be operated at a speed no more than several hundred MHz. Note that the reset signal won't be generated until a complete short pulse is created by the lagging signal. The minimum pulse width determines the maximum operation speed of a type-IV PFD, which is typically a few GHz in advance CMOS technologies. A commonly used CMOS version is depicted in Fig. 8.5(a).

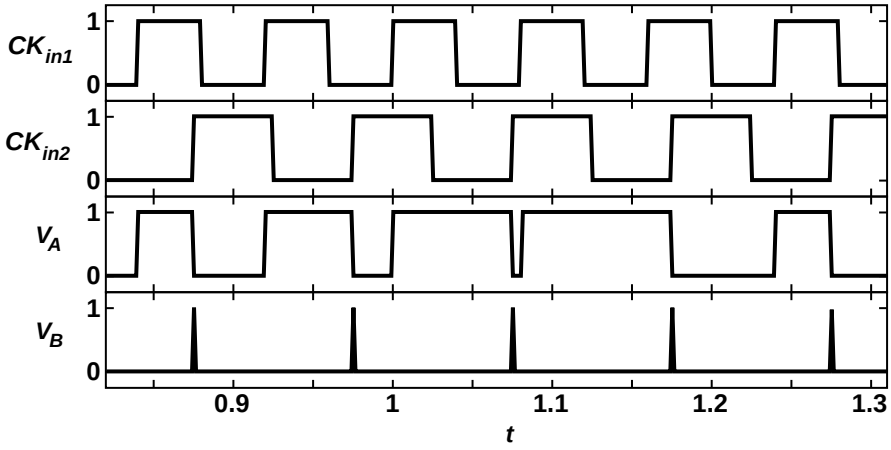
The asymmetric (with respect to y-axis) characteristic in Fig. 8.4 suggests an infinite frequency acquisition range. Indeed, type-IV PFDs maintain correct polarities as phase error exceeds $\pm 2\pi$. If a large frequency difference between CK_{in1} and CK_{in2} is presented, the leading path would "swallow" extra pulse and operate correctly afterwards [Fig. 8.5(b)]. Also known as "cycle slip", this behavior makes type-IV PFDs remarkably attractive owing to the capability of simultaneous phase and frequency tracking. Upon lock, CK_{in1} and CK_{in2} would have identical phase and frequency ($\Delta\phi = 0$, $\Delta\omega = 0$). A type-IV PFD presents no dead zone as it always need a complete pulse to reset the flip-flops.

Although the type-IV PFDs have tremendous advantage, the two pulses induce significant perturbation to the loop. Recognized as reference clock feedthrough, this periodic perturbation happens at every phase detection and causes ripples on the control line voltage. As a result, two spurs

would occur around the carrier in the spectrum (will be analyzed in detail later). The reasons to



(a)



(b)

Fig. 8.5 Type-IV PFD (a) realized in CMOS gates, (b) transient waveforms at large frequence difference.

cause this issue includes circuit mismatch, charge-pump current imbalance, skews, and other non-idealities. Many attempts have been made to minimize the reference spurs. For example, charge transfer technique spreads out the momentary (positive or negative) increment over longer period [1], [2]; analog phase detector utilizes current-mode logic to reduce swing [3], [4]; compensated charge-pump design balances the device mismatch [5], [6]; and distributed phase detector shortens the step of variation to avoid abrupt changes on the control voltage [7], [8]. However, none of

these approaches can really get rid of the pulse generation, so the control line ripple can never be removed entirely.

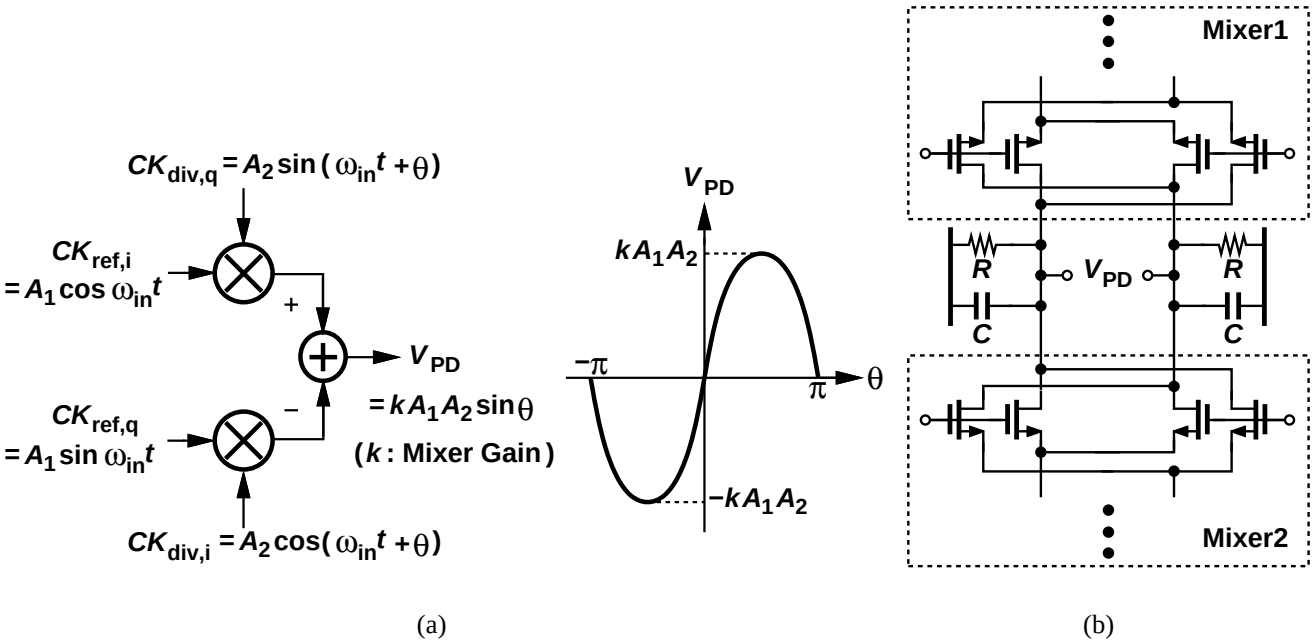


Fig. 8.6 Phase detector base on a SSB-mixer: (a) characteristic, (b) SSB-mixer with RC low-pass filter.

To avoid producing on-off pulses, the phase detection can be conducted by mixing two quadrature signals. One comes from the reference input (CK_{ref}, provided by the static divide-by-2 circuit) and the other from the last divider stage (CK_{div}). As illustrated in Fig. 8.6, an single sideband (SSB) mixer can distill the phase error of two synchronous signals, and reveals a sinusoidal input-output characteristic. Driven by the phase detector output, the V/I converter provides a continuous and proportional current, either positive or negative, to the loop filter and changes the control voltage accordingly. Since the characteristic can be approximately considered linear in the vicinity of origin and no pulse generation is involved, it achieves a truly ”quiet” phase examination and reference spurs are significantly reduced. It is important to know that the current imbalance in the V/I converter is no longer an issue here, since the phase detector would create an offset between the two inputs to compensate it perfectly. In the presence of mismatches, finite ”image” would be observed at twice the PD operation frequency $2\omega_{in}$. To suppress it, a low-pass filter must

be placed right after the SSB mixer. A clever realization is to load the mixer with RC networks, which generates a corner frequency to reject the image. For typical values of R and C, the corner could be 10 MHz or so to suppress the image by more than 40 dB. Note that the low-pass filtering has little impact on the overall loop bandwidth, which is designed to be much lower than that. The control line ripple can be dramatically reduced from mV to μV by this structure.

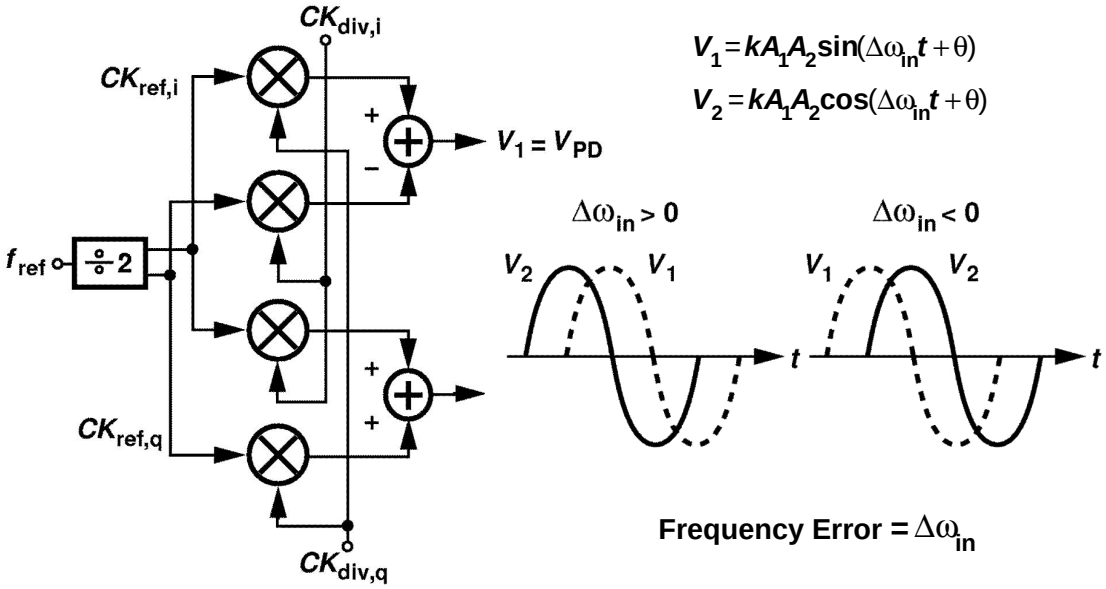


Fig. 8.7 Frequency detection.

The periodic characteristic of the phase detector implies a limited capture range. Fortunately, the frequency detection can be accomplished by introducing another SSB mixer, arriving at a wide operation range. As shown in Fig. 8.7, the two outputs and appear orthogonally in the presence of frequency error:

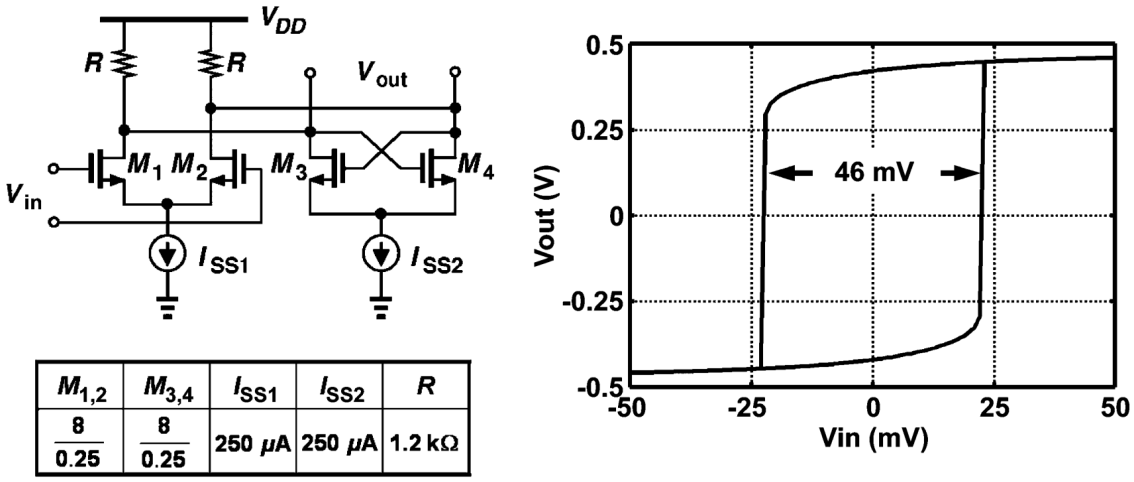
$$V_1 = V_{PD} = kA_1A_2\sin(\Delta\omega_{in}t + \theta) \tag{8.8}$$

$$V_2 = kA_1A_2\cos(\Delta\omega_{in}t + \theta) \tag{8.9}$$

Here, $\Delta\omega_{in}$ represents the frequency difference between CK_{ref} and CK_{div}. Obviously, whether V_1 is leading or lagging V_2 depends on the sign of $\Delta\omega_{in}$, and it can be easily obtained by using a flip-flop to sample one signal with the other [3]. Based on the flip-flop's output, the V/I converter

designated to the frequency detection loop [i.e., $(V/I)_{FD}$] injects a positive or negative current to the loop filter. This current is $3 \sim 4$ times larger than the peak current of $(V/I)_{PD}$ to ensure a smooth frequency acquisition.

To minimize the disturbance on VCO, the frequency acquisition should be turned off upon lock. Observing that V_2 stays low under phase locking, it can be used to automatically shut off the frequency detector. Here, we apply V_2 to $(V/I)_{FD}$ and have it disabled when the loop is locked. In other words, V/I converter activates for 50% of the time during tracking, and automatically switches off when the frequency acquisition is accomplished [4].



$M_{1,2}$	$M_{3,4}$	I_{SS1}	I_{SS2}	R
$\frac{8}{0.25}$	$\frac{8}{0.25}$	250 μA	250 μA	1.2 kΩ

Fig. 8.8 Hysteresis buffer.

The very slow sinusoids V_{PD} and V_2 may cause malfunction of FF_1 if they drive the flip-flop directly, because the transitions of V_{PD} and V_2 become extremely slow when the loop is close to lock. The fluctuation caused by unwanted coupling or additive noise would make the transitions ambiguous, possibly resulting in multiple zero crossings. To remedy this issue, hysteresis buffers are employed to sharpen the waveforms. Figure 8.8 depicts the buffer design, where the cross-coupled pair M_3-M_4 provides different switching thresholds for low-to-high and high-to-low transitions, and the positive feedback helps to create square waves as well. Here, $(W/L)_{1,2} = (W/L)_{3,4} = 8/0.25$, and a threshold difference of 46 mV is observed. Figure 8.9 shows the complete PFD design.

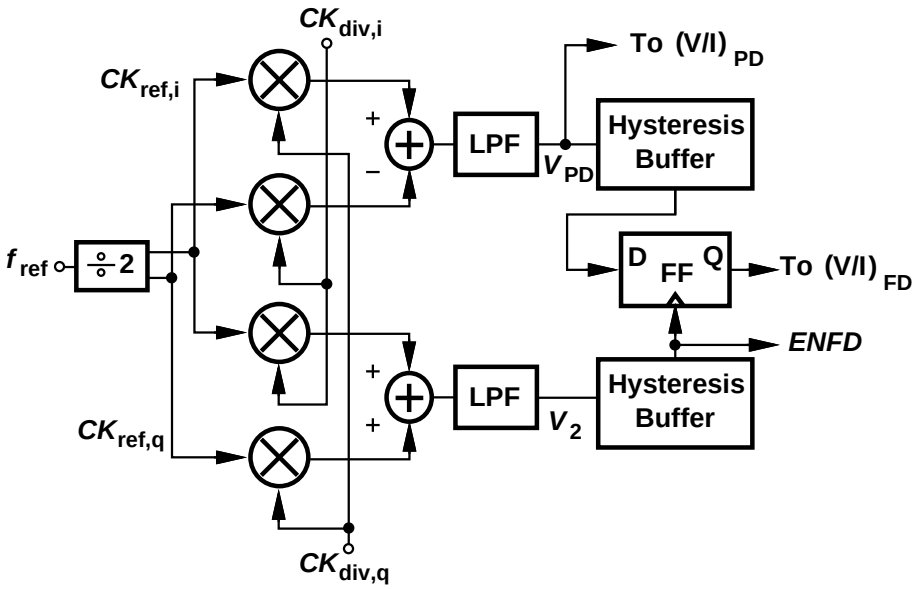


Fig. 8.9 Complete design of mixer-based PFD.

8.2.3 Charge Pump

Typical charge pump suffers from 3 issues: channel charge redistribution, (random) mismatch, and channel-length modulation. Turning on and off a switch involves channel formation and dismission. In charge pump operation, certain part of the channel changes would be injected into the loop filter. The unpredictable amount of injection is a function of control voltage V_c, device size, and process variation. Mismatch between current mirrors makes up and down current imbalanced, arriving at phase skew and control line ripple. Channel-length modulation causes the pumping current to vary as V_c changes, which deviates PLL's operation from optimal setting. While the last issue could be minimized by circuit techniques, charge redistribution and mismatch ultimately limit the charge pump performance.

Let's consider a charge pump structure shown in Fig. 8.10(a). The charging and discharging currents mirrored from I_1 and I_2 suffers from channel-length modulation. Besides, when both switches are off, M_2 and M_3 are in deep triode region and carry no current. This makes charge redistribution issue even worse, as the internal nodes participate in charge sharing. A modified version is depicted in Fig. 8.10(b), where all devices are in saturation when the charge pump is

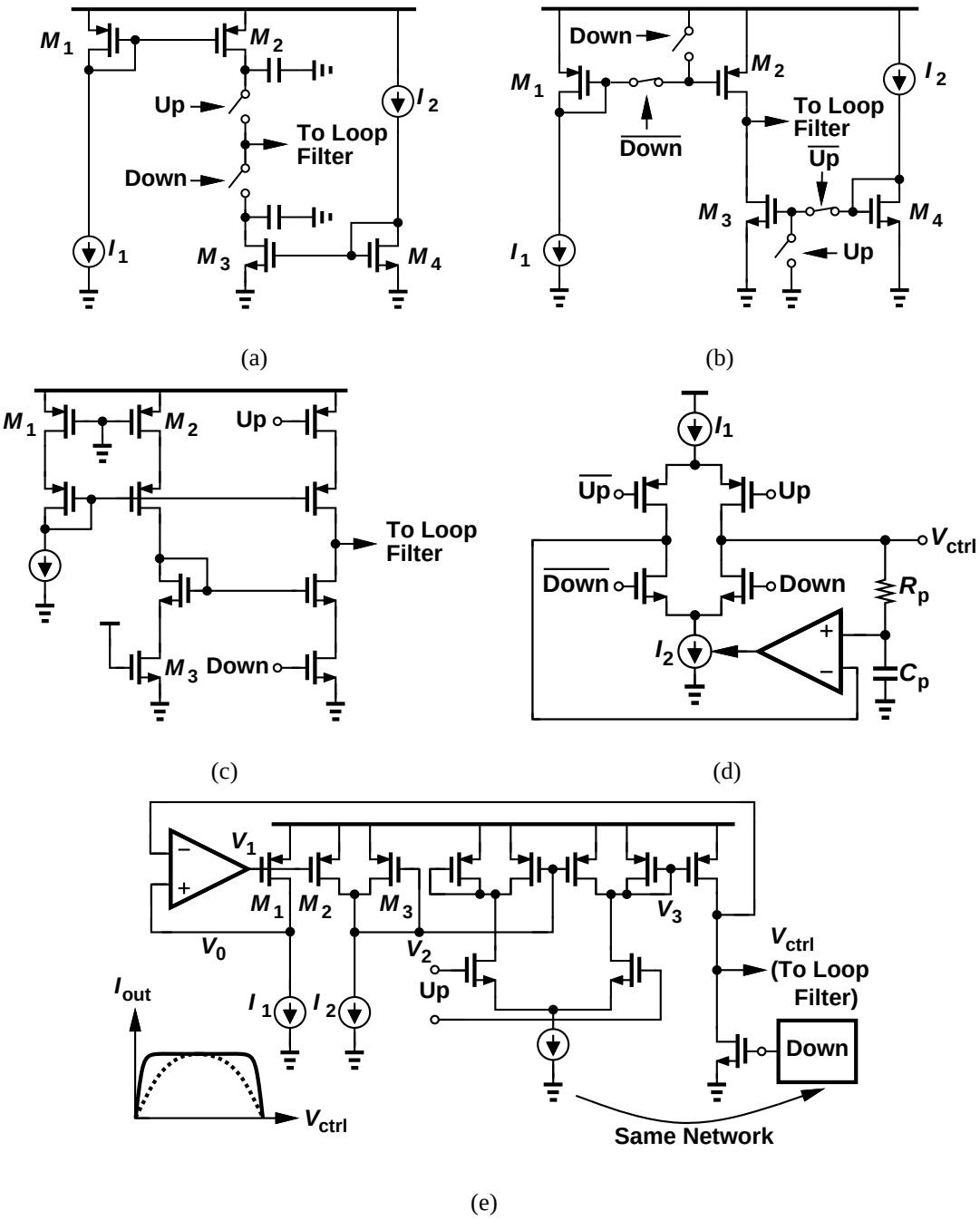


Fig. 8.10 Charge pumps: (a) direct switching, (b) gate switching, (c) cascode structure, (d) servo control, (e) linearization.

idle. However, charge redistribution issue still exists. Pumping current would vary because of channel-length modulation, even though we can ensure $I_1 = I_2$.

A cascode structure is illustrated in Fig. 8.10(c), where M_1, M_2 and M_3 mimic the (large signal) on-resistance of the Up/Down switches. Mismatch and channel-length could be reduced to some extent, but the voltage headroom forms another issue. A better way to remove the mismatch is to use a servo control loop, as illustrated in Fig. 8.10(d). The feedback loop dynamically adjust I_2 to be exactly equal to I_1, canceling out the effect of channel-length modulation. Since random mismatch of the circuit are removed at the same time, the pumping current is relatively constant as V_c changes. A similar approach can be found in Fig. 8.10(e), where the Up and Down paths are separated. For higher control voltage V_{ctrl}, V_0 follows because of the negative feedback around the Opamp. V_1 goes down to drive M_1 harder so as to maintain constant current I_1. Since M_2 is also driven by V_1, V_2 has to go up to keep I_2 constant. That is, V_3 goes down slightly to stay in balance. The other side for Down signal follows the same rules. As a result, the pumping current remains constant over a wide range of V_c. Such a linearization technique compensates the effect of channel-length modulation as well.

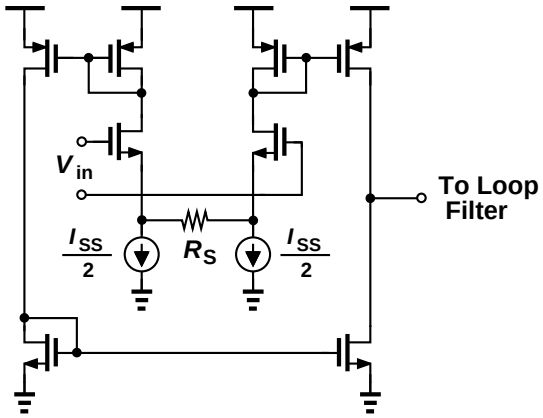


Fig. 8.11 V/I converter used in SSB-mixer based PFD.

The charge pump circuits in Fig. 8.10(a)-(e) share one thing in common — none of them can get rid of the effect of charge redistribution. It is simply because these charge pumps have switches to cooperate with a type-IV PFD. The SSB-mixer based PFD manifests itself in charge pump design again. Since no abrupt switching is involved, one can take a simple current steering circuit

as a charge pump (or more precisely, a V-to-I converter). Figure 8.11 presents one example. Here, the near-dc input is converted into current in proportion. The degeneration resistor R_s extends the linear region of operation. In the presence of mismatch, the PLL itself would create a small static offset appear in the PFD's two inputs to cancel it out. No control line ripple would be created. Such a simple V/I converter will work nicely as long as it bandwidth is much greater than the loop bandwidth, which is very easy to achieve in today's CMOS technologies.

8.2.4 Loop Filter

Higher order loop filters can be used to further suppresses the control line ripple. Figure 8.12 illustrates two popular approaches. Adding a small C_2 is the easiest solution, which absorbs a significant amount of ripple caused by pulse skews. However, owing to the additional pole introduced by C_2, the phase margin would be degraded.

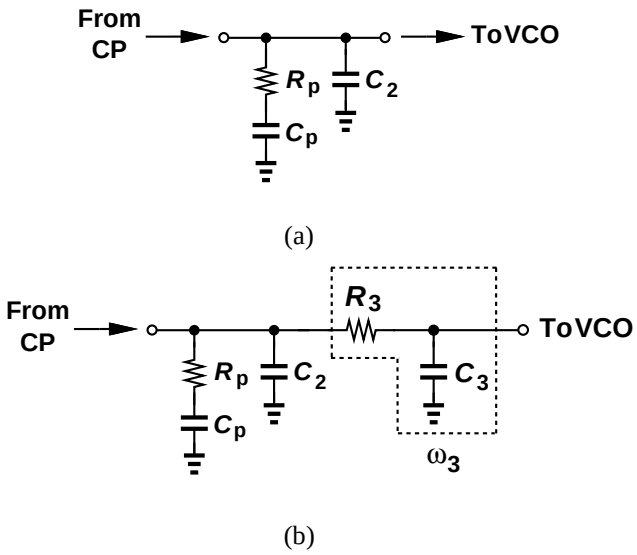


Fig. 8.12 (a) 2nd-order, (b) 3rd-order loop filter.

For xxxx, the optimal phase margin is given by [9]

$$PM \approx \tan^{-1}(4\zeta^2) - \tan^{-1}\left(4\zeta^2 \frac{C_{eq}}{C_p}\right). \tag{8.10}$$

where $C_{eq} = C_p C_2/(C_p + C_2)$. Usually C_2 is chosen to be less than or equal to $C_p/20$ as a mild averaging capacitor. Shown in Fig. 8.12(b) is a realization of 3rd-order filter, where R_3 and C_3

form another low-pass filter to damp the ripple. It is important to know that, the corner frequency $\omega_3[= 1/R_3C_3]$ must be greater than the loop bandwidth (to perform normal function) and less than the reference frequency (to suppress the ripple). In other words,

$$Loop\,BW < \omega_3 < \omega_{ref} \leq \omega_{osc}, \tag{8.11}$$

where ω_{osc} denotes the VCO (output) frequency. Note that these higher order loop filter are not necessary for the SSB-mixer based PFD, as its near-dc operation already contributes negligible ripple. It is another fact demonstrating the superiority of a SSB-mixer based PFD. It is worth

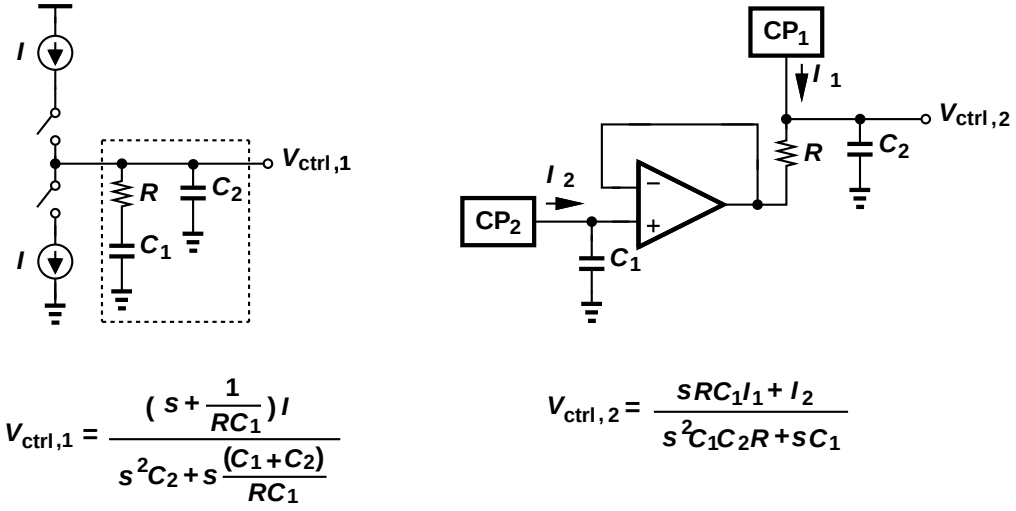


Fig. 8.13 Loop filter with 2 pumping currents.

noting that it is possible to use two pumping currents to achieve more flexible design. Figure 8.13 illustrates an example. Here, the loop filter is split into two parts, and each charge pump injects its own current into one of them. As compared with the conventional design, the transfer function still presents two poles and one zero. The key point is that the design with two current has one more parameter to optimize the tradeoffs among phase margin, clock feedthrough, and jitter performance. One serious issue of loop filter design is the large capacitance C_p. In low-supply environments, MOS capacitor is not a good choice for loop filter design. After all, neither the channel nor the capacitance will be established unless the voltage across it is greater than $|V_{th}|$. Other linear capacitors are not suitable as well if large capacitor is required. A 500 pF

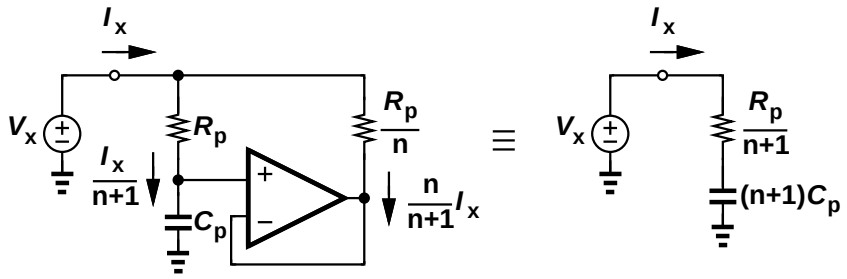


Fig. 8.14 Capacitance multiplication technique.

fringe capacitor, for example, would occupy an area as large as 0.16 mm^2! To overcome the area issue, a capacitance multiplication technique is usually adopted. As illustrated in Fig. 8.14, an additional resistor R_p/n is placed in parallel with R_p, and a unity-gain buffer copies the voltage. The impedance seen looking into the network can be calculated as

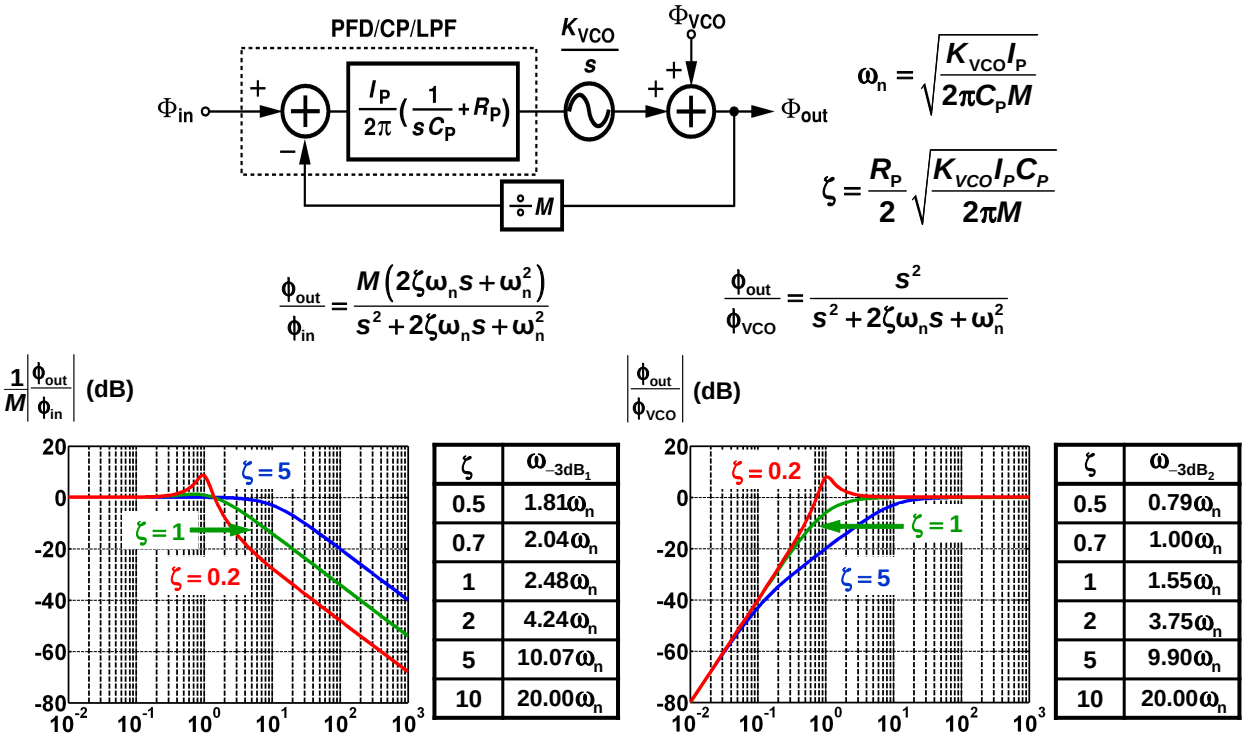
$$Z_{eq} \triangleq \frac{V_x}{I_x} = \frac{1}{n+1}\left(R_p + \frac{1}{sC_p}\right), \tag{8.12}$$

which is nothing more than an equivalent RC network of $R_p/(n+1)$ and $(n+1)C_p$. In other word, the effective capacitance has been enlarged by a factor of $n+1$.

8.2.5 Loop Bandwidth Optimization

Integer-N PLLs must be optimized with different tradeoffs for different applications. In many wireline systems, we are looking for clock generators which different frequencies and/or phases. For instance, a transmitter in data link needs a clock multiplication unit (CMU) to create clocks for different muxing stages. Such a PLL basically operates at a single frequency, and need not worry about settling time for frequency jumping. Reference spurs may not be an issue either if it locates out of the band of spectrum integration. Frequency synthesizers in wireless concern the phase noise and reference spurs. Spread-spectrum clock generators require a proper setting on loop bandwidth.

If the time-domain rms jitter (integrated from spectrum) is the main concern (which is true for most wireline PLLs), we have to determine the optimal loop bandwidth first. Consider the linear model shown in Fig. 8.15 again. The input noise (including noise from the reference, PFD/CP and



ζ	ω_{-3dB_1}
0.5	$1.81\omega_n$
0.7	$2.04\omega_n$
1	$2.48\omega_n$
2	$4.24\omega_n$
5	$10.07\omega_n$
10	$20.00\omega_n$

ζ	ω_{-3dB_2}
0.5	$0.79\omega_n$
0.7	$1.00\omega_n$
1	$1.55\omega_n$
2	$3.75\omega_n$
5	$9.90\omega_n$
10	$20.00\omega_n$

Fig. 8.15 Noise transfer function.

divider chain) presents a transfer function to the output:

$$\frac{\phi_{out}}{\phi_{in}}(s) = \frac{M(2\zeta\omega_n s + \omega_n^2)}{s^2 + 2\zeta\omega_n s + \omega_n^2}, \tag{8.13}$$

which is identical to Eq.(8.4). Peaking of this low-pass transfer function vanishes when $\zeta \geq 1$, and the -3-dB bandwidth (ω_{BW1}) increases as ζ grows. The transfer function eventually rolls off at -20 dB/dec no matter what ζ it has. In fact, for $\zeta \gg 1$, the transfer function degenerate to

$$\frac{\phi_{out}}{\phi_{in}}(s) = \frac{M \cdot 2\zeta\omega_n}{s + 2\zeta\omega_n}, \tag{8.14}$$

which possesses a -3-dB bandwidth of $2\zeta\omega_n$. Similarly, the VCO noise has its own transfer function:

$$\frac{\phi_{out}}{\phi_{VCO}}(s) = \frac{s^2}{s^2 + 2\zeta\omega_n + \omega_n^2}, \tag{8.15}$$

peaking disappear for $\zeta \geq 1$. The -3-dB bandwidth (ω_{BW2}) for different ζ is listed as well. Unlike ϕ_{out}/ϕ_{in}, the VCO noise transfer function presents different traces for underdamped and overdamped loop. For $\zeta \gg 1$, the climbing ramp bends from $+40$ dB/dec to $+20$ dB/dec at the same point, and gradually merges to flat line afterwards. For $\zeta \leq 1$, on the other hands, it keep the $+40$ dB/dec slope until the -3-dB bandwidth. For $\zeta \gg 1$, the VCO noise transfer function becomes

$$\frac{\phi_{out}}{\phi_{VCO}}(s) = \frac{s}{s + 2\zeta\omega_n}. \tag{8.16}$$

The difference between ω_{BW1} and ω_{BW2} diminishes. For simplicity, we follow the tradition and define the loop bandwidth $\omega_{BW} \triangleq 2\zeta\omega_n \cong \sqrt{\omega_{BW1}\omega_{BW2}}$.

Since the two noise sources are uncorrelated, their contribution to the output can be calculated separately. Recall the theory that for a linear time-invariant (LTI) system, the spectral density at the output is the product of the square of the transfer function and the spectral density at the input (S_X):

$$S_Y(\omega) = S_X(\omega)|H(\omega)|^2. \tag{8.17}$$

If the input and VCO noise spectrum are denoted as $S_{\phi,in}$ and $S_{\phi,VCO}$ respectively, the overall output spectrum of a PLL is given by be the combined effects of the two:

$$\begin{aligned} S_{\phi,out}(\omega) &= S_{\phi,out,in}(\omega) + S_{\phi,out,VCO}(\omega) \\ &= S_{\phi,in}(\omega)\left|\frac{\phi_{out}}{\phi_{in}}\right|^2 + S_{\phi,VCO}(\omega)\left|\frac{\phi_{out}}{\phi_{VCO}}\right|^2. \end{aligned} \tag{8.18}$$

Figure 8.16 illustrated the calculation. For simplicity, we assume $S_{\phi,VCO} \propto 1/\omega^2$ only. It does not lose generality in most cases, as the middle band of offset frequency dominates the jitter/noise performance. For example, OC-192 defines jitter generation (JG) as the integration of clock spectrum from 20 kHz to 80 MHz offset. Thus,

$$S_{\phi,VCO}(\omega) = S_{\phi,VCO}(\omega_0) \cdot \frac{\omega_0^2}{\omega^2}. \tag{8.19}$$

To effectively observe the bandwidth optimization, we divide our discussion into two parts.

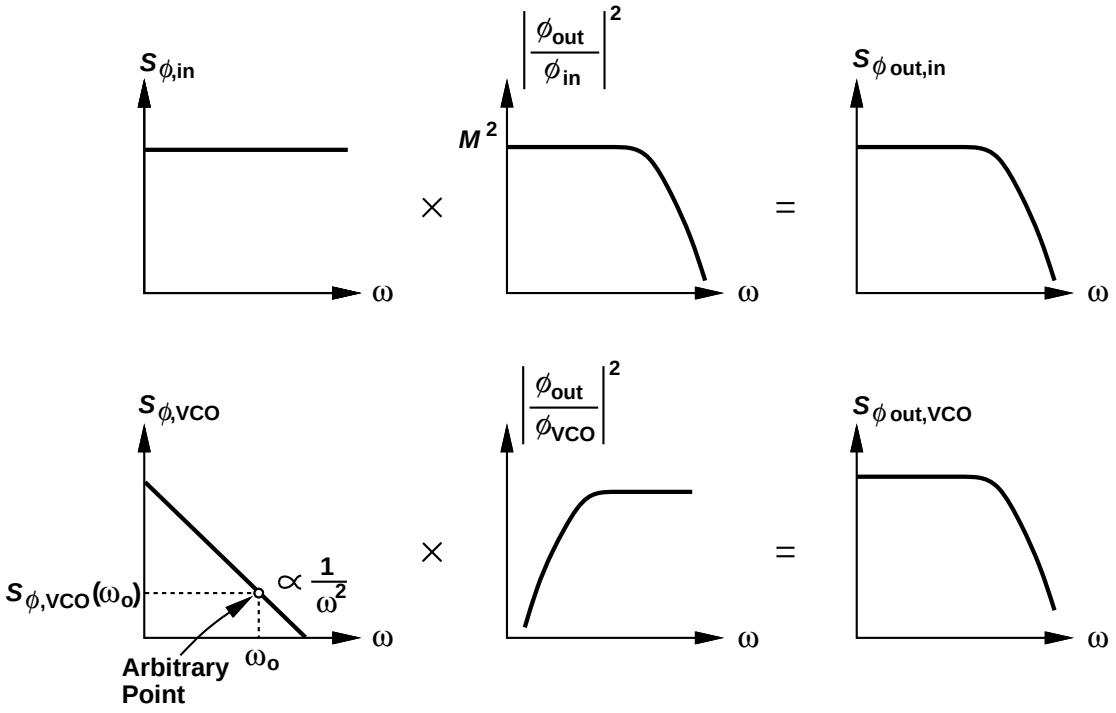


Fig. 8.16 Integer-N PLL spectrum calculation.

Overdamped PLLs Since the two transfer functions become first-order, the output spectrum is equal to

$$S_{\phi,out}(\omega) = S_{\phi,in} \cdot M^2 \frac{\omega_{BW}^2}{\omega^2 + \omega_{BW}^2} + S_{\phi,VCO}(\omega_0) \cdot \frac{\omega_0^2}{\omega^2} \cdot \frac{\omega^2}{\omega^2 + \omega_{BW}^2}. \tag{8.20}$$

Integration $S_{\phi,out}(\omega)$ yields the total noise, or equivalently, the rms jitter:

$$\begin{aligned} J_{rms,nor}^2 &= 2 \cdot \int_0^\infty S_{\phi,out}(\omega = 2\pi f) df \\ &= 2 \cdot \left[S_{\phi,in} \cdot M^2 \cdot \omega_{BW}^2 + S_{\phi,VCO}(\omega_0)\omega_0^2 \right] \cdot \int_0^\infty \frac{1}{4\pi^2 f^2 + \omega_{BW}^2} df \\ &= \frac{1}{2} \cdot \left[S_{\phi,in} \cdot M^2 \cdot \omega_{BW} + S_{\phi,VCO}(\omega_0) \cdot \frac{\omega_0^2}{\omega_{BW}^2} \right], \end{aligned} \tag{8.21}$$

which is a function of ω_{BW}. To find the optimal ω_{BW} that results in a minimum jitter, we have

$$\frac{\partial (J_{rms,nor}^2)}{\partial \omega_{BW}} = 0. \tag{8.22}$$

It turns out

$$S_{\phi,in} \cdot M^2 = S_{\phi,VCO}(\omega_0)\frac{\omega_0^2}{\omega_{BW,opt}^2}. \tag{8.23}$$

That is, the optimal loop bandwidth locates in the intersection of $M^2 \cdot S_{\phi,in}$ and $S_{\phi,VCO}$. Figure 8.17(a) illustrates the result. Note that the two noise sources contribute equal amount of rms jitter if the loop bandwidth is optimized.

Critical-damped or under-damped PLLs For the case of $\zeta \approx 1$, we can still use the first-order approximated transfer function of input noise. However, it no longer fits that of VCO noise. Instead, we adopt piecewise integration as a substitute. That is,

$$\frac{\phi_{out}}{\phi_{VCO}} = \begin{cases} \dfrac{s^2}{\omega_n^2}, & 0 \le \omega < \omega_{BW} \\ 1, & \omega_{BW} \le \omega < \infty. \end{cases} \tag{8.24}$$

Taking it into calculation, we arrive at

$$\begin{aligned} J_{rms,nor}^2 &= 2 \cdot \left[\int_0^{\infty} S_{\phi,in} \cdot \frac{M^2 \cdot \omega_{BW}^2}{4\pi^2 f^2 + \omega_{BW}^2} df + \int_0^{\frac{\omega_{BW}}{2\pi}} S_{\phi,VCO}(\omega_0) \cdot \frac{\omega_0^2}{4\pi^2 f^2} \cdot \frac{16\pi^4 f^4}{\omega_n^4} df \right. \\ &\quad \left. + \int_{\frac{\omega_{BW}}{2\pi}}^{\infty} S_{\phi,VCO}(\omega_0) \cdot \frac{\omega_0^2}{4\pi^2 f^2} \cdot 1 df \right] \\ &= \frac{1}{\pi} \cdot \left[\frac{\pi}{2} M^2 S_{\phi,in} \omega_{BW} + \left(1 + \frac{16}{3}\zeta^4\right) S_{\phi,VCO}(\omega_0) \frac{\omega_0^2}{\omega_{BW,opt}} \right]. \end{aligned} \tag{8.25}$$

Again, to minimize jitter, we make the derivative equal to 0. As a result,

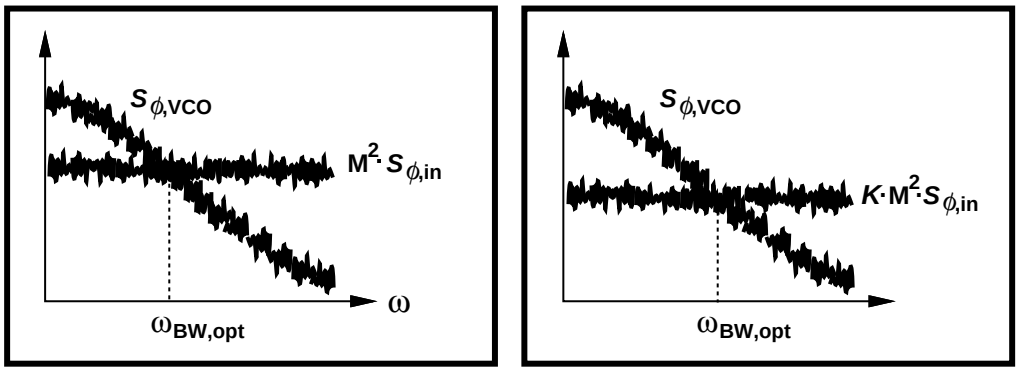
$$K \cdot M^2 \cdot S_{\phi,in} = S_{\phi,VCO}(\omega_0)\frac{\omega_0^2}{\omega_{BW,opt}^2}, \tag{8.26}$$

where

$$K = \frac{3\pi}{6 + 32\zeta^4}. \tag{8.27}$$

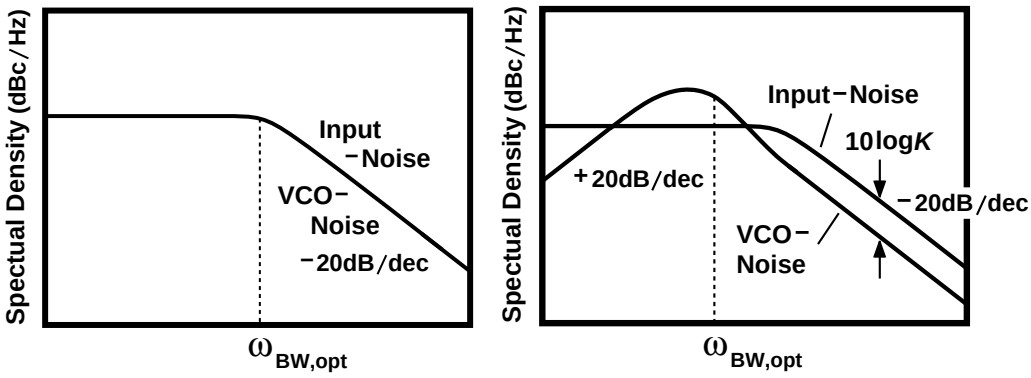
The $\omega_{BW,opt}$ now becomes the intersection of $K \cdot M^2 \cdot S_{\phi,in}$ and $S_{\phi,VCO}$. For $\zeta = 1$, $K \approx 0.25$. Figure 8.17(b) depicts the results. It can be shown that, under optimal loop bandwidth, the VCO and the input noise spectrums roll off at the same rate (-20 dB/dec) in the out-of-band region, but the former falls below the latter by $10 \log_{10} K$. For $K = 0.25$, it is a 6-dB gap. Figure 8.18 depicts

the simulated phase noise contribution under optimal loop bandwidth. The input noise spectrum is assumed flat in our previous discussion. In reality, it may not be true since the spectrum of reference clock (usually a crystal oscillator) is not white. To accurately model the noise/jitter performance, $S_{\phi,in}$ and $S_{\phi,VCO}$ need to be trimmed by measurement results before applying into calculation/simulation.



(a) (b)

Fig. 8.17 Optimal loop bandwidth for (a) overdamped, (b) critical- and under-damped PLLs.



(a) (b)

Fig. 8.18 Spectral density for (a) overdamped ($\zeta = 5$), (b) critical-damped ($\zeta = 1$) PLLs under optimal.

8.3 FRACTIONAL-N PLLS

Although widely used, integer-N PLLs apparently suffer from some issues. The loop bandwidth must be much smaller than the reference, which is usually equal to channel width ($\sim$ 1 MHz) in wireless communication. It makes VCO noise suppression very difficult. The finite frequency resolution also limits the application. Spread-spectrum clocks and FMCW radars necessitate (approximately) continuous tuning in frequency. The large divide ratio is another potential difficulty to improve performance. Consider that a VCO would be corrected only once every thousands of cycles. It is hard to imagine how good the phase noise would be.

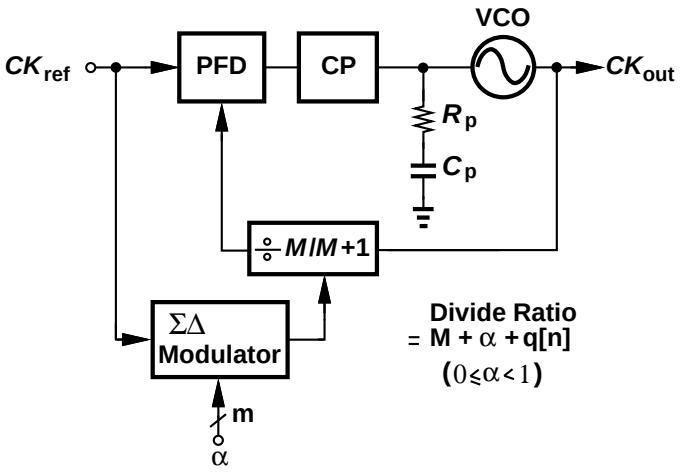


Fig. 8.19 Fractional-N PLL strcture.

As a result, fractional-N PLLs are created. It's divide modulus is not necessary an integer, but can be a fraction. Here, if the divide modulus can be randomly toggle between M (with probability $1-\alpha$) and $M+1$ (with probability α), we arrive at an average divide ratio of $M+\alpha$. Since the ratio is determined by an m-bit (e.g., $m = 16$) binary code, we expect a very fine frequency resolution at output. The best way to generate a randomized bit sequence with a given average value is to use a sigma-delta ($\Sigma\Delta$) modulator. Figure 8.19 illustrates the architecture. Usually driven by the reference clock, the $\Sigma\Delta$ modulator scrambles the output with desired ratio α. The $\Sigma\Delta$ modulator can have output more than one bit, depending on its order. For example, a 3rd-order $\Sigma\Delta$ modulator has output $\{-3, -2, -1, 0, 1, 2, 3, 4\}$, corresponding to divider modulus $\{M-3, M-2, M-1,$

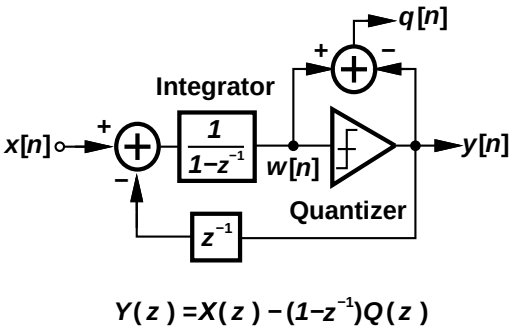
M, $M+1$, $M+2$, $M+3$, $M+4\}$, respectively. The average of this multi-bit output is still equal to α. Since the fraction is obtained by averaging randomized bits, quantization error becomes inevitable. Ideally, the output sequence should be random enough such that quantization error has a uniform distribution between $-1/2$ and $1/2$.

8.3.1 $\Sigma\Delta$ Modulator

Let us look at the operation of a 1st-order $\Sigma\Delta$ modulator. Shown in Fig. 8.20(a) is a typical structure, which can be entirely realized in digital circuits. The difference between input and delayed output are integrated and quantized to form the new output. The quantization error Q is taken from the difference between integrator output W and quantized output Y. The input-output relationship becomes

$$Y(z) = X(z) - (1 - z^{-1})Q(z). \tag{8.28}$$

Here, the signal has transfer function of unity, the quantization noise is shaped by $(1-z^{-1})$. Indeed, the output Y presents a pulse sequence, whose appearance probability of "1" is averagely equal to the input. Figure 8.20(b) reveals two cases of different inputs. While the case for $x = 0.38$



(a)

x	w	y	q
0.38	0	0	0
0.38	0.38	0	0.38
0.38	0.76	1	−0.24
0.38	0.14	0	0.14
0.38	0.52	1	−0.48
0.38	−0.10	0	−0.10
0.38	0.28	0	0.28
0.38	0.66	1	−0.34
0.38	0.04	0	0.04
0.38	0.42	0	0.42
0.38	0.80	1	−0.20

x	w	y	q
0.2	0	0	0
0.2	0.2	0	0.2
0.2	0.4	0	0.4
0.2	0.6	1	−0.4
0.2	−0.2	0	−0.2
0.2	0	0	0

(b)

Fig. 8.20 (a) First order $\Sigma\Delta$ modulator, (b) operation with different inputs.

looks more random, quantization error distribution becomes quite uniform between $[-1/2, 1/2]$, the case for $x = 0.2$ is obviously periodic. If $M = 4$ and $x = \alpha = 0.2$, for example, we arrive

at output frequency $\omega_0 = 4.2\,\omega_{ref}$. One the other hand, the divider provides four divide-by-4 and one divide-by-5 cycle (i.e., 21 VCO cycles). The phase error between reference input (CK_{ref}) and divider output (CK_{div}) occurs periodically, as illustrated in Fig. 8.21. Consequently, the voltage control line V_{ctrl} experiences a ripple. In fact, if α can be simplified to N_1/N_2, where N_1, N_2 are relatively prime, the ripple frequency would be equal to ω_{ref}/N_2. Similar to the effect of clock feedthrough, the periodic ripple also induces spurs.The first harmonic would locate at ω_{ref}/N_2 away from the carrier. Recognized as fractional spurs, this issue could be alleviated by increasing randomizes. The reader can prove that the output for the case of $x = 0.38$ in Fig. 8.20(b) still repeat itself every 50 cycles ($38/100 = 19/50$).

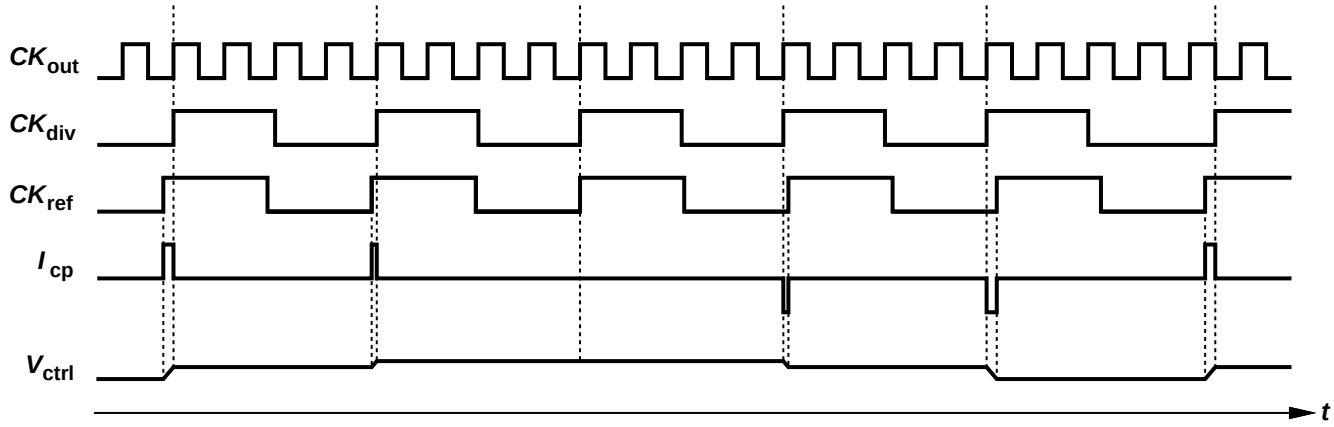
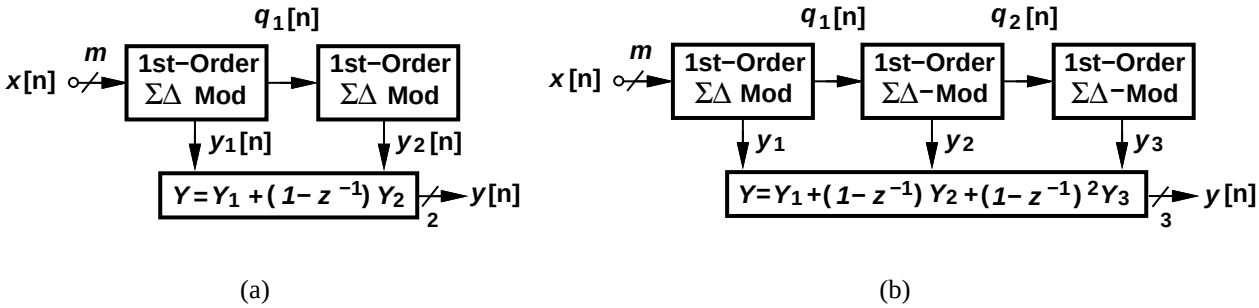


Fig. 8.21 Control line ripple due to periodic output of 1st-order $\Sigma\Delta$ modulator.

How to increase the randomized of the modulator? Or equivalently, how to shift the quantization noise to somewhere so that the overall phase noise is barely affected? The easiest way to do so is to use higher order modulators. Quite a few methods have been developed to implement higher order $\Sigma\Delta$ modulators. The most efficient solution is to cascade lower order sections and wisely adjust the final combination. It allows pipelining of identical blocks, achieving regularity, low power, and high speed realization in CMOS technologies. The all-digital implementation makes it immune from any sort of analog imperfection. The most popular cascade structure is the so called multi-stage-noise-shaping (MASH) topology.



(a) (b)

Fig. 8.22 MASH $\Sigma\Delta$ modulator : (a) 2nd-order, (b) 3rd-order.

Figure 8.22 illustrates the 2nd- and 3rd-order modulator, where the quantization error of each stage is applied to the next stage as input. The reader can prove the transfer function becomes

$$Y(z) = X(z) - (1 - z^{-1})^2 Q_2(z) \quad \text{(2nd-order)} \tag{8.29}$$

and

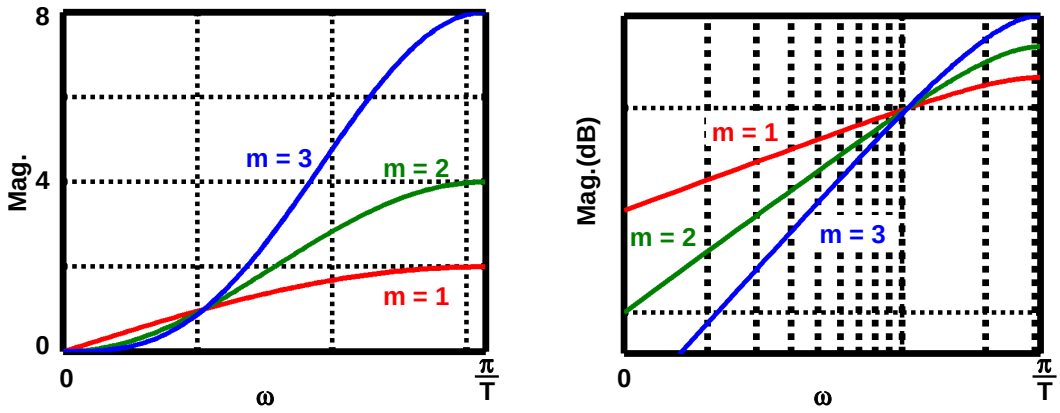
$$Y(z) = X(z) - (1 - z^{-1})^3 Q_3(z), \quad \text{(3rd-order)} \tag{8.30}$$

respectively. Name as "MASH 1-1" and "MASH 1-1-1", there higher order modulators have multiple bits of output as well. In general, the noise transfer function (NTF) for m-th order is equal to

$$\begin{aligned} \text{NTF} &= |1 - z^{-1}|^m = |1 - e^{-j\omega T}|^m \\ &= \left(2 \sin \frac{\omega T}{2}\right)^m . \end{aligned} \tag{8.31}$$

T denotes the period of its driving clock (usually the reference). Figure 8.23 reveals the magnitude. The m-th order NTF reaches a maximum of 2^m at $\omega = \pi/T$. The high-pass characteristic tends to "shift" quantization noise to out-of-band region. Note that $\sin\theta \approx \theta$ as $\theta \to 0$,[2] as a m-th order NTF has a slope of $+10 \cdot m$ dB/dec at low frequencies if log scales are used. The 2nd- and 3rd-order structures are known as "MASH 1-1" and "MASH 1-1-1", respectively. It is instructive to see the output in time domain (Fig. 8.24). For the same input of 0.2, the higher-order modulators blend the output bits more completely. We expect to see the fractional tones spread out and the noise

[2]Actually, $\sin\theta \approx \theta$ holds until $\theta \approx 30^\circ$, where the error is still less than 5%.



(a) (b)

Fig. 8.23 Noise transfer function for different orders in (a) linear, (b) log scale.

shifted to higher frequencies. In a proper design, the low-pass transfer function of PLL suppresses the quantization noise to an insignificant level.

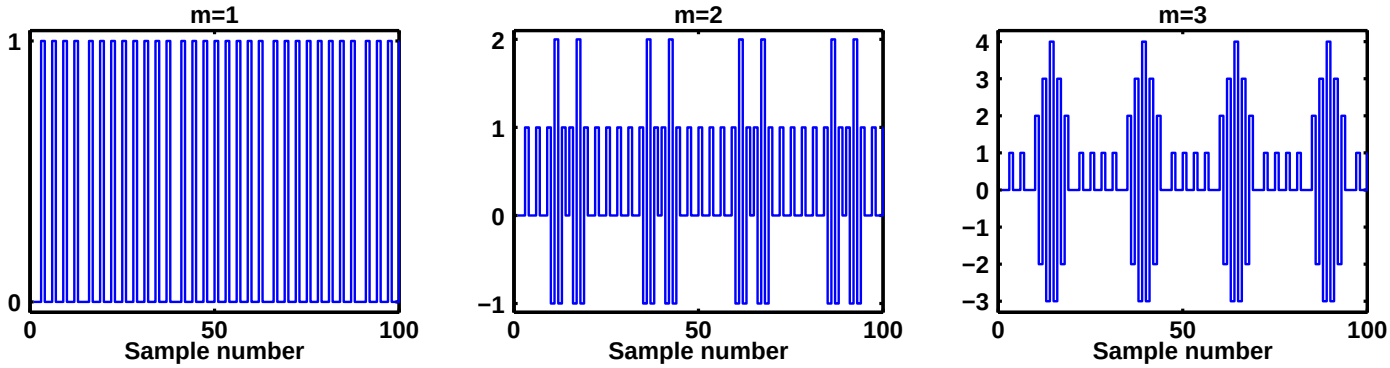


Fig. 8.24 Time-domain behavior of MASH $\Sigma\Delta$ modulator with differnt orders ($\alpha = 0.2$).

8.3.2 Noise Calculation

Let us calculate the noise contribution here. As illustrated in the first-order $\Sigma\Delta$ modulator, the quantization error in MASH structure can be approximately modeled as a uniformly distributed probability density function. It is equivalent to a pulse sequence with period T and random magnitude uniformly distributed from $-1/2$ to $1/2$. Of course the mean of the pulse magnitude is equal

to 0. By definition, the spectrum of this quantization error is given by

$$S_Q(\omega) = \frac{\sigma^2}{T}|p(\omega)|^2, \tag{8.32}$$

where σ^2 denotes the variance of magnitude, and $p(\omega)$ the Fourier transform of $q(t)$. with a uniform distribution, $\sigma^2 = 1/12$ and $|p(\omega)| = \sin(\omega T/2)/(\omega T/2)$ and

$$S_Q(\omega) = \frac{1}{12T}\left|\frac{\sin(\omega T/2)}{(\omega/2)}\right|^2. \tag{8.33}$$

Note that $S_Q(\omega)$ has nothing to do with input resolution m. The sinc function comes from pulse sequence. Recall that output $y[n]$ modulates the divide modulus (which is a direct frequency mod-

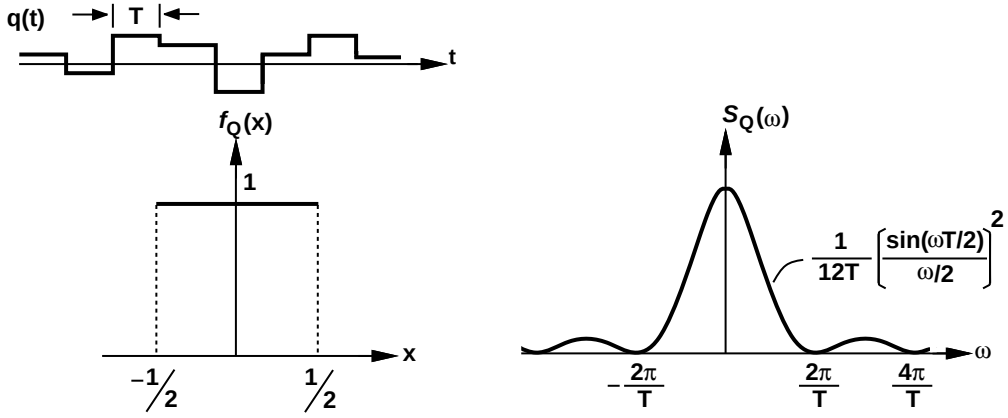


Fig. 8.25 Quantization noise of $\Sigma\Delta$ modulators.

ulation). To convert it to phase, we digitally integrate it as

$$\Phi(z) = \frac{1}{1-z^{-1}}Q(z). \tag{8.34}$$

Here, $\Phi(z)$ represents the z-transform of quantization phase error $\Phi_Q[n]$, appearing in the output of divider chain. Apparently this phase error is presented at one input of PFD. That is, the quantization noise has a transfer function identical to the input noise transfer function ϕ_{out}/ϕ_{in}. Since the power spectral density after digital integration becomes

$$S_\phi|_{@PFDInput} = \left|\frac{1}{1-z^{-1}}\right|^2 S_Q, \tag{8.35}$$

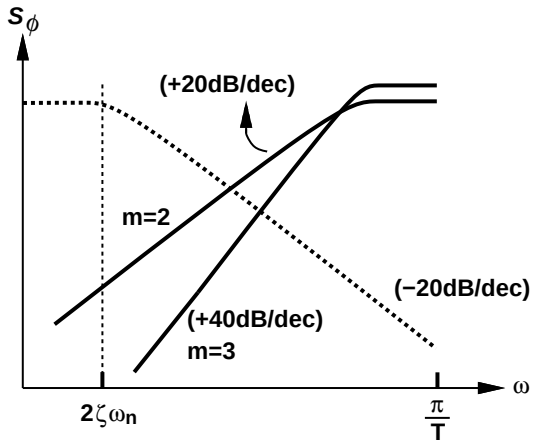


Fig. 8.26 Quantization noise spectrum calculation.

We arrive at the phase noise due to quantization at the output as

$$S_{\phi,\Sigma\Delta} = \frac{1}{12T}\left[\frac{\sin(\omega T/2)}{(\omega/2)}\right]^2 \cdot \left|\frac{1}{1-z^{-1}}\right|^2 \cdot |1-z^{-1}|^{2m} \cdot \left|\frac{\phi_{out}}{\phi_{in}}\right|^2 . \tag{8.36}$$

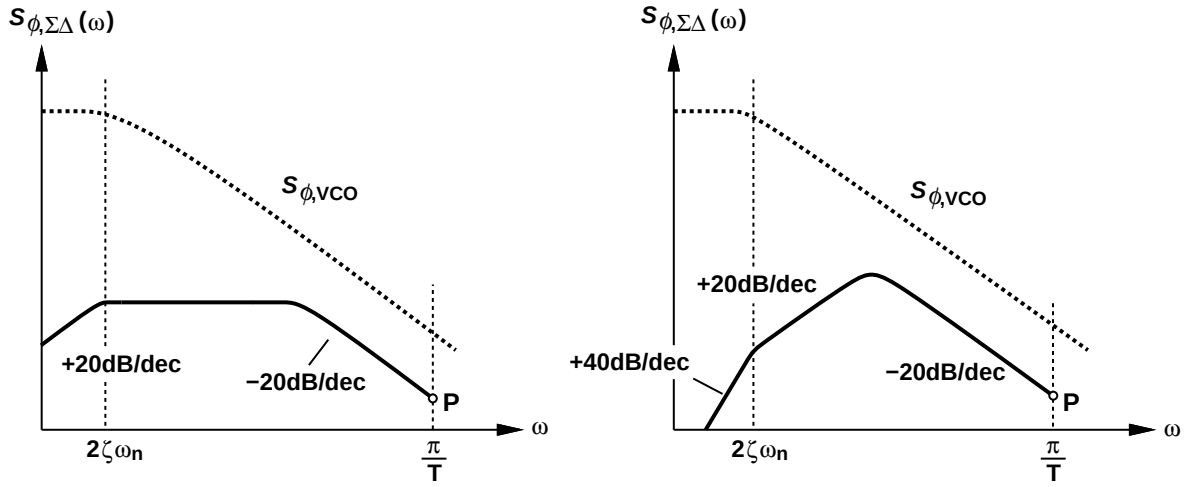
Be aware that $S_Y = |H|^2 \cdot S_X$ hold for both continuous and discrete modes. If the PLL structure of Fig. 8.19 is used, we have

$$S_{\phi,\Sigma\Delta} = \underbrace{\frac{1}{12T}\left[\frac{\sin(\omega T/2)}{(\omega/2)}\right]^2}_{①} \cdot \underbrace{\left(2\sin\frac{\omega T}{2}\right)^{2(m-1)}}_{②} \cdot \underbrace{\frac{4\zeta^2\omega_n^2}{\omega^2+4\zeta^2\omega_n^2}(M+\alpha)^2}_{③} . \tag{8.37}$$

Note that T denotes the period of operation clock (usually the reference clock). While the term ① stands for the quantization noise spectrum, term ② and term ③ reveals the effects of noise shaping and loop suppression. In our discussion, we focus on the range of $0 \le \omega \le \pi/T$, i.e., from dc to $f_{ref}/2$. It is indeed true because the loop bandwidth of a fractional-N PLL is still much less than its reference frequency. Spectrum beyond $f_{ref}/2$ contribute insignificant amount of rms jitter. In the region of $0 \le \omega \le \pi/T$, the term ① decreases slowly from $T/12$ to $T/\left(3\pi^2\right)$, i.e., 3.9 dB. such a gradual move has little influence on the overall spectrum shape, as compare with term ② and ③. It can be considered a flat response for simplicity.

How does the output spectrum of quantization noise $S_{\phi,\Sigma\Delta}$ look like? For $m = 2$, the term ② presents $+20$ dB/dec slope, which cancels the rolling-off rate of term ③. The overall $\Sigma\Delta$ spectrum

reveals a flat response in middle band. For $m = 3$, on the other hand, the $+40$ dB/dec slope of term ② interacting with term ③ yields a hill-shaped response. Figure 8.27 illustrates the plots.



(a) (b)

Fig. 8.27 Output noise spectrum of $\Sigma\Delta$ modulator : (a) 2nd-order, (b) 3rd-order.

How to properly design a fractional-N PLL after all? Since $\Sigma\Delta$ modulation has no influence on the input and VCO noise, we need to minimize $\Sigma\Delta$ modulation noise under the original PLL setting (e.g., optimal loop bandwidth). The rule of thumb is that quantization noise must be insignificant as compare with the other two noise $S_{\phi,in}$ and $S_{\phi,VCO}$. By doing so we could enjoy the benefit of fractional-N PLL without paying too much cost. Owing to the zeros of term ②, $S_{\phi,\Sigma\Delta}$ is quite low at low frequencies. However, it might possibly exceed $S_{\phi,in}$ or $S_{\phi,VCO}$ at high frequencies if it is not designed properly. Recall that fractional-N structures are most likely adopted in association with critical-damping applications (e.g., spread spectrum clocks). According to our discussion in 8.2, $S_{\phi,VCO}$ falls below $S_{\phi,in}$ in the out-of-band region. Thus, we need to ensure $S_{\phi,\Sigma\Delta}$ is well below $S_{\phi,VCO}$ at high frequencies. As $S_{\phi,\Sigma\Delta}$ ramps down at a rate of -20 dB/dec eventually, the point at $\omega = \pi/T$ is obviously the closest point to $S_{\phi,VCO}$. In other words, we have

$$S_{\phi,\Sigma\Delta}\left(\omega = \frac{\pi}{T}\right) < S_{\phi,VCO}\left(\omega = \frac{\pi}{T}\right). \tag{8.38}$$

as a criteria to judge the whether $\Sigma\Delta$ noise is low enough. Unfortunately, this condition is quite difficult to achieve.

Example 8.2

Consider a fractional-N PLL as shown in Fig. 8.19, where $M = 16$, $f_{ref} = 200$ MHz, $\alpha = 0.2$, $\zeta = 1$, and $2\zeta\omega_n = 2\pi \times 1$ MHz. Determine the phase noise contributed by $\Sigma\Delta$ modulator at $f_{ref}/2$.

Solution:

Since $T = 5$ ns and $f_{ref}/2$ is far away from 1 MHz, we have

$$S_{\phi,\Sigma\Delta}(\omega = \frac{\pi}{T}) \cong \frac{1}{12T} \cdot \frac{4T^2}{\pi^2} \cdot 2^{2(m-1)} \cdot (16.2)^2 \cdot \frac{4\zeta^2\omega_n^2}{(\pi/T)^2}$$

$$= \begin{cases} -101.5(dBc/Hz), \ for\ m = 3 \\ -107.5(dBc/Hz), \ for\ m = 2. \end{cases}$$

For a 3.2 GHz VCO, typical phase noise is well below -140 dBc/Hz at 100 MHz offset.

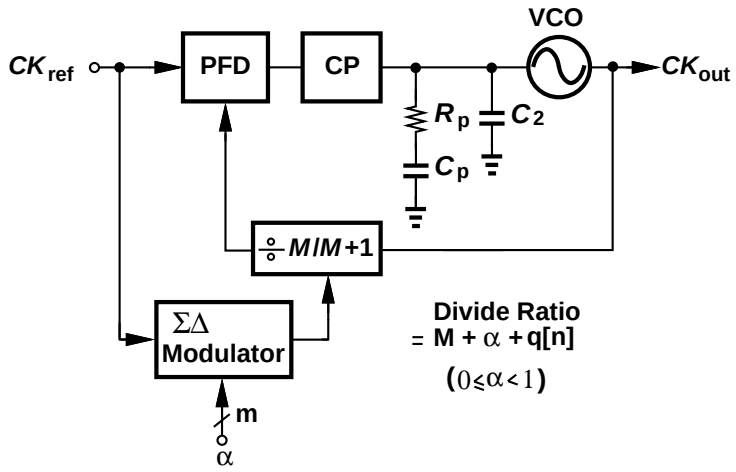


Fig. 8.28 Fractional-N PLL wuth 2nd-order loop filter.

The overwhelming quantization noise requires the PLL to perform higher-order suppression in the out-of-band region. Consider the same fractional-N PLL as shown in Fig. 8.19 but the 1st-order loop filter is replaced by 2nd-order one (Fig. 8.28). The closed loop transfer function (from

input to output) of this 3rd-order PLL is equal to

$$\frac{\phi_{out}}{\phi_{in}} = \frac{\frac{K_{VCO}I_p}{2\pi C_p}(sC_pR_p+1)}{R_pC_2s^3 + \frac{(C_p+C_2)}{C_p}s^2 + \frac{K_{VCO}I_pR_p}{2\pi M}s + \frac{K_{VCO}I_p}{2\pi MC_p}}, \tag{8.39}$$

where an additional pole has been introduced. For simplicity we neglect α and assume divide modulus is M. It is quite difficult to analyze the poles with exact solutions. However, if we let $C_p \gg C_2$ and $\omega_{p3} \triangleq (R_pC_2)^{-1}$, the above transfer function can be approximated as

$$\frac{\phi_{out}}{\phi_{in}} \approx \frac{M\cdot(2\zeta\omega_n s+\omega_n^2)}{\left(1+\frac{s}{\omega_{p3}}\right)(s^2+2\zeta\omega_n s)+\omega_n^2}. \tag{8.40}$$

In the vicinity of $\zeta \approx 1$. Note that parameters of 2nd-order PLL (i.e., ζ and ω_n) are preserved to facilitate the analysis. Now the transfer function has one zero $[\omega_z = (R_pC_p)^{-1}]$ and three poles.

Example 8.3

Show that the above approximation is valid with $\zeta = 1$, $C_p = 40C_2$.

Solution:

The poles and zeros become

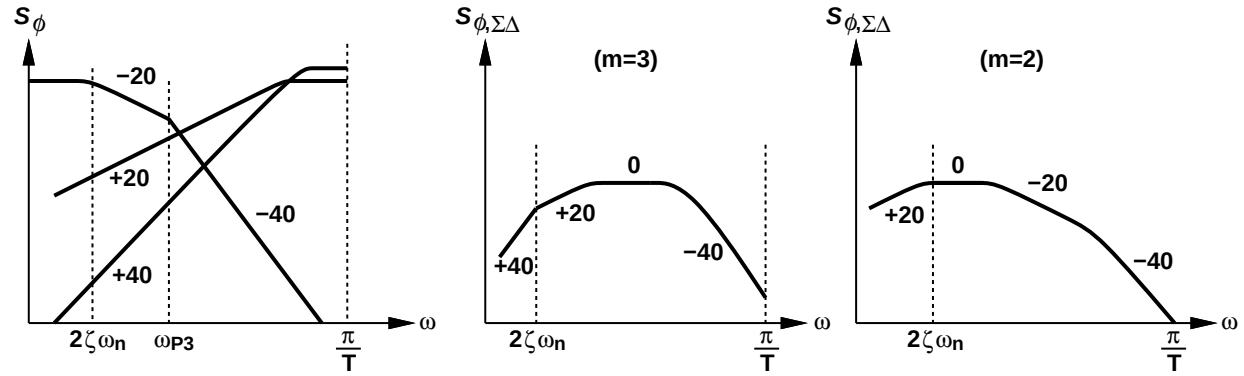
$$\omega_z = \frac{1}{R_PC_P}$$
$$\omega_{p1} = \omega_{p2} = \omega_n = \frac{2}{R_PC_P}$$
$$\omega_{p3} = \frac{40}{R_PC_P}$$

and its loop bandwidth $\approx 2\zeta\omega = 4/(R_PC_P)$. To make (8.39) and (8.40) approximately equal, the following expression must hold :

$$1+2\zeta\omega\cdot\frac{1}{\omega_{p3}} \approx \frac{C_P+C_2}{C_P}$$
$$2\zeta\omega+\frac{\omega_n^2}{\omega_{p3}} \approx \frac{I_PK_{VCO}R_P}{2\pi M}.$$

Both are indeed true in our setup.

In practical design, $C_p \geq 10C_2$ is a reasonable choice to maintain loop stability and out-of-band noise suppression simultaneously. For smaller ratio of C_p/C_2, the approximation of (8.39) may deviate from (8.40) but has little impact on overall analysis.



(a) (b)

Fig. 8.29 $\Sigma\Delta$ modulator in 3rd-order PLLs. Slopes (in dB/dec) of segments are marked.

With the help of the 3rd pole, the quantization phase noise can be further reduced at high offset frequencies. As can be seen in Fig. 8.29, the transfer function now presents an intermediate point ω_{p3} to bend itself from -20 dB/dec to -40 dB/dec. As a result, we obtain $S_{\phi,\Sigma\Delta}$ for $m = 2$ and $m = 3$ [Fig. 8.29(b)]. The reader can prove that

$$S_{\phi,\Sigma\Delta}(\omega = \frac{\pi}{T}) \cong \frac{T^3}{3} \cdot \frac{1}{\pi^2} \cdot 2^{2(m-1)} \cdot \frac{4\zeta^2\omega_n^2}{\pi^2} \cdot (M+\alpha)^2 \frac{1}{\left(\frac{\pi}{T} \cdot \frac{1}{\omega_{p3}}\right)^2}. \tag{8.41}$$

With the same parameters in example 8.3 and $\omega_{p3} = 2\pi \times 2.5$ MHz, $S_{\phi,\Sigma\Delta}$ at $f_{ref}/2$ is equal to -133.5 and -139.5 dBc/Hz for $m = 3$ and $m = 2$, respectively.

8.4 INJECTION LOCKING PLLS

8.4.1 Limitation of ω_{BW} Optimization

From our discussion in 8.2, it is well-known that the input noise (including the noise from the reference and the phase and frequency detector) and the VCO noise are shaped by a low-pass and a high-pass transfer function, respectively, when they are presented at the output. Generally

speaking, an optimal noise performance can be obtained by properly selecting the loop bandwidth. If the input noise is assumed flat (which is not exactly true in reality), the optimal bandwidth of the loop can be chosen as the intersection of VCO phase noise and N^2-times input noise.

The above approach, however, suffers from an intrinsic limitation. As the VCO frequency increases, its noise begins to dominate and becomes more difficult to suppress. To quantify this issue, let us consider two similar PLLs (with different VCOs inside), running at two frequencies ω_{c1} and ω_{c2}, respectively. Assuming $\omega_{c2}/\omega_{c1} = N$ and identical quality factor Q for the resonators, we recognize that the two VCO phase noise lines are vertically separated by $20\log_{10} N$ dB [10]. As shown in Fig. 8.30, we also assume the two loop bandwidths are ω_{BW1} and ω_{BW2}, and the corresponding VCO phase noise at these points are $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively. Now let's neglect the

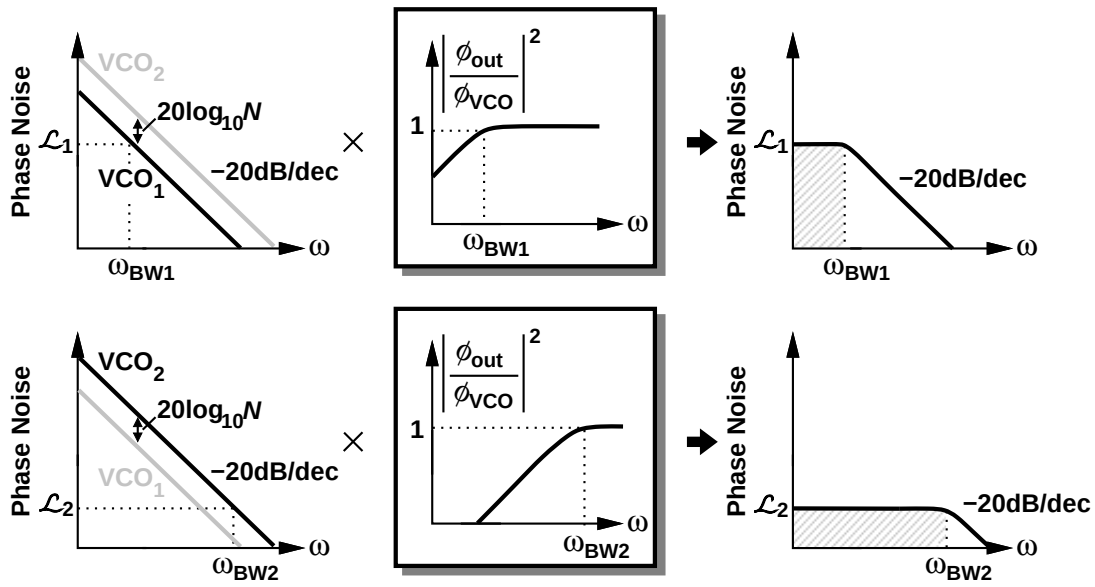


Fig. 8.30 VCO phase noise shaping with different loop bandwidths.

input noise and consider the phase noise contributed by VCO only. The PLL output spectrum can be readily available through multiplying the VCO phase noise by the high-pass transfer function $|\phi_{out}/\phi_{VCO}|^2$. That is, the output phase noise remains flat as $\mathcal{L}_1(\mathcal{L}_2)$ until $\omega_{BW1}(\omega_{BW2})$, and rolls off at a rate of -20 dB/dec beyond the loop bandwidth. On the other hand, the rms jitter is given

by integrating the phase noise [11], [12]

$$J_{rms}^2 = \left(\frac{1}{2\pi f_c}\right)^2 \cdot 2 \cdot \int_0^\infty S_\phi(f)df = \left(\frac{1}{2\pi f_c}\right)^2 \cdot 2 \cdot \int_0^\infty 10^{\frac{\mathcal{L}(f)}{10}} df \tag{8.42}$$

which can also be normalized to one clock period

$$J_{rms,nor}^2 = 2\int_0^\infty 10^{\frac{\mathcal{L}(f)}{10}} df \ \ (\text{rad}). \tag{8.43}$$

Now, if the two PLLs in Fig. 30 are designed to present the same jitter performance (i.e., identical normalized jitter), we must have

$$10^{\frac{\mathcal{L}_1}{10}} \cdot \omega_{BW1} = 10^{\frac{\mathcal{L}_2}{10}} \cdot \omega_{BW2}. \tag{8.44}$$

Here, only the in-band noise (the shadow area) is considered for simplicity. We also assume the loop damping factors are so high that the transfer curve can be modeled as a first-order function. Since $\mathcal{L}_1 = \mathcal{L}_2 + 20\log_{10}(\omega_{BW2}/\omega_{BW1}) - 20\log_{10} N$, we obtain

$$\frac{\omega_{BW2}}{\omega_{BW1}} = N^2 \tag{8.45}$$

from (8.44). That is, if we migrate from one standard to another that operates at a frequency N-times higher, the loop bandwidth needs to be raised up by a factor of N^2 in order to maintain the same VCO noise contribution. This requirement is difficult to achieve because 1) some standards pre-define the bandwidths mandatorily; 2) even with no restriction posed, the loop bandwidth still needs to be kept below approximately one twentieth of the reference frequency in order to ensure stability [13]; 3) a high loop bandwidth allows more noise from the phase and frequency detector (PFD) and the charge pump (CP) to come into the output. Nonetheless, at high frequencies, it gets more and more difficult to reduce the noise (or equivalently, jitter) solely by adjusting the loop bandwidth. We must resort to other techniques. At high frequencies, injection locking is believed to be the most powerful tool that suppresses phase noise jitter.

8.4.2 Noise-Shaping Phenomenon

Consider a free-running VCO undergoes a fundamental injection (Fig. 8.31(a)). It has been demonstrated that the VCO phase noise could be reduced to the same level of the injection signal's spectrum, given that the injection signal CK_{inj} comes from a low noise source. It is not difficult to

explain it as the VCO's oscillation has been corrected in each cycle. Since no phase noise is accumulated, the output is as clean as the injection signal. Interestingly, the same idea can be applied to subharmonic injection, if the injection occurs "frequently" enough. Shown in Fig. 8.31(b) is

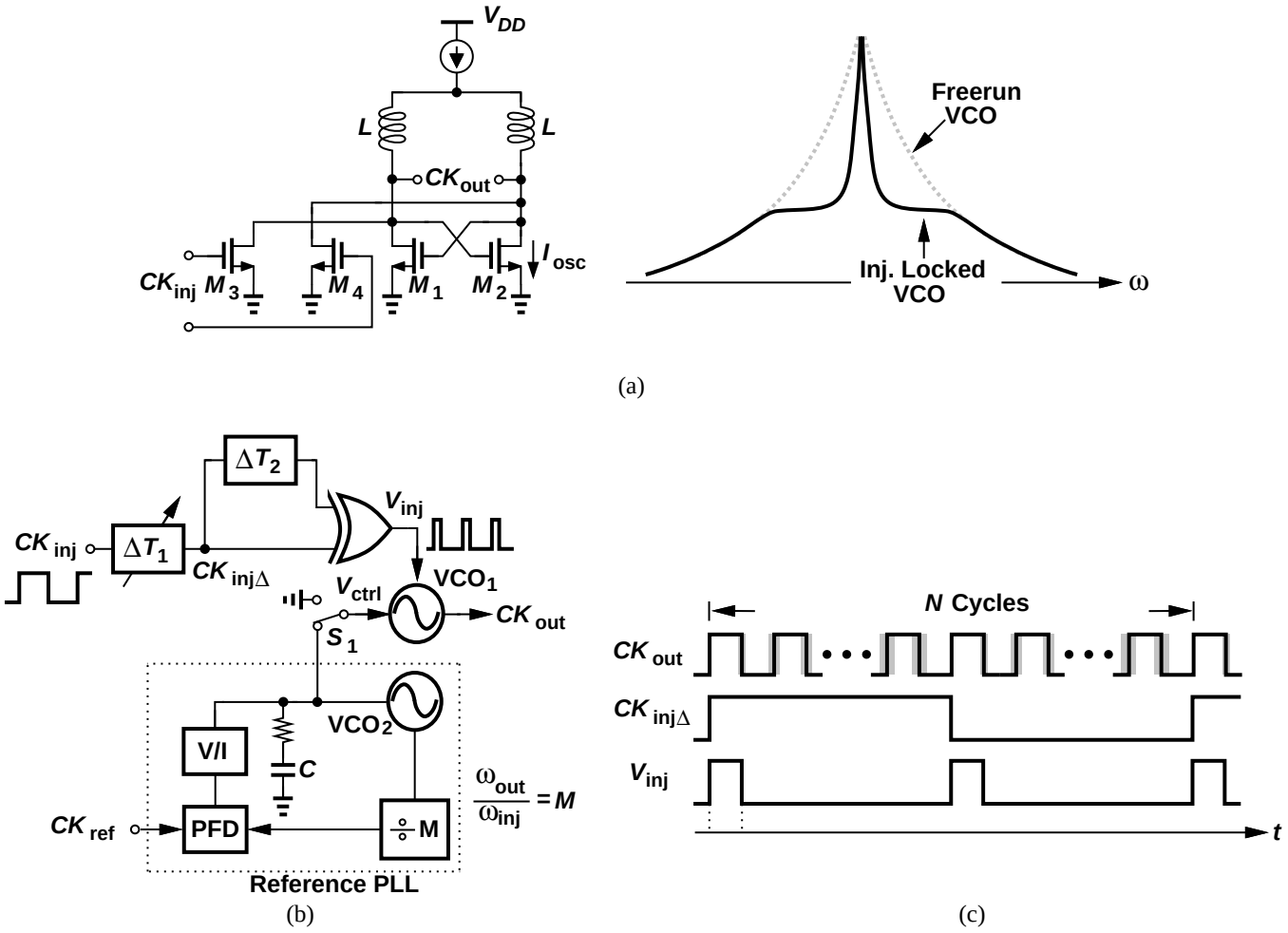


Fig. 8.31 (a) Fundamentally-injection-locked VCO (b) subharmonic-injection locked PLL, (c) the timing diagram.

a subharmonically-injection locked PLL. In addition to the normal phase locking, the VCO_1 is also injection-locked to the edges of an independent source CK_{inj}. Here we apply a subrate signal CK_{inj} as with different frequencies ($\omega_{inj} = \omega_{out}/N$) to investigate the properties of injection-locked PLLs. A constant delay ΔT_2 and an XOR gate are employed to generate pulses (V_{inj}) on

occurrence of CK_{inj} transitions, leading to a double-edge injection periodically appearing every $N/2$ cycles [Fig. 8.31(c)].

Let us first consider a typical phase-locked loop as shown in Fig. 8.32(a). It is well-known that the in-band phase noise of it ($\mathcal{L}_{PLL}$) is shaped from the free-running line of the VCO to a relatively flat response at moderate offset frequencies, and the turning point is roughly given by the loop bandwidth ω_{BW}. If the oscillator is under fundamental injection locking, it can be shown [14], [15] that the phase noise within the lock range ω_L will be suppressed to that of the injection signal. It is thus deducible that for a subharmonic locking with a frequency ratio N, the phase noise inside the lock range ω_L would be constrained to $\mathcal{L}_{inj} + 20\log_{10} N$, where $\mathcal{L}_{inj}$ denotes the phase noise of the subrate injection signal CK_{inj}. Fig. 8.32(b) illustrates this phenomenon. Certainly such a noise reduction only occurs when N is an integer. Since usually the lock range of an LC-tank VCO is not only small but sensitive to PVT variations, we must provide a proper control voltage such that the VCO natural frequency can always track the desired multiple of the injection frequency ω_{inj}. This task is accomplished by combining the injection locking technique with a PLL, as shown in Fig. 8.32(c) and (d). Here, we have two situations: if $\omega_L > \omega_{BW}$, the whole in-band noise is drawn down to $\mathcal{L}_{inj} + 20\log_{10} N$ (dB), leading to a significant jitter reduction [Fig. 8.32(c)]. With the help of the PLL, the noise suppression can always be maintained around the optimal position. If $\omega_L < \omega_{BW}$, on the contrary, the noise shaping becomes less effective because the turning point ω_{BW} is not covered within the range of suppression [Fig. 8.32(d)]. It is intuitive that the spectrum degenerates to that of an ordinary PLL if $\omega_L \ll \omega_{BW}$. Fortunately, in most cases, $\omega_L > \omega_{BW}$. Note that the noise suppression technique could never be practical for a standalone injection-locked oscillator without frequency-tracking PLL [e.g., Fig. 8.32(b)] because the PVT variations would cause substantial performance degradation or simply fail the locking.

The case in Fig. 8.32(c) is somewhat over-simplified because the $\mathcal{L}_{inj} + 20\log_{10} N$ and $\mathcal{L}_{PLL}$ lines need not intersect at ω_L. The former may be higher than the latter by a few dB at ω_L in reality. On the other hand, it is obvious that the phase noise would tightly follow $\mathcal{L}_{PLL}$ for the offset frequencies higher than ω_{inj}, since the subharmonic injection has little influence on it. Between ω_L and ω_{inj}, the spectrum deviates from the governance of $\mathcal{L}_{inj}$ and approaches $\mathcal{L}_{PLL}$ with a gradual

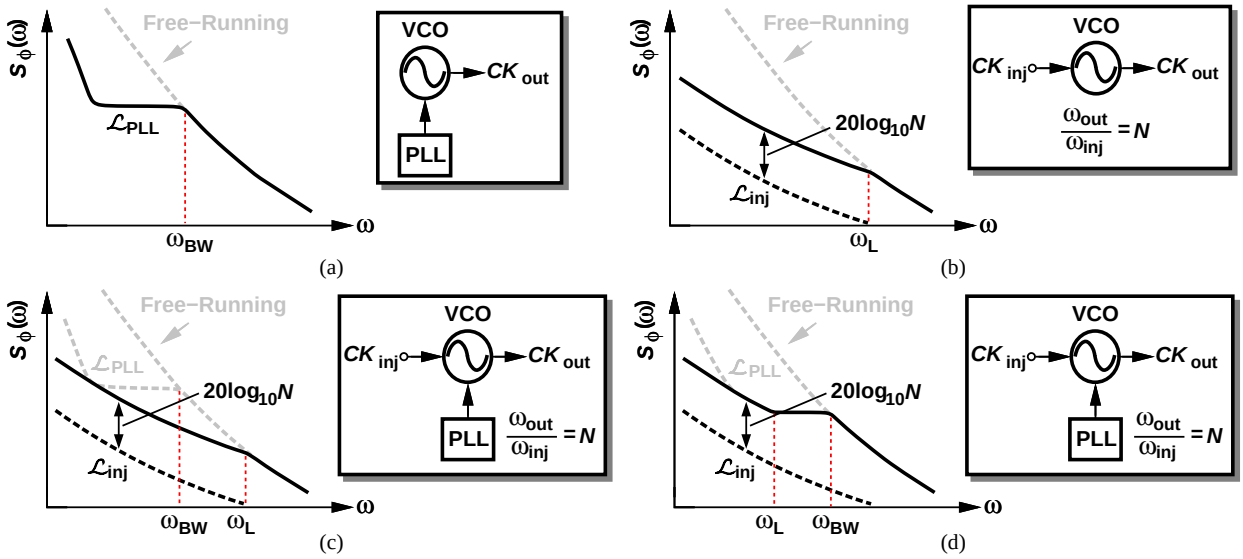


Fig. 8.32 Illustration of subharmonic locking: (a) typical PLL, (b) subharmonically injection-locked VCO, (c) subharmonically injection-locked PLL with $\omega_{BW} < \omega_L$ (d) subharmonically injection-locked PLL with $\omega_{BW} > \omega_L$.

and smooth transition. It should not be surprising because the influence from injection locking fades out as the offset frequency goes up.As a result, we model the phase noise in this region as a straight line (in log scale), as illustrated in Fig. 8.33. Overall speaking, the phase noise of a subharmonically injection-locked PLL is

$$\mathcal{L}(\omega) = \begin{cases} \mathcal{L}_{inj}(\omega) + 20\log_{10} N, & \text{for } \omega \le \omega_{\rm L} \ \ (\text{Region I}) \\ \mathcal{L}_{PLL}(\omega_{inj})\dfrac{\log_{10}(\omega/\omega_L)}{\log_{10}(\omega_{inj}/\omega_L)} & \\ \quad +[\mathcal{L}_{inj}(\omega_L) + 20\log_{10} N]\dfrac{\log_{10}(\omega_{inj}/\omega)}{\log_{10}(\omega_{inj}/\omega_L)}, & \text{for } \omega_{\rm L} \le \omega \le \omega_{\rm inj} \ \ (\text{Region II}) \\ \mathcal{L}_{PLL}(\omega), & \text{for } \omega \ge \omega_{\rm inj} \ \ (\text{Region III}). \end{cases} \tag{8.46}$$

The rms jitter is thus readily available through the integration of (8.46).

To further investigate the above analysis, we realize a 20-GHz PLL with subharmonic injection locking and measure the output spectrum. To prove the above analysis, we measure the output spectrum of the circuit in Fig. 8.31(b) for different N and plot the results in Fig. 8.34. Here, the phase noise of CK_{out} (bold line) is shown in company with that of the injection signal $\mathcal{L}_{inj}$.

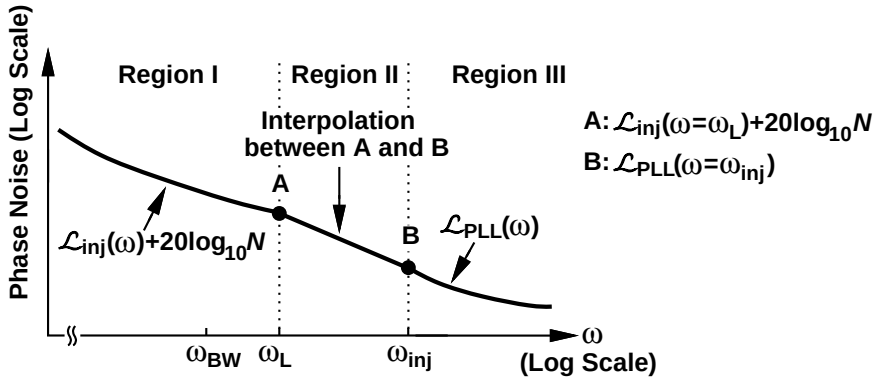


Fig. 8.33 Prediction of phase noise of injection-locked PLL.

The output phase noise without the injection (i.e., $\mathcal{L}_{PLL}$) is also depicted as a reference. It can be shown that the phase noise closely follows the $\mathcal{L}_{inj} + 20\log_{10} N$ line within the lock range and gradually returns back to $\mathcal{L}_{PLL}$ beyond ω_L. Due to limitation of the spectrum analyzer, the phase noise measurement is restricted to 1-GHz offset. Nonetheless, we can still observe the output phase noise merging into $\mathcal{L}_{PLL}$ at around 1 GHz in the case $N = 32$, and see a clear trend for $N = 2$ and $N = 8$. The noise shaping manifests itself for $N \leq 8$, and it gets degraded as N increases. Note that this testing circuit uses double-edge injection. In the cases with single-edge injection, we may further restrict the frequency ratio. As will be shown in Section (xxx), cascading can be applied to solve this issue. For large N (e.g., $N = 128$), the output phase noise degenerates to $\mathcal{L}_{PLL}$ as expected, because the injection appears so sparse that the noise profile is barely affected.

8.4.3 Lock Range

The lock range affects the noise shaping of an injection-locked PLL significantly. It is worth noting that the lock range ω_L degrades as N increases. Actually, if we define the oscillation and injection currents of the LC-tank VCO as I_{osc} and I_{inj}, the lock range of fundamental (full-rate) injection is given by [14], [16]

$$\omega_L = \frac{\omega_{out}}{2Q} \cdot \frac{I_{inj}}{I_{osc}} \cdot \frac{1}{\sqrt{1 - \dfrac{I_{inj}^2}{I_{osc}^2}}}. \tag{8.47}$$

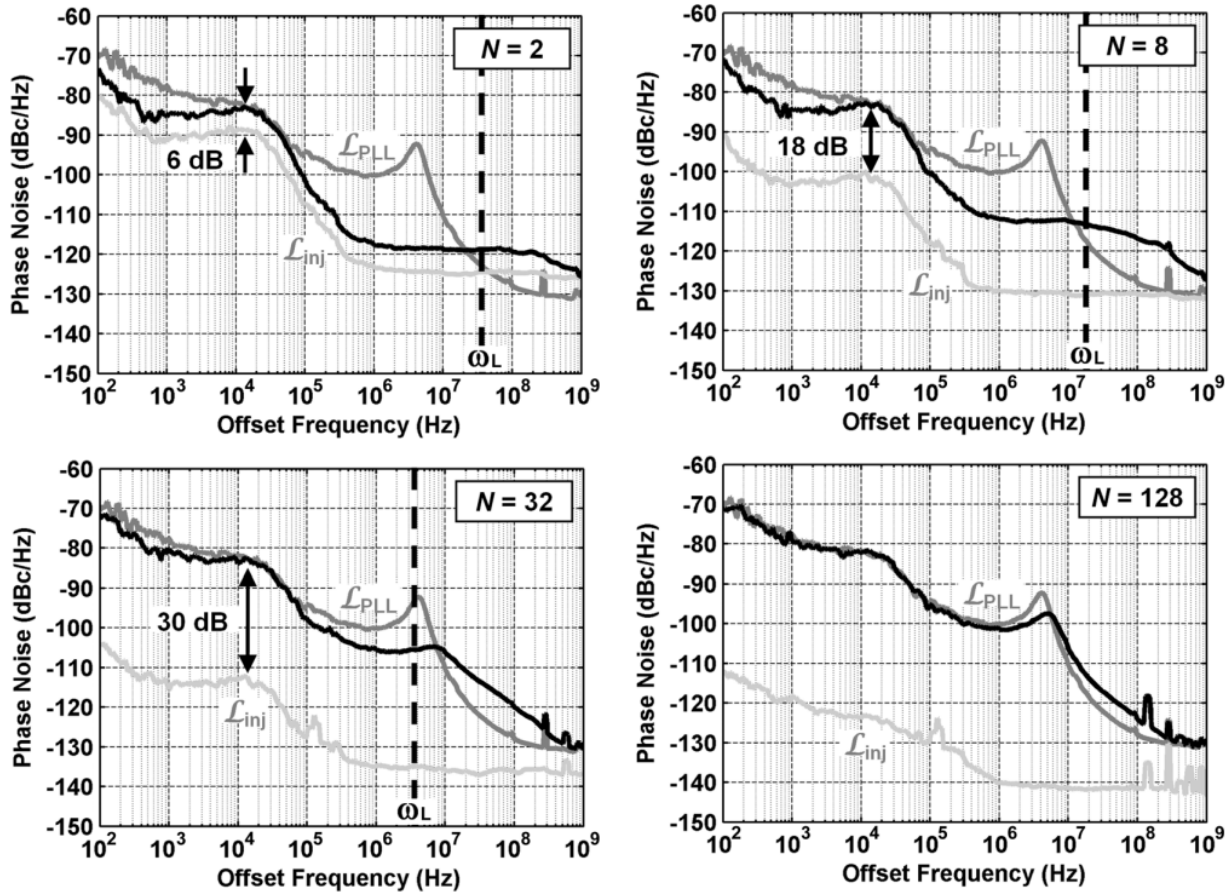


Fig. 8.34 Phase noise for different frequency ratios.

where Q represents the quality factor of the tank. Note that both I_{osc} and I_{inj} come from averaging of large signals. In subharmonic injection, I_{inj} needs to be modified as $I_{inj,eff} = I_{inj}/N$ if the injection occurs once every N cycles. It is because the effective current becomes $1/N$ in magnitude. The lock range therefore becomes

$$\omega_L = \frac{\omega_{out}}{2Q} \cdot \frac{I_{inj}}{I_{osc}} \cdot \frac{1}{N} \cdot \frac{1}{\sqrt{1 - \dfrac{I_{inj}^2}{I_{osc}^2 N^2}}} \approx \frac{\omega_{out}}{2Q} \cdot \frac{I_{inj}}{I_{osc}} \cdot \frac{1}{N}. \tag{8.48}$$

8.4.4 Tolerance to PVT Variations

As demonstrated in the above analysis, the subharmonic-locking PLLs achieve similar in-band phase noise performance as $\omega_L > \omega_{BW}$. It implies that a very stable clock generator can be achieved, given that a clean reference clock is applicable. Fig. 8.35 demonstrates the output spectra under different conditions with and without the subharmonic locking. Here, we change the supply

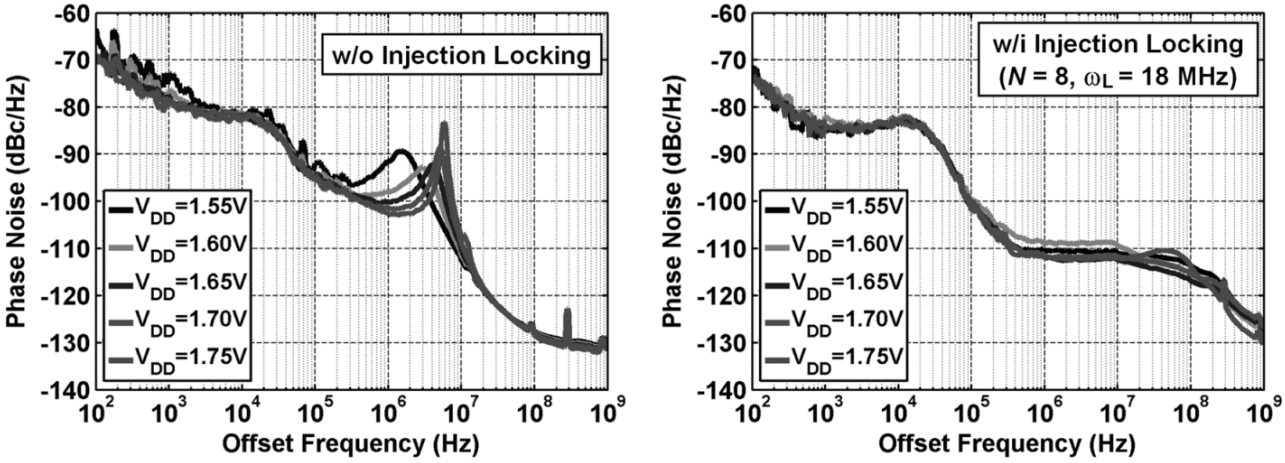


Fig. 8.35 Phase noise with different loop bandwidth.

voltage to create different loop bandwidths for the reference PLL in Fig. 8.31(b). It can be shown that even with a ratio of 8, the noise shaping presents almost identical results for different cases. That is, the PLL can be designed in a more relaxed way since it can tolerate a much wider range for variations. Note that the PVT deviation of ΔT_2 has negligible impact on the overall performance due to the injection locking mechanism. The injection locking technique also rejects the supply noise, if the locking can be maintained throughout the perturbation. To demonstrate this property, we provide a sinusoidal disturbance of 50 mV_{pp} with different frequencies onto the V_{DD} of the testing circuit. Fig. 8.36 shows the noise suppression of two cases. The coupled supply variation has little influence on the overall output phase noise if injection locking is imposed. Measurement suggests that, for $N \leq 8$, supply noise at any frequency below 100 MHz is substantially rejected.

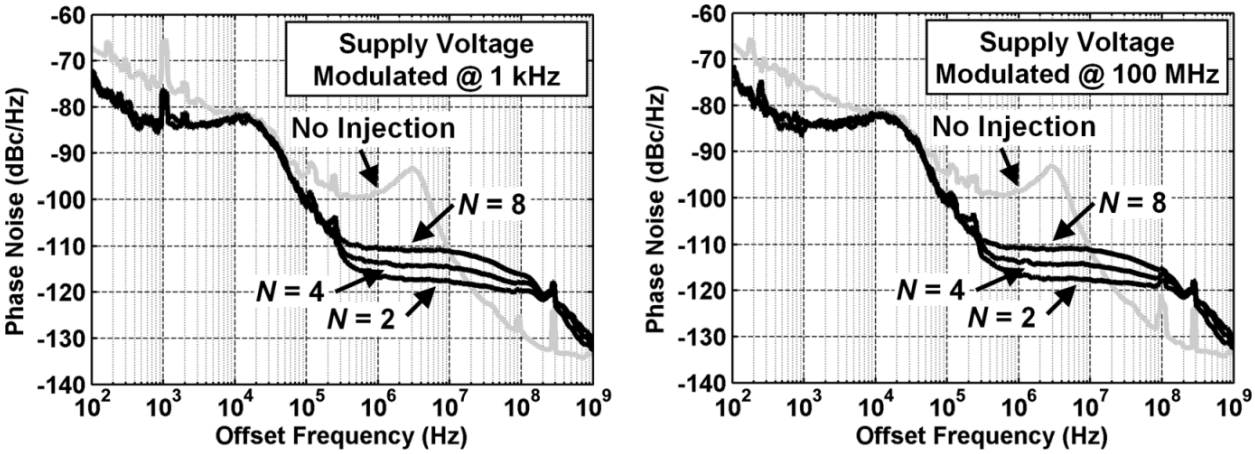


Fig. 8.36 Phase noise with different supply noise.

8.4.5 Locking Behavior

One issue hidden behind the beauty of the injection-locked PLLs is the pulling between the two locking forces, namely, the phase locking (from the reference PLL) and the injection locking (from the injection signal). Let's revisit the circuit in Fig. 8.31(b) again, and assume the injection clock CK_{inj} comes in after the reference PLL has already reached a steady locking. At this moment, the phase of CK_{out} is exclusively determined by the phase of CK_{ref}. As an independent CK_{inj} arrives, finite phase error may exist between CK_{inj} and CK_{ref}, i.e., V_{inj} need not coincide with the already existing CK_{out}. In other words, the two forces "fight" each other and probably pull the output phase. Such a conflict may lead to quite a few uncertainties. Up to this point, quite a few questions arise. How much phase error can it tolerate after all? What happens if the injection signal is totally ($180°$) out of phase with the intrinsic CK_{out}? Does such a destructive injection still suppress the phase noise? Or it simply destroys the loop locking?

To answer these questions, we must go back to the injection locking theories [14], [16], [17]. Surprisingly, if finite phase error exists between the regular phase locking and the injection locking, the LC tank of the VCO would create a shift on resonance frequency to accommodate the non-zero phase difference, even though ω_{out} is exactly a multiple of ω_{inj}. Following the analysis in [14], we redraw the equivalent half circuit of an injection-locked oscillator in Fig. 8.37.

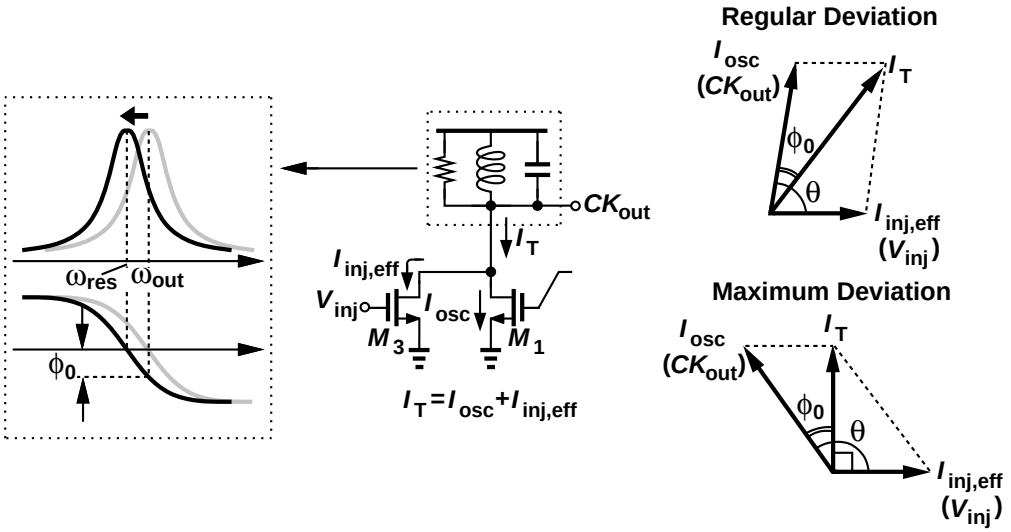


Fig. 8.37 Locking behavior analysis.

Indeed, for a subharmonically injection-locked PLL, the VCO core current I_{osc} (in phase with CK_{out}) and $I_{inj,eff}$ (in phase with V_{inj}) can be separated by an angle θ. Suppose in the absence of injection, the VCO steadily oscillates at ω_{out}. The LC tank would also resonate at ω_{out} without any phase shift. As the injection comes in, however, the resonance frequency will no longer stay in ω_{out}, but shift to some point ω_{res} as illustrated in Fig. 8.37. From the derivation in [14], we realize that the created phase ϕ_0 is the angle between I_{osc} and I_T (the total current driving the tank), and (the angle between V_{inj} and CK_{out}) reaches a maximum as I_T and $I_{inj,eff}$ form a right angle. That is, at steady state, an injection-locked PLL would automatically adjust the phase relationship to maintain the stability and accomplish the noise suppression. The maximum tolerable phase error is therefore given by

$$\theta_{max} = \frac{\pi}{2} + \sin^{-1}\left(\frac{I_{inj,eff}}{I_{osc}}\right). \tag{8.49}$$

In our testing circuit, for example, we set $N = 4$ and $I_{inj,eff} = I_{osc}/4$, obtaining $\theta_{max} = 105°$. That is, the maximum tolerable range for phase offset is about $210°(\pm 105°)$. This effect can be easily verified as follows.Gradually adjusting ΔT_1 in Fig. 8.31(b), we observe the change of the output spectrum. The recorded jitter for different ΔT_1 is shown in Fig. 8.38(a). As expected, the

rms jitter stays low (≈ 360 fs) for approximately $210°$, and goes up dramatically outside the stable region. It fully validates the prediction of (8.49).

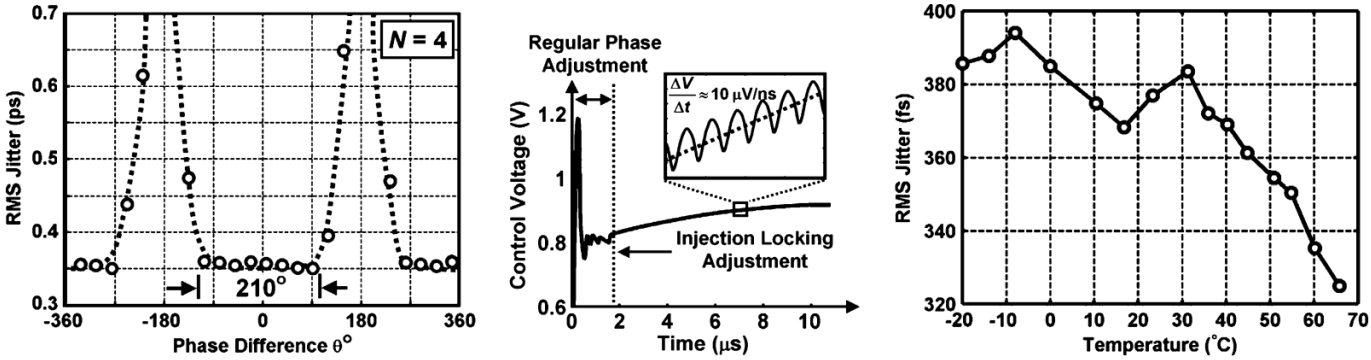


Fig. 8.38 Locking behavior analysis.

It is instructive to investigate the acquisition of locking. In the beginning, the phase difference between the two inputs of the PFD is very large. The reference PLL tries to neutralize this error through the normal phase locking process, regardless of the existence of injection signal. After this "coarse" locking is achieved, the injection then conducts the "fine" phase tuning, i.e., shifting the resonance frequency of the LC tank to create a proper θ. Note that the two PFD inputs are now roughly aligned, so the fine tuning would take a much longer time. It is because the phase difference for the 20-GHz CK_{out} (period $= 50$ ps) is very small with respect to the 312.5-MHz reference (period $= 3.2$ ns) in Fig. 8.31(b), making the available current from the V/I converter very small. In our testing circuit, for example, the maximum pumping current coming from the V/I converter is only 0.78% ($25ps \div 3.2$ ns) as large as its peak value. As a result, the loop presents a settling time at least 100 times longer than a regular PLL. Fig. 8.38(b) plots the simulated locking behavior. It can be clearly shown that the fine phase adjustment for injection locking draws a long tail (≈ 10 μs). Note that in many applications that require no frequency hopping, the long settling time is not a concern. The above analysis implies that a proper delay ΔT_1 must be maintained over the PVT variations. One would think of placing another delay-locked loop (DLL) around ΔT_1 to do so. However, such a solution is plausive because (1) judging from Fig. 8.38(a), the jitter performance is very constant within the tolerable range of $210°$; (2) adding another DLL may

induce more noise and consume more power and area, let along the possible instability issue. To evaluate the robustness of the loop, we apply a fixed ΔT_1 in Fig. 8.31(b) and measure the rms jitter under different conditions. As depicted in Fig. 8.38(c), for a temperature variation from $-20°C \sim 65°C$, the rms jitter deviates no more than 69 fs. Thus, a simple fixed delay (at most with manual tuning capability) is well sufficient in most applications.

8.4.6 Pseudo Locking Phenomenon

What happens if the desired θ exceeds θ_{max}? Imagine a fully destructive case as shown in Fig. 8.39(a), where the positive pulse V_{inj} aligns with the valley of CK_{out}. In such a case, the required θ is $180°$. From (8.49), we realize that the only possible way to sustain the loop stability is to set $I_{inj,eff} = I_{osc}$, which is difficult to achieve in sub-rate injection. As a result, the loop could never find a solution to satisfy the phase relationship, and the resonance frequency of the VCO would wander back and forth across the lock range. The output frequency is therefore modulated, creating multiple tones around the carrier. Note that it is the case even though the two inputs (CK_{ref} and CK_{inj}) are perfectly lined up in frequency. Called "pseudo locking", this state can never reach a real locking either in phase or frequency.

To further explain this phenomenon, we illustrate the circuit behavior in detail in Fig. 8.39(b). Suppose the resonance frequency of the tank, ω_{res}, locates at position ① initially. Attempting to correct the residual phase, the loop pushes it toward one end of the lock range (i.e., position ②) by lifting the control voltage. Since the desired θ can never be achieved, the VCO becomes out of lock momentarily at some frequency slightly higher than $N\omega_{inj} + \omega_L$.The PFD soon accumulates enough phase errors, changing the polarity of the pumping current and moving ω_{res} to position ③ . Note that the progress from ② to ③ is relatively fast: if $\omega_L/\omega_{out} = 1\%$, it takes only 25 cycles of CK_{out} to create a $90°$ phase difference. Subsequently, the loop continues to adjust the phase by lowering ω_{res} until it hits the other end of the lock range $N\omega_{inj} - \omega_L$, which is position ④. Again, the VCO stays in free run temporarily and the resonance frequency goes back to position ① afterwards. The process repeats itself if the situation continues. Note that throughout the durations of ① $\rightarrow$ ② and ③ $\rightarrow$ ④, the VCO is prone to injection locking and the output frequency is very

close to $N\omega_{inj}$. Utilizing the control voltage variation, it is possible to estimate the cyclic period

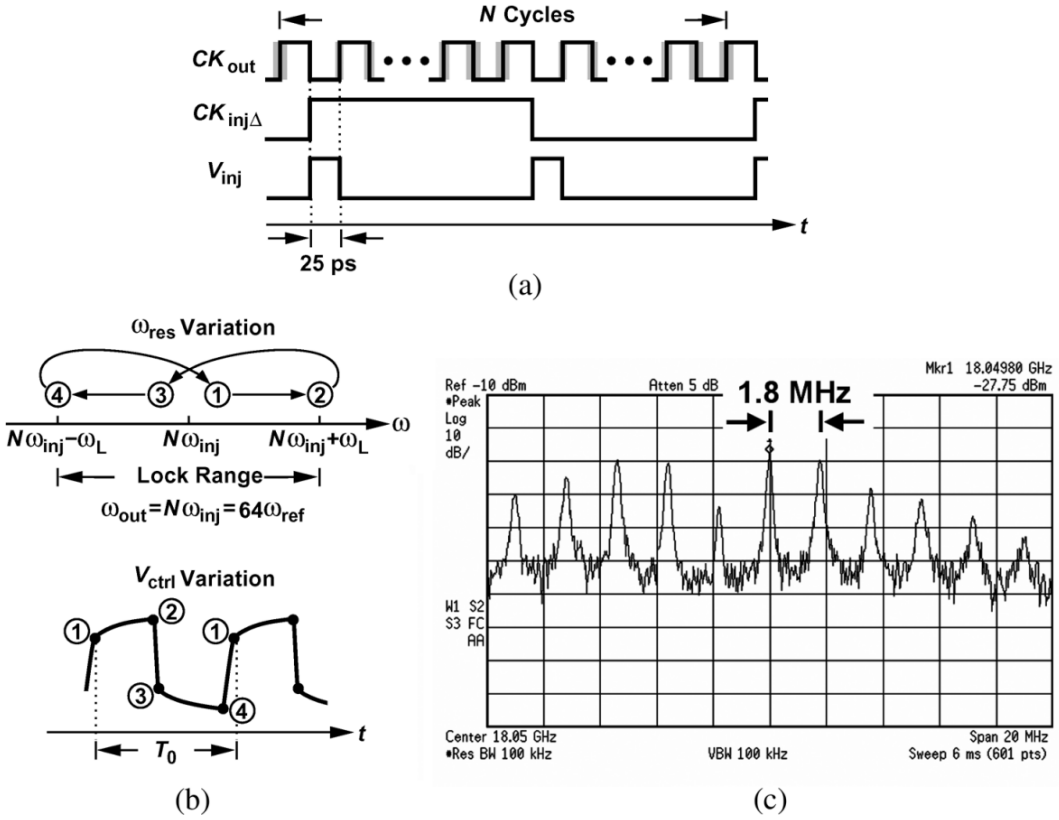


Fig. 8.39 (a) Timing diagram of fully destructive case. (b) Variation of VCO resonance frequency during pseudo lock and the corresponding control voltage. (c) The measured spectrum under pseudo-lock mode.

T_0 of the circulation. Neglecting the sharp transitions of ② → ③ and ④ → ①, we recognize that T_0 is primarily determined by time for the loop capacitor C [in Fig. 8.31(a)] to charge or discharge. The pumping current under pseudo locking, however, is hard to determine, because it depends on many other factors. Simulation shows that the effective current I'_p is about 20% to 40% of the peak current. Overall, we calculate T_0 as

$$T_0 \approx \frac{C}{I'_p} \times \frac{\omega_L}{K_{VCO}} \times 2. \tag{8.50}$$

In the testing chip, we have $I'_p = 30$ μA, $K_{VCO} \approx 2\pi \times 1$ Grad/sec·V, and $C = 120$ pF, resulting in $T_0 \approx 0.48$ μs. With the periodic modulation imposed on the control voltage, the output spectrum

reveals multiple tones around the desired frequency with a spacing of $1/T_0$. Figure 8.37(c) shows the measured output spectrum under pseudo-locking operation. The spacing between adjacent tones is approximately 1.8 MHz, which is 13% lower of the estimation from (8.50). Such an error is reasonable for our over-simplified calculation. For example, the loop filter here is modeled as a big capacitor. The actual charging and discharging currents are subject to mismatch as well, because V_{ctrl} experiences a large swing here. It also causes the different heights for the peaks in Fig. 8.39(c). Nonetheless, (8.49) still quantifies this issue with moderate accuracy.

8.5 ALL-DIGITAL PLLS

So far we have discussed different type of PLLs which is suitable for various applications. With nature theoretical and practical developments, one can easily think of a standardize PLL solution in digital domain. Indeed, if one optimized PLL can be migrated from one technology node to another (just like most digital circuits), significant cost and time could be saved. Another important advantage for a digitized PLL is that the area of its loop filter can be dramatically reduced. The performance of digital circuits is basically immune from PVT variations. We introduce all digital phase-locked loops (ADPLLs) in this section.

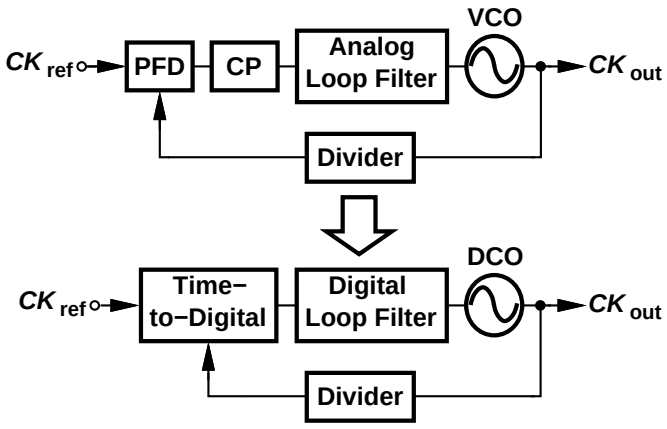


Fig. 8.40 PLL migration from analog to digital domain.

To digitize its blocks as much as possible, we look at the comparison between analog and digital PLLs (Fig. 8.40). If the output of PFD can be digitized to numbers, the charge pump is no

longer necessary. The subsequent analog loop filter can be realized in digital format. Similarly, a continuous-tuning VCO is replaced with a digital-controlled oscillator (DCO). As a result, the frequency tuning is now in discrete mode (with a very fine step). All blocks have been digitized except the frequency divider, which deals with analog clock waveforms.

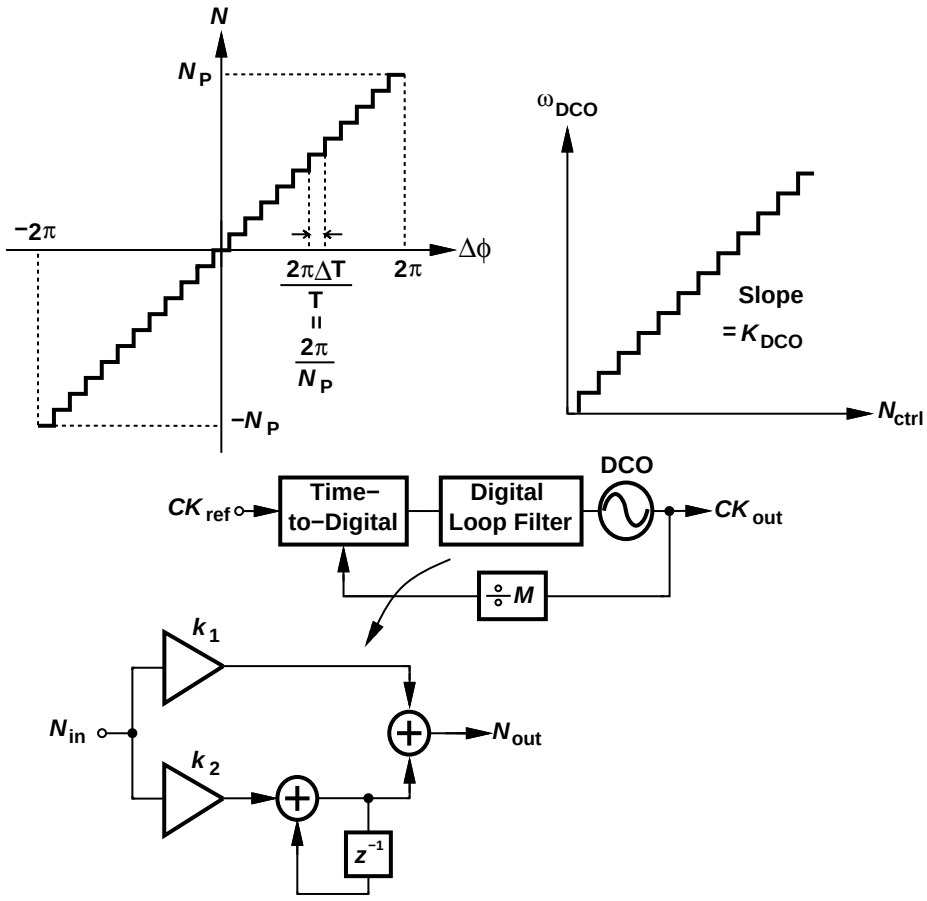


Fig. 8.41 All digital PLL behavior model.

All digital PLLs can be analogous to type-II PLLs. As illustrated in Fig. 8.41, the time-to-digital converter (TDC) provides a quantized output N, which is proportional to the phase difference between its two inputs. Denoting TDC's resolution as ΔT, we have

$$\Delta T \cdot N_p = T, \tag{8.51}$$

where $T = 1/f_{ref}$ is the reference period also the sampling period. It follows that each step is equal to $2\pi\Delta T/T = (2\pi/N_p)$. The DCO's frequency has also been quantized with a gain of K_{VCO}, and

integration of frequency gives rise to a transfer function of K_{VCO}/s. A first-order digital loop filter is adopted here. The input-output transfer function is given by

$$\frac{N_{out}}{N_{in}} = K_1 + \frac{K_2}{1 - z^{-1}}. \tag{8.52}$$

Recognizing $z = e^{sT}$, we have $z^{-1} = e^{-sT} \approx 1 - sT$. It is indeed true because phase variation rate (around or less than the loop bandwidth) is much lower than the reference frequency, i.e, $sT \ll 1$. That is,

$$\frac{N_{out}}{N_{in}} \cong K_1 + \frac{K_2}{sT}. \tag{8.53}$$

This s-domain expression has exactly the same form as an first-order RC loop filter in Fig. 8.2. Combining all blocks together, we obtain

$$\frac{\phi_{in} - \dfrac{\phi_{out}}{M}}{2\pi} \cdot N_p \cdot \left(K_1 + \frac{K_2}{sT}\right) \cdot \frac{K_{DCO}}{s} = \phi_{out}. \tag{8.54}$$

Since CK_{ref} and CK_{out} are in continuous mode, we arrive at a closed-loop transfer function

$$\frac{\phi_{out}}{\phi_{in}} = \frac{M \cdot (2\zeta\omega_n s + \omega_n^2)}{s^2 + 2\zeta\omega_n s + \omega_n^2}, \tag{8.55}$$

where

$$\omega_n = \sqrt{\frac{N_p K_{DCO} K_2}{2\pi T M}} \tag{8.56}$$

$$\zeta = \frac{K_1}{2}\sqrt{\frac{K_{DCO} N_p T}{K_2 \cdot 2\pi M}}. \tag{8.57}$$

As expected, and ADPLL behavior just like a regular type-II (charge-pump) PLL. The reader can easily prove that they have similar properties of stability, loop behavior, and oscillator noise contribution. The major difference is that the TDC presents significant quantization noise. To quantify it, consider the noise model shown in Fig. 8.42. It is clear that TDC's quantization error is equivalent to a periodic random phase error in its input. In other words, it can be modeled as a pulse sequence with period T and random magnitude uniformly distributed from $-\pi/N_p$ to π/N_p.

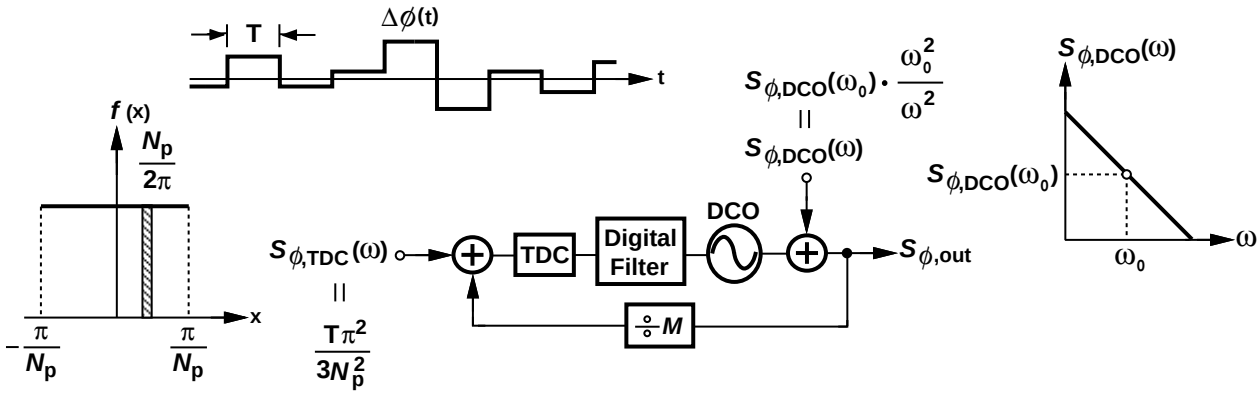


Fig. 8.42 ADPLL noise model.

Recall that the power spectral density of a zero mean random pulse sequence is equal to

$$\begin{aligned} S_{\phi,TDC}(\omega) &= \frac{\sigma^2}{T}|p|^2 \\ &= \frac{1}{T}\cdot\int_{-\pi/N_p}^{\pi/N_p}\frac{N_p}{2\pi}\cdot x^2 dx\cdot\left[\frac{\sin(\omega T/2)}{\omega/2}\right]^2 \\ &\approx \frac{T\pi^2}{3N_p^2}. \end{aligned} \tag{8.58}$$

Multiplying its transfer function ϕ_{out}/ϕ_{in} we obtain the output phase noise coming from TDC's quantization. The DCO's noise has its own transfer function

$$\frac{\phi_{out}}{\phi_{DCO}} = \frac{s^2}{s^2+2\zeta\omega_n s+\omega_n^2}, \tag{8.59}$$

which is has the same format as a VCO. Together with the DCO's contribution, the overall phase noise at the ADPLL's output is

$$S_{\phi,out}(\omega) = S_{\phi,TDC}\cdot\left|\frac{\phi_{out}}{\phi_{in}}\right|^2 + S_{\phi,DCO}\cdot\left|\frac{\phi_{out}}{\phi_{DCO}}\right|^2 \tag{8.60}$$

Unfortunately, TDC noise is usually much higher than DCO's noise. (Reference and other building blocks contribution is negligible, too). The quantization noise forms a bottleneck in ADPLL's performance.

Example 8.4

Consider a ADPLL with 1st-order loop filter. If $\zeta = 5$, $2\zeta\omega_n = 2\pi \times 10$ MHz, $f_{out} = 10$ GHz, $f_{ref} = 100$ MHz, TDC resolution $\Delta T = 10$ ps, and DCO has a phase noise of -105 dBc/Hz at 1-MHz offset. Sketch the output phase noise components.

Solution:

With an overdamped loop, both $S_{\phi,TDC} \cdot |\phi_{out}/\phi_{in}|^2$ and $S_{\phi,DCO} \cdot |\phi_{out}/\phi_{DCO}|^2$ have the same outline with $2\zeta\omega_n$ corner frequency and -20 dBc/dec slope for out-of-band noise. For TDC noise, the low-frequency magnitude would be

$$S_{\phi,TDC} \cdot \left|\frac{\phi_{out}}{\phi_{in}}\right|^2\Bigg|_{f=0} = \frac{T\pi^2}{3N_p^2} \cdot M^2 = -94.8dBc/Hz.$$

The DCO has -125 dBc/Hz phase noise at loop bandwidth (10 MHz). The noise contribution is plotted in Fig. 8.43.

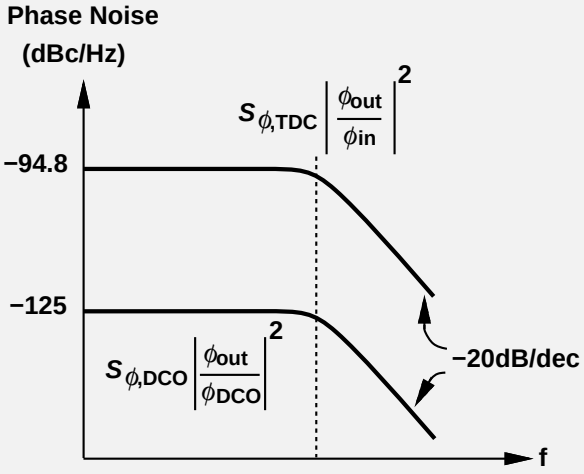


Fig. 8.43 Noise contribution of TDC and DCO.

The above example reveals that the fact the TDC's quantization noise dominates over other noise sources. We need to reduce it by circuit techniques. From Eq. (8.58) we realize that $S_{\phi,TDC}$ could be reduced by enlarging f_{ref} and N_p. For a given reference frequency, we need to measure the TDC resolution. A conventional TDC realization is illustrated in Fig. 8.44(a). The divider's input is delayed and sampled by the reference clock. Obviously the resolution depends on the

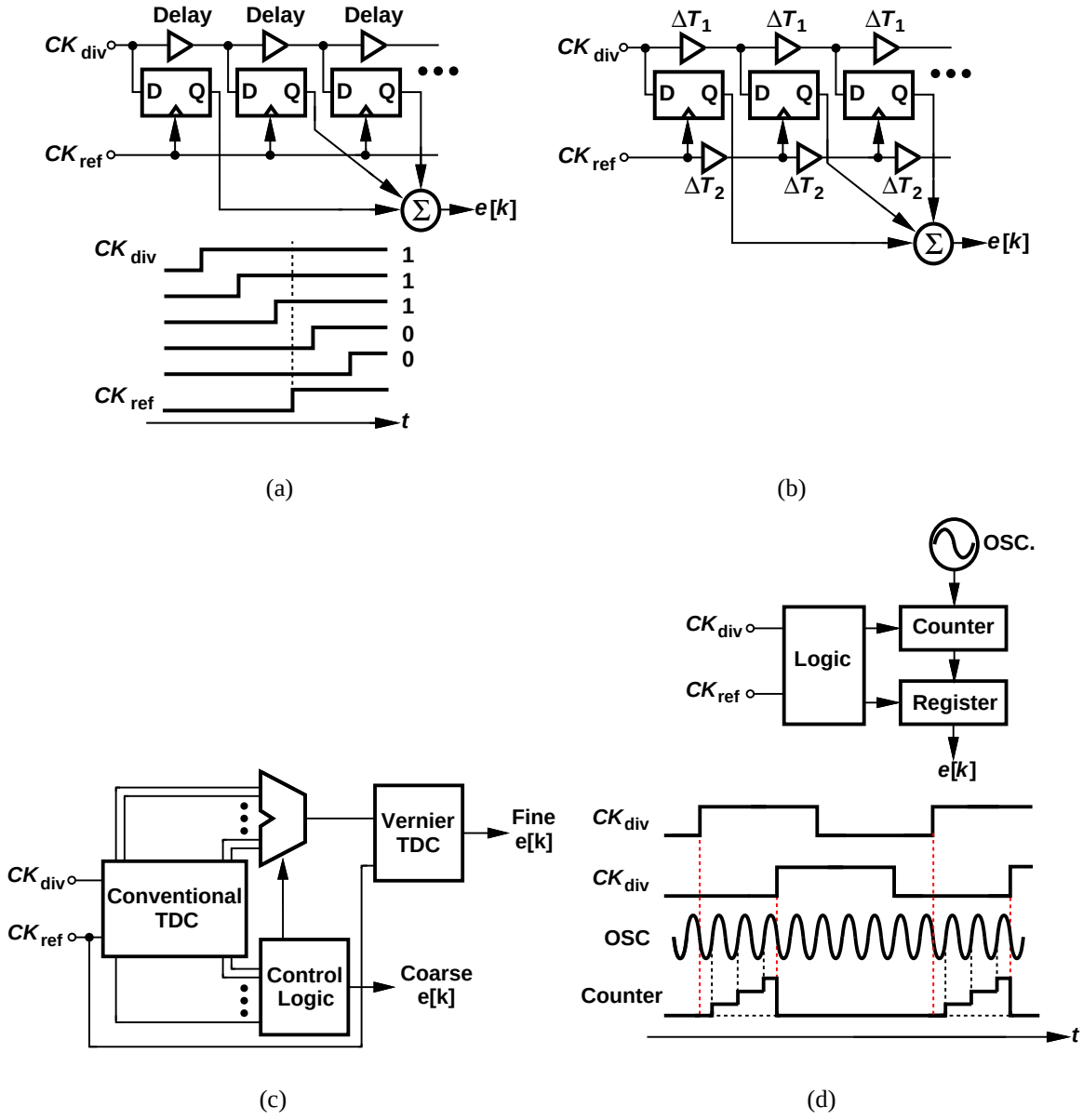


Fig. 8.44 TDC structure: (a) conventional, (b) Vernier, (c) two-step, (d) oscillator-based.

delay itself. In 40nm CMOS technology, for example, a fan-out-of-4 inverter present's a delay of $7.5 \sim 10$ ps. To improve precision, a Vernier technique can be applied [Fig. 8.44(b)]. Instead of using one delay, here we have two delays for each sample. The resolution now between the difference, i.e, $\Delta T_1 - \Delta T_2$, significantly better than that in Fig. 8.44(a). However, much more area and power would be consumed in this approach. The mismatch among segments becomes more severe as well. A compromised yet efficient approach is to combine the two methods. Recognized as "two-step" architecture, it utilizes single-delay TDC to generate coarse result and Vernier TDC to obtain the fine output [Fig. 8.44(c)]. Since TDC would only move around a small region (e.g., the vicinity of origin) when the loop is locked, it is a good approach to increase (effective) N_p for a given power and area budget.

The above approach share same issue. First, the area/power consumption is still large even the TDCs are implemented with digital circuit. Moreover, they suffer from deterministic error owing to the stair-shaped characteristic, no matter how fine the resolution is. A clever method is to take advantage of an oscillator. As depicted in Fig. 8.44(d), one can count the number of cycles within one phase error rather than using delays. By doing so, a great amount of chip area and power could be saved. Moreover, the quantization error is actually averaged out, as it occurs randomly and uncorrelatedly at beginning and end of each counting interval. As a result, effective resolution is improved as well. In reality, ring oscillators are good candidates for such oscillator-based TDCs. Their multiple phases are essential for mismatch shaping, too.

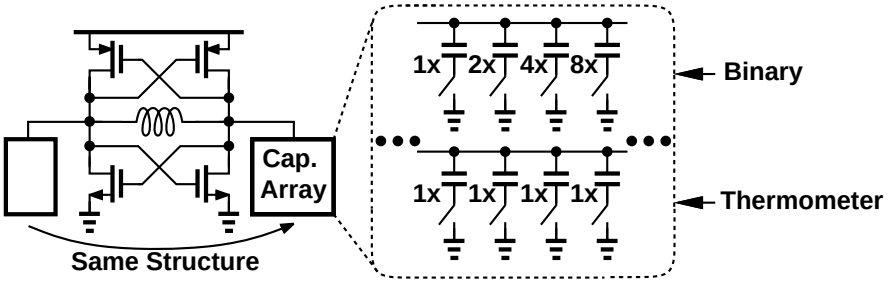


Fig. 8.45 Digital control oscillator.

A typical DCO can be found in Fig. 8.45, where an LC-tank oscillator with switched capacitor array are incorporated. Again, coarse and fine tuning can be achieved separately by binary

and thermometer controls. A DCO usually needs 12-bit resolution for frequency tuning, and the switching capacitors linearity matters. Note that due to rounding, a DCO would inevitable toggle between 2 (or more) states (N_{ctrl}) upon locking. The finite frequency resolution leads to deterministic jitter. Capacitor dithering could provide some remedy, but it provides additional quantization noise.

ADPLL has quite a few on going development. For instance, the all-digital operation can easily merge with $\Sigma\Delta$ modulator, resulting in a all-digital fractional-N synthesizer. DCO itself could possibly be modulated with the same manner. Calibration techniques would be added into it as well. Nonetheless, because of some nature limitations, performance of ADPLLs can not compete with that of type-II (charge-pump) PLLs in high-end applications at this moment. Its highly migratable property makes a new design trend for consumer ICs.

8.6 DIRECT-DIGITAL FREQUENCY SYNTHESIZERS

REFERENCES

[1] A. Maxim *et al*., "A low jitter 125-1250 MHz process independent and ripple-poleless 0.18-μm CMOS PLL based on a sample-reset loop flter," *IEEE J. Solid-State Circuits*, vol. 36, no. 11, pp. 1673-1683, Nov. 2001.

[2] J. Kim, J.-K. Kim, B. Lee, N. Kim, D. Jeong, and W. Kim, "A 20-GHz phase-locked loop for 40-Gb/s serializing transmitter in 0.13-μm CMOS," *IEEE J. Solid-State Circuits*, vol. 41, no. 4, pp. 899-908, Apr. 2006.

[3] J. Lee, "High-speed circuit designs for transmitters in broadband data links," *IEEE J. Solid-State Circuits*, vol. 41, no. 5, pp. 1004-1015, May 2006.

[4] R. C. H. van de Beek *et al*., "A 2.5-10-GHz clock multiplier unit with 0.22-ps RMS jitter in standard 0.18-μm CMOS," *IEEE J. Solid-State Circuits*, vol. 39, no. 11, pp. 1862-1872, Nov. 2004.

[5] R. Gu *et al*., "A 6.25 GHz 1 V LC-PLL in 0.13-μm CMOS," in *IEEE ISSCC 2006 Dig. Tech. Papers*, Feb. 2006, pp. 594-595.

[6] A. Ng *et al*., “A 1V 24GHz 17.5mW PLL in 0.18/spl mu/m CMOS,” in *IEEE ISSCC 2005 Dig. Tech. Papers*, Feb. 2005, vol. 1, pp. 158-590.

[7] D. Park and S. Mori, “Fast acquisition frequency synthesizer with the multiple phase detectors,” in *1991 IEEE Pacific Rim Conf. Communications, Comput. Signal Process. Conf. Proc.*, Victoria, BC, Canada, May 1991, vol. 2, pp. 665-668.

[8] T. Lee and W. Lee, “A spur suppression technique for phase-locked frequency synthesizers,” in *IEEE ISSCC 2006 Dig. Tech. Papers*, Feb. 2006, pp. 592-593.

[9] B. Razavi, *RF Microelectronics.*, Second Edition, Prentice-Hall, 2012.

[10] D. B. Leeson, “Simple model of a feedback oscillator noise spectrum,” *Proc. IEEE*, vol. 54, no. 2, pp. 329-330, Feb. 1966.

[11] K. Kundert, “Predicting the phase noise and jitter of PLL-based frequency synthesizers,” [Online]. Available: http://www.designers-guide.org

[12] “Clock jitter and phase noise conversion,” Maxim IC, Sunnyvale, CA [Online]. Available: http://www.maxim-ic.com/appnotes.cfm/an_pk/3359

[13] F. Gardner, “Charge-pump phase lock loop,” *IEEE Trans. Commun. Electron.* vol. 28, no. 11, pp. 1949-1858, Nov. 1980.

[14] B. Razavi, “A study of injection locking and pulling in oscillators,” *IEEE J. Solid-State Circuits*, vol. 39, no. 9, pp. 1415-1424, Sep. 2004.

[15] K. Kurokawa, “Noise in synchronized oscillators,” *IEEE Trans. Microw. Theory Tech.*, vol. MTT-16, pp. 234-240, Apr. 1968.

[16] R. Adler, “A study of locking phenomena in oscillators” *Proc. IEEE*, vol. 61, no. 10, pp. 1380-1385, Oct. 1973.

[17] A. Mirzaei *et al*., “The quadrature *LC* oscillator: A complete portrait based on injection locking,” *IEEE J. Solid-State Circuits*, vol. 42, no.9, pp. 1916-1932, Sep. 2007.

CLOCK AND DATA RECOVERY (I)

9

Perhaps clock and data recovery circuits (CDRs) are the most significant parts among all the transceiver building blocks. Tremendous amount of brilliant idea has been proposed on this subject over the past decades, resulting in at least a dozen of mainstream CDR structures for different applications. Primarily based on the operation of phase detectors (PDs), CDR circuits are categorized as PLL-based, DLL-based, phase-interpolator (PI) based, burst-mode, over-sampling, and others. We study these architectures in this chapter.

9.1 INTRODUCTION

A CDR circuit bears a few functions in the receiver. It extracts the clock of data rate from the incoming NRZ data stream, which does not exist in the data spectrum. Some sub-rate CDRs generates lower-speed clocks in the same manner. A CDR also needs to clean up the jitter/noise from the input data, providing retimed or even demultiplexed data for the subsequence circuits. Modern CDR circuits usually require co-design with DFEs. While the former needs the latter to provide input data with clean-enough eyes, the latter depends on the former to create one-bit delay. Figure 9.1 illustrates a typical SerDes receiver, where the CDR is actually located in the center. Like the heart to a human being, a CDR pumps clock (instead of blood) to the whole system. The recovered clock would be further divided so as to deserialize the data.

CDR circuits rely on a phase detector to align the clock with incoming data. There are so many different type of phase detectors, and CDRs are usually classified accordingly (Fig. 9.2).

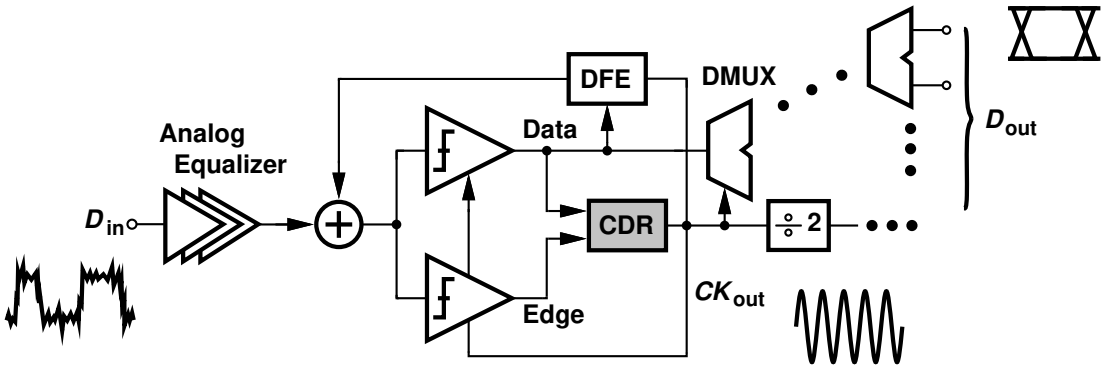


Fig. 9.1 Position and function of clock and data recovery circuits.

The analog PLL-based CDRs basically follows the rules of a phase-locked loop, where both linear and binary phase detection circuits have been thoroughly developed. These circuits can be highly digitized, forming so called ”all-digital” CDRs. It is very much like the all-digital PLLs we introduced in chapter 8. In applications where a global reference clock is available or the frequency offset between TX and RX is tolerable, the DLL-based phase detectors could be used. A more efficient approach can be found in the PI-based structures, which simply track the phase by rotating the synthesized clock. Again, this popular architecture can be realized in either analog or digital format.

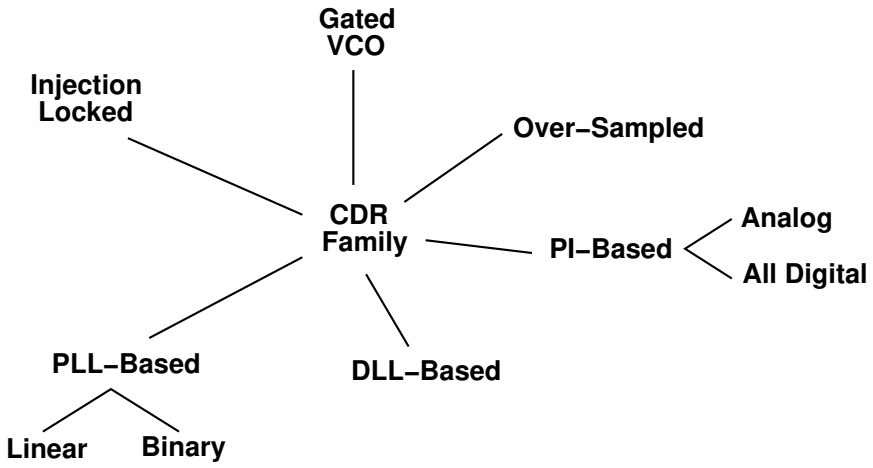


Fig. 9.2 Modern CDR classification.

Over-sampling technique could be also adopted in CDR circuit. Also known as blind-sampling, this approach picks out the "most correct" data sampling by examining the data transition. Once again, if the frequency offset between clock and data rate is small enough, the lop maintains lock. Other CDR formats such as gated VCO or injection-locked are primarily used in burst-mode applications.

Recall from chapter 1 that NRZ data stream contains no spectral signal at the frequency of data rate and its harmonics. How do we generate such a clock from nothing? The transition detector provides a simple solution. As shown in Fig. 9.3, the data transition can be detected by XORing the input data and the delayed version. The created pulse stream reveals a power spectrum of squared sinc function together with spectral lines at data rate and harmonics. Simulated results for a 20 Gb/s data input. It can be proven that the breadth of the lobes and the appearance of the clock lines vary significantly for different pulse widths: for a half bit-period $(T_b/2)$ pulse, the spectrum nulls at $2/T_b$ [Fig. 9.4(a)], whereas for a quarter bit-period $(T_b/4)$ pulse [Fig. 9.4(b)], it expands as twice wider but with lower magnitude. The impulses still locate at the harmonics of $1/T_b$ except the nulls. As a matter of fact, the magnitude of spectral line at data rate can be normalized as

$$P(f = \frac{1}{T_b}) = \frac{\sin(x\pi)}{\pi}, \tag{9.1}$$

where x is defined as

$$0 < x = \frac{\Delta T}{T_b} < 1. \tag{9.2}$$

In other words, one can extract the clock of data rate from NRZ data by placing an arbitrary delay. The line at $1/T_b$ could be further distilled to obtain a pure clock. We discuss more details in section 9.xx.

The above clock extraction method needs additional sampler (e.g., a flipflop) to retime the data. At high speed, it also requires alignment circuits to ensure the phase relationship between the created clock and input data. We introduce mainstream phase detectors in the following sections, which resolve these issues more efficiently.

The reader must be aware that CDRs are dealing with random data instead of periodic reference. Unfortunately, phase detectors for CDR circuits have limited capture range (approximately

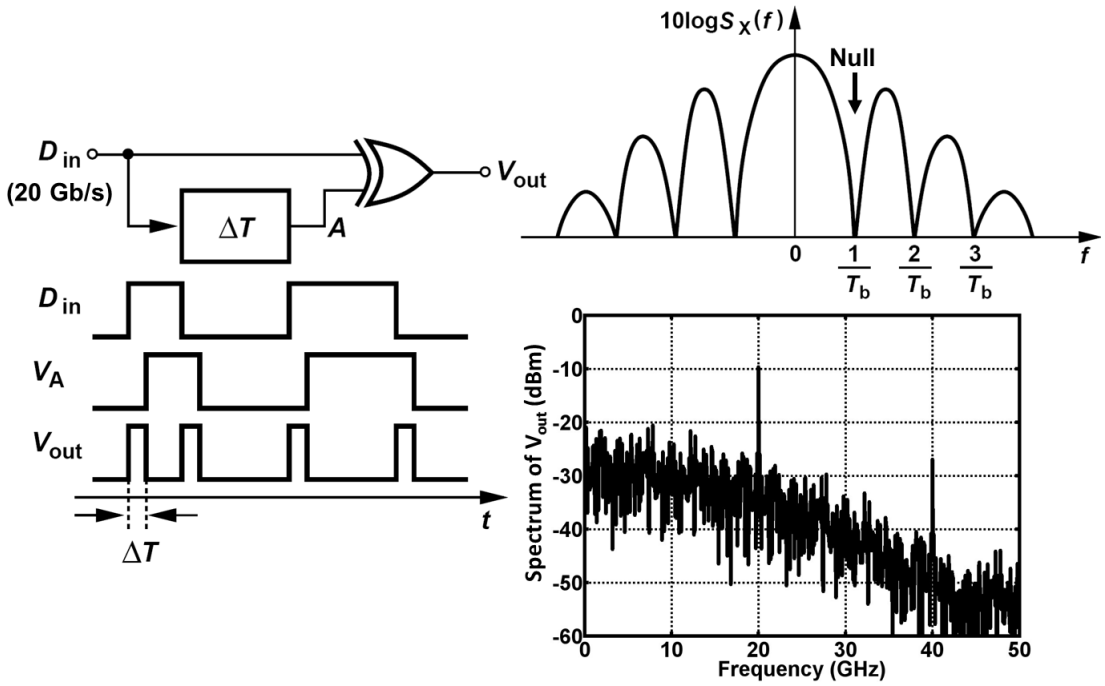


Fig. 9.3 Clock extraction from NRZ data using transition detector.

on the order of the loop bandwidth, see 9.xx), and frequency acquisition at power up becomes mandatory. Unlike type-IV PFDs in PLLs, there is no all-in-one solution for both phase and frequency detection in CDR circuits. In other words, we need a frequency detector (FD). The phase detection and frequency tracking are usually done in separate loops. Figure 9.5 illustrates two examples of dual-loop architecture for analog, PLL-based CDRs. Figure 9.5(a) presents an intuitive approach, where a type-IV PFD analog with a local reference forms a frequency acquisition loop, calibrating the VCO frequency before performing phase detection. Advanced frequency detector (e.g., Pottbacker FD, see 9.xx) can extract the frequency information directly from data stream, obviating the use of local crystal oscillators [Fig. 9.5(b)]. Built-in or additional lock detectors are recommended to provide out-of-lock alarm. The FD loops are switch off in most cases in order to minimize disturbance.

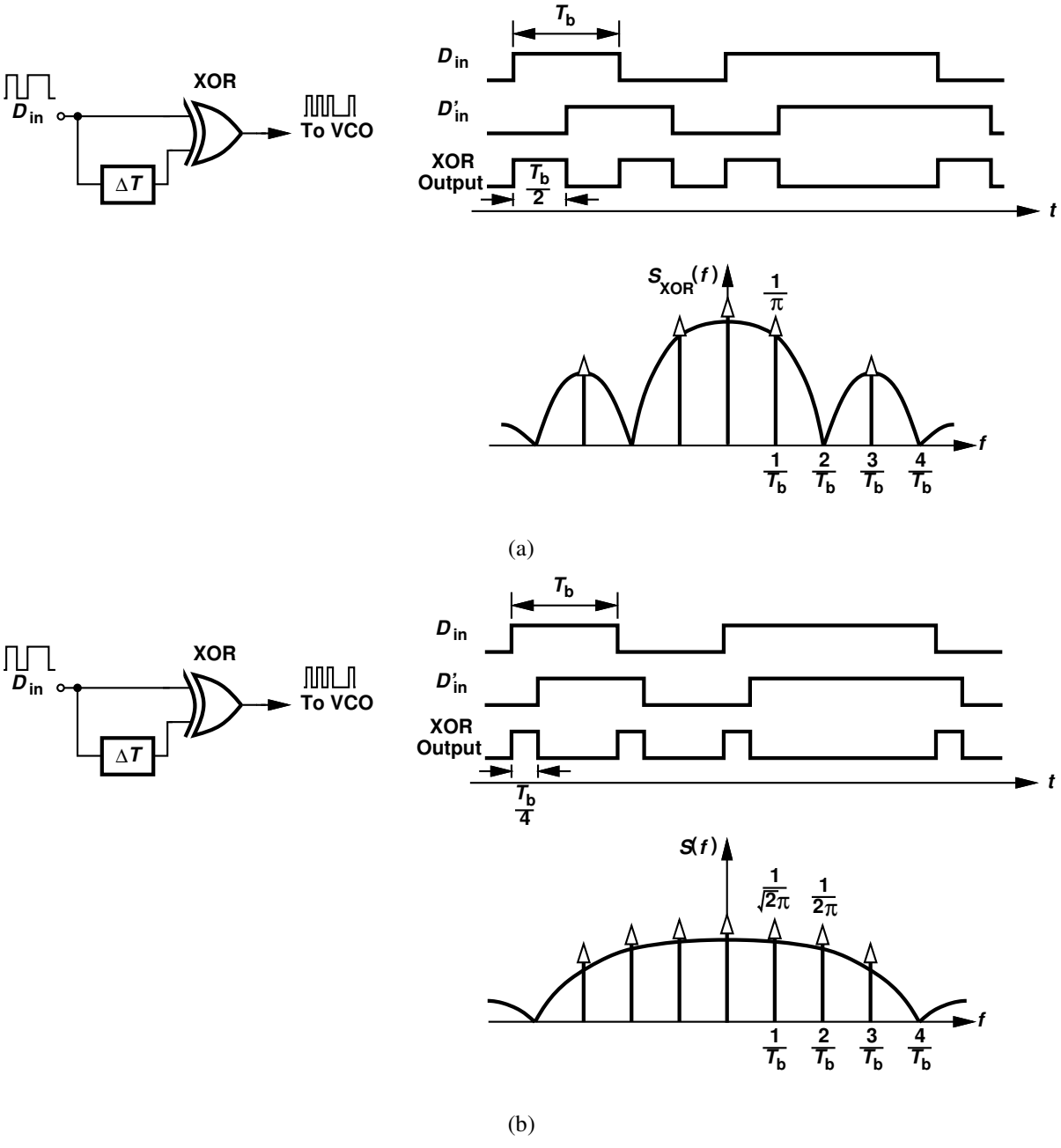


Fig. 9.4 Spectra of transition detector with different pulsewidths: (a) $\Delta T = T_b/2$, (b) $\Delta T = T_b/4$.

9.2 LINEAR, PLL-BASED CDRS

Linear, PLL-based CDRs can be easily analogous to traditional linear PLLs. Two main PD types, namely, Hogge and purely linear, are introduced here. Their modifications are included in our discussion as well.

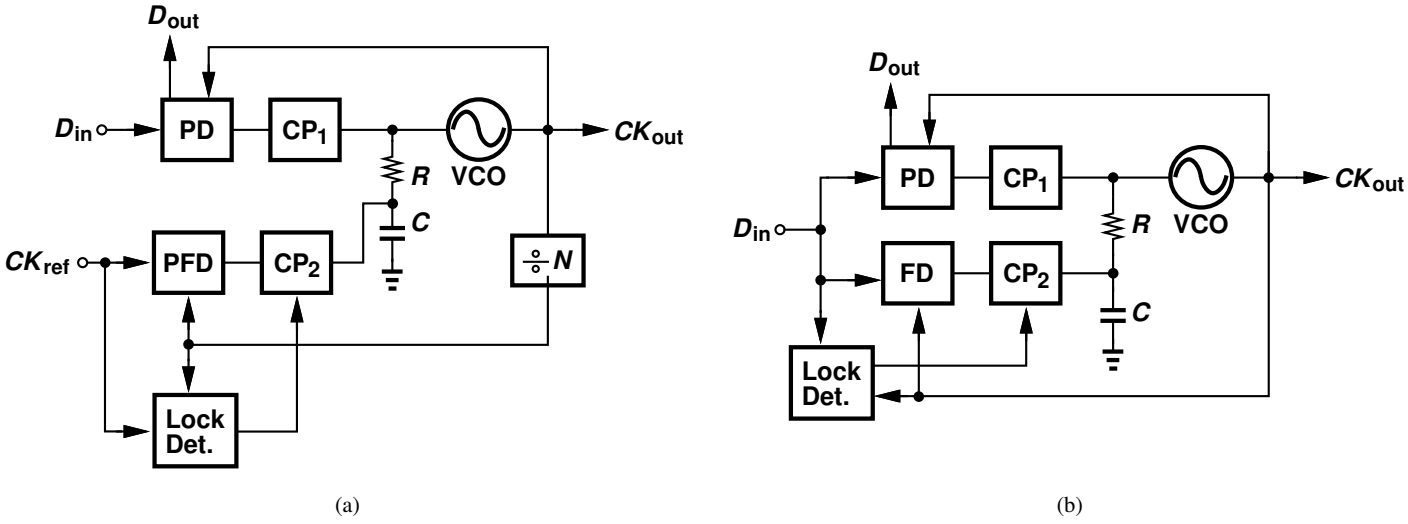


Fig. 9.5 Examples of analog, PLL-based CDR architecture: (a) with, (b) without local reference clock.

9.2.1 Hogge PD

One representative linear PD structure is proposed in [xx]. It can be easily studied by the well-known linear PLL model. As shown in Fig. 9.6, if the Hogge PD and charge pump present an average output current I_{av} proportional to the phase difference (between clock and incoming data), we obtain the closed-loop transfer function in phase (or frequency) domain as

$$\frac{\phi_{out}}{\phi_{in}}(s) = \frac{2\xi\omega_n s + \omega_n^2}{s^2 + 2\xi\omega_n s + \omega_n^2}, \tag{9.3}$$

where

$$\omega_n = \sqrt{\frac{I_P K_{VCO}}{2\pi C_P}} \tag{9.4}$$

$$\xi = \frac{R_P}{2}\sqrt{\frac{I_P C_P K_{VCO}}{2\pi}}. \tag{9.5}$$

It is exactly the same as Eq. (8.4) $\sim$ (8.6) with $M = 1$. Note that the effective pumping current for Hogge PD is scaled down by a factor of 2 as the chance of data transition between bits is $1/2$. All characteristics of a linear PLL can be directly applied to Hogge PD based linear CDRs, given that granularity approximation sustains. For regular wireline applications, CDRs have large ξ ($\gg$1) to

safely stabilize the loop, degenerating the transfer function as

$$\frac{\phi_{out}}{\phi_{in}}(s) = \frac{2\xi\omega_n}{s + 2\xi\omega_n}. \tag{9.6}$$

The loop bandwidth is thereforce given by

$$\omega_{-3dB} = 2\xi\omega_n = \frac{K_{VCO}I_P R_P}{2\pi}. \tag{9.7}$$

Like analog PLLs, higher-order loop filters can be applied here to minimize disturbance on the control line.

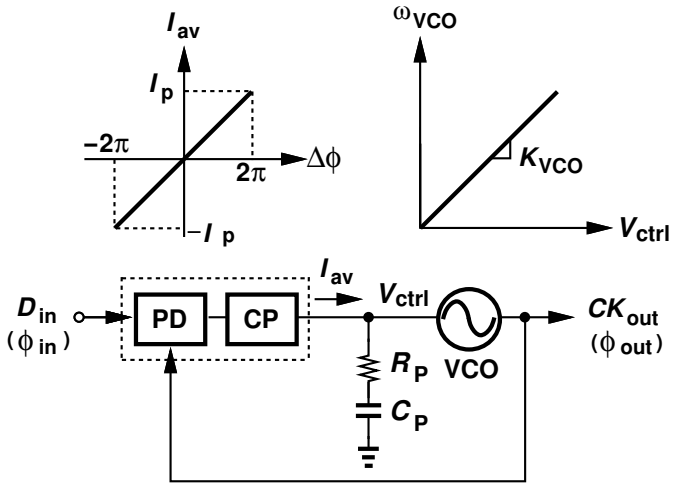


Fig. 9.6 PLL-based analog CDR model.

How to implement a Hogge PD? A typical realization can be found in Fig. 9.7(a), where the input data is sampled by two flipflops in a row with opposite driving clocks. As can be shown in Fig. 9.7(b), the first XOR gate produces V_X pulses whose width is proportional to the phase difference between D_{in} and CK_{in}. The second XOR gate, on the other hand, generates a pulse sequence with constant width $(= T_b/2)$. As a result, the average output $\overline{V_X}$ and $\overline{V_Y}$ provide adequate information about the phase error. As illustrated in Fig. 9.7(b), $\overline{V_X}$ - $\overline{V_Y}$ prevents linear characteristic in most of the range. At 5-Gb/s data rate, for example, the linear (operation) range is approximately xx° in 65-nm process. Note that during long runs, no pulse is created, leaving the control line undisturbed. Recognized as "tri-state" operation, this function is an important feature of decent PDs. Retimed data could be taken out from either V_A or V_B.

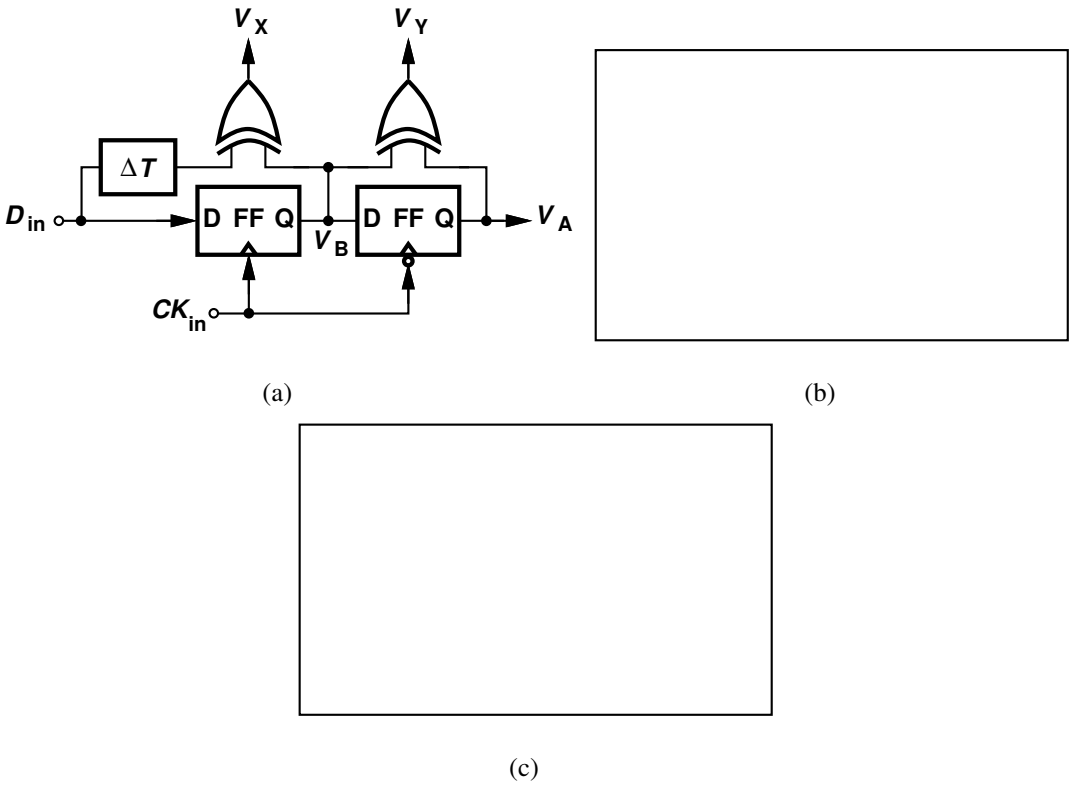
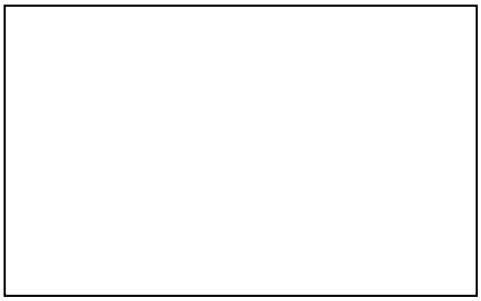


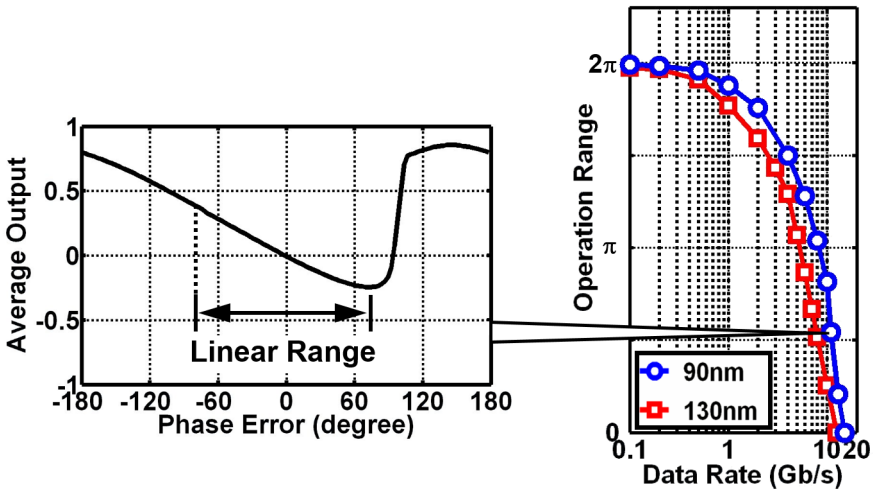
Fig. 9.7 Hogge phase detector: (a) architecture, (b) waveforms, (c) $\overline{V_X}$ and $\overline{V_Y}$ (designd for 5 Gb/s data in 65-nm CMOS).

A few issues come along with such a pulse detection. For example, the operation of a realistic flipflop always accompanies finite clock-to-Q delay. Thus, an artificial delay ΔT must be added so as to compensate foe the delay. Unfortunately these two delays do not match well over PVT variations. More seriously, the limited rise/fall time makes it very difficult to generate a complete pulse at high speed. Since the Hogge PD involves pulse generation and pulsewidth comparison, it is very challenging to reach data rate beyond 10 Gb/s. We plot the rise/fall time ($10 \sim 90\%$) for different technology nodes in Fig. 9.8(a). In 90-nm CMOS, for instance, the rise/fall time of a fan-out-of 4 inverter is as large as 33 ps. Current-mode logic (CML) can speed up the operation to some extent, but the fundamental issue still remains. Fig. 9.8(b) shows the simulated linear range of a standard full-rate Hogge PD designed with CML topologies in 90-nm and 0.13-μm CMOS technologies. Even though the circuits are optimized (e.g., adding proper delays to compensate for

the skew), the operation range still drops dramatically after 1 Gb/s. This is because at high speed, the finite transition times compress the width of the reference pulses, making the PD characteristic imbalanced.



(a)



(b)

Fig. 9.8 (a) Rise/fall time of fan-out-of-4 inverter chain, (b) operation range.

The speed limitation can be somewhat relaxed by sub-rate operation. Figure 9.9 illustrates an example, which operates at half rate. Here, half-rate clock drives 4 latches (equivalent to 2 flipflops as well), and outputs X_1, X_2, Y_1, and Y_2 are XORed accordingly. Since X_1 and X_2 contain phase error information, the width of pulse sequence Err is linearly proportional to phase difference. On the other hand, Y_1 and Y_2 are split by fixed, one-bit delay, which results in a Ref pulse sequence twice as wide as the nominal Err pulse sequence. In other words, phase error is still linear proportional to the difference between twice the pulsewidth of Err and the pulsewidth of Ref.

A linear phase detection can be achieved of two fold pumping current is allocated to Err signal. Good features of Hogge PDs such as tri-state output and data retiming are preserved here. Effect of rise/fall time of Err and Ref pulses and mismatch associated with current mirroring may degrade the performance inevitably. Overall speaking, the operation speed can be improved to some extent ($\approx$ 10 Gb/s) by parallelism. Linear CDRs require novel architecture to make a breakthrough.

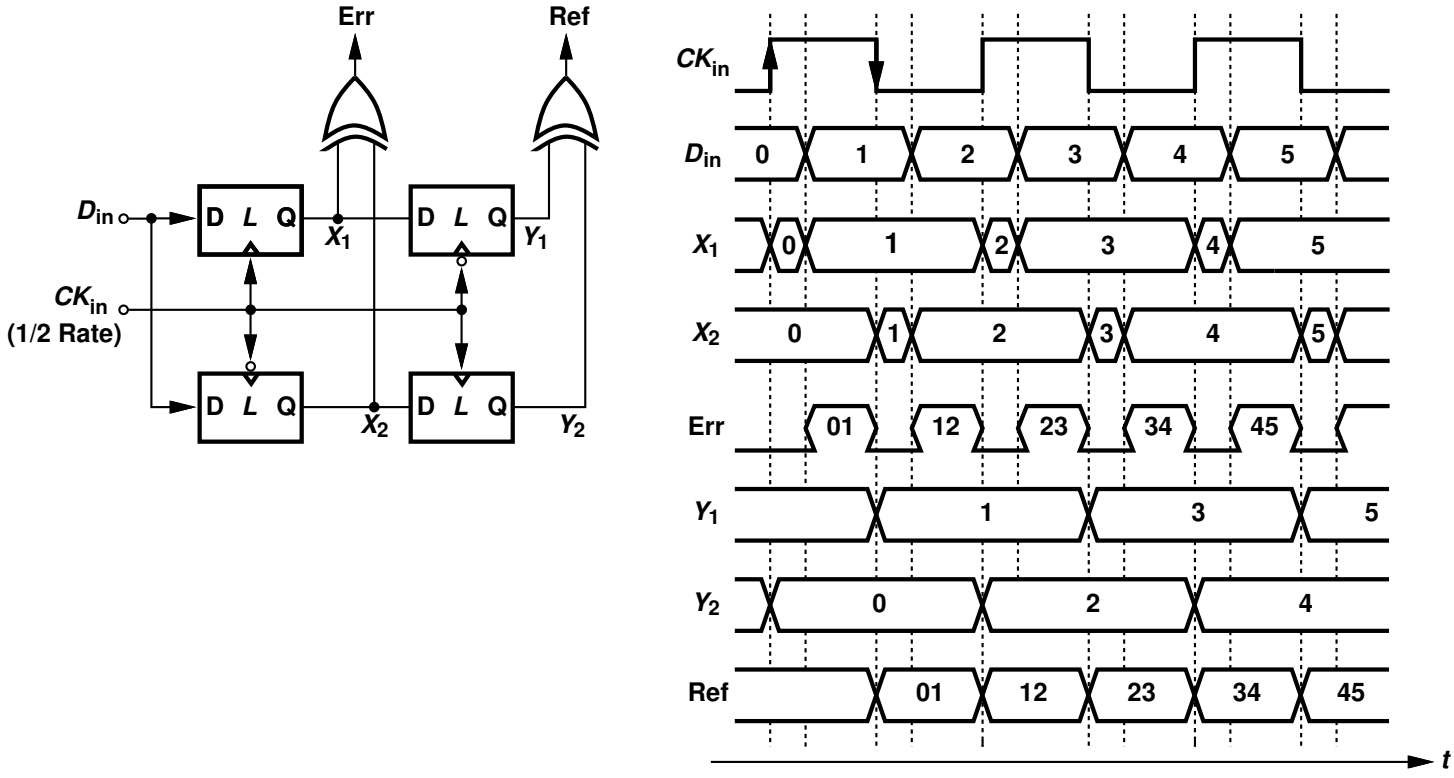


Fig. 9.9 Half-rate Hogge PD.

9.2.2 Purely Linear PD

As mentioned in 9.2.1, the pulse generation and comparison involved in Hogge PD limit the speed because of the long rise and fall signal edges of the XOR gates and finite CK-to-Q delay in the flip-flop. These issues can be alleviated by using a mixer-based phase detector. As shown in Fig. 9.10(a), the input data passes through a chain of delay cells, providing a total delay (from V_A to V_E) approximately equal to half a bit. An XOR gate examines this fixed phase difference,

creating a pulse nominally equal to 25 ps upon occurrence of data transitions. Acting as a reference for phase detection, this pulse sequence is mixed with the clock from the VCO.

In order to illustrate the operation easily, let us assume the pulses are ideally sharp for the time being. We plot conceptual waveforms of important nodes under locked condition in Fig. 9.10(b). When a data edge is present, the mixer produces an output pulse whose width is proportional to the phase difference between the XOR output and the clock. This result can be used for phase alignment. During consecutive bits, on the other hand, the mixer generates a periodic signal which is in phase with the clock. This signal has a zero average, given that the duty cycle of the clock is 50%. In other words, for random data the mixer provides an average output voltage proportional to the phase error between the two inputs. A V-to-I converter (or equivalently, a charge pump) then translates the voltage into current and injects it into the loop filter. As a result, the center tap V_C always aligns with the clock, and data sampling can be accomplished in the retiming flip-flop using the falling clock edges.

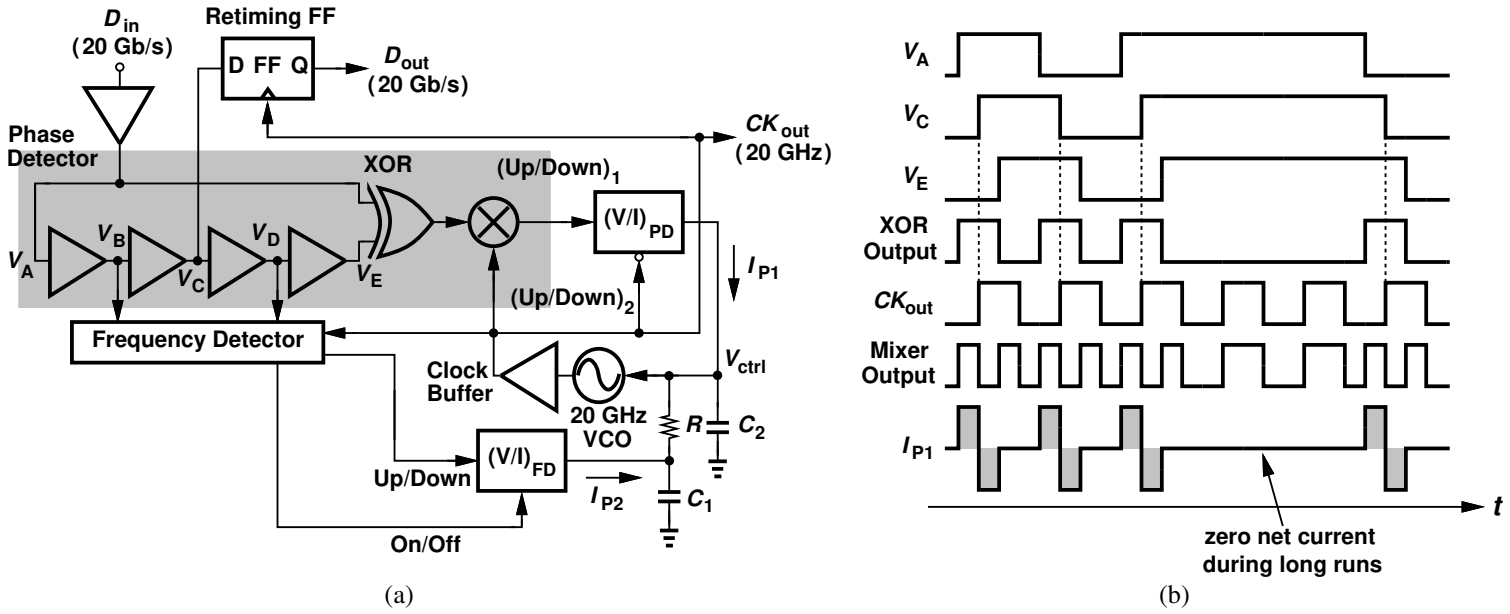


Fig. 9.10 (a) Purely linear PD, (b) conceptual phase relationship.

What happens if the clock duty cycle deviates from 50%? If $(V/I)_{PD}$ were solely driven by mixer, the distortion would lead to finite residue current and modulate the control voltage.

Fortunately, we can apply the complement of the clock ($\overline{CK}$) into $(V/I)_{PD}$ to overcome this difficulty. Since the clock and the mixer's output are in phase, it completely cancels out the periodic disturbance for consecutive bits. Note that I_{P1} [the output current of $(V/I)_{PD}$] reveals pure zero output during long runs.

In reality, pulse in Fig. 9.10(b) can never be so sharp for high data-rate inputs. How does this PD structure perform linear phase detection? Unlike Hogge PDs, this work need not generate narrow pulses at all. For data rate greater than 20 Gb/s, the clock and XOR outputs are quite "round" instead of "square", simply because the higher order harmonics are suppressed. Thus, the phase detection is nothing more than mixing two sinusoidal signals. As illustrated in Fig. 9.11, if the delay from V_A to V_E is exactly 0.5 UI, the XOR gate and the clock outputs can be simply modeled as

$$V_{XOR} = \begin{cases} A\cos(\omega t + \pi), & \text{for data transition} \\ -A, & \text{for long runs} \end{cases} \tag{9.8}$$

$$CK_{out} = B\cos(\omega t + \theta) \tag{9.9}$$

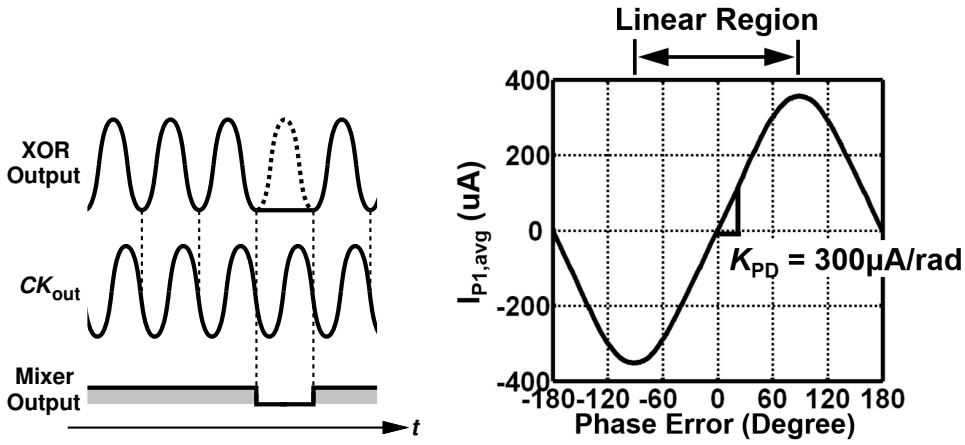
where A and B denote the magnitudes of two signals, respectively, and θ the phase error. The mixer's output thus becomes

$$V_{mixer} = \begin{cases} \dfrac{AB}{2}[\cos(\theta - \pi) + \cos(2\omega t + \theta + \pi)], & \text{for data transition} \\ -AB\cos(\omega t + \theta), & \text{for long runs.} \end{cases} . \tag{9.10}$$

In other words, when the data edge is presented, the phase difference is obtained as a near-dc output $(AB/2)\cos(\theta - \pi)$. The second-order term is filtered out by means of intrinsic parasitics. The fundamental modulation during consecutive bits is eliminated by $\overline{CK}$ as described before.

A transistor-level design in 90-nm CMOS of such a purely linear PD reveals more details. Figure 9.11(b) depicts the average output current as a function of phase error. As expected, it presents a sinusoidal characteristic with a PD gain [together with $(V/I)_{PD}$] of 300 μA/rad in the vicinity of origin. A linear operation region of about 180° is obtained. To compare with the Hogge PD (Fig. 9.12). We see that the proposed structure achieves a large operation bandwidth all the way from dc to 40 Gb/s. Note that the sharp pulses of I_{P1} in Fig. 9.10(b) do not exist in reality

either. Upon phase locking, the output current I_{P1} would only be modulated by a small amount of current because of the low-pass filtering. It gets rejected by the limited loop bandwidth of CDR anyway.



(a) (b)

Fig. 9.11 (a) Actual operation of purely-linear PD, (b) its characteristic curve.

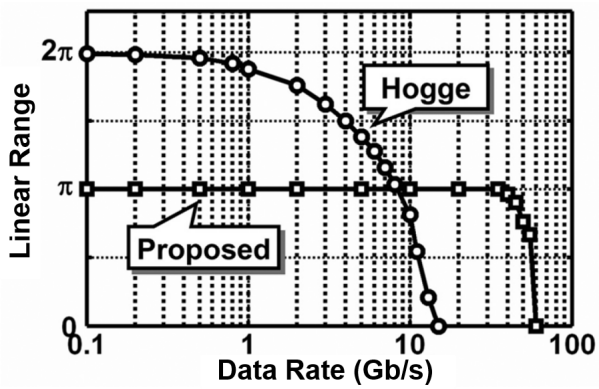


Fig. 9.12 Operation bandwidth comparison between Hogge and purely-linear PD.

Another important advantage is that under locked condition, the clock edges always align with the center of the generated pulses, whether or not the delay from V_A to V_E ($\Delta T_{A\rightarrow E}$) is exactly 0.5 UI. Fig. 9.13 reveals two cases where the delay is longer and shorter than half bit period. Obviously, V_C still coincides with the clock, keeping an optimal phase for data retiming. Note that the buffered

clock CK_{out} directly drives the mixer, $(V/I)_{PD}$, and the retiming flip-flop simultaneously, so no phase error is expected. Other sources of misalignment such a XOR gate delay have insignificant influence on the overall performance.

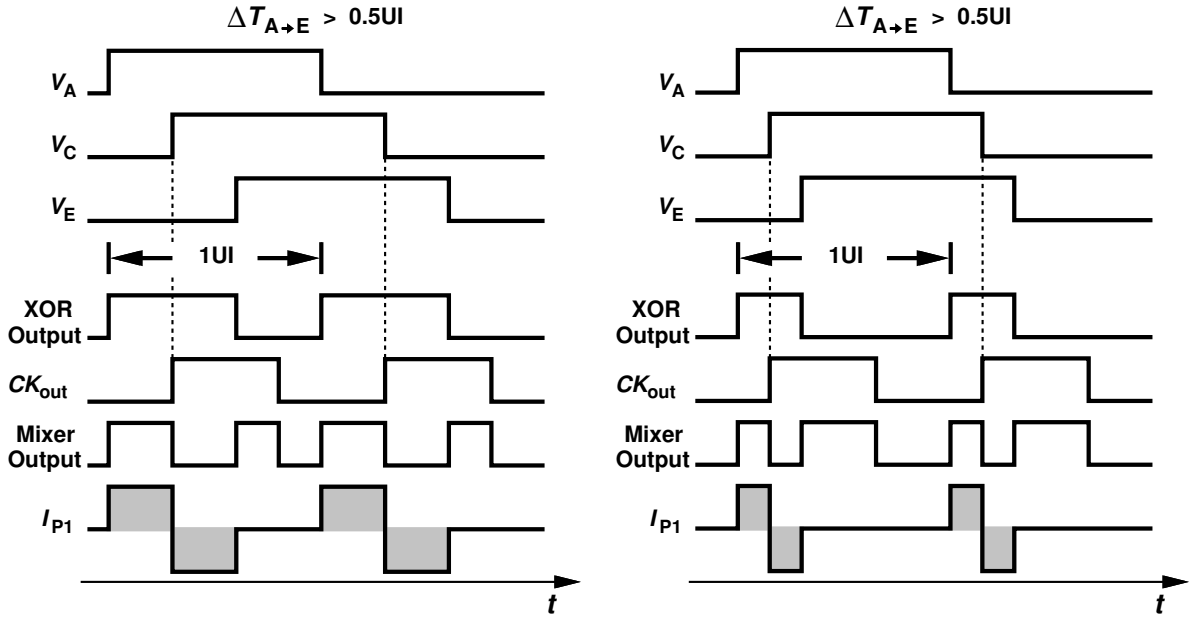


Fig. 9.13 Waveforms of important nodes as $\Delta T_{A\rightarrow E}$ deviates from 0.5 UI.

9.3 BINARY, PLL-BASED CDRS

9.3.1 Bang-Bang PD

In contrast to linear phase detection, a bang-bang PD provides binary operation. Proposed by Alexander in 1978 [xx], this type of PD only produces the polarity information (or sign bit) of the phase difference. As shown in Fig. 9.14, it can be modeled as a binary PD/CP combination injection either $+I_P$ or $-I_P$ current into the loop filter, depending on the phase error. Other components behave the same. It is obvious that such a non-linear system can not be investigated by the s-domain analysis we did for linear CDRs. The loop bandwidth, lock range, phase tracking behavior, and parameter setting are significantly different from their linear counterparts. We leave loop analysis details to chapter xx.

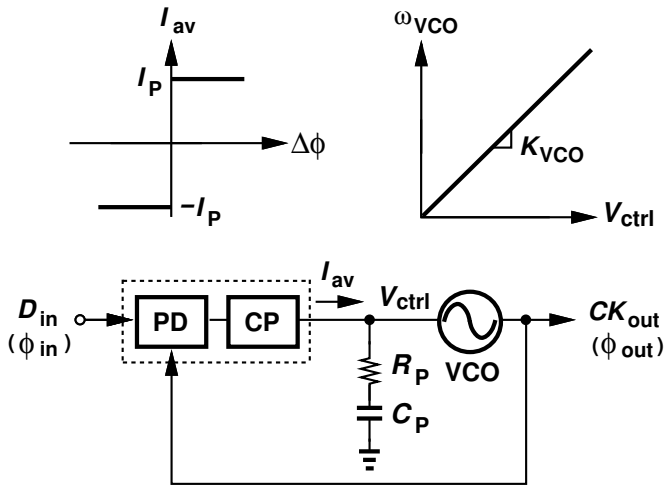


Fig. 9.14 Bang-bang CDR model.

The implementation of a binary PD is actually quite straightforward. Shown in Fig. 9.15(a) is standard Alexander PD, where 3 flipflops plus 1 latch (i.e., 7 latches in total) from a sampling sequence. Known as Nyquist sampling, the phase relationship of clock and data can be determined if there are exactly two samples on each bit. If there consecutive samples A, B ,and C [Fig. 9.15(a)] are compared, one can determine whether CK is early or late. That is, if A and B are on the same bit, CK_{in} (falling edge) is early. Otherwise, if B and C are the same, CK_{in} is late. The physical circuit implementation shifts the sampled points A and B by one and half bits, respectively, to line up with sample C. Figure 9.15(b) and (c) illustrate the details of operation. The two XOR gate outputs V_X and V_Y present (0, 1) for CK early and (1, 0) for CK late. During long runs, they are (0, 0) for sure. With proper design, using such a phase detector in a feedback loop would force the falling edge of CK_{in} to align with the transition of D_{in}, placing rising edge right in the center of data eye.

The bang-bang PD preserves quite a few good manners. It provides tri-state outputs, leaving control line undisturbed. Data retiming is automatically accomplished as well. However, the nonlinear PD behavior makes the loop analysis here quite different from the linear one that we are familiar with. The loop filter design is quite different too, requiring large loop capacitors in some applications.

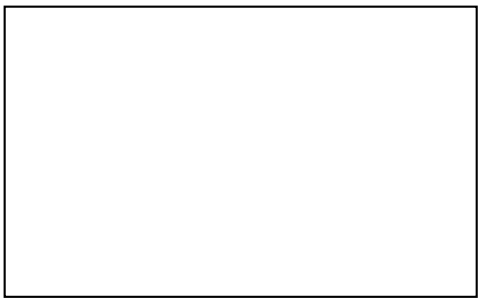


Fig. 9.15 (a) Bang-bang PD. Phase detection for (b) CK early, and (c) CK late.

The bang-bang PDs could be extended to sub-rate operation as well. Indeed, parallel processing relaxes the stringent speed requirement for samplers. The key point here is that we keep the principle of Nyquist sampling $-$ two samples per bit. Multi-phase clocks are therefore mandatory. Shown in Fig. 9.16 is an example of half-rate operation. Here, if both quadrature clocks CK_I and CK_Q are double-edge samplers, the above requirement can be fully satisfied. Figure 9.17 illustrates two examples. In Fig. 9.17(a), D_{in} is sampled in sequence by CK_I and CK_Q as A_1, A_2, and A_3 points. Giving proper delays T_b and $T_b/2$ to A_1 and A_2, respectively, the PD arrives at the same function as the full-rate one in Fig. 9.15. Such a simple realization actually skips 50% of data transitions, examining the phase relationship every $2T_b$. Figure 9.17(b) presents another example, where double-edge-triggered flipflop sample D_{in} alternately. Again, sampling points 1, 2, 3, and 4 are taken sequentially. Instead of using XOR gates, this work adopts flipflop to tell the polarity of phase error. Note that all transitions are covered in this design. The reader can prove that the PD in Fig. 9.17(b) does not provide tri-state output, potentially incurring larger jitter.

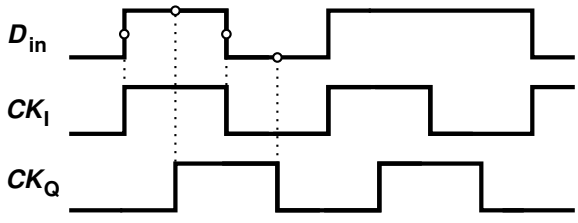
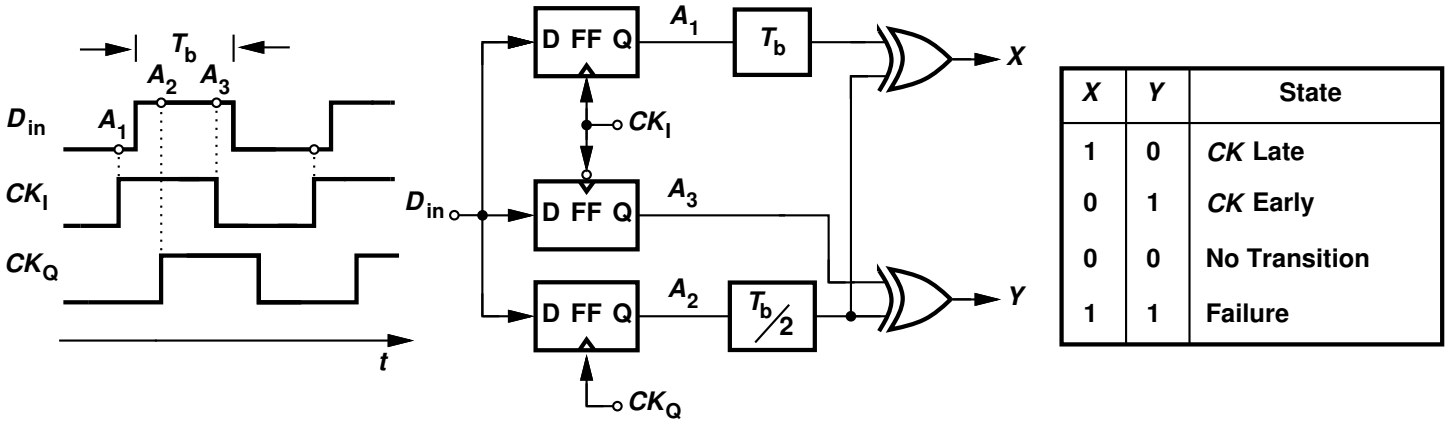
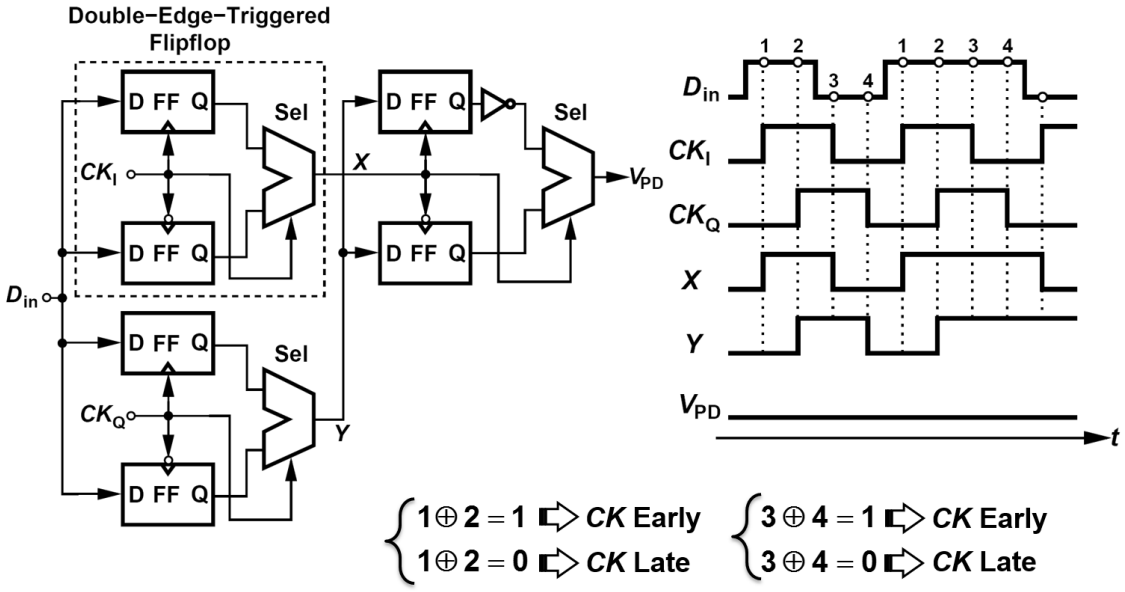


Fig. 9.16 Principle of half-rate bang-bang PD.



X	Y	State
1	0	*CK* Late
0	1	*CK* Early
0	0	No Transition
1	1	Failure

(a)



(b)

Fig. 9.17 Example of half-rate bang-bang PDs: (a) with XOR gates, (b) with a FF decider.

We introduce a quarter-rate phase detector to close the discussion of analog, PLL-based PDs. Depicted in Fig. 9.18 is a rotatory design presenting bang-bang PD operation in quarter rate. Semi-quadrature clocks separated by 45° are used, driving the 8 flipflops respectively. Originally designed to achieve 40-Gb/s operation, their work places the 8 sampled outputs $Q_1 \sim Q_8$ in a row and puts them through XOR gates un pairs. To determine the polarity of the phase error from three

consecutive samples, the outputs of two XORs are applied to a V/I converter, which produces a net current if its inputs are unequal. In lock, every other sample serves as a retimed and demultiplexed output.

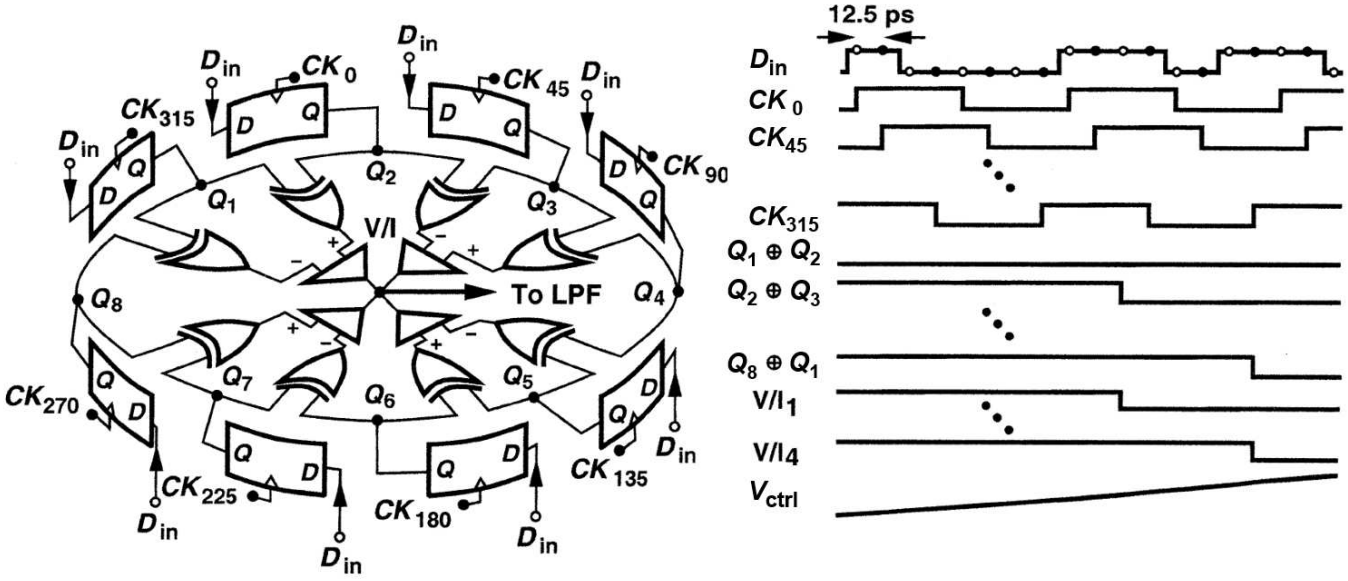


Fig. 9.18 Quarter-rate bang-bang PD.

It is important to note that, in the absense of data transitions, the FFs generate equal outputs, and each V/I converter produces a zero current, in essence presenting a tristate (high) impedance to the oscillator control.

The early-late phase detection method used here exhibits a bang-bang characteristic, forcing the CDR circuit to align every other edge of the clock with the zero crossings of data after the loop is locked.[1]

How do we analyze such an nonlinear loop? To do this, we must realize the fact that there is a finite sampling time for any sampler. As the sampling point approaches data transition, the inadequate regeneration time and noise would lead to metastability. That is, a bang-bang PD together with the charge pump actually presents a transfer function as shown in Fig. 9.19(a). There exists a finite linear region $\pm\phi_m$, and the average output current I_{av} is approximately linear inside

[1]Whether the odd-numbered or even-numbered samples are metastable depends on the polarity of the feedback around the CDR loop.

this region. We use this model to derive the closed-loop transfer function, and will revisit the detailed analysis of it in chapter xx.

Let us apply a phase variation (i.e., jitter), $\phi_{in}(t) = \phi_{in,p}\cos(\omega_\phi t)$, into the loop. If $\phi_{in,p} < \phi_m$, then the PD operates in the linear region, yielding a standard second-order system. On the other hand, as $\phi_{in,p}$ exceeds ϕ_m, the phase difference between the input and output may also rise above ϕ_m, leading to nonlinear operation. At low jitter frequencies, ϕ_{out} still tracks ϕ_{in} closely, $|\Delta\phi| < |\phi_m|$, and $|\phi_{out}/\phi_{in}| \approx 1$. As ω_ϕ increases, so does $\Delta\phi$, demanding that the V/I converter pump a larger current into the loop filter. However, since the available current beyond the linear PD region is constant, large and fast variation of ϕ_{in} results in "slewing."

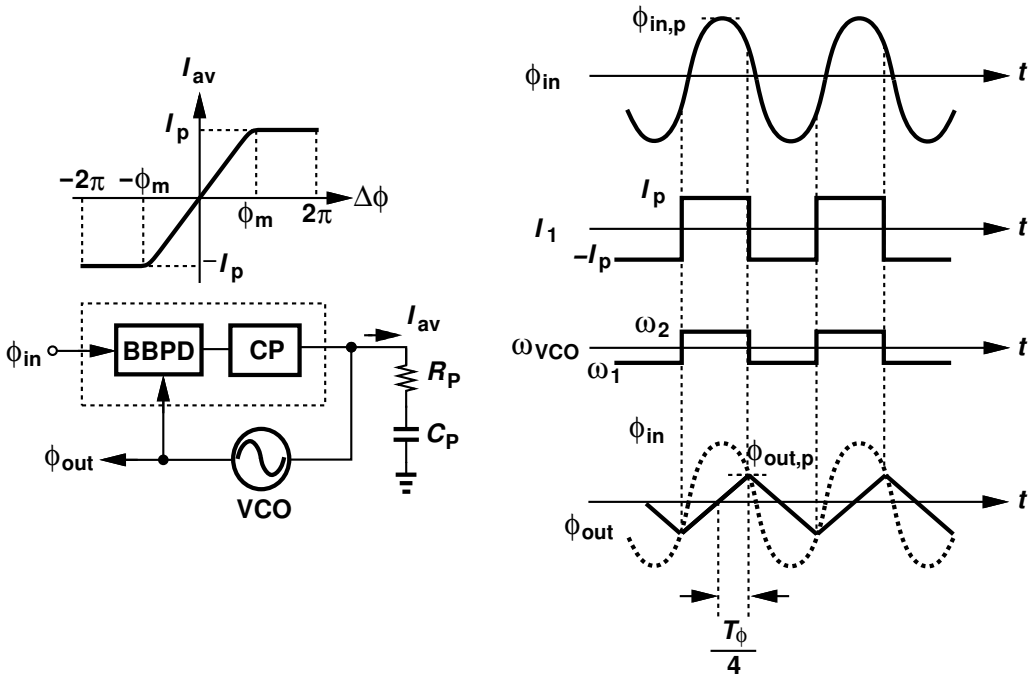


Fig. 9.19 slewing in bang-bang CDR loop.

To study this phenomenon, let us assume $\phi_{in,p} \gg \phi_m$ as an extreme case so that $\Delta\phi$ changes polarity in every half cycle of ω_ϕ, requiring that I_1 alternately jump between $+I_P$ and $-I_P$ (Fig. 9.19). Since the loop filter capacitor is typically large, the oscillator control voltage tracks $I_1 R_p$, leading to binary modulation of the VCO frequency and hence triangular variation of the output phase.

The peak value of ϕ_{out} occurs after integration of the control voltage for a duration of $T_\phi/4$, where $T_\phi = 2\pi/\omega_\phi$; that is,

$$\phi_{out,p} = \frac{K_{VCO} I_p R_p T_\phi}{4}, \tag{9.11}$$

and

$$\left| \frac{\phi_{out,p}}{\phi_{in,p}} \right| = \frac{\pi K_{VCO} I_p R_p}{2\phi_{in,p}\omega_\phi}. \tag{9.12}$$

Expressing the dependence of the jitter transfer upon the jitter amplitude, $\phi_{in,p}$, this equation also reveals a 20-dB/dec roll-off in terms of ω_ϕ. Of course, as ω_ϕ decreases, slewing eventually vanishes, Eq. (9.12) is no longer valid, and the jitter transfer approaches unity. As depicted in Fig. 9.20(a), extrapolation of linear and slewing regimes yields an approximate value for the -3-dB bandwidth of the loop:

$$\omega_{-3dB} = \frac{\pi K_{VCO} I_p R_p}{2\phi_{in,p}}. \tag{9.13}$$

It is therefore possible to approximate the entire closed-loop transfer function as

$$\frac{\phi_{out,p}}{\phi_{in,p}} = \frac{1}{1 + \dfrac{s}{\omega_{-3dB}}}. \tag{9.14}$$

Also known as "jitter transfer" in most technical documents, this transfer function is of great importance in CDR design. Fig. 9.20(b) plots the jitter transfer for different input jitter amplitudes. The transfer approaches that of a linear loop as $\phi_{in,p}$ decreases toward ϕ_m.

It is interesting to note that the jitter transfer of slew-limited CDR loops exhibits negligible peaking. Due to the high gain in the linear regime, the loop operates with a relatively large damping factor in the vicinity of ω_{-3dB}. In the slewing regime, as evident from the ϕ_{in} and ϕ_{out} waveforms in Fig. 9.19, $\phi_{out,p}$ can only fall monotonically as ω_ϕ increases because the slew rate is constant. We address this issue in chapter xx.

9.3.2 All Digital, PLL-Based CDRs

The binary-operation of bang-bang PDs manifests itself in all-digital implementation. Indeed, like all-digital PLLs, CDRs can be digitized by replacing building blocks to gain more resistance

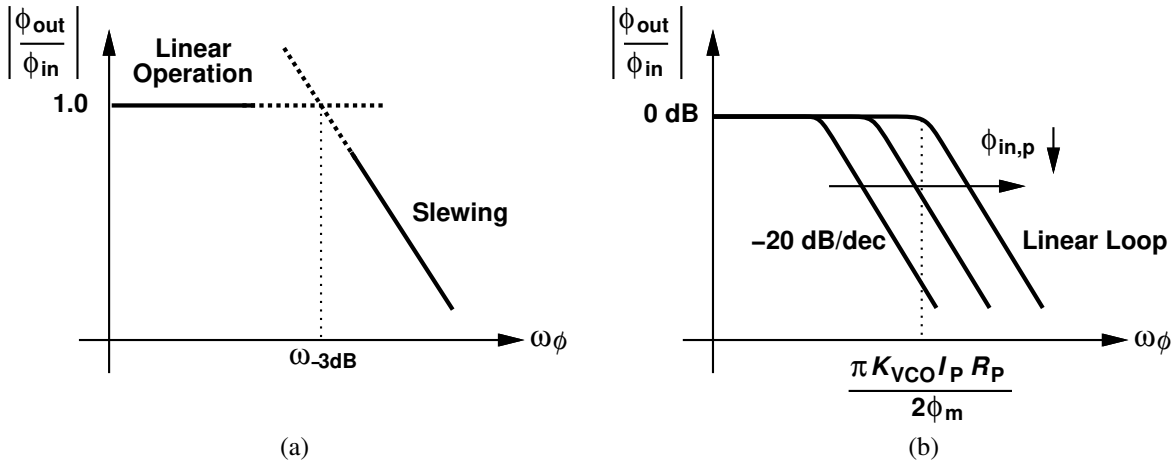


Fig. 9.20 (a) Calculation of -3dB. (b) Closed-loop transfer function of a bang-bang CDR.

to PVT variations. Standardized design minimizes the effort of migrating design from one process to another.

The output of a bang-bang PD is naturally in digital format, which can be easily processed by a digital loop filter. A straight-forward approach is to use a DAC, which converts the digital filter's output back to analog domain so as to tune the VCO. As illustrated in Fig. 9.21(a), such a approach still need an FD loop to acquire proper VCO frequency before performing phase locking. The DACs may induce quite a few issues, such as linearity, power consumption, area, and reliability. It is preferable to combine the DAC and VCO, forming a DCO [Fig. 9.21(b)].

The output of full-rate bang-bang PD is usually too fast for the subsequent digital loop filter to handle. A better approach is to deserialize the input right in the PD. Figure 9.21(c) depicts an example, where D_{in} gets sampled by sub-rate, multi-phase clocks. It still accomplish Nyquist sampling, obtaining information of data and edge for each bit. The parallelized results could be further demultiplexed, leading to a final data rate of a few hundred Mb/s. These low-speed data streams can be processed by a majority voting logic (i.e., a finite-state machine) to determine the polarity of phase error. The result feeds into DLF in parallel, which in turn drives the DCO. Multi-frequency, multi-phase clock generators must be included in the feedback loop in order to provide proper clocks for the sub-rate PD. The averaging effect alleviates possible jitter caused by clock

mismatch, and the demultiplexed data outputs are intrinsically ready for the subsequent blocks. The CDR architectures in Fig. 9.21(b) and (c) behave in the same manner as an analog bang-bang CDR.

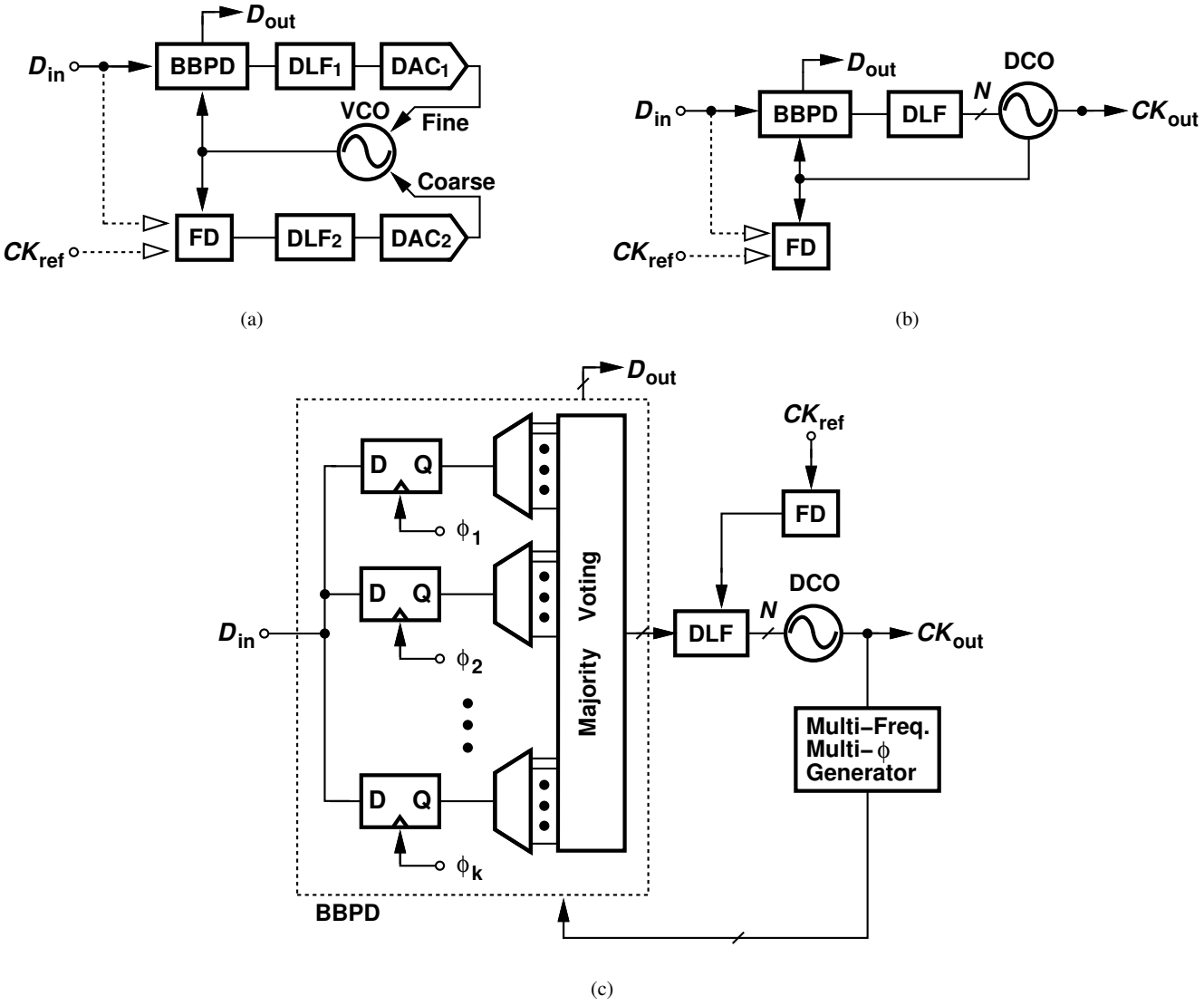


Fig. 9.21 Digital PLL-based CDRs with (a) VCO + DAC, (b) DCO, (c) interleaved BBPD.

Example 9.1

Describe the analog between the all-digital BB CDR in Fig. 9.21(c) and an analog BB CDR.

Solution: We draw the models for analog and digital CDRs in Fig. 9.22

Example 9.1(*Continued*)

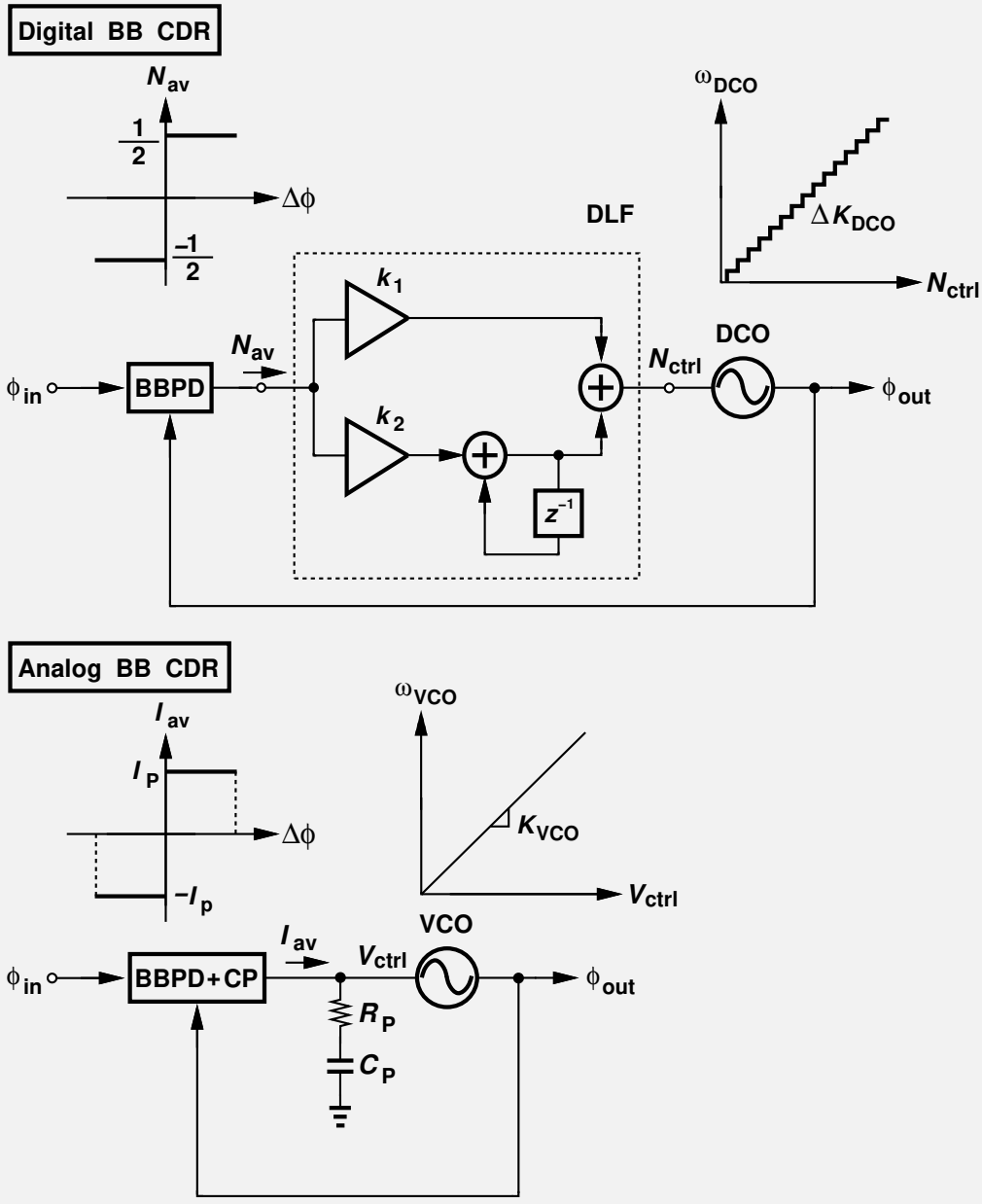


Fig. 9.22 All-digital binary CDR in analogy with its analog counterpart.

as an analogy and comparison. Note that random data sequence has 50% chance of transition between bits. The averaged characteristic of digital bang-bang PD looks the same way as the analog one, but N_{av} locates at $\pm 1/2$. Here we denote the operation period in DLF as T, and the

Example 9.1(*Continued*)

phase variation rate is much lower than its reciprocal. For the analog BB CDR, the excessive ϕ_{out} associated with the phase error is given by

$$\phi_{out} = \pm I_P \cdot (R_P + \frac{1}{sC_P}) \cdot \frac{K_{VCO}}{s}. \tag{9.15}$$

By the same token, that for the digital BB CDR is equal to

$$\begin{aligned} \phi_{out} &= \pm\frac{1}{2} \cdot (K_1 + \frac{K_2}{1 - z^{-1}}) \cdot \frac{K_{VCO}}{s} \\ &\approx \pm\frac{1}{2} \cdot (K_1 + \frac{K_2}{sT_{DLF}}) \cdot \frac{K_{VCO}}{s}, \end{aligned} \tag{9.16}$$

where the approximation holds because $ST_{DLF} \ll 1$. T_{DLF} denotes the operation (cycle) period of the digital loop filter. The two circuits in Fig. 9.22 are interchangeable.

There are some issues which must be solved in all-digital CDRs. The relatively longer latency in the digital logics may cause problems in applications where dynamic phase and frequency tracking is necessary (e.g., spread-spectrum clocks). The limited resolution for DAC or DCO could possibly introduce wandering jitters. The power consumption is not guaranteed to be smaller than its analog counterparts. Overall speaking, all-digital CDRs are still of great interest owing to the robustness and many on-going researches are under way.

The reader may wonder if linear CDRs could be digitized as well. The lack of very high-speed time-to-digital converters (specially dedicated to random data) makes the realization of all-digital linear CDRs very difficult.

9.4 DLL- AND PHASE-INTERPOLATOR BASED CDRS

In applications where data rate is given, one can simplify the CDR design by removing the frequency detection loop. The VCOs (or DCOs) could be replaced by delay lines or phase interpolators (PIs) since the frequency is known, substantially reducing the complexity. It can be found

in some backplane systems, where a global reference clock is provided to both TX and RX. Such a situation of zero frequency offset allows the use of DLL- or PI-based CDRs.

Let us consider the linear CDR shown in Fig. 9.23(a), where a voltage-controlled delay line (VCDL) is adopted to tune the clock phase. With the reference PLL providing accurate frequency, the PD loop is responsible for phase alignment only. Here, the PD + CP combination presents a linear relationship between average output current and phase error, and the VCDL is also linear with a gain of K_{VCDL}. Since the VCDL is tracking phase instead of frequency, the 1/s term for integration is removed. Thus, we simplify the loop filter as a capacitor while maintaining the stability of it. The input/output phase relationship is now given by

$$\frac{\phi_{in} - \phi_{out}}{2\pi} \times I_P \times \frac{1}{SC} \times K_{VCDL} = \phi_{out}. \tag{9.17}$$

It yields

$$\frac{\phi_{out,p}}{\phi_{in,p}} = \frac{1}{1 + \dfrac{s}{\omega_{-3dB}}}, \tag{9.18}$$

where

$$\omega_{-3dB} = \frac{K_{VCDL} I_P}{2\pi C}. \tag{9.19}$$

As expected, it is a first-order loop, which is unconditionally stable. Note that such a DLL-based CDR can not tolerate any frequency offset. The tuning range of VCDL must be wide enough to cover the whole bit period, otherwise the PD loop may become out of lock. The DLL-based CDR is rarely used individually. We look at a modified version of it in chapter xx.

A much more popular CDR structure that performs phase tracking only is to use a phase interpolator. Fortunately, with the knowledge of DLL-base CDRs, we can easily build up a model for it [Fg. 23(b)]. The linear operation of PD + CP remains the same, whereas the VCDL is replaced by a PI. The reader can prove this structure has identical transfer function as that in Fig. 9.23(a), except the PI. Again, a reference PLL must be incorporated to provide clocks.

The above PLL- or PI-based CDR architecture can also be applied to binary operation. Depicted in Fig. 9.24 is an example, where bang-bang PD is employed. Providing either $+I_P$ or $-I_P$ to the loop filter C, the BBPD + CP combination discloses only the polarity of phase error. With

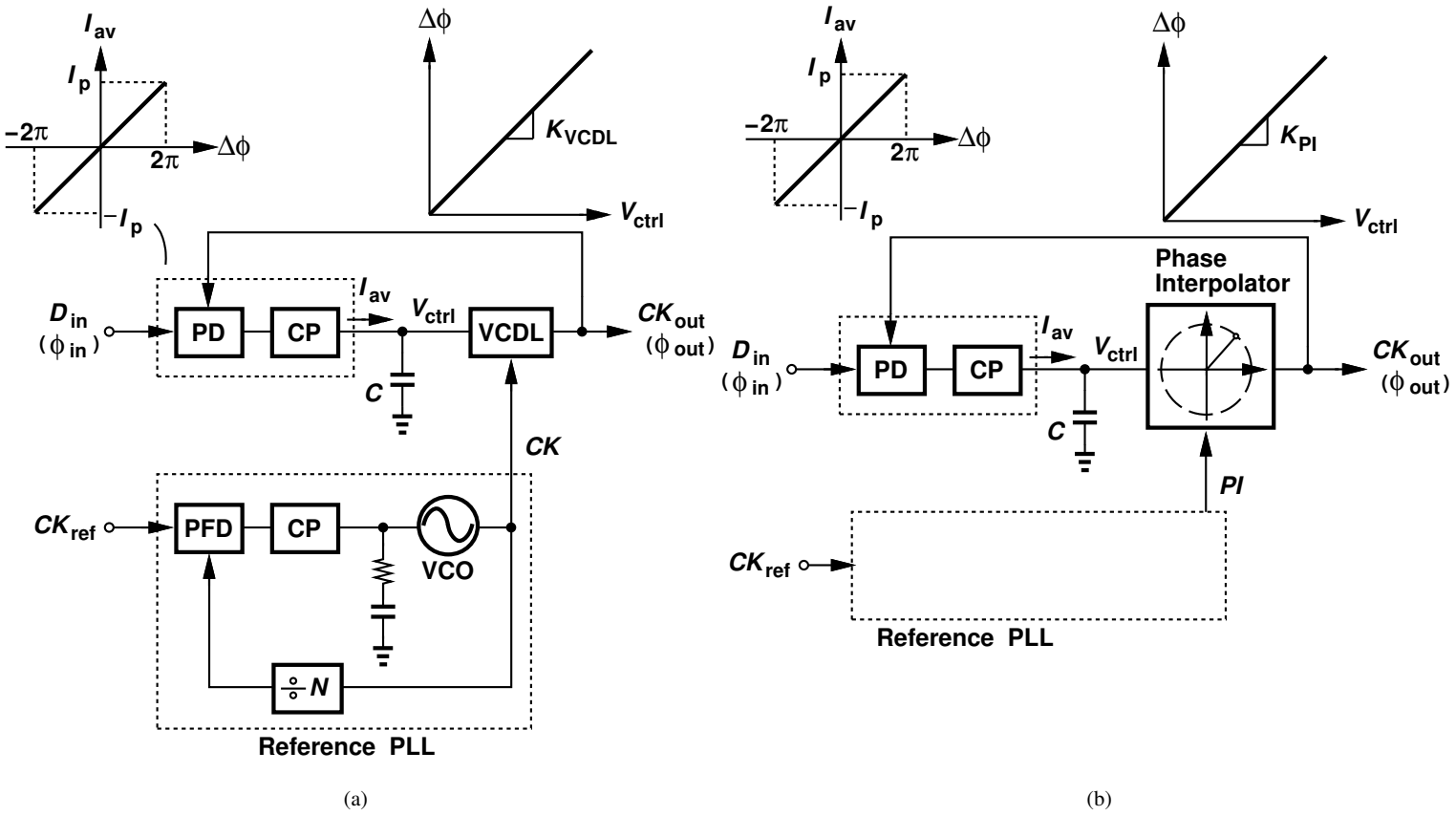


Fig. 9.23 Linear CDRs based on (a) DLL, (b) phase interpolator.

the linear relationship of PI, this loop would still lock and align the phase of data and clock as a PLL-based bang-bang CDR. Note that BBPD need not be full rate. Sub-rate PDs can serve here as well.

It is instructive to investigate the loop behavior of the binary PI-based CDR in Fig. 9.24. We assume a sinusoidal phase modulation ϕ_{in} applies to the input data. Similar to that of bang-bang, PLL-based CDRs, the output phase ϕ_{out} follows ϕ_{in} tightly as the modulation is slow. As the modulation frequency increases, ϕ_{out} gradually fails to follow ϕ_{in} and slewing occurs. We also denote ϕ_{in} magnitude as $\phi_{in,p}$, where $\phi_{in,p}$ is much greater than the linear region (i.e., $\phi_{in,p} \gg \phi_m$).

At slewing, an equivalent $\pm I_P$ continuously pumps into C, leading to V_{ctrl} changing rate as

$$\frac{\mathrm{d}V_{ctrl}}{\mathrm{d}t} = \pm\frac{I_P}{C}. \tag{9.20}$$

It corresponds to output phase magnitude $\phi_{out,p}$ as

$$\phi_{out,p} = K_{PI} \cdot \frac{I_P}{C} \cdot \frac{T_\phi}{4}, \tag{9.21}$$

where T_ϕ represents the modulation period. As a result,

$$\left|\frac{\phi_{out}}{\phi_{in}}\right| = \frac{\pi K_{PI} I_p}{2C\phi_{in,p}\omega}, \tag{9.22}$$

in the slewing region. Once again, the intersection point of 0 dB and $1/\omega$ lines stands for the -3-dB bandwidth:

$$\omega_{-3dB} = \frac{\pi K_{PI} I_p}{2C\phi_{in,p}\omega}, \tag{9.23}$$

which is also inversely proportional to the amplitude of input.

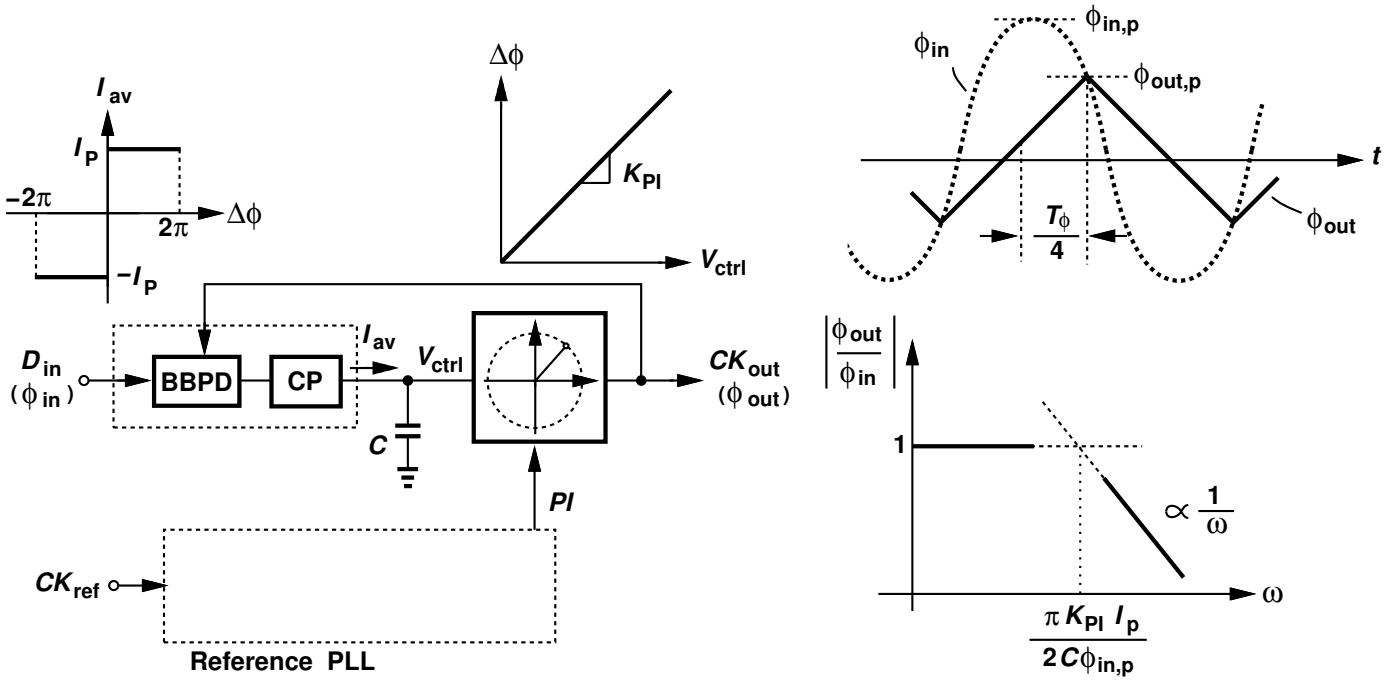


Fig. 9.24 Bang-bang CDR based on phase interpolator.

The foregoing structures relies on analog PIs, which intrinsically has finite tracking range (after all, $0 \leq V_{ctrl} \leq V_{DD}$). Digitizing the loop filter and PI would lead to infinite tracking range. The DLF can simply serve as a counter, which restarts from zero at overflow. Figure 9.25 illustrates such an all-digital design. Now the PI is tuned is discrete mode, whose conversion gain can still be

denoted as K_{PI}. Since the PI's input is a unit-less number (N_{ctrl}), K_{PI} is in the unit of rad. The DLF also degenerates to an accumulator, in which K_1 path is removed. At slow phase modulation, ϕ_{out} flows ϕ_{in} nicely. Depending on the PI's resolution, temporary phase error can be found. In typical designs, PIs would be realized in $6 \sim 8$ bits, limiting the error to a few degrees. Under slewing, ϕ_{out} would track ϕ_{in} with its best effort, arriving at a shape of stairway (Fig. 9.25). The following example derives more details of it.

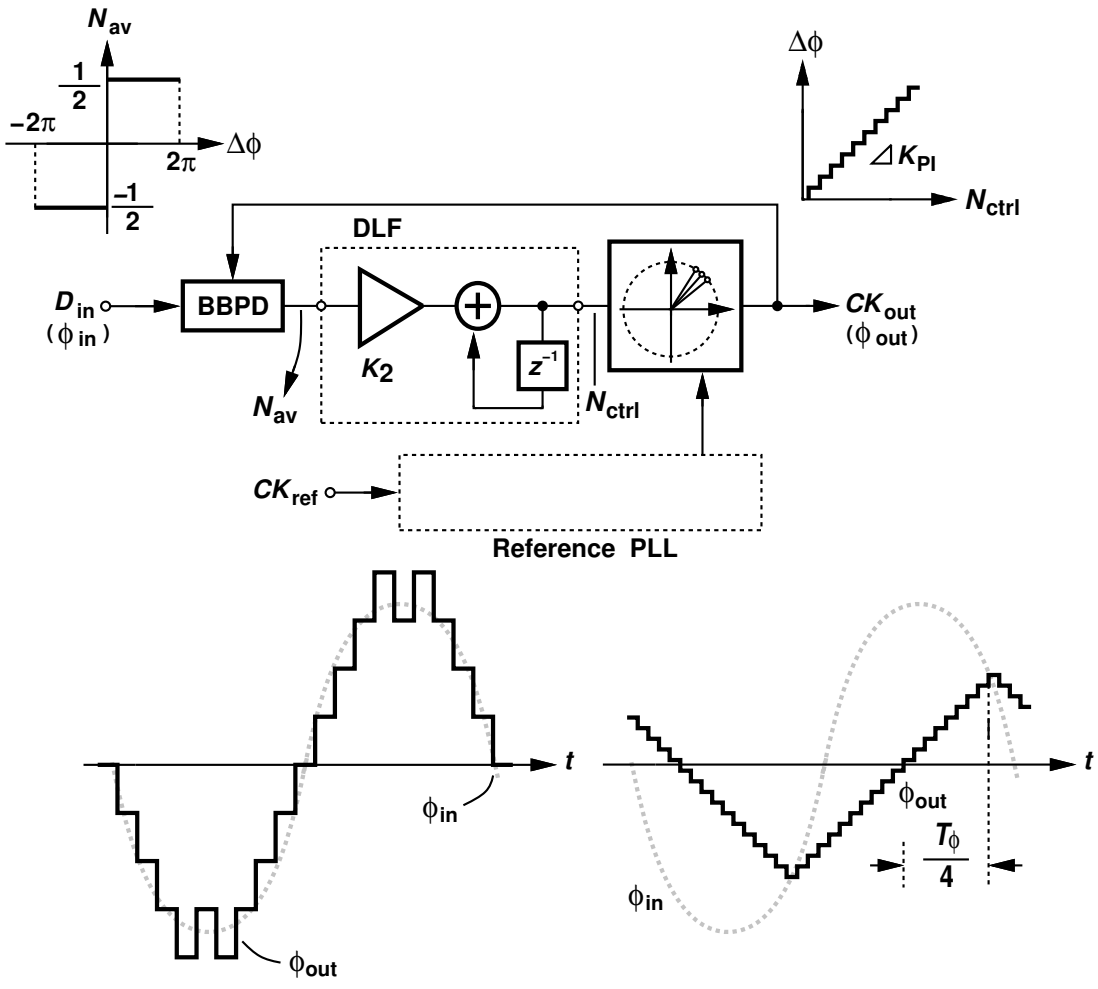


Fig. 9.25 All-digital bang-bang CDR based on phase interpolator.

Example 9.2

Suppose the refresh rate of DLF in Fig. 9.25 is $1/T_{DLF}$ determine the loop bandwidth of it.

Solution: In slewing, the DLF's output N_{ctrl} increases or decreases itself by K_2 every T_D seconds. For a period of $T_\phi/4$, it translates to a total phase accumulation of $T_\phi \cdot K_2 \cdot K_{PI}/(8T_{DLF})$. As a result, the rolling-off region is given by

$$\left|\frac{\phi_{out}}{\phi_{in}}\right| = \frac{\pi K_{PI} K_{PI}}{4\phi_{in,p} T_{DLF} \omega}, \tag{9.24}$$

implying the loop bandwidth as

$$\omega_{-3dB} = \frac{\pi K_{PI} K_{PI}}{4\phi_{in,p} T_{DLF}}. \tag{9.25}$$

Figure 9.26 depicts the transfer function, which is similar to that in Fig. 9.24.

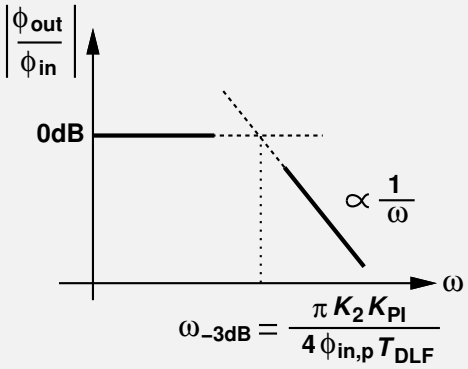


Fig. 9.26 Calculation of closed-loop transfer function of all-digital, binary, PI-based CDRs.

In reality, the finite resolution of digital PIs would lead to phase wandering around the right position. Although the data jitter performance might not be degraded significantly, clock itself suffers from larger jitter inevitably. Methods to alleviate this issue can be found in [xx].

Example 9.3

Consider the PI-based all-digital CDR shown in Fig. 9.27, which employs a second-order DLF. Derive the transfer function (or equivalently, JTRAN) of it.

Example 9.3(*Continued*)

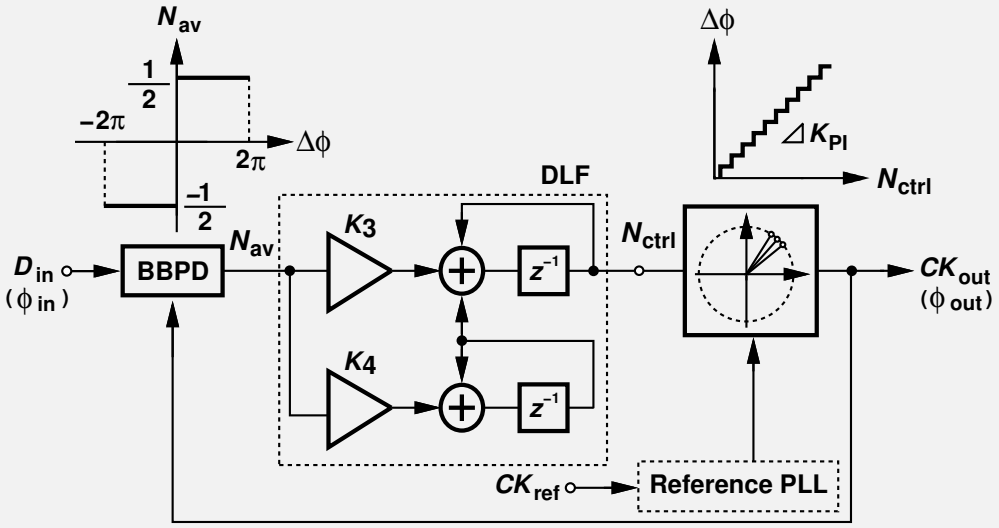


Fig. 9.27 Calculation of closed-loop transfer function of all-digital, binary, PI-based CDRs.

Solution: The 2nd-order DLF presents a transfer function of

$$\frac{N_{out}}{N_{in}} = \frac{K_3 z^{-1}}{1 - z^{-1}} + \frac{K_4 z^{-2}}{(1 - z^{-1})^2}, \tag{9.26}$$

which can be approximated to s-domain as

$$\begin{aligned} \frac{n_{out}}{n_{in}}(s) &= \frac{K_3(1 - sT_{DLF})}{sT_{DLF}} + \frac{K_4(1 - sT_{DLF})^2}{s^2 T_{DLF}^2} \\ &\approx \frac{K_3}{sT_{DLF}} + \frac{K_4}{s^2 T_{DLF}^2}. \end{aligned} \tag{9.27}$$

Again, T_{DLF} denotes the operation time of DLF and $ST_{DLF} \ll 1$. While the K_3 term still serves as a linear phase movement, the K_4 term implies a 2nd-order (parabolic) phase tracking. Drawing ϕ_{in} and ϕ_{out} under slewing in Fig. 9.28, we obtain the output phase by taking inverse Laplace transform:

$$\phi_{out} = \frac{1}{2} \cdot [\int \frac{K_3}{T_{DLF}} dt + \int \int \frac{K_4}{T_{DLF}^2} d\tau dt] \cdot K_{PI}. \tag{9.28}$$

It follows

$$\begin{aligned} \phi_{out,p} &= \frac{1}{2} \cdot \frac{1}{2} \cdot K_{PI} \cdot (\frac{K_3}{T_{DLF}} t + \frac{1}{2} \cdot \frac{K_4}{T_{DLF}^2} t^2) \Big|_0^{T_\phi/2} \\ &= \frac{\pi K_3 K_{PI}}{4 T_{DLF} \omega_\phi} + \frac{\pi^2 K_4 K_{PI}}{8 T_{DLF}^2 \omega_\phi^2}. \end{aligned} \tag{9.29}$$

Example 9.3(*Continued*)

Thus, for a given $\phi_{in,p}$, $\phi_{out,p}$ remains equal to $\phi_{in,p}$ until -3-dB bandwidth ω_{-3dB}. The rolling-off region starts with -40 dB/dec and migrates to -20 dB/dec afterwards. The -3-dB bandwidth is approximately given by

$$\omega_{-3dB} = \frac{\pi}{2T_{DLF}}\sqrt{\frac{K_4 K_{PI}}{2\phi_{in,p}}}, \tag{9.30}$$

which is inversely proportional to $\phi_{in,p}^{1/2}$. The intersection of -40 and -20 dB/dec regions ω_1 is also available by equating the two terms of Eq. (9.29):

$$\omega_1 = \frac{\pi K_4}{2T_{DLF}K_3}. \tag{9.31}$$

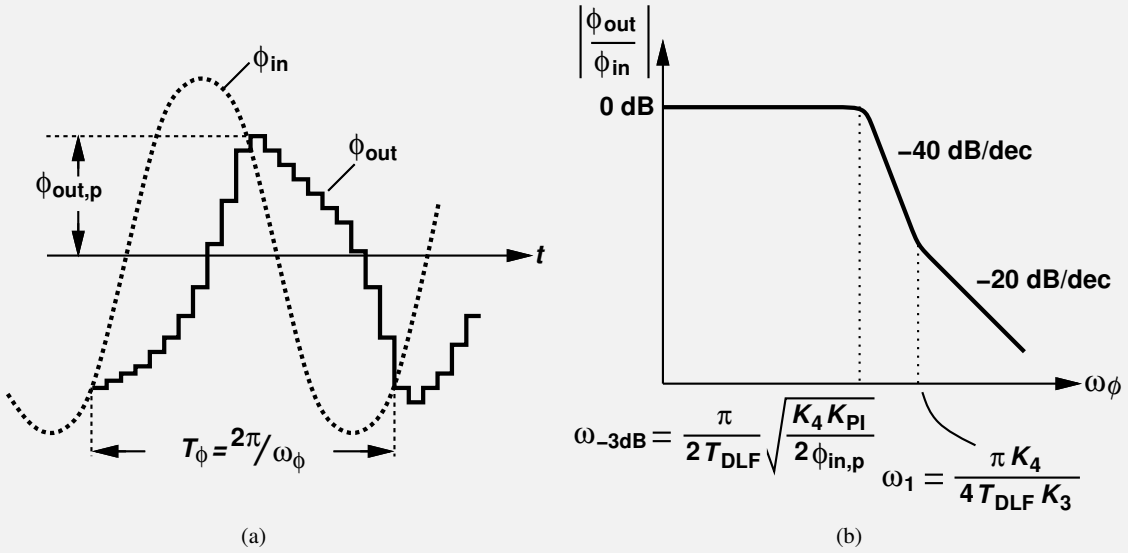


Fig. 9.28 Calculation of closed-loop transfer function of all-digital, binary, PI-based CDRs.

9.5 OVER-SAMPLING CDRS

The CDR architectures introduced in the foregoing sections are classified as Ntquist sample CDRs. If more than two samples are available per bit, we arrive at an oversampling structure. Figure 9.29 illustrates a standard design of it. Driven by a multiphase clock generator, the multiphase

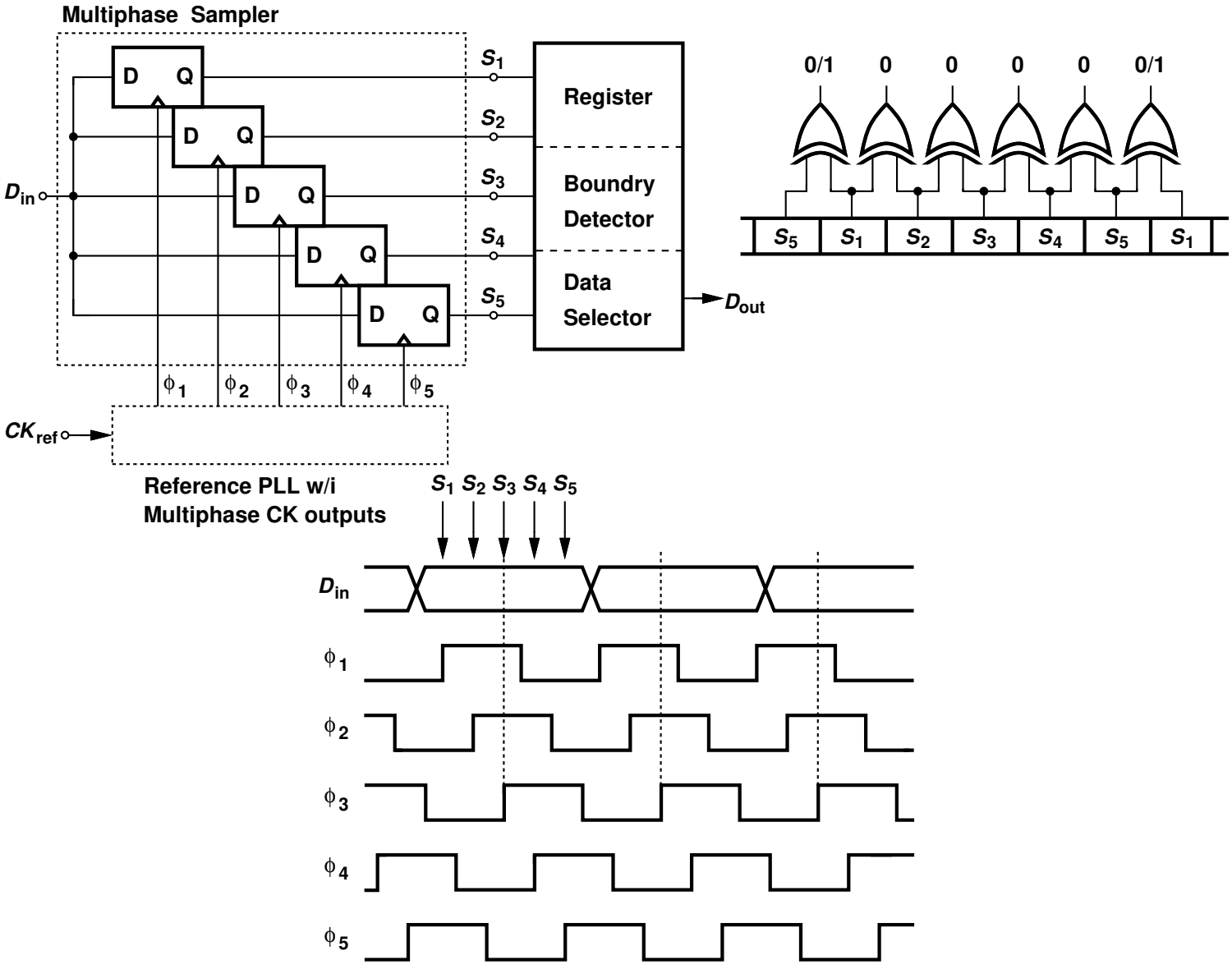


Fig. 9.29 Over-sampling CDR.

sampler fronted samples the incoming data at least 3 times per bit.[2] Here, we have 5 samples per bit as an example. The outputs are sent to a large first-in-first-out (FIFO) register, where a series of logics examines the data transition, i.e., the boundaries of a bit. For example, S_1 and S_5 may experience different results, causing the corresponding XOR gate to produce a 1. The middle one

[2]Odd number is preferable.

S_3 naturally represents the final data output, as it stays far away from the edges on both sides. The clock generator can be realized as a PLL with multiphase clock outputs.

The over-sampling CDRs present several advantages. It's feedforward structure avoids the use of feedback loop, arriving at fast acquisition and inherent stability. The processor behind the fronted sampling head provides wide bandwidth of operation. However, it bears a few issues as well. It requires a long FIFO register or data buffer in order to store the sampled data. CML: logics must be performed to deal with high-speed data, which consumes significant power. Parallelism can be applied here, but the circuit complexity increases as well. Finally, finite frequency offset may occur if the TX and RX are driven by different clock sources. It is obvious that the frequency offset between data rate and clock leads to continuous changing of the samplers positions. If the logics can not accommodate high-speed operation, errors may occur.

In applications where reference clock is not available, the CDR itself must explore the data rate as well. Illustrated in Fig. 9.30(a) is an example realized with digital blocks. A TDC examines the data transition, allowing the data-rate detector to calculate the data rate. Due to the finite TDC resolution, the detected data rate may still have finite residue (i.e., offset). This offset must be small

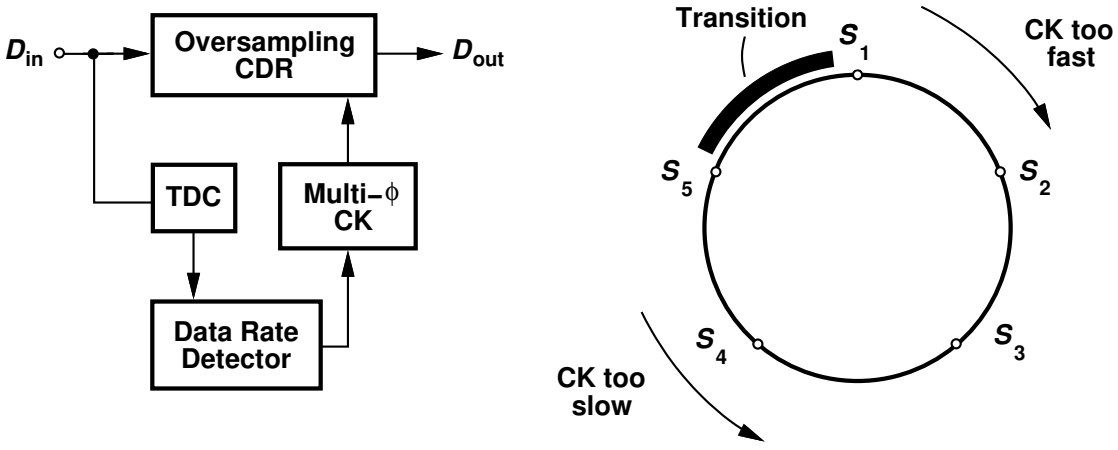


Fig. 9.30 Data-rate detection for over-sampling CDRs.

enough in order to prevent sampling errors in the final output data. The frequency acquisition can also be implemented in analog domain, just as we will see in section xx.xx. Nonetheless, Fig. 9.30(b) reveals the frequency tracking principle. If the transition (gradually) shifts from S_5-S_1 to S_1-S_2, the clock is too fast. Otherwise, it is too slow. Frequency acquisition can be accomplished simply by knowing the polarity of frequency error.

9.6 BURST-MODE CDRS

The over-sampling CDRs can lock the phases of clock an data in a rapid way, given that the accurate data rate is available. For long-distance applications, however, the RX can only generate a frequency locally (e.g., with a cryatal) with finite offset. If the RX is required to present instantaneous phase locking, we usually resort to other technique, i.e., burst-mode CDRs. The most typical application for immediate phase and frequency locking is the so-called passive optical networks (PONs). Figure 9.31 illustrates such a system, where the optical line terminal (OLT) must deal with asynchronous packets with different amplitudes and lengths during upstream mode. It needs clock and data recovery (CDR) circuits with immediate clock extraction and data retiming. Unlike synchronous optical network (SONET) systems that impose strict specification on jitter transfer, the above applications have no or few repeaters in their data paths. It allows us to trade the loop bandwidth with fast locking in phase and frequency. We introduce two approaches in this section.

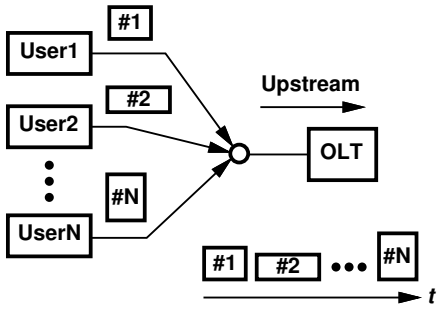


Fig. 9.31 PON system.

9.6.1 Gated-VCO Technique

Burst-mode operation means the incoming data sequences are interleaved by very long idle time (on the order of us). A burst-mode CDR is required to respond and lock to the incoming data within a few bits (i.e., "preamble"). Structures with gated-VCOs are popular in recent development. Shown in Fig. 9.32 is an example, in which two identical ring oscillators VCO_1, and VCO_2 are incorporated. Governed by the same control voltage (V_{ctrl} and V_{ctrl}'), the two oscillators run at the same frequency. With the local reference PLL, they are oscillating at the frequency of data rate nominally. The unity-gain buffer provides good isolation and prevents V_{ctrl} from disturbance. The delay cell and XOR gate produce a pulse sequence upon arrival of D_{in}. During long runs, $V_{XOR} = 0$, making $CK_{out} = 1$. As a pulse appears at V_{XOR}, CK_{out} falls after a certain period of time (i.e., gate delay of NAND gate and two inverters). Since the delay cell provides a delay

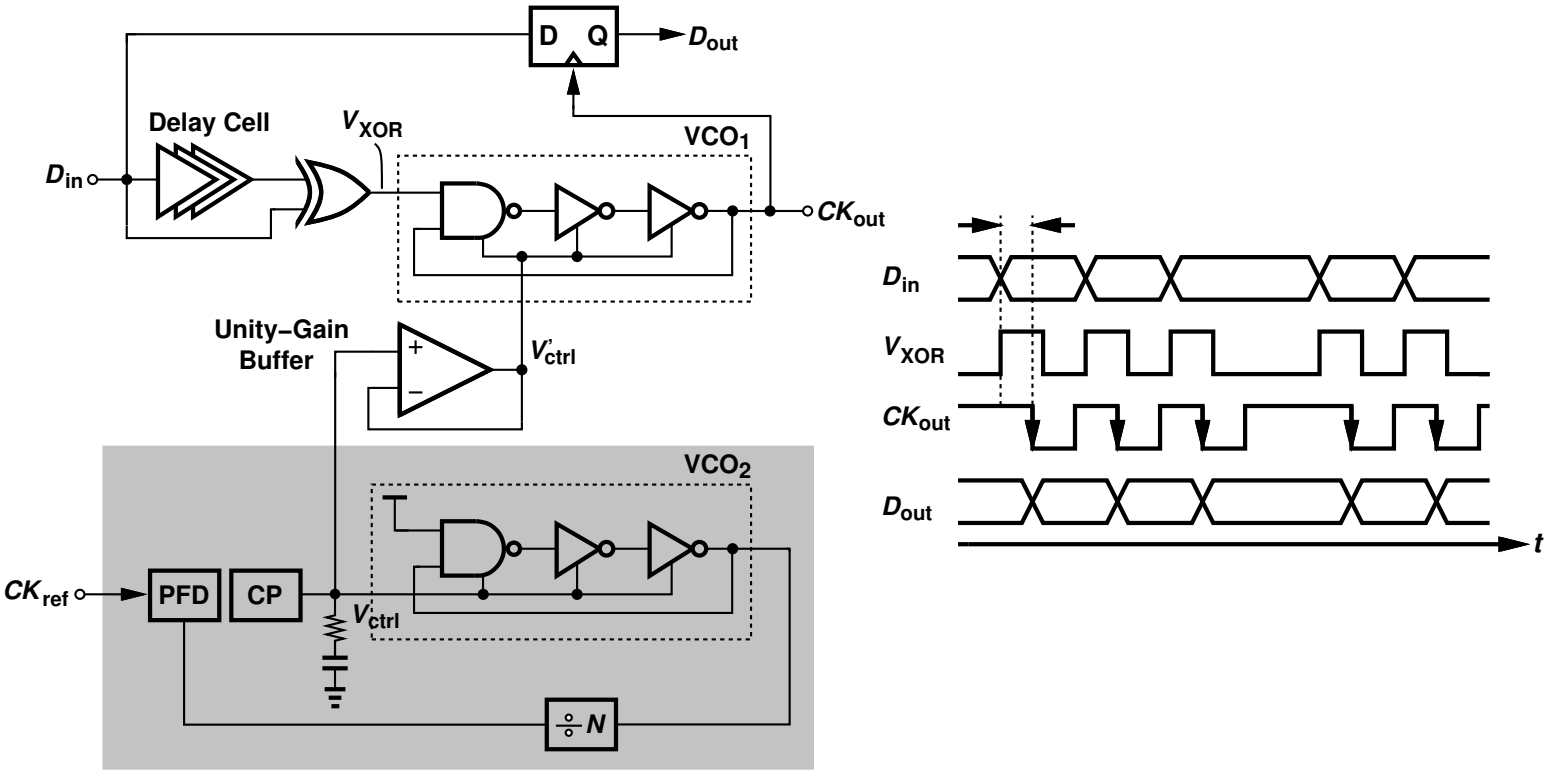


Fig. 9.32 Burst-mode CDRs with gated VCO.

roughly 0.5 UI, a cycle repeats itself at CK_{out} as long as data transit. In other words, the NAND gate "blocks" the clock during long runs, and "admits" it at data transition. The falling edge of CK_{out} samples D_{in}, arriving at retimed output D_{out}.

The burst-mode CDRs using gated VCOs inevitably suffers from some issues. The ring oscillator results in higher phase noise and lower operation speed. More seriously, the gating behavior would cause momentary fluctuation on the recovered clock, potentially incurring undesired jitter and intersymbol interference (ISI). In addition, the truncation or prolongation of the clock cycle during phase alignment induces other uncertainties such as locking (settling) time.

9.6.2 Injection-Locking Technique

Injection-locking technique has found tremendous usage in many applications of communication, as we described in the previous chapter. Here we introduce a CDR architecture utilizing injection locking to achieve ultra-fast locking. Figure 9.33 illustrates such a design, where the input D_{in} and its delayed replica D'_{in} are XORed to create pulses upon occurrence of data transition.

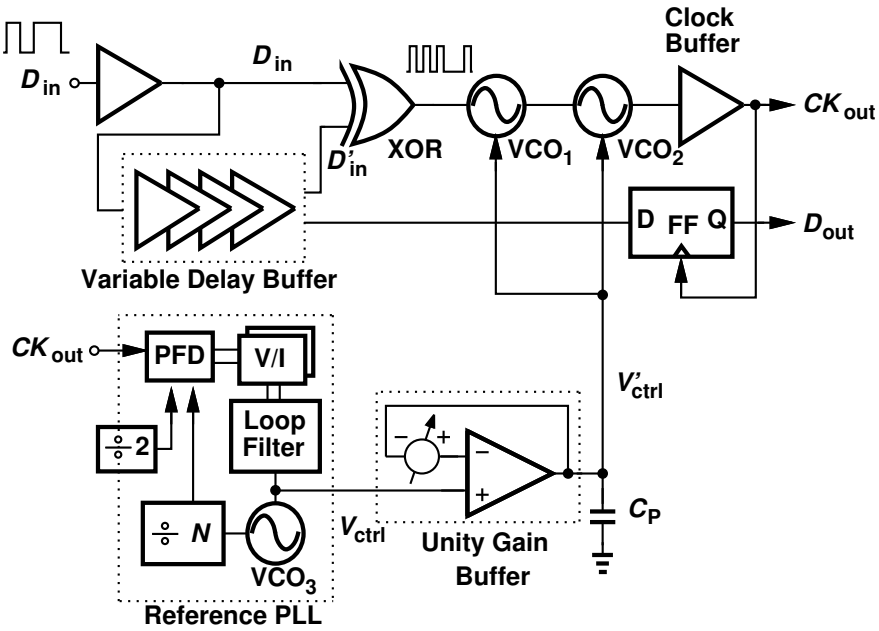
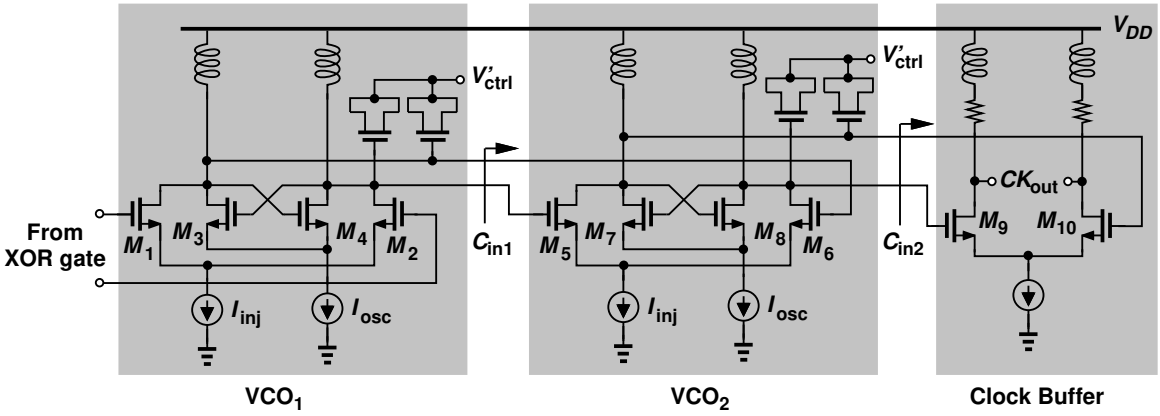


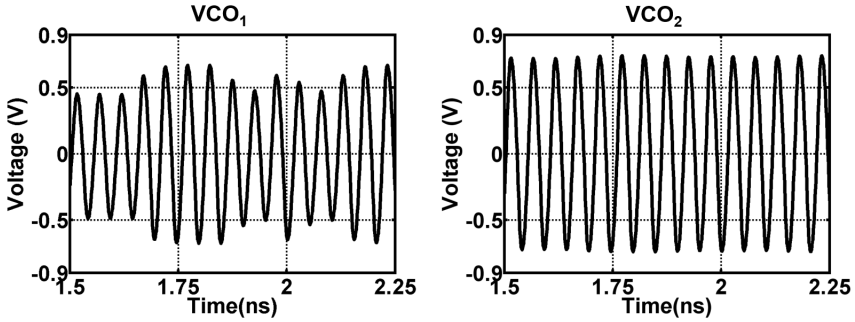
Fig. 9.33 Burst-mode CDRs using injection-locking technique.

Again, a pulsewidth of half bit period is generated to achieve an optimal injection to VCO_1. Two identical oscillators, VCO_1 and VCO_2, are coupled in cascade to purify the clock. In contrast to the gating circuits, this two-stage coupling ensures a constant amplitude in output clock CK_{out}, and suppresses more noise by the filtering nature of the LC tanks. The reference PLL, consisting of another duplicated VCO (i.e., VCO_3) and a divider chain of modulus N, produces a control voltage V_{ctrl} for VCO_1 and VCO_2.

For our discussion 9.1, we realize that spectral line at data rate would be created at output of XOR gate. Applying this pulse sequence into $VCO_{1,2}$ forces them to be injection locked to the exact data rate momentarily. As a result, it provides an instantaneous clock locking and retimes the data without any latency. It can be demonstrated the locking time is less than 1 UI.



(a)



(b)

Fig. 9.34 (a) Cascading LC-tank VCOs, (b) clock purification.

It is important to know that without the servo PLL, PVT variations would easily deviate the VCO natural frequency from the desired value and make the CDR out of lock. The VCO and buffer design is shown in Fig. 9.34(a), where the injection pairs $M_{1,2}$ and $M_{5,6}$ translate the input signal into current to lock the oscillator. Fig. 9.34(b) shows the output waveform of the two VCOs injection-locked to a PRBS of $2^7 - 1$. VCO_2 oscillates with almost uniform magnitude, since VCO_1 still swings during long runs.

Example 9.4

Suppose a purely random data sequence is fed into the burst-mode CDR shown in Fig. 9.33. Calculate the deterministic clock rms jitter duo to finite frequency offset.

Solution: To quantify the jitter, we define the frequency deviation Δf as

$$\Delta f = f_b - M \cdot f_{ref}, \tag{9.32}$$

where $f_b = 1/T_b$ denotes the data rate, f_{ref} the reference frequency, and M the corresponding divide ratio. Since Δf is typically much less than f_b, the clock zero crossing shifts $\Delta f/f_b$ UI per bit period during long runs [positions 3, 6, and 7 in Fig. 9.35]. Here we assume the clock zero crossing aligns to data transition immediately whenever it occurs (positions 1, 2, 4, 5, and 8). For N consecutive bits, the phase error accumulates up to $(N-1)\Delta f/f_b$ in the last bit, and a bit error would occur if it exceeds 0.5 UI. That is, in the presence of frequency offset, the maximum tolerable length of consecutive bits is given by

$$N_{max} = \frac{1}{2} \cdot \frac{f_b}{\Delta f} + 1. \tag{9.33}$$

It is of course an optimistic estimation since VCO's phase noise would deteriorate the result considerably.

Moreover, for a random sequence, the probability of occurring a phase deviation of $n\Delta f/f_b$ is equal to $2^{-(n+1)}$. Fig. 9.35 illustrates the probability distribution. That is, the clock zero-crossing points accumulate at equally-spaced positions with different probabilities, and the average position is therefore given by

$$\frac{\Delta f}{f_b} \sum_{n=0}^{\infty} \frac{n}{2^{n+1}} = \frac{\Delta f}{f_b}. \tag{9.34}$$

Example 9.4(*Continued*)

The rms jitter due to this effect can be obtained as

$$\begin{aligned} J_{rms} &= [(-1)^2 \cdot \frac{1}{2} + 0^2 \cdot \frac{1}{4} + 1^2 \cdot \frac{1}{8} + 2^2 \cdot \frac{1}{16} + \cdots]^{1/2} \cdot \left|\frac{\Delta f}{f_b}\right| \\ &= (\frac{1}{2} + 0 + \sum_{n=1}^{\infty} n^2 \cdot \frac{1}{2^{n+2}})^{1/2} \cdot \left|\frac{\Delta f}{f_b}\right| \\ &= \sqrt{2} \cdot \left|\frac{\Delta f}{f_b}\right|. \end{aligned} \tag{9.35}$$

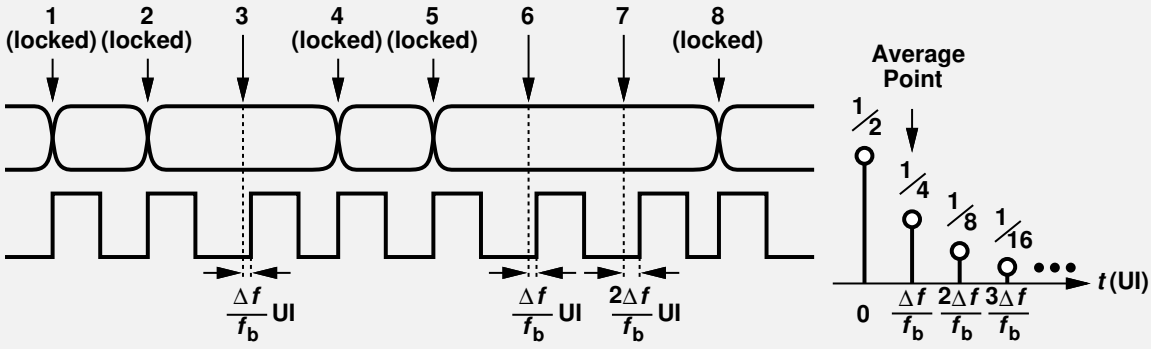


Fig. 9.35 Calculation of .rms clock jitter due to frequency offset.

CLOCK AND DATA RECOVERY (II) 10

With the fundamental CDR knowledge developed in Chapter 9, we are now ready for advanced features. We begin our discussion on capture range of different PDs, and investigate frequency acquisition techniques. Three important jitter specifications, namely, jitter transfer (JTRAN), jitter tolerance (JTOL), and jitter generation (JG) would be studied thoroughly.

10.1 CAPTURE RANGE

As mentioned in Chapter 9, all of the existing PD solutions have finite operation range due to the random data input. We investigate the capture range in this section. By definition, the capture range of a phase locking CDR is the range of input data rate over which the CDR can capture and lock itself to the input data. It refers to the capability of capturing the input data rate from an initial clock frequency which deviates by a certain offset. It is of great importance for a PLL-based CDR design, as the frequency detection loop needs to know how close it should bring the clock frequency to the data rate. The reader should not confuse the capture range with lock range. Also known as tracking range, the lock range of a CDR is defined as the range of data rate over which the CDR can gradually track the data rate and remain in lock. The lock range is roughly the operation range that we mention for a CDR, and it is usually wider than the capture range (Fig. 1). Note that it is not difficult to check the capture range in testing. By putting the CDR in stable locking condition, we could jump the data rate and see how much it can tolerate. Similarly, by gradually tuning the data rate, the lock (tracking) range can be obtained. Frequency acquisition loop must be switched off while conducting these experiments.

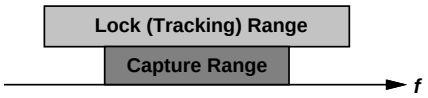


Fig. 10.1 Definition of Capture range and Lock range.

We first look at the capture range of linear, Pll-based CDRs. Fig. 2(a) redraws the model for the sake of convenience. Recall the linear transfer function is given by

$$\frac{\Phi_{out}}{\Phi_{in}}(s) = \frac{2\zeta\omega_n s + \omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}, \tag{10.1}$$

where

$$\omega_n = \sqrt{\frac{I_p K_{VCO}}{2\pi C_p}} \tag{10.2}$$

$$\zeta = \frac{R_p}{2}\sqrt{\frac{I_p C_p K_{VCO}}{2\pi}}. \tag{10.3}$$

Suppose the loop is locked properly for t < 0, and the data rate ($\omega_{DR} = 2\pi R_D = 2\pi/T_b$) jumps abruptly from ω_{DR} to $\omega_{DR} + \Delta\omega$ at t = 0. The output phase tracks the sharper curve of $(\omega_{DR} + \Delta\omega)t$ immediately in order to minimize the phase error and get back to lock [Fig. 2(b)]. However, for the loop to relock, the maximum phase deviation $\Delta\Phi_{max}$ must not exceed 2π. Since $\Phi_{in}(t) = (\omega_{DR} + \Delta\omega)t$, we obtain $\Phi_{out}(t)$ by taking the inverse Laplace transform

$$\Phi_{out}(t) = \frac{\Delta\omega}{2\omega_n\sqrt{\zeta^2 - 1}}(e^{k1t} - e^{k2t}) + (\omega_{DR} + \Delta\omega)t, \tag{10.4}$$

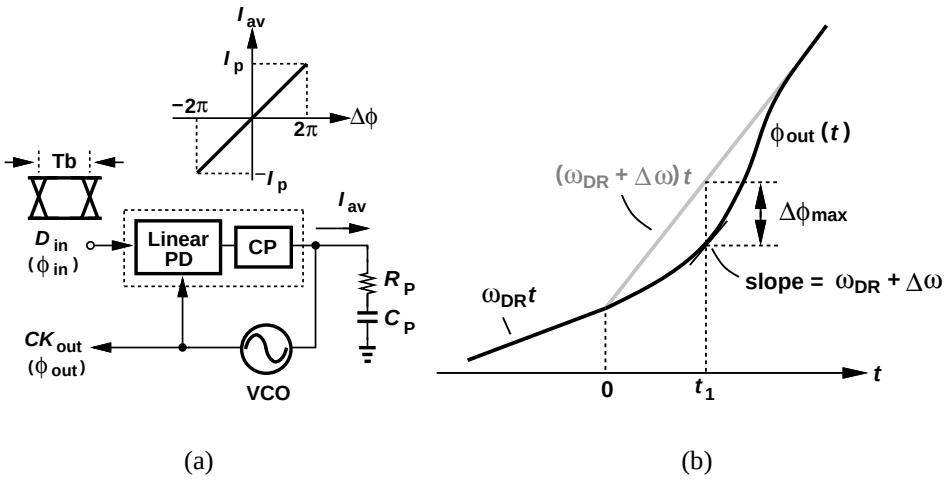
where $k_1 = -\omega_n(\zeta + \sqrt{\zeta^2 - 1})$ and $k_2 = -\omega_n(\zeta - \sqrt{\zeta^2 - 1})$.

It can be clearly shown that $\Delta\Phi_{max}$ occurs at t_1, where $d\Phi_{out}/dt = \omega_{DR} + \Delta\omega$.
It follows that

$$t_1 = \frac{1}{2\omega_n\sqrt{\zeta^2 - 1}} \ln\left[\frac{\zeta + \sqrt{\zeta^2 - 1}}{\zeta - \sqrt{\zeta^2 - 1}}\right]. \tag{10.5}$$

To ensure relocking, we must have $\Delta\Phi_{max} = (\omega_{DR} + \Delta\omega)t_1 - \Phi_{out}(t_1) < 2\pi$.
that is,

$$\Delta\Phi_{max} = \frac{\Delta\omega_n}{2\omega_n\sqrt{\zeta^2 - 1}}\underline{\left[\frac{\zeta - \sqrt{\zeta^2 - 1}}{\zeta + \sqrt{\zeta^2 - 1}}\right]^{\frac{\zeta - \sqrt{\zeta^2-1}}{2\sqrt{\zeta^2-1}}}} < 2\pi. \tag{10.6}$$



(a) (b)

Fig. 10.2 Capture range calculation of linear, PLL-based CDRs.

For regular wireline systems (especially long haul), we have $\zeta \gg 1$ and then $\sqrt{\zeta^2 - 1} \approx \zeta$. The underline part of the above inequality approaches unity. The capture range is therefore given by

$$\begin{aligned} |\Delta\omega| &< 2\pi \cdot 2\zeta\omega_n \\ &= 2\pi \cdot \omega_{-3dB}. \end{aligned} \tag{10.7}$$

That is, the capture range of linear PLL-based CDRs is on the order of the loop bandwidth. The absolute value accounts for both sides of deviation. In reality, the capture range would be somewhat smaller than that as PDs can hardly reach $\pm 2\pi$ operation region.

Example 10.1

Prove the underline part of Eq. (10.6) approaches 1 as $\zeta \to \infty$

Solution:

For $\zeta \to \infty$, $\sqrt{\zeta^2 - 1} \to \zeta$. We simplify the problem by defining

$$t \triangleq \frac{\zeta - \sqrt{\zeta^2 - 1}}{2\zeta}. \tag{10.8}$$

The above statement proves since

$$\lim_{t \to 0} t^t = 1. \tag{10.9}$$

How do we calculate the capture range of a bang-bang CDR? The nonlinear behavior of the loop presents us from doing s-domain analysis. However, judging from the model in Fig. 3, we can still estimate the capture range. Suppose the data rate suddenly jumps by $\Delta\omega$, the edge sampling point would no longer stay at transition point 1, but rather shift to point 2 in next bit [Fig. 3(b)]. The bang-bang PD as well as the loop reacts immediately, creating a temporary boost of I_pR_p on VCO control line. Note that C_p for bang-bang CDR is typically very large, so the voltage across C_p barely changes for a period of time as short as a few bits. The control line voltage step I_pR_p translates itself to VCO frequency step by K_{VCO}, arriving at a frequency step of $K_{VCO}I_pR_p$. Whether the loop can go back to lock depends on the relationship between $\Delta\omega$ and $K_{VCO}I_pR_p$. That is, for the loop to relock, we must have

$$|\Delta\omega| < K_{VCO}I_pR_p. \tag{10.10}$$

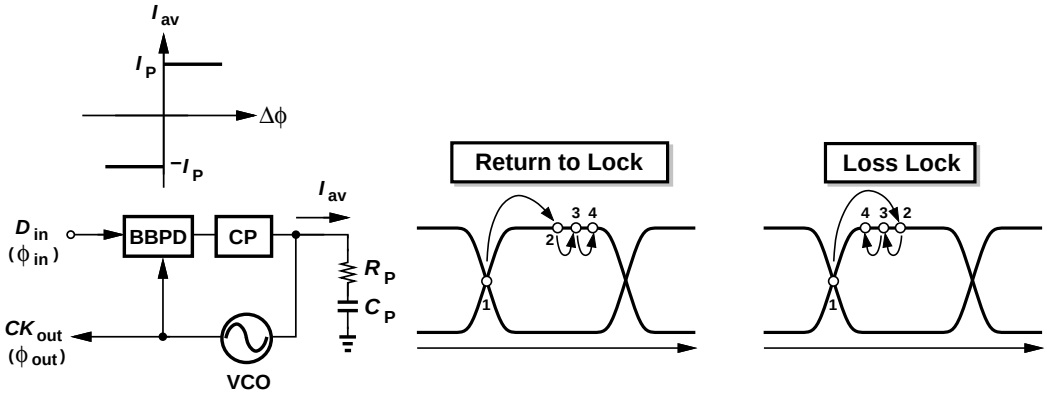


Fig. 10.3 Capture range calculation of bang-bang PLL-based CDRs.

As illustrated is Fig. 3(b), the sampling point would gradually move toward the right position(i.e.,point 1). On the other hand, if the data rate deviation exceeds $K_{VCO}I_pR_p$, the loop would eventually become out of lock (and the sampling point keeps moving away from the right position). We conclude the capture range of such a bang-bang CDR is given by $K_{VCO}I_pR_p$. Recall from our discussion in Chapter 9. This capture range is still commensurate with the loop bandwidth of a bang-bang CDR. Note that the effect of random data has been included in the

model (e.q.,BBPD's characteristic), which presents average results only. Actual case may vary to some extent based on data patterns.

Example 10.2

Determine the capture range of all digital bang-bang CDR shown in Fig. 9.20.

Solution:

For an instant data rate change, the loop creates $0.5 \cdot K_1 \cdot K_{DCO}$ instant boost on frequency, which must be greater than $\Delta\omega$ in order to relock. The capture range is simply given by

$$|\Delta\omega| < \frac{1}{2} \cdot K_1 \cdot K_{DCO}. \tag{10.11}$$

It could also be verified from Example 9.1, where $I_p R_p$ of analog design is replace by $0.5 \cdot K_1$ of all-digital approach.

Other than the PLL-based CDRs, it is instructive to look at the frequency offset issue for PI-based CDRs. Taking the all-digital PI-based CDR in Fig. 9.23 for example. If a finite frequency offset exists between data rate (R_D) and reference PLL clock, the phase interpolator would keep rotating either clockwisely or counter-clockwisely so as to "track" the phase. Once again, the loop may go out of lock if the frequency offset exceeds a certain value. This phenomenon resemble the behavior of capture range in PLL-based CDRs.

To analytically estimate the maximum tolerable frequency offset, we follow the same notation of Fig. 9.23. In the present of frequency offset, the phase deviation in each bit would be $2\pi \times \Delta\omega/\omega_{DR}$, where $\omega_{DR} = 2\pi R_D = 2\pi/T_b$. On the other hand, if the digital loop filter takes T_{DLF} to update one output, the maximum phase difference it could pursue in each bit is $0.5K_2K_{PI}T_b/T_{DLF}$. To relock the loop, we need

$$\frac{\Delta\omega}{\omega_{DR}} \times 2\pi < \frac{0.5K_2K_{PI}T_b}{T_{DLF}}. \tag{10.12}$$

It follows that

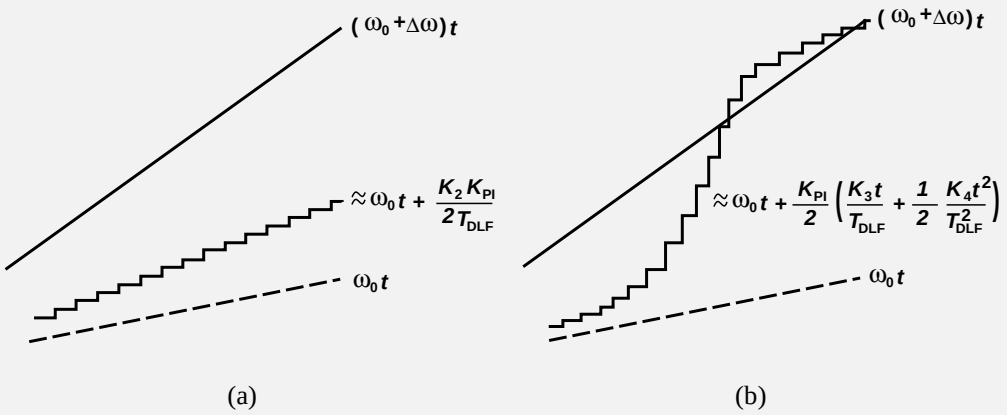
$$|\Delta\omega| < \frac{K_2K_{PI}}{2T_{DLF}}. \tag{10.13}$$

It is of course an optimistic estimation. CDRs with crystal oscillators as references would present a few hundred ppm of frequency offset in practice, which is harmless in most applications. The reader can prove that the frequency offset issue in oversampling CDRs can be analyzed with similar approach.

Example 10.3

Explain why a second-order DLF improves the tolerance of frequency offset in PI-based CDRs.

Solution:



(a) (b)

Fig. 10.4 (a) Linear, (b) parabolic phase tracking on occurrence of frequency offset.

Figure 4(a) illustrates the linear tracking behavior, which can tolerate frequency offset (between the reference PLL clock and input data) up to $K_2 K_{PI}/(2T_{DLF})$. The PI-based CDR with a 2nd order DLF, on the other hand, provides a parabolic phase tracking as depicted in Fig. 4(b). That allows a much larger offset in frequency, given that the proportional and the integral terms are properly chosen.

10.2 FREQUENCY DETECTORS

Now that we realize the limitation of phase detectors, we focus our study on frequency detectors (FDs) in this section. There are several mainstream methods to acquire frequency (data rate) information, namely, dual loop, Pottbacker, direct-dividing, and all-digital. We introduce their operation as well as properties.

10.2.1 Dual-Loop Frequency Acquisition

The most straightforward and perhaps the most popular way to capture the right data rate is to use a reference PLL as we described in Chapter 9. To make it more specific we redraw such a dual loop approach is Fig. 5. The frequency tracking loop is actually unconditionally stable, even though the loop filter is as simple as a single capacitor. Thus, its corresponding charge pump (CP_2) can be driving the main capacitor C_1. In most cases, the FD loop would be turned off once the proper frequency is obtained. It not only saves power but minimizes disturbance. The frequency acquisition loop is usually accompanied by a lock detector, which monitors the loop states and activates the FD loop once the PD loop is out of lock.

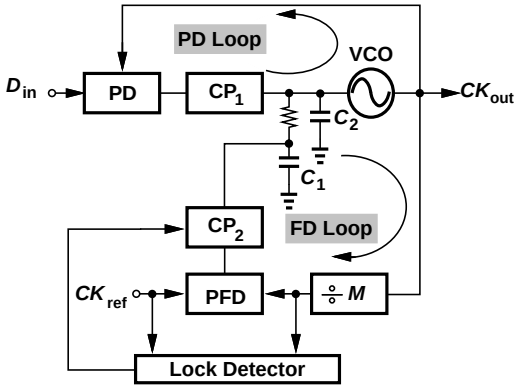


Fig. 10.5 Typical dual loop CDR.

How do we implement a lock detector? A simple way to examine the frequency error between two clocks is to use a mixer. As illustrated in Fig. 6, CK_{ref} and CK_{out}/M are mixed up to get the beat frequency f_b.* [1] If the two clocks are not synchronous, finite f_b would occur. Using a counter and some logic control allows us to determine whether the two clock frequencies are close enough. If the beat frequency is within a preset threshold, the loop is considered "locked". The lock detector thus switches off the corresponding charge pump (i.e., CP_2 in Fig. 5) to disable the FD loop. Otherwise, the FD loop would remain active until the frequency acquisition is achieved.

[1] A low-pass-filter should be added behind the mixer if the sum frequency is a concern.

It is not surprising that some tricks can be added in lock detector to enhance its flexibility and reliability. For example, different thresholds can be imposed such that the “pull-in” and “out-of-lock alarm” have different ranges. The mixer could be replaced by digital circuits as well. Other calibration techniques may be included in all-digital implementation.

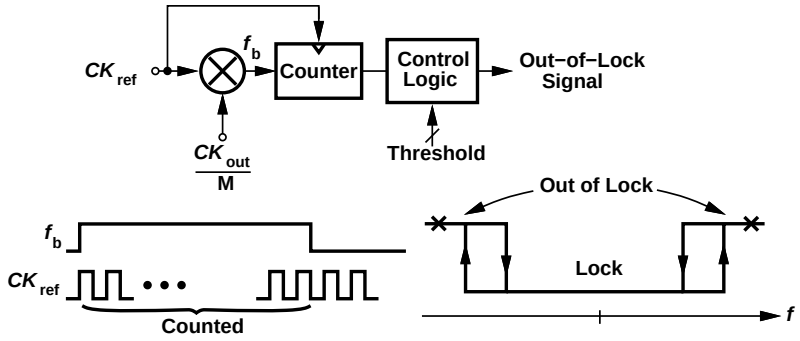


Fig. 10.6 Lock detector design.

10.2.2 Pottbacker FD

The use of external crystal oscillators as reference requires at least one more pad on chip and extra space and cost on board. Here, we introduce an elegant way to distill the frequency information from random data without a reference. First proposed by Pottbacker [8], this type of FD mandates quadrature clocks in full rate. Figure 7 reveals the idea of operation. Instead of sampling data with clocks, here we use input data (D_{in}) to sample clocks (CK_I and CK_Q). If the clock frequency is less than the data rate, the sampling points would gradually shift to the left (i.e., forward). Similarly, if it is greater than the data rate, we see the sampling points moving to the right (i.e., backward). As a result, we obtain two slow waves Q_1 and Q_2 roughly in quadrature. Whether Q_1 is leading or lagging depends on the polarity (sign) of the frequency error. In other words, frequency detection is accomplished by examining the phase relationship between Q_1 and Q_2, which can be easily obtained with an additional flipflop. We get the final polarity result at Q_3.

The Pottbacker FD could be further modified to turn itself off upon lock. As illustrated in Fig. 8(a), the rising or falling edge of D_{in} always aligns with the valley of CK_Q when the PD loop

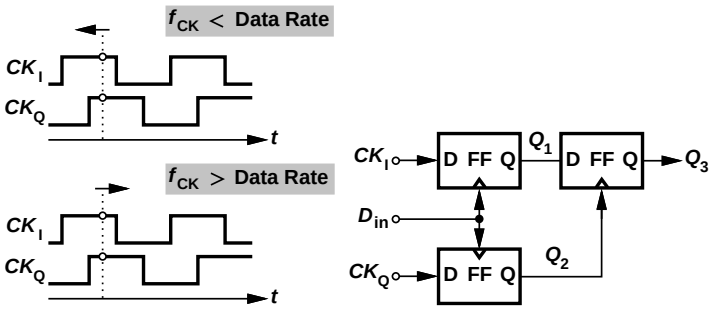
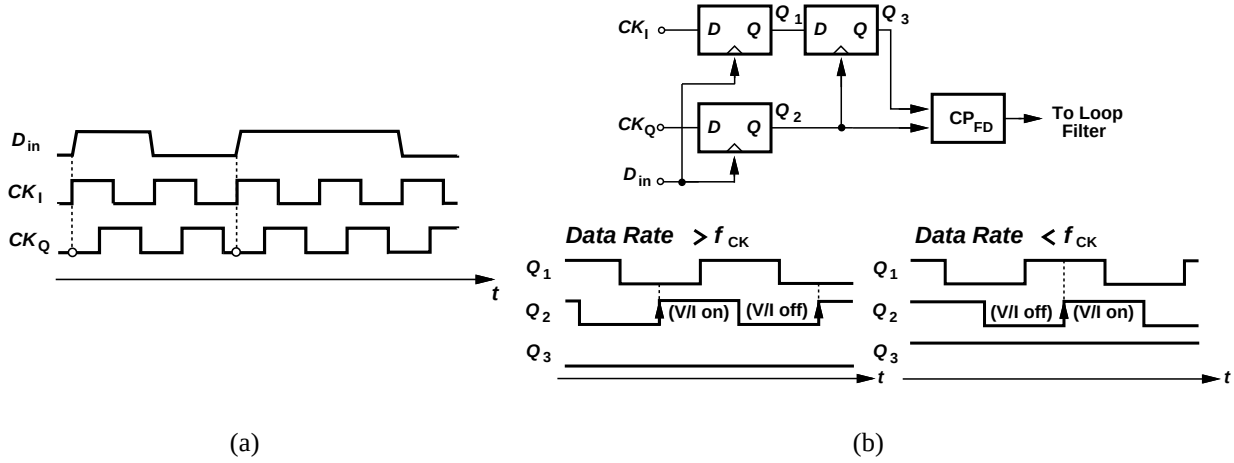


Fig. 10.7 Pottbacker frequency detector.

is locked. That means Q_2 would stay low upon phase locking. As a result, we can apply Q_2 to the corresponding charge pump (CP_{FD}) directly, arriving at the circuit shown in Fig. 8(b). Here, only about 50% of the frequency tracking time is active as CP_{FD} turns on and off periodically. It presents little influence on the overall performance, since frequency acquisition time in wireline systems is not that critical. A major advantage here is that we do not need a lock detector any more. All functions are implemented in analog domain.



(a) (b)

Fig. 10.8 (a) Pottbacker FD under phase locking, (b) modified Pottbacker FD with automatic shutoff.

At high data rates, generating quadrature clocks may not be trivial. The purely linear CDR introduced in Chapter 9.2 provides an alternative solution. Recall from Fig. 9.10 that the input data are first applied to a series of buffers to create 0.5 UI delay. The nominal 0.5 UI delay

from V_A to V_E implies a 0.25 UI delay from V_B to V_D, which allows us to extract the frequency difference. Indeed, the 0.25 UI data delay corresponds to a 90° phase shift[2] of a full-rate clock, making it possible to realize a rotational frequency detector without using quadrature clocks. The proposed FD is shown in Fig. 9. Here, the clock is sampled by using the PD's by-product V_B and V_D, producing two outputs Q_1 and Q_2, respectively. Similar to the Pottbacker FD, Q_1 is further sampled by Q_2 through another flip-flop. The polarity of frequency error Q_3 is therefore obtained. Like the Pottbacker FD, the $V_B \rightarrow V_D$ delay need not be exactly 0.25 UI. Simulation shows that a range of more than 25% on the delay variation is tolerable for the FD to function properly. The automatic switching off function here works in the same way as that of Pottbacker FD does.

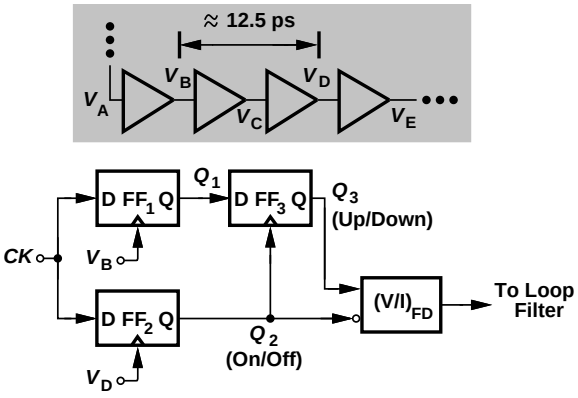


Fig. 10.9 Modified Pottbacker FD with differential clock only.

It is instructive to examine the FD operation in detail and quantize the operation range. The states of Q_1 and Q_2 can be characterized in Fig. 10(a), where the rotating direction indicates the sign of the beat frequency. For example, a clockwise rotation suggests the clock frequency (f_{CK}) is less than the data rate (R_D). Of course, the rotation rate represents the beat frequency. For such an FD to make a right decision on every sampling, we must require the states of Q_1 and Q_2 to jump no more than one step at a time. That is, the average output current I_{av} remains fixed (either positive or negative) for low frequency error, forming a binary characteristic. This situation continues until

[2] As a matter of fact, a precise 90° separation on adjacent phases is not mandatory. A looser condition (such as 80° or 100°) would still allow an FD to achieve similar performance, given that the initial frequency deviation stays within a certain range.

the above condition is violated. To determine the points where I_{av} begins to drop, we study one worst case as illustrated in Fig. 10(b). Here, without loss of generality, we assume f_{CK} is less than R_D and the transition of V_B (and thus Q_1) is already very close to the clock edge. Starting from (1, 1), the state either stays at (1, 1) or moves to (0, 1) in the next sampling. As we know, for a PRBS of $2^N - 1$, the longest run length between transitions is N bits. Since the longest run accumulates the most error, we can determine the largest beat frequency at which the average output current begins to degrade. That is, after N bits, the sampled Q_2 remains high. The boundary condition gives

$$N \cdot |\frac{1}{f_{CK}} - \frac{1}{R_D}| = \frac{1}{4f_{CK}}. \tag{10.14}$$

It follows that the deviation is given by

$$\Delta f_1 \triangleq |f_{CK} - R_D| = \frac{R_D}{4N}. \tag{10.15}$$

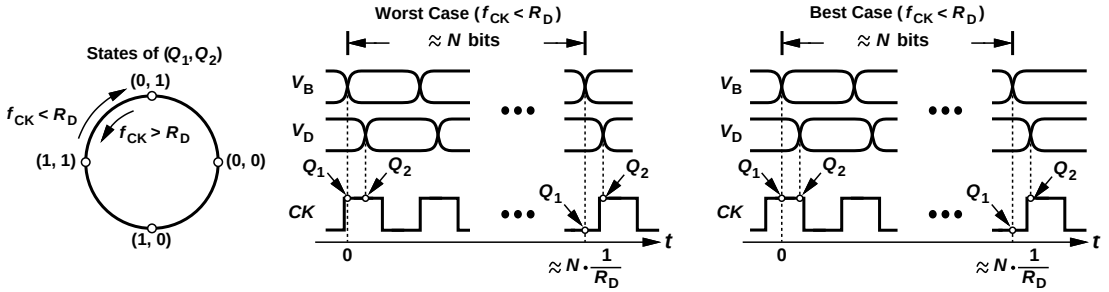
If N = 7, for example, the binary range is equal to $\pm 3.6\%$. It can be easily proven that Δf is symmetric with respect to the origin. Strictly speaking, the use of N bits as the longest period of error accumulation is not exactly correct because the flip-flops in the FD are single-edge triggered. The actual accumulation time would be longer than $N \cdot (1/R_D)$. For example, the longest distance between two adjacent rising edges in a $2^7 - 1$ PRBS is 13 bits, so the binary characteristic begins to roll off at around $R_D/(4 \cdot 13)$.

The above analysis is based on the worst-case scenario. In practice, Q_1 may stay far away from the clock edge before the N-bit long run. The best-case scenario is also shown in Fig. 10(a), where the phase error accumulated over N bits must be less than a half rather than a quarter of a clock cycle in order to maintain a saturated $I_{P2,avg}$. Thus, the widest binary range would be twice as large as that in Eq. (10.15):

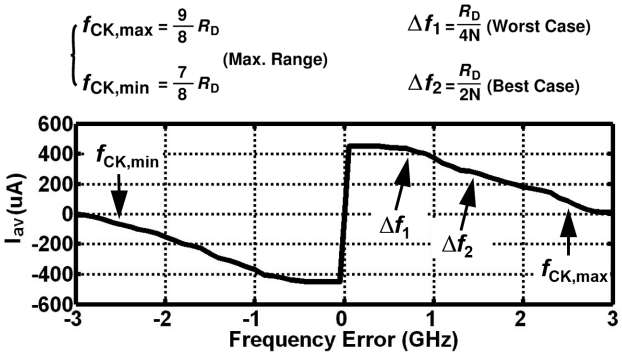
$$\Delta f_2 = \frac{R_D}{2N}. \tag{10.16}$$

Depending on the initial phase relationship, the binary range in reality lies between the two extremes Δf_1 and Δf_2.

The FD performance begins to degrade beyond the binary range, as the sequence of Q_1 and Q_2 becomes chaotic and the erroneous samplings occur in FF_3. It is expected to see the average



(a)



(b)

Fig. 10.10 (a) Determing operation range of Pottbacker FD, (b) simulated FD characteristic (data rate=20 Gb/s).

output eventually approaching zero as the sequential states of Q_1 and Q_2 become totally random, i.e., no reliable average on Q_3 can be obtained. The vanishing point can be roughly estimated as follows. For random data, the expected interval between two adjacent transitions is two bits. Since FF_1 and FF_2 are single-edge triggered, V_B and V_D on average sample the clock every four bits. Now, if the frequency error is so significant that (Q_1, Q_2) steps more than one state in each sampling, the beat-frequency sequences become totally corrupted and the FD has no way to judge the polarity. Under such a circumstance, we have

$$4 \cdot |\frac{1}{f_{CK}} - \frac{1}{R_D}| \geq \frac{1}{2f_{CK}}. \tag{10.17}$$

It follows that

$$f_{CK,max} = \frac{9}{8}R_D \ \ and \ \ f_{CK,min} = \frac{7}{8}R_D. \tag{10.18}$$

In other words, the capture of the FD is about $\pm 12.5\%$. In fact, the vanishing point is slightly larger than the prediction of Eq. (10.18) because of the finite rising and falling times. Fig. 10(b) reveals the simulated FD characteristic for a $2^7 - 1$ input data sequence of 20 Gb/s.

10.2.3 Direct-Dividing FD

Another interesting approach to distill the frequency information from data stream counts on the randomness of the bit sequence. Consider a purely random data stream. The chance for a transition to occur between two consecutive bits is actually 50%. As a result, the probability of run length = 1 bit is 1/2, that of run length = 2 bits is 1/4, and so on. The average run length is therefore given by

$$A_{vg}.Run = \sum_{k=1}^{\infty} k \cdot (\frac{1}{2})^k = 2\ (bits), \tag{10.19}$$

which matches our intuition.

The above observation inspires us to capture the clock directly from the data stream. As depicted in Fig. 11, we can feed D_{in} directly into a divider chain. Since in average D_{in} has one transition every two bits, it is equivalent to a quarter-rate clock ($R_D/4$) for long term operation. That is, the first divider produces something similar to ($R_D/8$), and the N_{th} divider provides a "clock" CK_{out} containing frequency information of ($R_D/2^{N+2}$). After the low-pass filter, one can expect a fairly clean low-speed clock as a reference.The full-rate clock could be easily restored by an auxiliary PLL. Note that even though CK_{out} is very close to an ideal reference, it is not a real clock. After all, it comes from division of random data. However, in regular cases CK_{out} is capable of serving as a competent reference if N $\geq$ 10.

Example 10.4

Calculate the standard deviation of run length.

Solution:

$$\sigma^2 = \frac{1}{2} \cdot 1^2 + 0 + \frac{1}{8} \cdot 1^2 + \frac{1}{16} \cdot 2^2 + \frac{1}{32} \cdot 3^2 + \cdots \tag{10.20}$$

$$= 2. \tag{10.21}$$

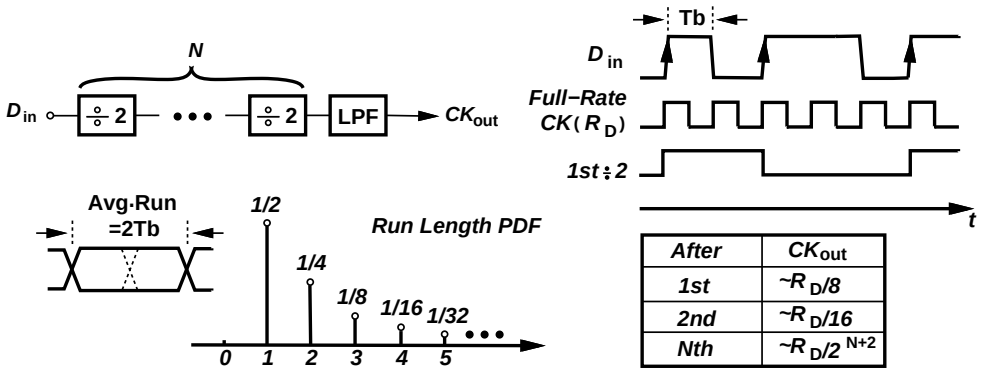


Fig. 10.11 Direct dividing FD.

Example 10.4 ***(Continued)***

The standard deviation of random bit sequence is equal to $\sqrt{2}$ bits.

It goes without saying, that the period of CK_{out} varies from time to time. A robust and reliable lock detector would be necessary to ensure smooth transition from frequency acquisition to phase locking.

10.3 JTRAN OF LINEAR CDRS

Starting from this section, we will study jitter specifications extensively adopted in difference standards. We look at jitter transfer (JTRAN) of linear CDRs in this section.[3]

Jitter transfer is defined as the response of a CDR loop to input jitter, which is actually nothing more than the transfer function we derived in Chapter 9. It is because the s-domain analysis exactly reflects the loop's behavior in response to a sinusoidal variation of input phase. For convenience we redraw the linear CDR model in Fig. 12(a), which presents a closed-loop transfer function equivalent to (JTRAN) of

$$JTRAN \triangleq \frac{\Phi_{out}}{\Phi_{in}}(s) = \frac{2\zeta\omega_n s + \omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}, \tag{10.22}$$

[3]Note that jitter transfer is usually dedicated to long-haul optical links. Thus, we only discuss the case of PLL-based CDRs.

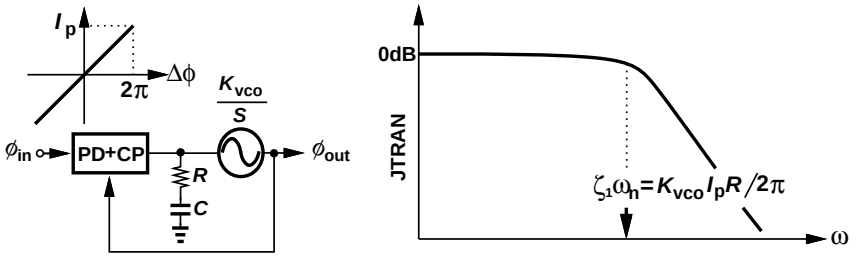
$$\omega_n = \sqrt{\frac{I_p K_{VCO}}{2\pi C_p}} \tag{10.23}$$

$$\zeta = \frac{R_p}{2}\sqrt{\frac{I_p C_p K_{VCO}}{2\pi}}. \tag{10.24}$$

For most long-haul applications, $\zeta \gg 1$, arriving at

$$JTRAN \approx \frac{2\zeta\omega_n}{s + 2\zeta\omega_n}. \tag{10.25}$$

Figure 12(b) illustrates the jitter transfer of overdamped CDRs. As we will see in section 10.5, this is the shape we see on the phase noise plot of the recovered clock.



(a) (b)

Fig. 10.12 Jitter transfer of PLL-based linear CDRs.

An important specification regarding JTRAN is the possible peak around the -3-dB bandwidth f_0. Shown in Fig. 13(a) is an example of $\check{S}$ONET, which asks the peaking to be less than 0.1 dB. It is because tens or even hundreds of repeaters may be deployed, accumulating a huge peak in the far-end side even each repeater contribute only 0.1 dB of peaking. Figure 13(b) illustrates the effect.

10.4 JTRAN OF BANG-BANG CDRS

The binary characteristic of bang-bang PDs in practice exhibits a finite slope across a narrow range of the input phase difference. That is, small phase errors lead to linear operation whereas large phase errors introduce “slewing” in the loop, as we discussed in Chapter 9. Two main phenomena

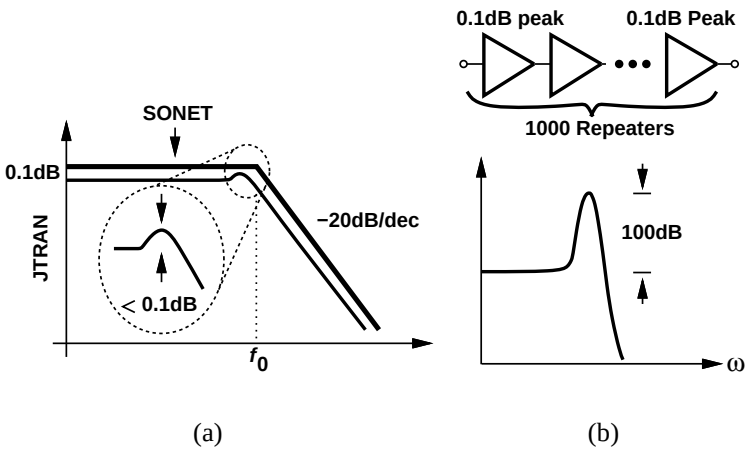


Fig. 10.13 (a) SONET jitter transfer specification, (b) peaking effect in long-haul systems.

cause such a characteristic smoothing. The first is the effect of metastability. When the zero-crossing points of the recovered clock fall in the vicinity of data transitions, the flipflops comprising the PD may experience metastability, thereby generating an output lower than the full level for some time.

To quantify the effect of metastability, we first consider a single latch consisting of a preamplifier and a regenerative pair (Fig. 14), assuming a gain of A_{pre} for the former and a regeneration time constant of τ_{reg} for the latter.[4] We also assume a slope of 2k for the input differential data and a sufficiently large bandwidth at X and Y so that $V_X - V_Y$ tracks D_{in} with the same slope.

Fig. 15 illustrates distinct cases that determine certain points on the PD characteristic. If the phase difference between CK and D_{in}, ΔT, is large enough, the output reaches the saturated level, $V_F = I_{SS}R_C$, in the sampling model [Fig. 15(a)], yielding an average approximately equal to V_F. For the case $2k\Delta T A_{pre} < V_F$, the circuit regeneratively amplifies the sampled level [Fig. 15(b)], providing $\overline{V_{PD}} < V_F$. Finally, if $\Delta T$is sufficiently small, the regeneration in half a clock period does not amplify $2k\Delta T A_{pre}$ to V_F [Fig. 15(c)], leading to an average output substantially less than V_F. Since the current delivered to the loop filter is proportional to the area under $V_X - V_Y$ and

[4]For the sake of brevity, the regenerative gain is included in τ_{reg}, allowing an expression of the form exp(t/τ_{reg}) for the positive feedback growth of the signal.

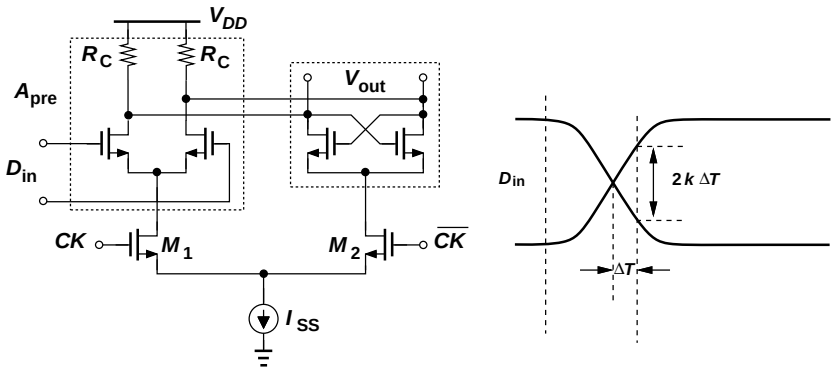


Fig. 10.14 CML latch with input data waveform.

since the waveform in this case begins with an initial condition equal to $2k\Delta TA_{pre}$, we have

$$\overline{V_{PD,meta}}(\Delta T) \approx \frac{1}{T_b}\int_0^{T_b/2} 2k\Delta TA_{pre}exp\frac{t}{\tau_{reg}}\mathrm{dt} \tag{10.26}$$

$$\approx 2k\Delta TA_{pre}\frac{\tau_{reg}}{T_b}exp\frac{T_b}{2\tau_{reg}}. \tag{10.27}$$

Thus, the average output is indeed linearly proportional to ΔT. The linear regime holds so long as the final value at t = $T_b/2$ remains less than V_F,[5] and the maximum phase difference in this regime is given by

$$2k\Delta T_{lin}A_{pre}exp\frac{T_b}{2\tau_{reg}} = V_F \tag{10.28}$$

and hence

$$\Delta T_{lin} = \frac{V_F}{2kA_{pre}exp\frac{T_b}{2\tau_{reg}}}. \tag{10.29}$$

For phase differences greater than ΔT_{lin}, the slope of the characteristic begins to drop, approaching zero if the preamplified level reaches V_F:

$$\Delta T_{sat} = \frac{V_F}{2kA_{pre}}. \tag{10.30}$$

Fig. 15(d) summarizes these concepts.

The binary PD characteristic is also smoothed out by the jitter inherent in the input data and the oscillator output. Even with abrupt data and clock transitions, the random phase difference

[5]Since the regeneration time is in fact equal to $T_b/2 - \Delta T$, the PD characteristic displays a slight nonlinearity in this regime.

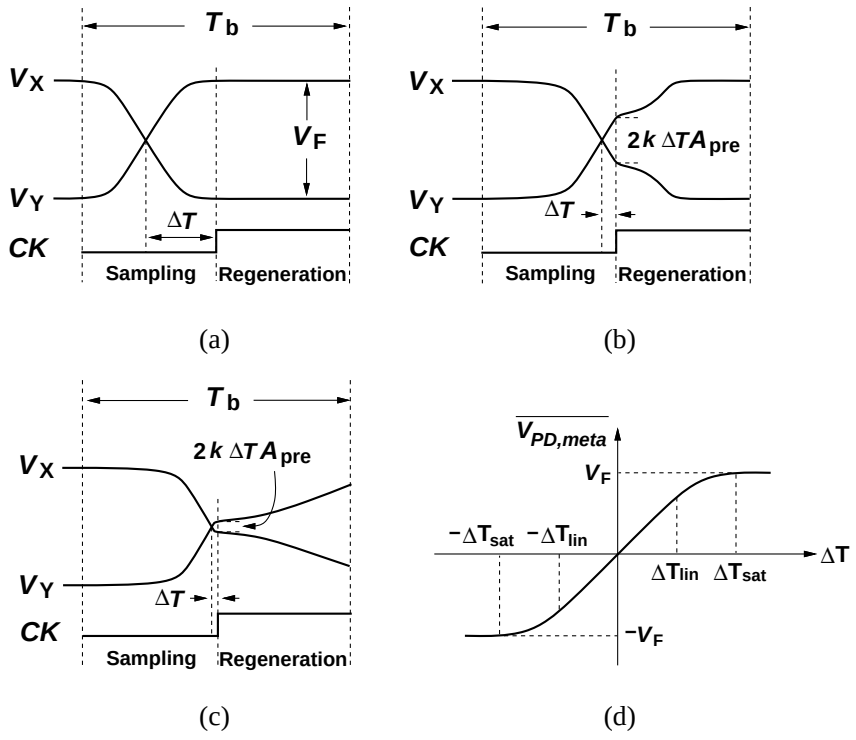


Fig. 10.15 Average PD output for (a) complete switching, (b) partial switching, (c) incomplete regeneration, (d) typical bang-bang characteristic.

resulting from jitter leads to an average output lower than the saturated levels. As illustrated in Fig. 16(a), for a phase difference of ΔT, it is possible that the tail of the jitter distribution shifts the clock edge to the left by more than ΔT, forcing the PD to sample a level of $-V_0$ rather than $+V_0$. To obtain the average output under this condition, we sum the positive and negative samples with a weighting given by the probability of their occurrences:

$$\overline{V_{PD}}(\Delta T) = -V_0 \int_{-\infty}^{-\Delta T} p(x)\, dx + V_0 \int_{-\Delta T}^{+\infty} p(x)\, dx \tag{10.31}$$

where $p(x)$ denotes the probability density function (PDF) of jitter. Since the PDF is typically even-symmetric, this result can be rewritten as

$$\overline{V_{PD}}(\Delta T) = -V_0 \int_{+\Delta T}^{+\infty} p(x)\, dx + V_0 \int_{-\infty}^{+\Delta T} p(x)\, dx \tag{10.32}$$

which is equivalent to the convolution of the bang-bang characteristic and the PDF of jitter. Illustrated in Fig. 16(b), $\overline{V_{PD}}$ exhibits a relatively linear range for $|\Delta T| < 2\sigma$ if the PDF is Gaussian with a standard deviation of σ.

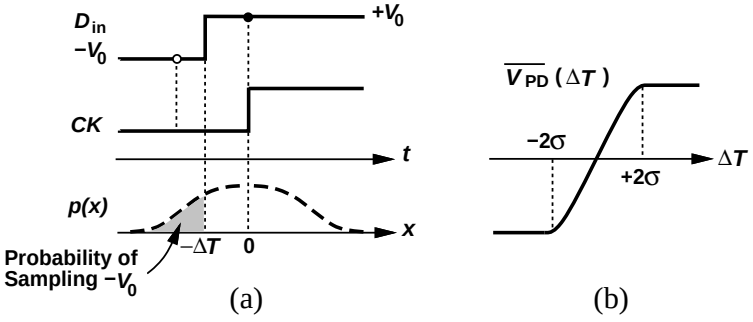


Fig. 10.16 Smoothing of PD characteristic due to jitter.

Combining the two effects, it is not difficult to obtain the resulting model in Fig. 17, where the BBPD+CP presents a linear region of $\pm\Phi_m$ and saturated pumping current $\pm I_p$. Suppose $\Phi_{in}(t) = \Phi_{in,p}\cos\omega_\Phi t$. If $\Phi_{in,p} < \Phi_m$ then the PD operates in the linear region, yielding a standard second-order system. On the other hand, as $\Phi_{in,p}$ exceeds Φ_m, the phase difference between the input and output may also rise above Φ_m, leading to nonlinear operation. At low jitter frequencies, $\Phi_{out}(t)$ still tracks $\Phi_{in}(t)$ closely, $|\Delta\Phi| < |\Phi_m|$, and $|\Phi_{out}/\Phi_{in}| \approx 1$. As ω_Φ increases, so does $\Delta\Phi$, demanding that the V/I converter pump a larger current into the loop filter. However, since the available current beyond the linear PD region is constant, large and fast variation of Φ_{in} results in "slewing".

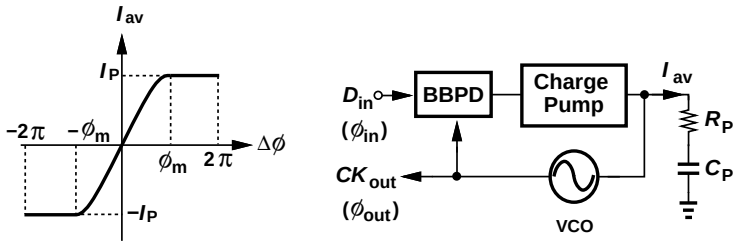


Fig. 10.17 PLL-based bang-bang CDR model.

To study this phenomenon, let us assume $\Phi_{in,p} \gg \Phi_m$ as an extreme case so that $\Delta\Phi$ changes polarity in every half cycle of ω_Φ, requiring that I_1 alternately jump between $+I_p$ and $-I_p$ (Fig. 18). Since the loop filter capacitor is typically large, the oscillator control voltage tracks $I_1 R_p$, leading to binary modulation of the VCO frequency and hence triangular variation of the output phase. The peak value of Φ_{out} occurs after integration of the control voltage for a duration of $T_\Phi/4$, where $T_\Phi = 2\pi/\omega_\Phi$; that is,

$$\Phi_{out,p} = \frac{K_{VCO} I_p R_p T_\Phi}{4} \tag{10.33}$$

and

$$\left|\frac{\Phi_{out,p}}{\Phi_{in,p}}\right| = \frac{\pi K_{VCO} I_p R_p}{2\Phi_{in,p}\omega_\Phi}. \tag{10.34}$$

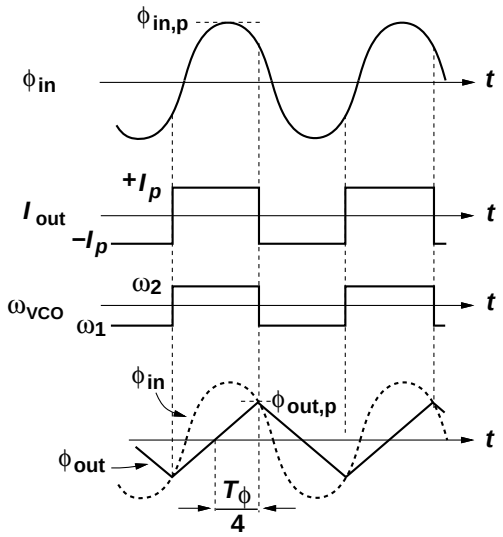


Fig. 10.18 Slewing of PLL-based bang-bang CDR.

Expressing the dependence of the jitter transfer upon the jitter amplitude, $\Phi_{in,p}$, this equation also reveals a 20-dB/dec roll-off in terms of ω_Φ. Of course, as ω_Φ decreases, slewing eventually vanishes, Eq. (10.34) is no longer valid, and the jitter transfer approaches unity. As depicted in Fig. 19(a), extrapolation of linear and slewing regimes yields an approximate value for the -3-dB bandwidth of the jitter transfer:

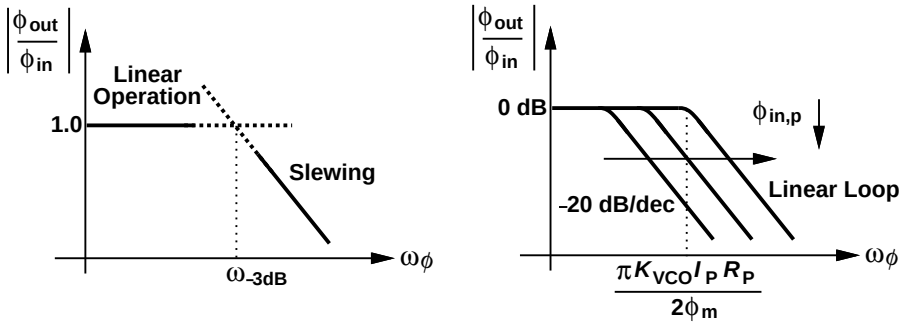
$$\omega_{-3dB} = \frac{\pi K_{VCO} I_p R_p}{2\Phi_{in,p}}. \tag{10.35}$$

It is therefore possible to approximate the entire jitter transfer as

$$\frac{\Phi_{out,p}}{\Phi_{in,p}}(s) = \frac{1}{1+\frac{s}{\omega_{-3dB}}}. \tag{10.36}$$

Fig. 19(b) plots the jitter transfer for different input jitter amplitudes. The transfer approaches that of a linear loop as $\Phi_{in,p}$ decreases toward Φ_m.

It is interesting to note that the jitter transfer of slew-limited CDR loops exhibits negligible peaking. Due to the high gain in the linear regime, the loop operates with a relatively large damping factor in the vicinity of ω_{-3dB}. In the slewing regime, as evident from the Φ_{in} and Φ_{out} waveforms in Fig. 18, $\Phi_{out,p}$ can only fall monotonically as ω_Φ increases because the slew rate is constant. Bang-bang CDR's loop bandwidth must specify what input jitter level is to be used.



(a) (b)

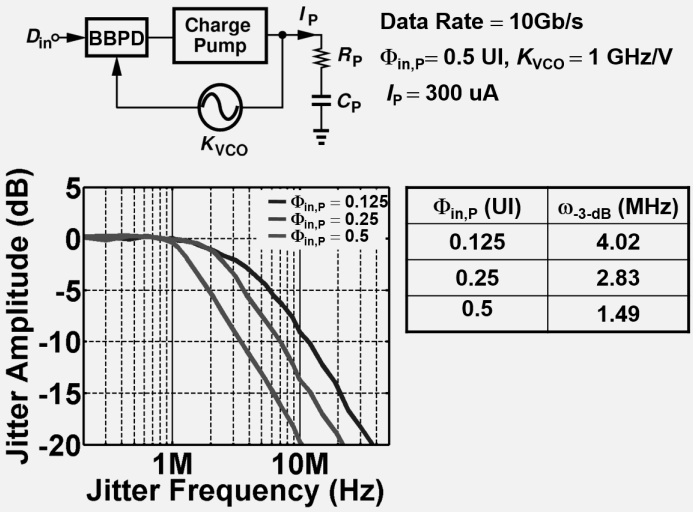
Fig. 10.19 (a) Calculation of -3-dB, (b) jitter transfer of PLL-based bang-bang CDRs.

Example 10.5

Consider the JTRAN measured results of a 10-Gb/s bang-bang CDR as shown in Fig. 20, where three different input jitter magnitudes are tested estimate the linear region boundary Φ_m.

Solution:

Example 10.5 ***(Continued)***



$\Phi_{in,P}$ (UI)	$\omega_{-3\text{-}dB}$ (MHz)
0.125	4.02
0.25	2.83
0.5	1.49

Fig. 10.20 Measured JTRAN of a 10-Gb/s bang-bang CDR for different Φ_{in}.

The loop bandwidth is inversely proportional to $\Phi_{in,p}$ as $\Phi_{in,p}$ varies from 0.25 to 0.5 UI. It obviously saturates as $\Phi_{in,p}$ drops to 0.125 UI. Since all other parameters are fixed, we have two equations to predict Φ_m:

$$4.02 \cdot \Phi_m = 2.83 \cdot 0.25 \tag{10.37}$$

$$4.02 \cdot \Phi_m = 1.49 \cdot 0.5. \tag{10.38}$$

Φ_m is given by 0.176 and 0.185, respectively. By averaging, we estimate Φ_m to be 0.18 UI.

Example 10.6

With the same setup of Fig. 20, now we fix $\Phi_{in,p}$ = 0.5 UI and change C_p. The result is shown in Fig. 21, where three cases give roughly the same curves of ω_{-3dB} =2.75 MHz. Calculate what R_p we use here.

Solution:

Example 10.6 *(Continued)*

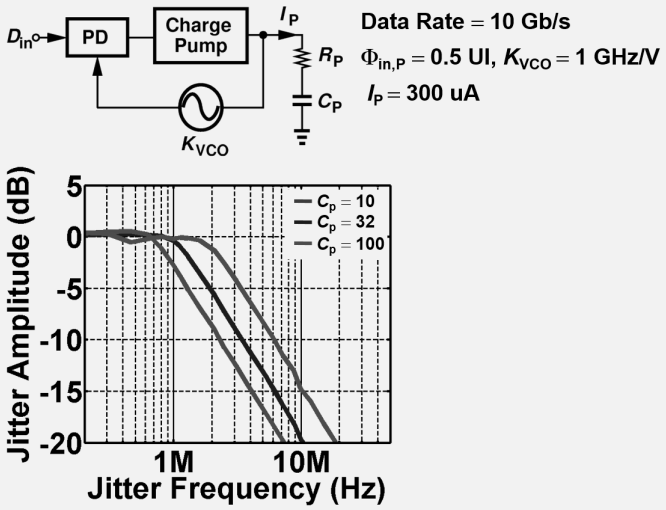


Fig. 10.21 Measured JTRAN of a 10-Gb/s bang-bang CDR for different C_p.

Recognizing that C_p has no effect on JTRAN, we calculate R_p by Eq. (10.35) as 18.3 Ω.

The above two examples are based on real measurement results of a 10-Gb/s CDR with a standard Alexander PD realized in 90-nm CMOS technology.

10.5 JTOL OF LINEAR CDRS

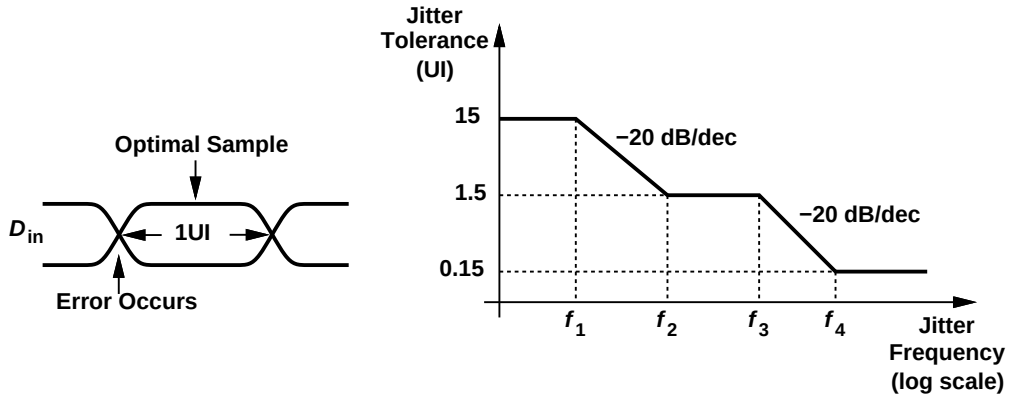
Jitter tolerance (JTOL) is defined as the maximum input jitter that a CDR loop can tolerate without increasing the bit error rate at a given jitter frequency. As the phase error, $\Phi_{in} - \Phi_{out}$, approaches π = 0.5 UI, BER rises rapidly [Fig. 22(a)].

It is straight forward to derive JTOL from JTRAN for linear CDRs. Since in theory, an error would occur if

$$|\Phi_{in} - \Phi_{out}| \geq 0.5\ (UI). \tag{10.39}$$

That is,

$$\Phi_{in}(1 - \frac{\Phi_{out}}{\Phi_{in}}) \geq 0.5. \tag{10.40}$$



(a) (b)

Fig. 10.22 (a) Jitter tolerance calculation, (b) jitter tolerance mask.

Jitter tolerance is therefore available

$$JTOL = \frac{0.5}{1 - \frac{\Phi_{out}}{\Phi_{in}}} = \frac{0.5}{1 - JTRAN}. \tag{10.41}$$

Usually a mask is imposed as a specification. Fig. 22(b) reveals an example. The mask is defined by 4 corner frequencies f_1, f_2, f_3 and f_4. The device under test (DUT, could be a CDR or a complete RX) experience jittery data input under different modulation frequencies and check the bit error rate (BER). For a given threshold (e.g., BER=10^{-12}), the JTOL curve could be obtained. If the JTOL curve is above the mask for all jitter frequencies, we say the DUT passes the corresponding JTOL test. Generally speaking, f_4 is the most critical point for a CDR to pass JTOL test, as the available jitter margin is much smaller at high frequencies.

For overdamped systems. JTOL can be further derived as

$$JTOL = \frac{0.5}{1 - JTRAN} \tag{10.42}$$

$$= \frac{0.5(s + 2\zeta\omega_n)}{s}. \tag{10.43}$$

That is, JTOL rolls off at a rate of -20 dB/dec starting from the origin, and flattens after the zero $2\zeta\omega_n$. The JTRAN and JTOL of linear CDRs actually share the same turning point, which is $\omega_{-3dB} = 2\zeta\omega_n$ (Fig. 23). In other words, JTRAN and JTOL of linear CDRs are bound together. It causes some trouble if the linear CDR is designed for certain applications. For example, as

illustrated in Fig. 24, the ITU defines the loop bandwidth on JTRAN to be 120 kHz, where as the major corner f_4 is as high as 4 MHz. A dilemma is created here, as a traditional linear CDR can never satisfy both specifications. More sophisticated CDR architecture must be adopted to overcome this difficulty.

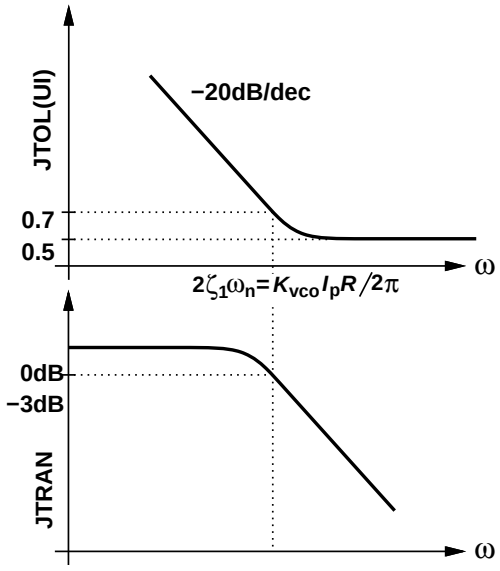


Fig. 10.23 Jitter transfer and jitter tolerance of PLL-based linear CDRs.

JTRAN		JTOL	
f_0	120kHz	f_1	2kHz
		f_2	20kHz
		f_3	400kHz
		f_4	4MHz

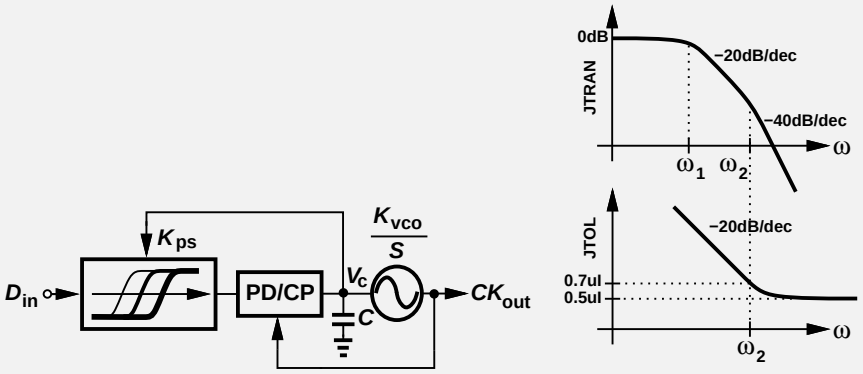
Fig. 10.24 JTRAN and JTOL of OC-192.

Example 10.7

A linear CDR combining DLL and PLL has been proposed to untie the coupling between JTRAN and JTOL. As shown in Fig. 25(a), this structure uses a simple capacitor as the loop filter. Analyze the circuit and determine its JTRAN and JTOL. The voltage-controlled delay line (i.e., phase shifter) presents a gain of K_{ps}.

Example 10.7 ***(Continued)***

Solution:



(a) (b)

Fig. 10.25 (a) D/PLL based linear CDR, (b) its JTRAN and JTOL.

Since D_{in} experiences a phase shifting before entering PD, we have

$$\frac{\Phi_{in} - V_c K_{ps} - \Phi_{out}}{2\pi} \cdot I_p \cdot \frac{1}{sc} = V_c = \frac{s\Phi_{out}}{K_{VCO}}. \tag{10.44}$$

Solving the equation, we obtain JTRAN as

$$JTRAN = \frac{\Phi_{out}}{\Phi_{in}} = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}. \tag{10.45}$$

Now the natural frequency and damping factor become

$$\omega_n = \sqrt{\frac{K_{VCO} I_p}{2\pi c}} \tag{10.46}$$

$$\zeta = \frac{K_{ps}}{2}\sqrt{\frac{I_p}{2\pi c K_{VCO}}}. \tag{10.47}$$

For overdamped loop, $\zeta \gg 1$, JTRAN has two real poles. Namely,

$$JTRAN = \frac{\omega_n^2}{(s + \omega_{p1})(s + \omega_{p2})} \tag{10.48}$$

where

$$\omega_{p1} \cdot \omega_{p2} = \omega_n^2 \tag{10.49}$$

$$\omega_{p1} + \omega_{p2} = 2\zeta\omega_n. \tag{10.50}$$

Example 10.7 *(Continued)*

Suppose $\omega_{p1} < \omega_{p2}$, we arrive at

$$\omega_{p1} \cong \frac{\omega_n}{2\zeta} = \frac{K_{VCO}}{K_{ps}} \tag{10.51}$$

$$\omega_{p2} \cong 2\zeta\omega_n = \frac{K_{ps}I_p}{2\pi c}. \tag{10.52}$$

Fig. 25(b) illustrates JTRAN of such a design. JTOL can be derived with the same approach. Setting the critical condition

$$|\Phi_{in} - V_c K_{ps} - \Phi_{out}| = 0.5\ (UI), \tag{10.53}$$

we arrive at

$$JTOL = \Phi_{in,max} \tag{10.54}$$

$$= \frac{0.5}{1 - (1 + \frac{K_{ps}}{K_{VCO}} \cdot s) \cdot JTRAN} \tag{10.55}$$

$$= \frac{0.5(s + \omega_{p2})}{s}. \tag{10.56}$$

That is, JTOL's corner point now moves to ω_{p2}. The two specifications are now decoupled as JTRAN and JTOL can be designed separately.

10.6 JTOL OF BANG-BANG CDRS

Now we look at JTOL of binary CDRs. As we described in 10.4, a bang-bang CDR loop slews if it fails to follow the input phase modulation tightly.

It is important to recognize that a bang-bang loop must slew if it incurs errors. With no slewing, the phase difference between the input and output falls below $\Phi_m(\ll \pi)$, and the data is sampled correctly. Fig. 26(a) shows the case where Φ_{out} slews and $\Phi_{in,p}$ is chosen such that $\Delta\Phi_{max} = \pi$. It can be shown that $\Delta\Phi_{max}$ occurs at some point t_1, but $\Delta\Phi$ at t_0 is close to $\Delta\Phi_{max}$ and much simpler to calculate. If Φ_{out} slews for most of the period, t_0 is approximately equal to $T_\Phi/4$.

Assuming $\Phi_{in} = \Phi_{in,p}\cos(\omega_\Phi t + \delta)$,[6] we arrive at

$$K_{VCO}I_pR_p\frac{T_\Phi}{4} = \Phi_{in,p}\cos\delta \tag{10.57}$$

and

$$\delta = \tan^{-1}\frac{\sqrt{4{\omega_\Phi}^2\Phi_{in,p}^2 - \pi^2K_{VCO}^2I_p^2R_p^2}}{\pi K_{VCO}I_pR_p}. \tag{10.58}$$

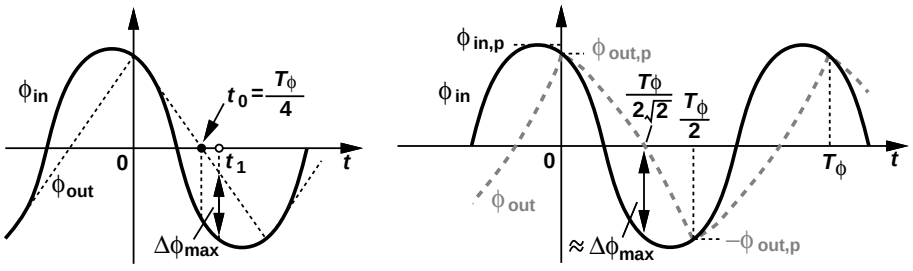
It follows that

$$\Delta\Phi_{max} \approx \Delta\Phi(t_0) = |\Phi_{in,p}\cos(\frac{\pi}{2} + \delta)| \tag{10.59}$$

$$= \frac{\sqrt{4{\omega_\Phi}^2\Phi_{in,p}^2 - \pi^2K_{VCO}^2I_p^2R_p^2}}{2\omega_\Phi}. \tag{10.60}$$

Equating $\Delta\Phi_{max}$ to 0.5 UI yields the maximum tolerable input jitter $\Phi_{in,p} = JTOL$:

$$JTOL = 0.5\sqrt{1 + \frac{K_{VCO}^2I_p^2R_p^2}{4\omega_\Phi^2}}. \tag{10.61}$$



(a) (b)

Fig. 10.26 JTOR calculation for bang-bang CDRs: (a) slewing, (b) non-linear slewing.

As expected, JTOL falls at a rate of 20 dB/dec for low ω_Φ, approaching π at high ω_Φ. A corner frequency, ω_1, can be defined by equating Eq. 10.61 to 0.7 UI

$$\omega_1 = \frac{K_{VCO}I_pR_p}{2}. \tag{10.62}$$

The above analysis has followed the same assumptions as before, namely, the change in the control voltage is due to I_1R_p and the voltage across C_p remains constant. At jitter frequencies below

[6]The angle δ is chosen such that the output peak occurs at t=0, simplifying the algebra.

$(R_pC_p)^{-1}$, however, this condition is violated, leading to "nonlinear slewing" at the output. In fact, for a sufficiently low ω_Φ, the (linear) voltage change across C_p far exceeds I_1R_p, yielding a parabolic shape for Φ_{out} [Fig. 26(b)]. Thus

$$\Phi_{out}(t) = -\int K_{VCO}\frac{I_p}{C_p}t\,dt + \Phi_{out,p} \qquad 0 < t < \frac{T_\Phi}{2} \tag{10.63}$$

$$= -\frac{1}{2}\frac{K_{VCO}I_p}{C_p}t^2 + \Phi_{out,p}. \tag{10.64}$$

Since Φ_{out} reaches $-\Phi_{out}$ at t = $T_\Phi/2$, we have

$$\Phi_{out}(\frac{T_\Phi}{2}) = -\Phi_{out,p} = -\frac{1}{2}\frac{K_{VCO}I_p}{C_p}\frac{T_\Phi^2}{4} + \Phi_{out,p} \tag{10.65}$$

and hence

$$\Phi_{out,p} = \Phi_{in,p}\cos\delta = \frac{K_{VCO}I_p\pi^2}{4C_p\omega_\Phi^2}. \tag{10.66}$$

Note that the zero-crossing point of Φ_{out} occurs at t = $T_\Phi/(2\sqrt{2})$. Adopting the same technique used for the linear slewing case, we approximate $\Delta\Phi_{max}$ with $|\Phi_{in}(T_\Phi/(2\sqrt{2})|$ and obtain

$$\Delta\Phi_{max} \approx |\Phi_{in,p}\cos(\omega_\Phi\frac{T_\Phi}{2\sqrt{2}} + \delta)| \tag{10.67}$$

$$= -\Delta\Phi_{in,p}\cos\frac{\pi}{\sqrt{2}}\cos\delta + \Delta\Phi_{in,p}\sin\frac{\pi}{\sqrt{2}}\sin\delta \tag{10.68}$$

$$= 0.61\frac{K_{VCO}I_p\pi^2}{4C_p\omega_\Phi^2} + 0.8\frac{\sqrt{16C_p^2\omega_\Phi^4\Phi_{in,p}^2 - K_{VCO}^2I_p^2\pi^4}}{4C_p\omega_\Phi^2}. \tag{10.69}$$

Again, equating $\Delta\Phi_{max}$ to 0.5 UI yields the jitter tolerance, JTOL = $\Phi_{in,p}$

$$JTOL = 0.5\sqrt{\frac{(1 - 0.61\frac{K_{VCO}I_p\pi}{4C_p\omega_\Phi^2})^2}{0.64} + \frac{K_{VCO}^2I_p^2\pi^2}{16C_p^2\omega_\Phi^4}}, \tag{10.70}$$

which is too complicated to analyze. Fortunately, at very low jitter frequency, we have

$$0.61\frac{K_{VCO}I_p\pi}{4C_p\omega_\Phi^2} \gg 1, \tag{10.71}$$

which simplifies JTOL as

$$JTOL = 0.63\,(\frac{K_{VCO}I_p\pi}{4C_p\omega_\Phi^2}). \tag{10.72}$$

In this region, JTOL falls at a rate of 40 dB/dec. Fig. 27 depicts the complete JTOL curve of bang-bang CDRs. The corner frequency ω_2 between the two regions can be calculated by extrapolation. Assuming $\omega_2 \ll \omega_1$,we have

$$\omega_2 = \frac{0.63\pi}{R_p C_p}. \tag{10.73}$$

The reader can also show that the above assumption is valid for most cases.

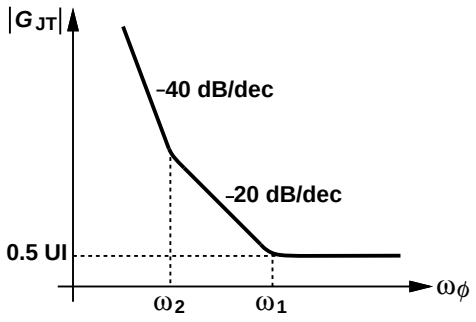


Fig. 10.27 Complete JTOR of bang-bang CDRs.

Example 10.8

For a certain 10-GB/s long-haul data link we have JTRAN bandwidth corner of 8 MHz and JTOL major corner (i.e.,f_4 in Fig. 24) of 4 MHz. Now design a bang-bang CDR and determine R_p to satisfy both JTRAN and JTOL. K_{VCO} =1.2 GHz/V, I_p = 600 μA, and $\Phi_{in,p}$ = 2 UI.

Solution:

From JTRAN and JTOL definitions we require

$$\frac{\pi K_{VCO} I_p R_p}{2\Phi_{in,p}} < 2\pi \times 8MHz \tag{10.74}$$

$$2\pi \times 4MHz < \frac{K_{VCO} I_p R_p}{2}. \tag{10.75}$$

It follows that

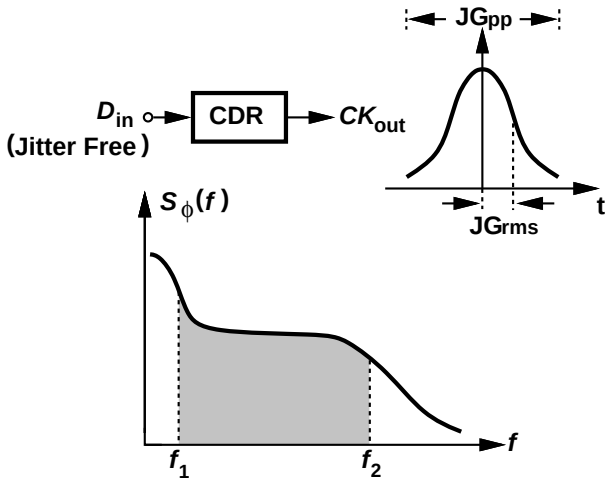
$$70\ \Omega < R_p < 89\ \Omega. \tag{10.76}$$

It is worth nothing that the JTOL of an ideal CDR approaches 0.5 UI as the phase modulation frequency ω_Φ keeps going up. In the presence of noise, jitter, offset, and/or other nonidealities,

JTOL would be further degraded. Thus, it is fair enough to set the mask of 0.15 UI boundary at high frequencies.

10.7 JITTER GENERATIONS

Jitter generation (JG) is defined as the jitter entirely produced by the CDR itself. The JG measurement is straightforward: apply a clean input data to the CDR under testing and collect the jitter distribution of the recovered clock. Using the clean clock synchronized with input data as the trigger signal, the statistical jitter results can be obtained in most digital oscilloscopes. Such a time-domain measurement requires the sample number to be at least 10,000 in order to get meaningful results. Fig. 28 shows the required rms and peak-to-peak jitters for different Optical Carrier (OC) levels. For example, in OC-192 (data rate ≈10 Gb/s) the recovered clock jitter must be less than 1 ps,rms and 10 ps,pp, respectively.



	f_1	f_2
OC−48	5kHz	20MHz
OC−192	20kHz	80MHz
OC−768	20kHz	320MHz

	JGrms	JGpp
OC−48	0.01UI	0.1UI
OC−192	0.01UI	0.1UI
OC−768	0.01UI	0.1UI

Fig. 10.28 Jitter generation definition.

A more strict definition of jitter generation can be found in frequency domain. By integrating the phase noise of recovered clock from dc to infinity, we would obtain the same rms jitter in theory. However, a completely jitter (noise) free data stream does not exist. The phase noise of a clean data stream still depends on that of its clock source ultimately. Shown in Fig. 29 is a typical phase noise plot of the recovered clock from a 20-Gb/s PLL-based linear CDR. The output phase

noise is governed by the input data profile at low frequency offsets, and gradually migrated to that of the free-running VCO. Thus, the integration must be restricted by boundaries. The lower limit f_1 excludes the low-frequency influence from the input data, and the high limit f_2 avoids the offset of undesired coupling at high frequencies.

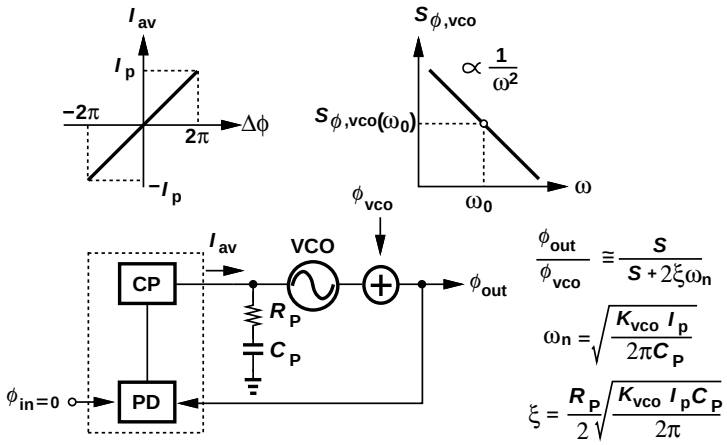


Fig. 10.29 Model of PLL-based linear CDRs in JG calculation.

Example 10.9

Derive the relationship between noise spectrum and rms jitter.

Solution:

By definition, the root-mean-square jitter ΔT_{rms} is equal to

$$\Delta T_{rms} = \left[\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \Delta T_j^2 \right]^{\frac{1}{2}} \tag{10.77}$$

$$= \left[\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \left(\frac{\Delta\Phi_j}{2\pi} \cdot T_b \right)^2 \right]^{\frac{1}{2}} \tag{10.78}$$

$$= \frac{T_b}{2\pi} \left[\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \Delta\Phi_j^2 \right]^{\frac{1}{2}}. \tag{10.79}$$

Example 10.9 ***(Continued)***

The term inside the brackets is the noise power, which is exactly the integration of spectrum S_Φ. Thus,

$$\Delta T_{rms} = \frac{T_b}{2\pi}\left[\int_{-\infty}^{\infty} S_\Phi(f)\,df\right]^{\frac{1}{2}} \tag{10.80}$$

$$= \frac{T_b}{2\pi}\left[2\cdot\int_{0}^{\infty} S_\Phi(f)\,df\right]^{\frac{1}{2}} \tag{10.81}$$

$$= \frac{T_b}{2\pi}\left[2\cdot\int_{0}^{\infty} 10^{\frac{\mathcal{L}(f)}{10}}\,df\right]^{\frac{1}{2}}, \tag{10.82}$$

where $\mathcal{L}(f)$ denotes the phase noise with the unit dBc/Hz. Jitter generation is available by changing integration limits:

$$JG_{rms} \triangleq \frac{\Delta T_{rms}}{T_b} = \frac{1}{2\pi}\left[2\cdot\int_{f_1}^{f_2} S_\Phi(f)\,df\right]^{\frac{1}{2}}\ (UI) \tag{10.83}$$

$$= \frac{1}{2\pi}\left[2\cdot\int_{f_1}^{f_2} 10^{\frac{\mathcal{L}(f)}{10}}\,df\right]^{\frac{1}{2}}\ (UI). \tag{10.84}$$

To be more specific, let us conduct the derivation of JG. For a PLL-based linear CDR, we re-draw its model in Fig. 30. As evidenced by Fig. 29, the input-referred noise of PD/CP is negligible as compared with input data noise. Therefore, the only major noise source is VCO.

For typical overdamped cases, the noise transfer function from VCO to output is given by

$$\frac{\Phi_{out}}{\Phi_{in}} \cong \frac{s}{s+2\zeta\omega_n}, \tag{10.85}$$

where

$$\omega_n = \sqrt{\frac{K_{VCO}I_p}{2\pi C_p}} \tag{10.86}$$

$$\zeta = \frac{R_p}{2}\sqrt{\frac{K_{VCO}I_pC_p}{2\pi}} \tag{10.87}$$

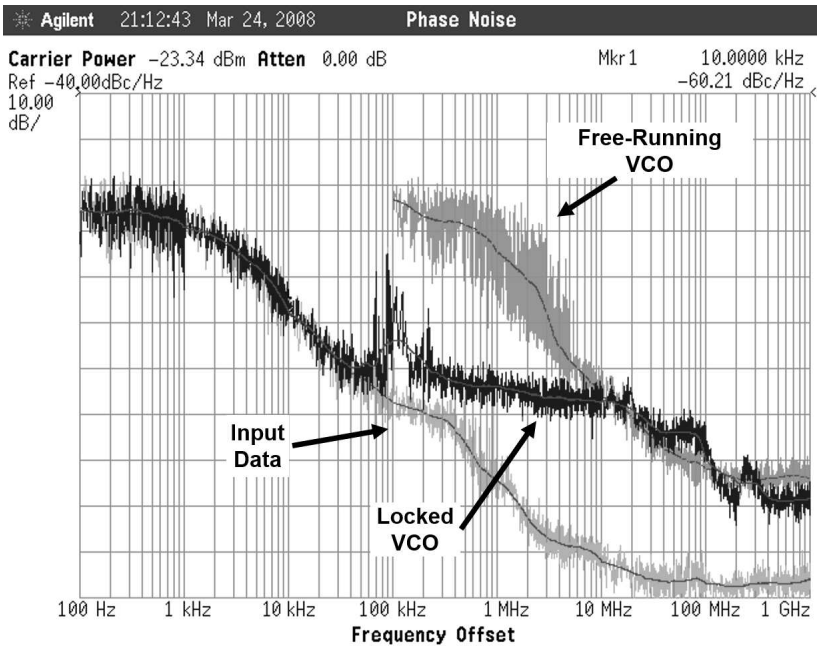


Fig. 10.30 Typical phase noise of recovered clock (PLL-based linear CDR, data rate=20 Gb/s).

and the loop bandwidth $\omega_{BW} = 2\zeta\omega_n = 2\pi f_{BW}$. Follow the derivation of Chapter 8, we define VCO's noise spectrum as

$$S_{\Phi,VCO} = S_{\Phi,VCO}(\omega_o) \cdot \frac{\omega_o^2}{\omega^2}. \tag{10.88}$$

Again, $\omega_0 = 2\pi f_0$ is an arbitrary frequency point along the -20 dB/dec spectrum. The output noise now becomes (Fig. 31)

$$S_{\Phi,out}(\omega) = S_{\Phi,VCO}(\omega_o) \cdot \frac{\omega_o^2}{\omega^2} \cdot \frac{\omega^2}{\omega^2 + \omega_{BW}^2}. \tag{10.89}$$

From the above example, we calculate jitter generation in UI directly

$$JG_{rms} = \frac{1}{2\pi} \cdot \left[2 \cdot \int_{f_1}^{f_2} S_{\Phi,VCO}(f_0) \frac{f_0^2}{f^2 + f_{BW}^2}\, df \right]^{\frac{1}{2}} \tag{10.90}$$

$$= \frac{f_0}{2\pi} \left\{ 2 \cdot \frac{S_{\Phi,VCO}(f_0)}{f_{BW}} \cdot \left[\tan^{-1}(\frac{f_2}{f_{BW}}) - \tan^{-1}(\frac{f_1}{f_{BW}}) \right] \right\}^{\frac{1}{2}}. \tag{10.91}$$

In most cases, the finite integration limits can be removed (i.e.,$f_2 \to \infty, f_1 \to 0$) to simplify the calculation:

$$JG_{rms} = \frac{f_o}{2} \sqrt{\frac{S_{\Phi,VCO}(f_o)}{\pi f_{BW}}}\, (UI). \tag{10.92}$$

Accuracy would be degraded only by an insignificant amount (usually $< 3\ \%$).

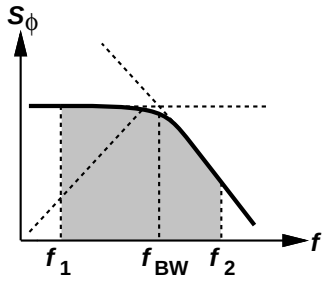


Fig. 10.31 Jitter generation calculation.

Example 10.10

For a 10-Gb/s linear CDR with f_{BW} = 10 MHz. Determine the minimum required VCO phase noise of the CDR if it is to be used in an OC-192 system.

Solution:

Let's pick f_o = 1 MHz, $S_{\Phi,VCO}$ is given by

$$S_{\Phi,VCO}\,(1MHz)\ = 1.25 \times 10^{-8}\,(Hz^{-1}). \tag{10.92}$$

Or equivalently, the VCO most present a phase noise $\mathcal{L}$ less then -79 dBc/Hz at 1-MHz offset.

How about the JG of bang-bang CDRs? The output jitter is still dominated by the VCO noise. Once we obtain the transfer function Φ_{out}/Φ_{VCO} of a binary loop, JG becomes readily available. The question is that, we need to know the operation mode of bang-bang PD under locked condition in the presence of VCO noise. Does the BBPD stay in the linear region of $\pm\Phi_m$ most of the time? Or it slews from time to time as the case for JTRAN and JTOL?

To answer this question, we go back to the definition of JG. Recall that JG_{rms} = 0.01 UI. Even with a very narrow linear region, say, $\pm$0.03 UI, the BBPD can still find that 99.9$\%$ of the sampled phase errors locate within the linear region! In other words, it is fair enough to say that the VCO phase noise experience linear operation around the loop. With the same notation as Fig. 10.29, we recalculate the noise transfer function. The transfer function is still given by

$$\frac{\Phi_{out}}{\Phi_{VCO}} \cong \frac{s}{s + 2\zeta\omega_n}. \tag{10.93}$$

Now the nature frequency and damping factor become

$$\omega_n = \sqrt{\frac{K_{VCO} I_p}{\Phi_m C_p}} \tag{10.94}$$

$$\zeta = \frac{R_p}{2}\sqrt{\frac{K_{VCO} I_p C_p}{\Phi_m}}, \tag{10.95}$$

simply because the equivalent PD+CP gain here is I_p/Φ_m rather than $I_p/2\pi$. The loop bandwidth is thus equal to

$$\omega_{BW,BB} = 2\pi f_{BW,BB} = \frac{K_{VCO} I_p R_p}{\Phi_m}. \tag{10.96}$$

With the same token, we can estimate the jitter generation for bang-bang CDRs as

$$JG_{rms,BB} = \frac{f_o}{2}\sqrt{\frac{S_{\Phi,VCO}(f_o)}{\pi f_{BW,BB}}}\,(UI). \tag{10.97}$$

It is worth noting that there are other sources causing jitter on the recovered clock. For example, the undesired coupling from data, supply noise, etc. Building a sophisticated model is necessary for designers to accurately estimate the overall jitter generation performance.

REFERENCES

[1] C. R. Hogge, A Self-Correcting Clock Recovery Circuits, IEEE J. Lightwave Tech., vol. 3, pp.1312-1314, Dec. 1985.

[2] J. D. H. Alexander, Clock Recovery from Random Binary Data, Electronics Letters, vol. 11, pp. 541-542, Oct. 1975.

[3] J. Savoj and B. Razavi, A 10-Gb/s CMOS Clock and Data Recovery Circuit with a Half-Rate Linear Phase Detector, IEEE Journal of Solid-State Circuits, vol. 36, pp. 761-768, May 2001.

[4] Jri Lee and Behzad Razavi, A 40-Gb/s Clock and Data Recovery Circuit in 0.18-μm CMOS Technology, IEEE Journal of Solid-State Circuits, vol. 38, pp. 2181-2190, Dec. 2003.

[5] Rodoni et al., 5.75 to 44Gb/s quarter rate CDR with data rate selection in 90nm bulk CMOS, Proc. ESSCIRC, 2008, pp. 166-169.

[6] T. Toifl, C. Menoifl,et al., A Low-Power 40 Gbit/s Receiver Circuit Based on Full-Swing CMOS-Style Clocking, Compound Semiconductor Integrated Circuit Symposium, 2007, pp.1-4, Oct. 2007.

[7] Jri Lee and Shanghann Wu, Design and Analysis of a 20-GHz Clock Multiplication Unit in 0.18-μm CMOS Technology, Digest of Symposium on VLSI Circuits, pp. 140-143, June 2005.

[8] A. Pottbacker, U. Langmann, and H.-U. Schreiber, A Si Bipolar Phase and Frequency Detector for Clock Extraction up to 8Gb/s, IEEE Journal of Solid-State Circuits, vol. 27, no. 12, pp. 1747-1751, Dec. 1992.

[9] Jri Lee, Ken Kundert and Behzad Razavi, Analysis and Modeling of Bang-Bang Clock and Data Recovery Circuits, IEEE Journal of Solid-State Circuits, vol. 39, pp. 1571-1580, Sept. 2004.

[10] Jri Lee and M. Liu, A 20-Gb/s Burst-Mode CDR in 90-nm CMOS, Digest of International Solid-State Circuits Conference, pp. 46-47, Feb. 2007.

[11] Jri Lee and M. Liu, A 20-Gb/s Burst-Mode Clock and Data Recovery Circuit Using Injection-Locking Technique, IEEE Journal of Solid-State Circuits, vol. 43, pp. 619-630, Mar. 2008.

[12] Jri Lee and K. Wu, A 20Gb/s Full-Rate Linear CDR Circuit with Automatic Frequency Acquisition, Digest of International Solid-State Circuits Conference, pp. 366-367, Feb. 2009.

PAM4 SERDES DESIGN

11

Owing to its twofold bandwidth efficiency, pulse-amplitude modulation (PAM) signaling becomes popular recently as data rate goes higher and higher. For example, a 400-Gb/s Ethernet system may require 8-lane data channels, which needs 50+ Gb/s bandwidth for each channel. If PAM4 is adopted, one can achieve 50-Gb/s data rate while keeping the 25-GHz optical components. Other applications such as backplane and chip-to-chip data links have similar tradeoffs. We study PAM4 SerDes in detail here.

11.1 GENERAL CONSIDERATION

In chapter 1, we have looked at the fundamental characteristics of PAM4 signal. We investigate its advanced properties in this section.

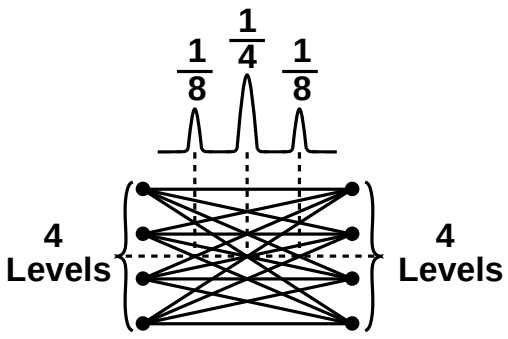


Fig. 11.1 Multiple crossover of PAM4 signaling.

Multiple Crossover The transition between 4 levels of PAM4 signal intrinsically reveals multiple zero-crossing points. If the middle line is taken as a threshold, we observe 3 cross-over points

as shown in Fig. 11.1. Among the 16 possible transitions (between adjacent symbols), 1/4 of them cause “middle crossover” points. Each of the two “side crossover” points has 1/8 chance of occurrence. This behavior inevitably leads to CDR design difficulty and large jitter. After all, the random wandering is nothing more than a broadband phase modulation on the input. While the high-frequency part would be rejected by the limited loop bandwidth of CDR, the low-frequency modulation drags the recovered clock phase and results in large jitter. It is intuitive to predict that a linear CDR would perform better than its bang-bang counterpart due to the proportionality. We address this issue again in the discussion of PAM4 CDR design.

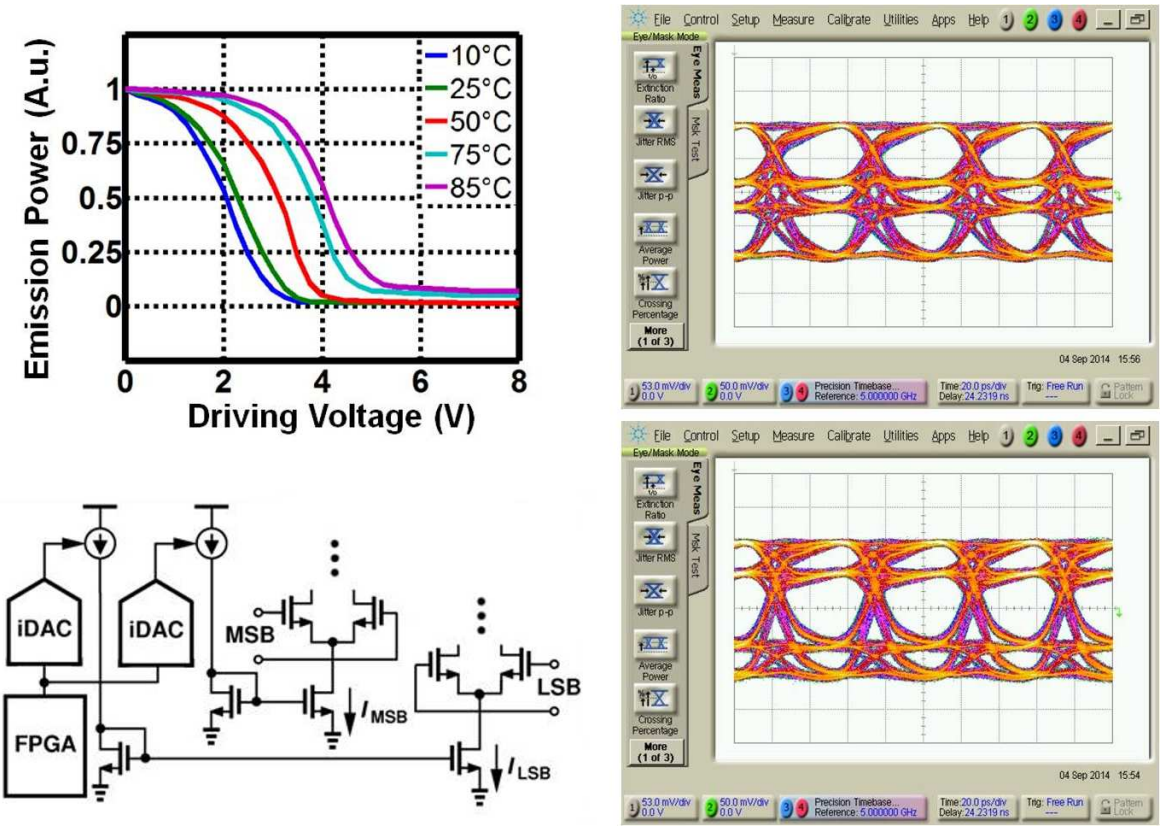


Fig. 11.2 EML nonlinearity.

EML Nonlinearity In optical applications, a typical electroabsorption-modulated laser (EML) would present a transfer characteristic as illustrated in Fig. 11.2. The nonlinear transfer function degrades the RX’s SNR and sensitivity. To obtain 4 uniformly-distributed levels at the input of RX,

the TX's output must be pre-distorted, i.e., squeezing the two middle levels. It is usually done by introducing a current-steering combiner, deviating the current ratio between I_{MSB} and I_{LSB} away from 2:1. For a given temperature, the two iDACs provide corresponding tail currents to the two data paths of the combiner, generating necessary pre-distortion. By doing so, the level-adjustable range can be as large as $\pm 100\%$, well-beyond any possible EML distortion. Two measured predistortion cases are illustrated in Fig. 11.2 as well.

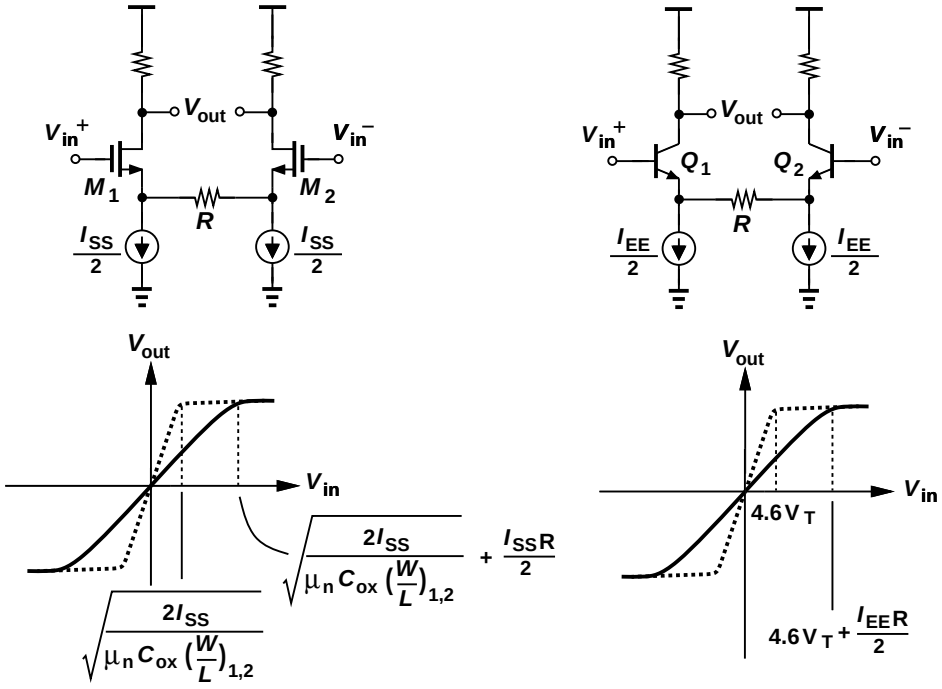


Fig. 11.3 Linearization of resistive degeneration.

Linearity One major difference between PAM4 and NRZ data is that the former needs to maintain linearity along the whole data path. In addition to the optical nonlinearity described above, amplifiers in the RX suffers from the same issue. Limiting amplifiers are no longer suitable for PAM4 obviously. Resistive degeneration in differential pairs serves as one major technique for linear amplifier.

Example 11.1

Determine the extended linear regions for the source and emitter degeneration pairs shown in Fig. 11.3.

Solution:

The linear region for CMOS differential pair would be extended by $\pm I_{SS}R/2$ as all of I_{SS}/R flows through R. Similarly, the linear region for bipolar differential pair would be enlarged by $\pm I_{EE}R/2$.

DFE Another issue arisen from the decision-feedback equalizer. For a NRZ data path, DFE is placed between the CTLE and DMUX with CDR providing the clock. For PAM4 signal, however, it is somewhat equivocal. As we know, a PAM4 signal must be first decomposed by a 3-threshold comparator to get the thermometer code. The subsequent thermometer-to-binary decoder and DMUXes restore the signal to NRZ format. Thus, it is worth thinking about the right position to place a DFE. As will be disclosed in section 11.3, putting the DFE in front achieves the best performance at a cost of 3X hardware and power consumption.

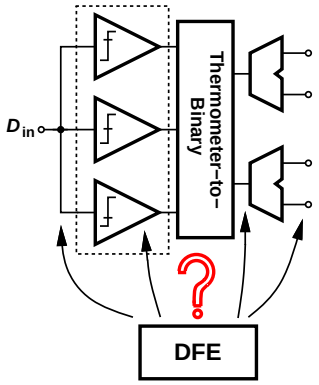


Fig. 11.4 Placing DFE along the PAM4 data path.

Other issues such as SNR and BER degradation and low-supply decoder design have been discussed in the previous chapters. A typical PAM4 SerDes structured is depicted in Fig. 11.5. At high speed, DLL or phase aligner may be incorporated in the TX to line up the phase between data and clock for the very last combiner stage. Adaptive equalization would be necessary for multi-purpose SerDes chips, and advanced CDR as well as eye monitoring circuit would be employed.

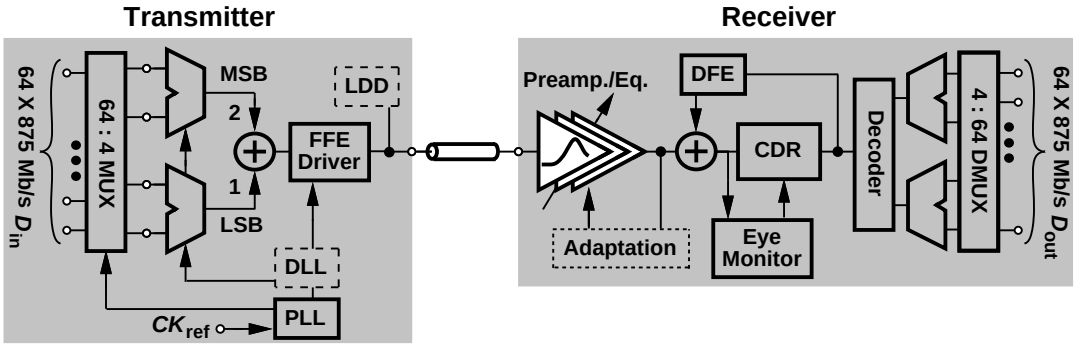
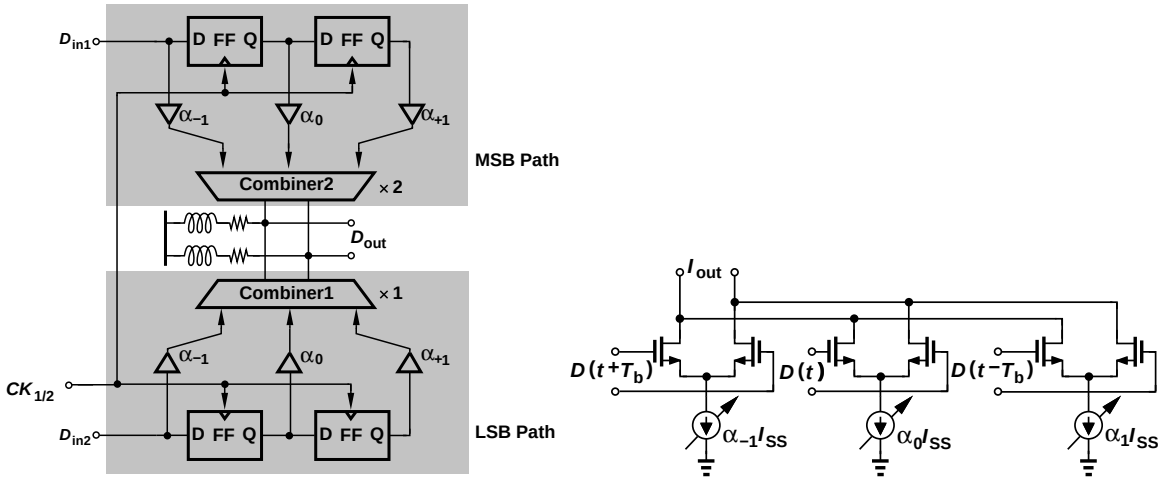


Fig. 11.5 High-speed PAM4 SerDes.

Nonetheless, the goal here is to overcome the difficulties of PAM4 signaling, and to make the full use of advantages of it. We start PAM4 SerDes design from the next section.

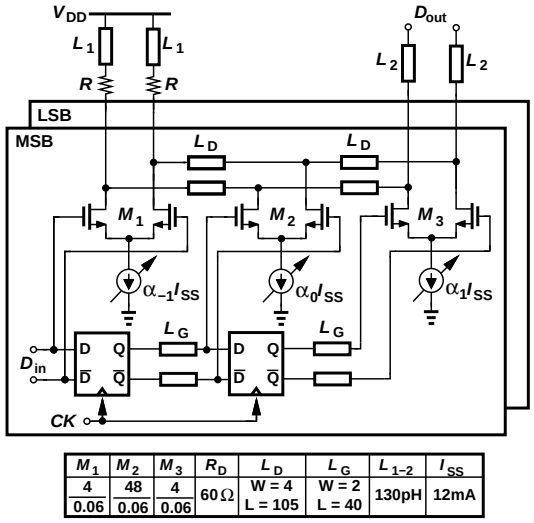
11.2 PAM4 OUTPUT DRIVER

Figure 11.6(a) illustrates a typical PAM4 driver with FFE. Two signal paths (MSB and LSB) are incorporated to provide pre-emphasis independently, serving as a 3-tap FFE with identical coefficients α_{-1}, α_0, α_1 on both sides. Recall from chapter 1 [Fig. 1.8(a)]The two preemphasis



(a) (b)

Fig. 11.6 (a) PAM4 Combiner/driver, (b) combiner details.



M_1	M_2	M_3	R_D	L_D	L_G	L_{1-2}	I_{SS}
4 / 0.06	48 / 0.06	4 / 0.06	60 Ω	W = 4 L = 105	W = 2 L = 40	130pH	12mA

Fig. 11.7 PAM4 output combiner with mm-wave technique.

results are combined together (with the MSB twice as large as the LSB) in current mode and converted to voltage output by means of the inductively-peaked terminations.The combiner design is depicted in Fig. 11.6(a), where the weighting factor tuning is realized by the tail currents. At tens of GHz, large elements such as inductors can no longer be considered lumped components, but instead distributed devices. In that sense, the peaking and signal-traveling circuits must be combined as a distributed network so as to minimize skews, reflection, and other non-idealities. Fig. 11.7 reveals the combiner design. Here, peaking inductors L_D and L_G are inserted between taps to absorb the gate and drain capacitance. These peaking inductors also sharpen the data transitions and reduce the skews to some extent. Design parameters for a 56-Gb/s PAM4 driver are also listed in Fig. 11.7. It is worth noting that the peaking inductors L_1 and L_2 steepen the rising and falling transitions by extending the bandwidth. However, these peaking inductors must be made more precisely than those in NRZ applications. It is because both overshoot (under-damped) or long tail (over-damped) responses would introduce deterministic ISI and further deteriorate the SNR.

$$P_{e,PAM4} = (1+2+2+1) \times \frac{1}{4} \times \int_{V_{pp}/(6\sigma_n)}^{\infty} \frac{1}{\sqrt{2\pi}} exp \frac{-x^2}{2} dx = 1.5Q\left(\frac{V_{pp}}{6\sigma_n}\right). \tag{11.1}$$

For an error rate less than 10^{-12}, the eye SNR [$=V_{pp}/(6\sigma_n)$] must be greater than **7.1** (= **17 dB**). The eye closes and SNR is severely degraded when there is either ringing or a long tail present in the received signal [Fig. 11.8(b)]. Inaccurate modeling on peaking inductors may cause significant degradation. Note that this impairment may not be repairable in the RX.[1]Optical drivers may contribute additional $2 \sim 3$ dB noise onto it. Other than the intrinsic half-rate structure, quarter-rate output driver for PAM4 signal is also feasible. We see the following example.

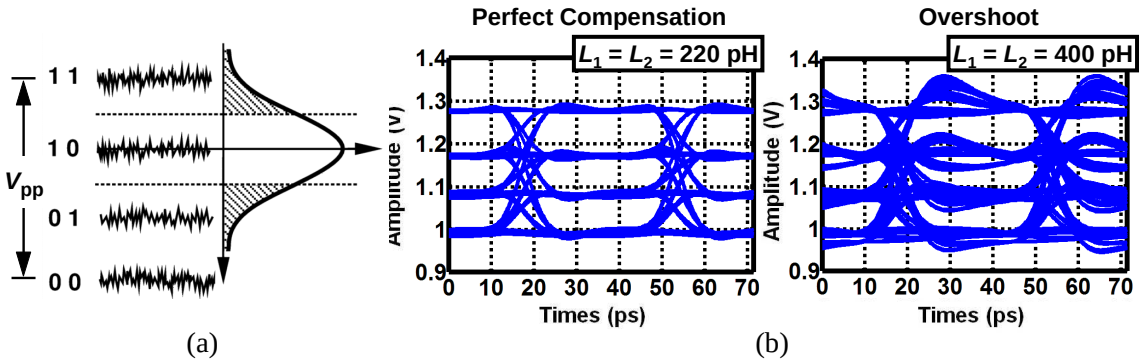


Fig. 11.8 (a) Effect of additive noise for PAM4 signal, (b) waveforms under different peaking inductance.

Example 11.2

Design a quarter PAM4 driver with 3-tap FFEs.

Solution:

Figure 11.9 reveals a design example, where 4 signal paths delivering 4 $\times$ 14-Gb/s signals through the driver. A 14-GHz PLL provides necessary quarter-rate clocks (in quadrature, CK_I and CK_Q). Here, CK_I drives all latches, which provides half-bit delay for 14-Gb/s data stream. In the case where clock-to-Q delay is negligible, one can apply CK_Q to the 2:1 selectors to achieve perfect timing for sampling. Two data paths are emphasized with identical coefficients α_{-1}, α_0, α_1. Finally, the 4 data streams are joined together by the two combiners to deliver 56-Gb/s output in PAM4 format.

[1]For example, DFE can handle only post cursors.

Example 11.2 ***(Continued)***

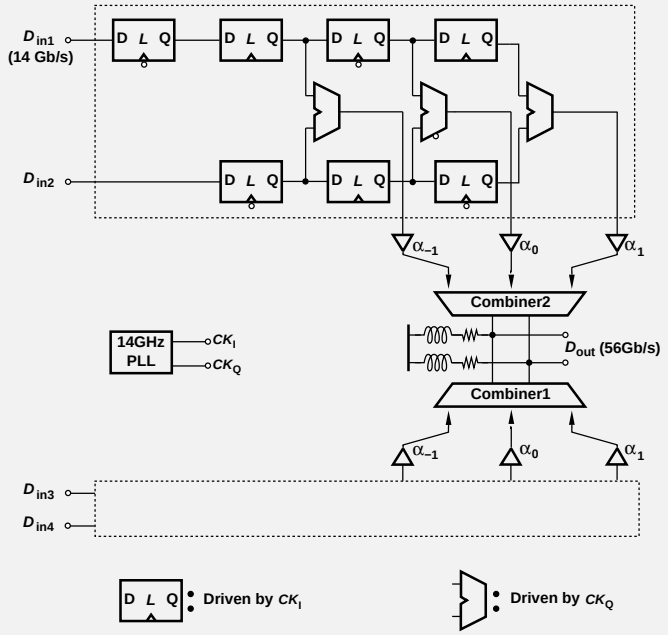


Fig. 11.9 Quater-rate PAM4 TX with 3-tap FFEs.

11.3 PAM4 RF FRONT-END

Implementing a PAM4 RX is more complicated than realizing a PAM4 TX. A fundamental PAM4 RX front-end must include a pre-amplifier and/or three comparators (or slicers) to discriminate the 4 levels. Limiting buffers (such as hysteresis buffers) must be used to create thermometer codes in full scale of swing. A PAM4 decoder thus converts the thermometer codes to binary codes. A CDR is definitely essential to synchronize all the half-rate binary bits before and after the decoder. The whole receiver works as a 2-bit ADC with the exception that the sampling clock is always synchronized with the input signal transition. Figure 11.10 illustrates a general realization of such a receiver.

The three preamplifiers can actually be combined as one circuit to save power. As depicted in Fig. 11.11, the switching quad $M_1 - M_4$, loading resistor R_1 and R_2, and tunable current source I_{SSA} and I_{SSB} produce three output $V_{out,1} - V_{out,3}$ with three different threshold levels.

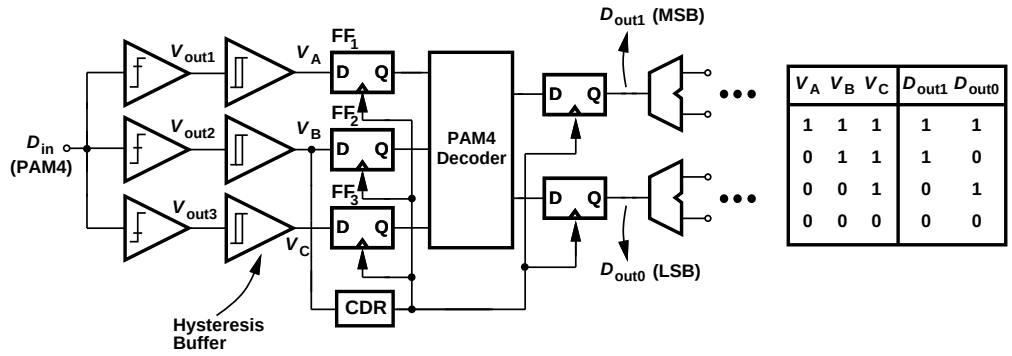


Fig. 11.10 DFE for PAM4 signal.

The upper and lower ones are symmetric with respect to the middle one, i.e., the input common-mode level. Note that the total current of I_{SSA} and I_{SSB} is kept constant so as to minimize the output common-mode variation. Most application would require an adaptive CTLE in front of the preamplifier which leads to an uncertain signal magnitude. The preamplifiers, however, necessitate constant input swing so as to obtain obtain the correct thermometer codes. As a result, we need an automatic gain control loop to fix the signal magnitude. Shown in Fig. 11.12(a) is an analog approach, where the PAM4 signal swing is detected by means of a power detector. On the other hands, a reference signal with constant swing (could be easily obtained by bandgap reference) is examined by another identical power detector. The difference between the two power detector is fed back to the control voltage of a variable gain amplifier (VGA), which may present a tunable range as large as 10 dB. In other words, the negative feedback loop forces the VGA to create a

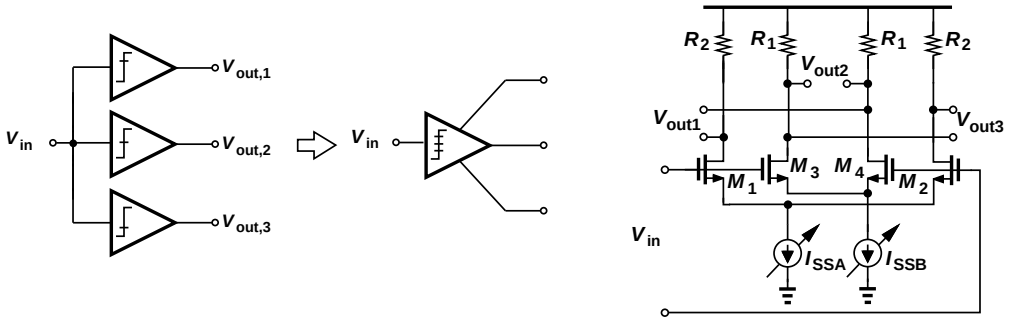


Fig. 11.11 PAM4 preamplifier obtaining 3 outputs in one shot.

constant PAM4 signal swing for the following preamplifier. Same approach can be implemented in digital domain, where the gain control is fully conducted in logics [Fig. 11.12(b)]. Note that comparing the DC power is not the only way to fulfill the gain control. Other techniques such as peak detecting could also serve in this applications. It is important to know that the DFE's feedback should be applied to the summer as well. As will be shown in section 11.4, each of the 3 thermometer output has to send a corresponding amount of feedback, arriving at 3 feedback paths for each tap. The AGC loop here must accommodate the adaptive tuning of DFE so as to ensure a constant input swing for the preamplifier.

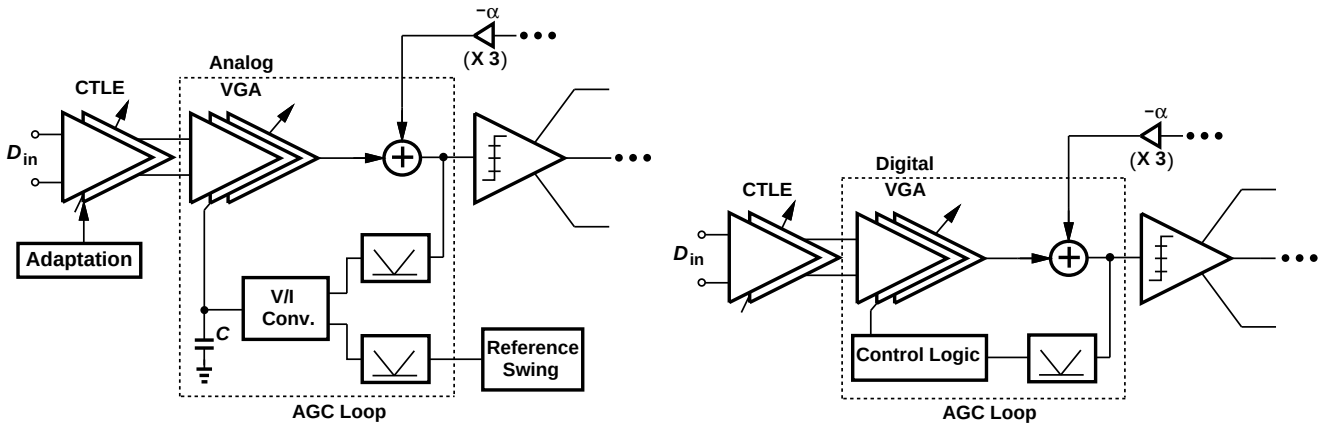


Fig. 11.12 Automatic gain control for PMA4 signal realized in (a) analog, (b) digital domain.

While providing a straightforward solution, the simple realization shown in Fig. 11.10 may suffer from a series of issue. First of all, in real applications, channel loss would be as high as 25-30 dB at Nyquist frequency. It is actually mandatory rather than optional to incorporate both CTLE and DFE. Meanwhile, with high channel loss and severe signal distortion, analog CDR may not be a good choice. The CDR needs to be modified to minimize jitter and power consumption. Advanced techniques such as eye opening monitor circuitry are recommended to further reduce BER. We investigate the DFE, CDR, and eye monitoring techniques for PAM4 signal in the following section.

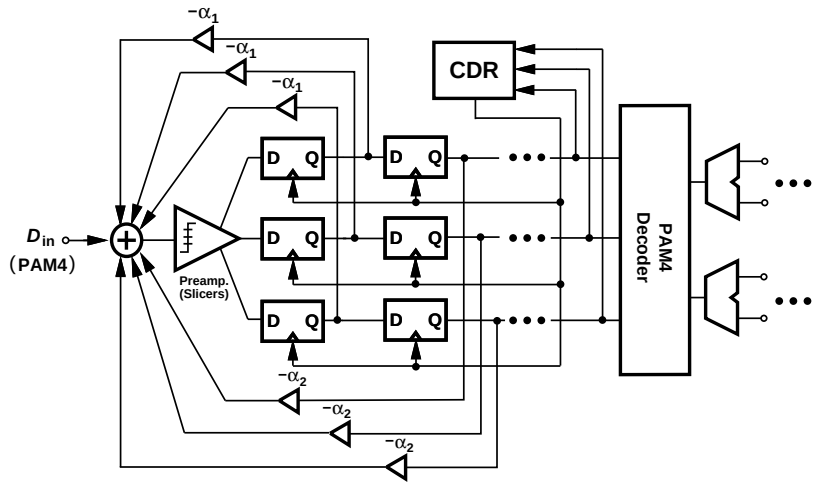


Fig. 11.13 DFE for PAM4 signal.

11.4 DFE FOR PAM4 SIGNAL

Recall from chapter 5 that a NRZ DFE is realized as all the feedbacks aggregated at the summer in front of the slicer. In PAM4, the preamplifiers together with subsequent comparators or hysteresis buffers serve as slicers. The difference is that there are 3 outputs of thermometer codes, and each of them deserves a feedback path. As a result, a DFE for PAM4 signal is implemented as shown in Fig. 11.13. Each tap has 3 feedbacks with identical coefficient $-\alpha_1$, $-\alpha_2$, $\cdots$ and so on. In that sense we need 3 times more flipflops to take care of the delay for each thermometer output. The power consumption dedicated to PAM4 DFE is roughly 3 times larger than that of NRZ DFE. Once again, the equalization is accomplished by sacrificing the low-frequency power, and the CDR is responsible for proper clocking of the flipflops.

Example 11.3

Consider the channel we described in example 5.5, which presents a single-pulse response of [0.7, 0.2, 0.1]. Now a PAM4 signal is applied into this channel (Fig. 11.14). Design a two-tap PAM4 DFE for it.

Example 11.3 ***(Continued)***

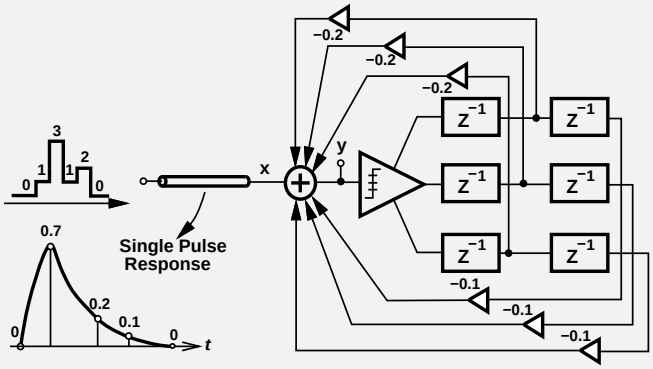


Fig. 11.14 PAM4 DFE with 2 taps.

Solution:

Based on the study in chapter 5, we realize the optimal DFE coefficients are the post-cursors. It is still the case for PAM4 signal. A complete design is depicted in Fig. 11.13.

To verify it, we draw the discrete waveform at node x and y in Fig. 11.15. It can be shown the equivalent data at node y are perfectly compensated with magnitude degraded by 30%.

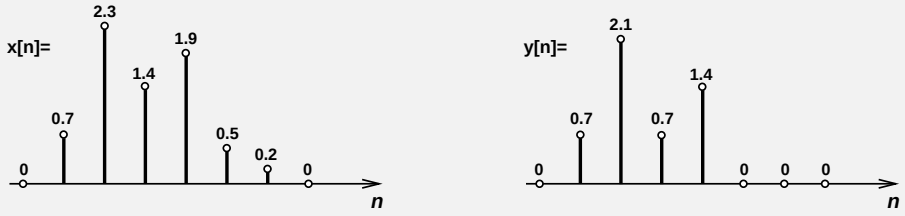


Fig. 11.15 Response at node x and y.

It is not difficult to modify the PAM4 DFEs with the known techniques developed for NRZ DFEs. For example, loop unrolling and sub-rate structures can be adopted. The intrinsic bandwidth efficiency of PAM4 allows the DFE to operate at half data rate. Further paralleling the data path actually benefits the power performance, as more digital flipflops could be used. For example, for a 56-Gb/s PAM4 SerDes, it would be desirable to adopt a quarter-rate structures, in which the quantized signals are running at only 14 Gb/s.

How do we make a PAM4 DFE adaptive? To dynamically adjust the coefficients, we need to know the dc power of the PAM4 signal right before the preamplifiers (slicers). However, it is even

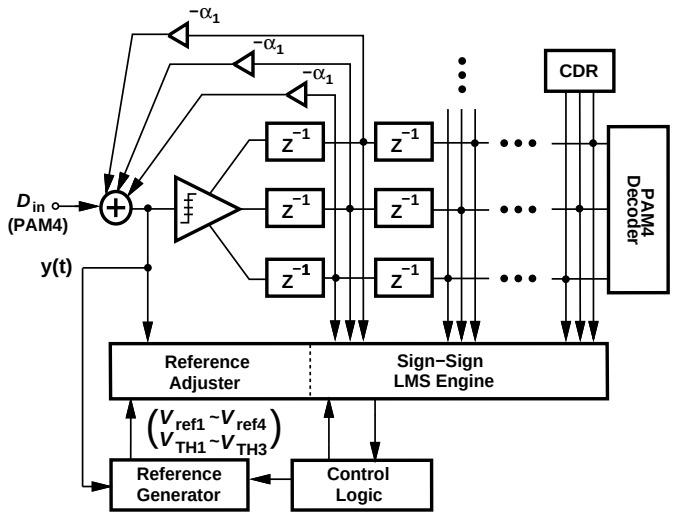


Fig. 11.16 Adaptive PAM4 DFE with dynamic level tracking technique.

harder than NRZ to detect the dc power level of a PAM4 signal by simple analog circuits. Thus, we resort to dynamic level tracking technique. Figure 11.16 illustrates the realization. Compared with Fig. 5.xx, now we need to create 4 signal levels (V_{ref1}, V_{ref2}, V_{ref3} and V_{ref4}, representing 00, 01, 10 and 11) and 3 threshold levels (V_{TH1}, V_{TH2}, and V_{TH3}). The differences between adjacent levels are equal. The easiest way to implement it is to use a voltage ladder as that in Fig. 5.xx, in which 7 instead of 3 reference voltages need to be created. Figure 11.17 shows the design of reference generator for PAM4 DFE.

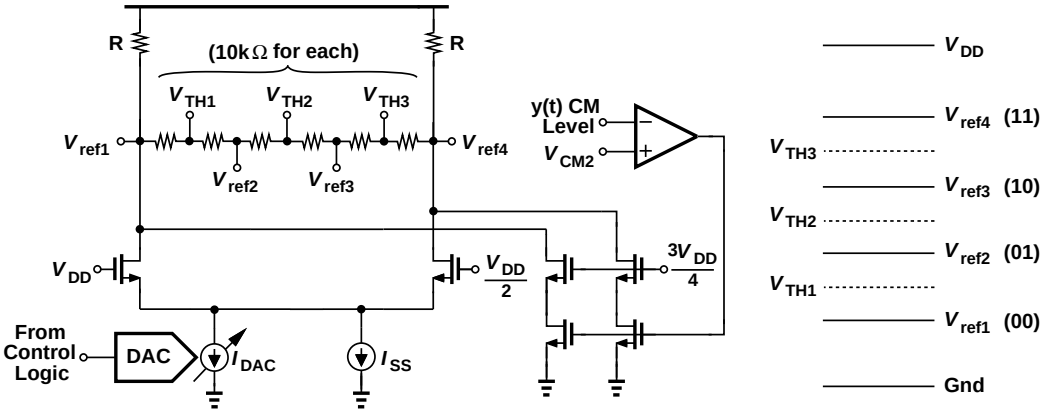


Fig. 11.17 Reference generator.

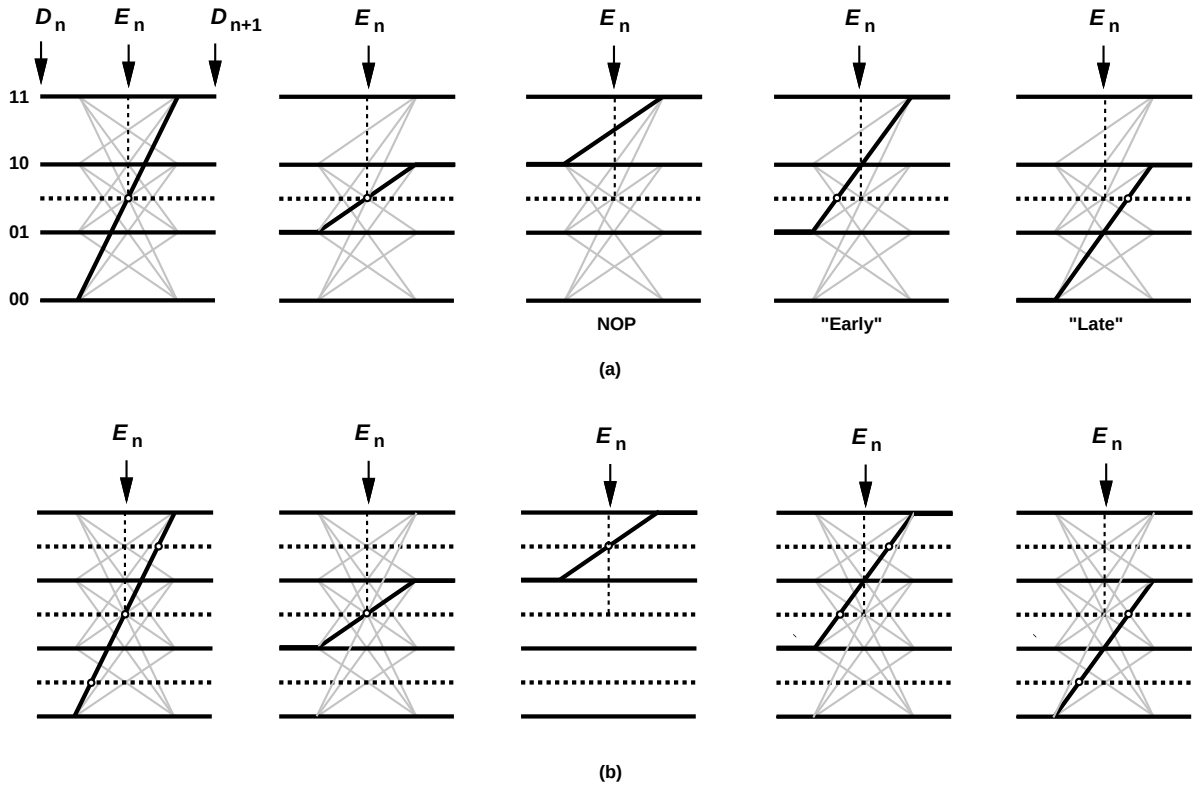


Fig. 11.18 PAM4 CDR with (a) single PD, (b) 3 PDs.

The calibration procedure is similar to that for NRZ data. First, we line up V_{TH2} with the common mode level of $y(t)$ by means of the Opamp loop. Then, we stretch or compress the reference and threshold levels (by means of I_{DAC}) until $V_{ref1} \sim V_{ref2}$ match the average signal levels. Finally, we turn on the sign-sign LMS engine to optimize DFE coefficients.

11.5 CDR FOR PAM4 SIGNAL

From our previous discussion, we realize that proceeding V_B in Fig. 11.10 with a traditional NRZ CDR is possible to lock the clock frequency to Baud rate. The jitter of recovered clock, however, is expected to be higher, simply due to multiple zero-crossing points. As depicted in Fig. 11.18(a), transitions between levels cause 3 different crossover points if only one threshold (i.e., the common-mode level) is used. The middle crossover occurs when the transition goes from “00”

to “01” and from “01” to “10” (and vice versa). If the transition goes from “01” to “11” or from “00” to “10” (and vice versa), early or late crossover would appear. It is desirable to remove this effect by circuit techniques. Figure 11.18(b) illustrates a great example. It can be clearly shown that if all 3 thermometer code outputs (V_A, V_B, and V_C in Fig. 11.10) are examined by PDs, the early or late crossover always happen concurrently and cancel out each other. It leads to zero net effect on the clock adjustment, arriving at much better jitter performance. Furthermore, since side transition such as from “10” to “11” and from “00” to “01” can also be examined, more phase comparison could be made to help improve the CDR performance. Both linear and binary PDs can be adopted here. The latter is preferable for a highly parallelized RX structure with all-digital implementation.

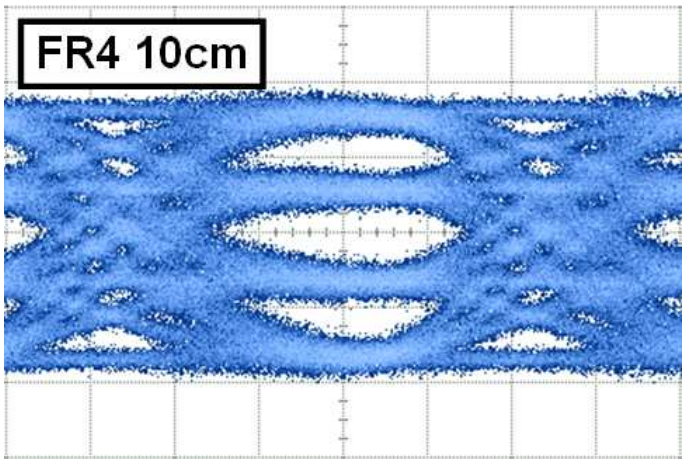


Fig. 11.19 Example of measured PAM4 data eye at 20-Gb/s after a 10-cm trace on FR4 board.

The CDR itself might not be adequate for the SerDes to achieve the optimal performance. For example, the asymmetric opening of the 3 eyes may need a shift on the data sampling point in order to optimize BER. Figure 11.19 reveals a typical PAM4 waveform after a 10-cm trace on FR4 board. The eye in the middle is obviously bigger than the other two. The optimal data sampling point may not be right in the eye center, but rather to the right slightly. An eye opening monitor circuit could be helpful. We look at it in the next section.

11.6 EYE MONITORING FOR PAM4 SIGNAL

PAM4 signal can be monitored in real time to achieve the lowest BER. Unlike the case for NRZ data, a PAM4 eye monitor needs to examine 3 data eyes for a given clock phase. Figure 11.20 illustrates the operation. Suppose the CDR provides a clock phase CK_{data} for data sampling, which is obtained from the above 3-level phase detection. Obviously CK_{data} falls in the nominal center of data eyes. Now, a variable clock CK_{ϕ} is created by introducing a tunable delay ΔT to check eye openings at different position. Meanwhile, a variable threshold level V_{TH} is added in the front-end, resulting in a 2-D eye opening monitor. The three black little box (■) represent the nominal sampling results from the eye centers, and the white little box (□) stands for the present checking point. If a certain checking point is error free,[2] its result must be coincident with one of the black box. In other words, one XOR gate will always produce logic zero whereas the other two have logic Ones.

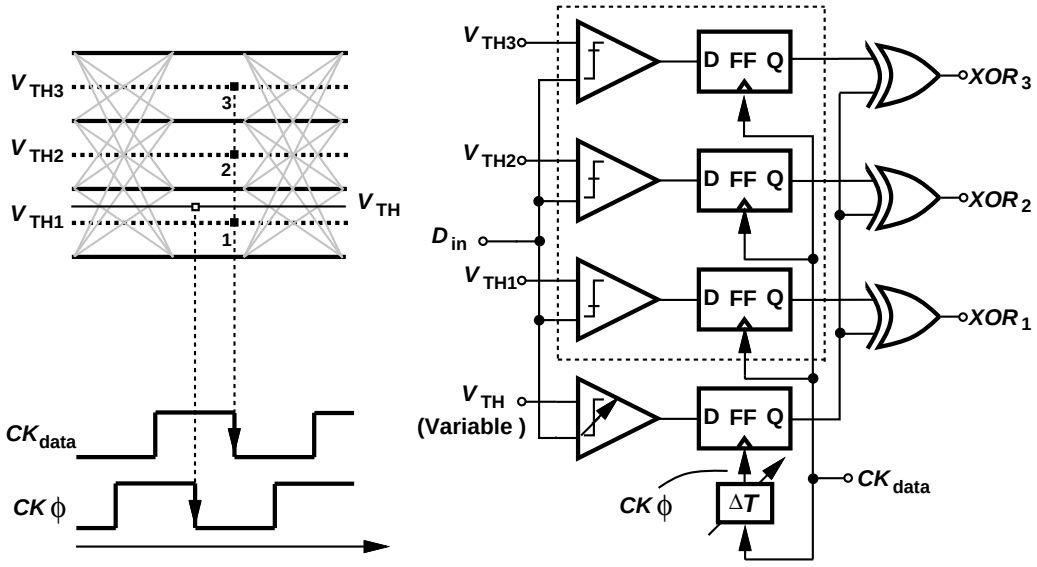


Fig. 11.20 Eye monitoring for PAM4 signal.

Defining a certain testing length, say, 1000 bits, we can determine whether this checking point is inside the opening area and which eye it belongs to. By sweeping V_{TH} and CK_{ϕ}, one can obtain

[2]Error free here is defined with respect to testing

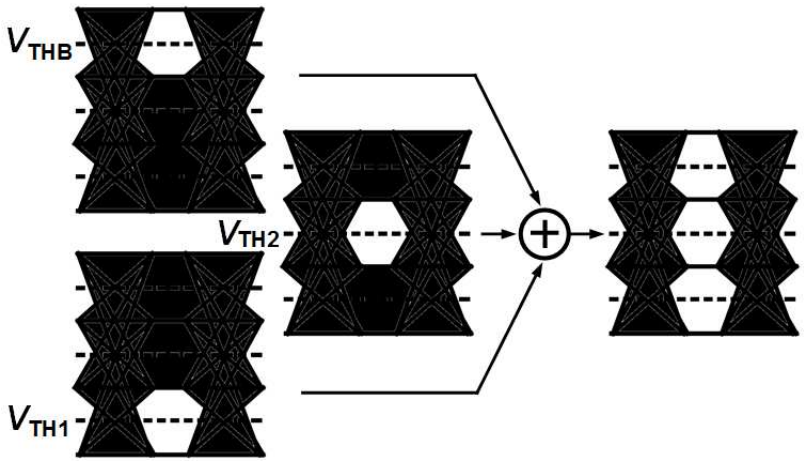
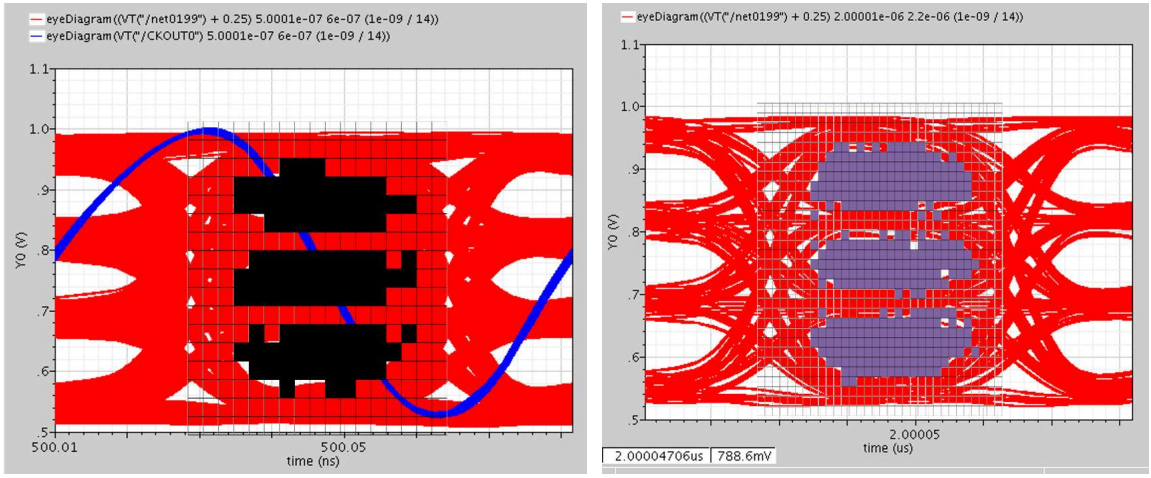


Fig. 11.21 Eye opening reconstruction.

a complete 2-D eye opening map as illustrated in Fig. 11.21. Depending on the available testing time, we plot eye monitoring results with different resolution. Figure 11.22 depicts the cases for 16×16 and 32×32 pixels.



(a) (b)

Fig. 11.22 Simulated eye opening monitor for (a) 16×16, (b) 32×32 pixels per symbol.

A complete PAM4 Rx design including a 3-level PD, a digital loop filter, and eye monitoring circuits can be studied as an example to close our discussion. As shown in Fig. 11.23, such a PI-based CDR loop employs a 2nd order DLF to accommodate frequency offset (chapter 9).

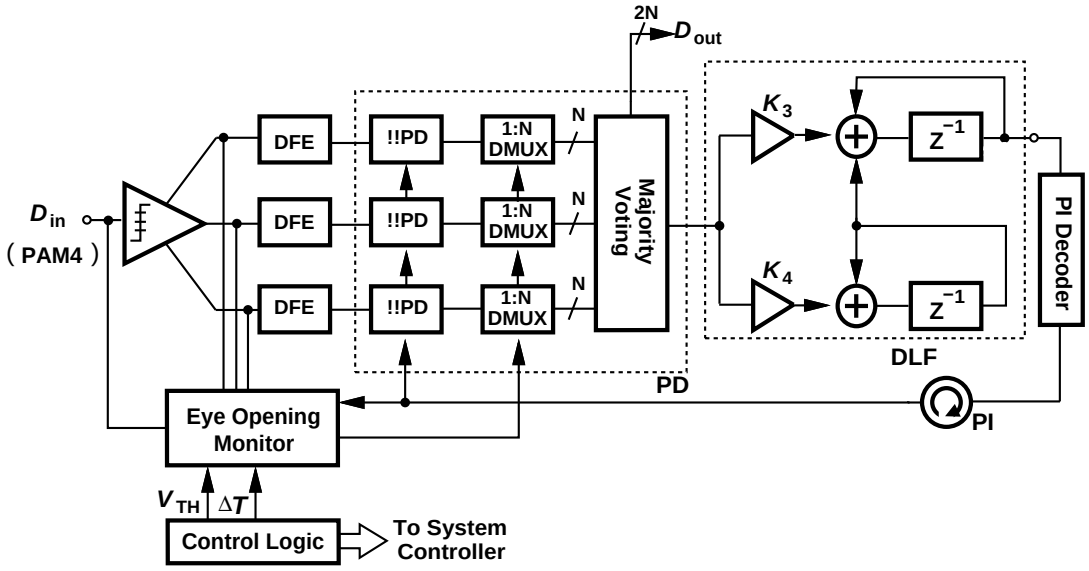


Fig. 11.23 Complete PAM4 RX including 3-way bang-bang PD, DLF, and eye monitoring.

Bang-bang phase detectors are followed by demultiplexes to slow down the data processing rate, facilitating digital implementation on the majority voting machine and other circuit. Phase interpolator is expected to have $7 \sim 8$ bit of resolution, and special techniques such as dithering [1] can be introduced to further reduce the jitter. Eye monitoring is included as well, which dynamically optimizes the sampling points of demultiplexers to minimize BER. The real-time eye situation can be sent to the system controller for monitoring.

11.7 DUOBINARY TX

In chapter 1 we have studied the fundamental operation of duobinary signal. We look at the physical design of duobinary transceivers in 11.7 and 11.8.

The duobinary signal is actually implemented by utilizing the low-pass channel response to fulfill a $1 + z^{-1}$ transfer function. For convenience we replot the conceptual TRX diagram in Fig. 11.24. The TX-side FFE, channel, and RX-side CTLE work together to form a response approximation equal to $1 + z^{-1}$. To restore the signal to NRZ, a decoder of response $1/(1 + z^{-1})$ must be added as well. Such an IIR filter contains a feedback loop, making itself vulnerable if an

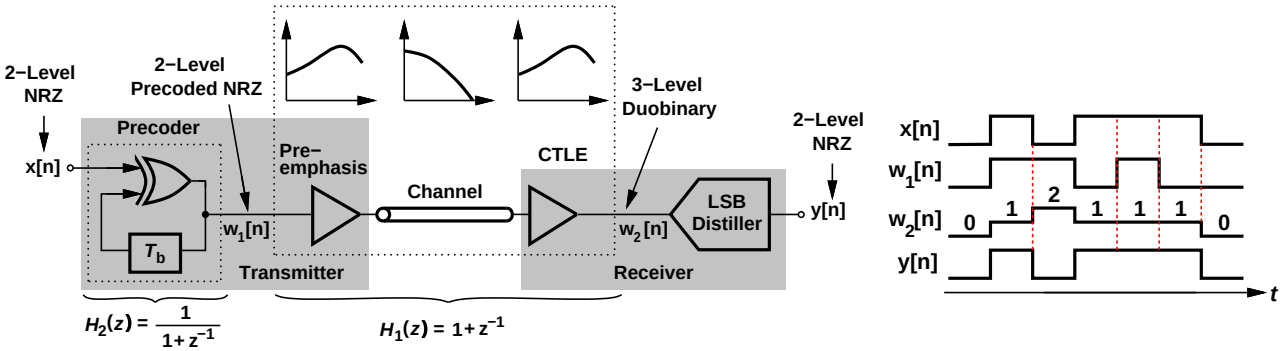


Fig. 11.24 Simplified duobinary TRX with precoder.

error occurs. That is, the error bit circulates along the loop, demolishing subsequent bits. Figure 11.25 illustrates such a phenomenon. If one bit of $y[n]$ is incorrectly decoded, the following bits are all wrong. In other words, the error bit propagates like a domino. Therefore, it is preferable to put the decoder in the TX side rather than the RX side, as the former contains complete, well behaved bits. Renamed as a precoder, this block is also designed for mod 2 operation. That is, the XOR gate behaved as half adder, arriving at a 2-level precoded NRZ of $W_1[n]$.

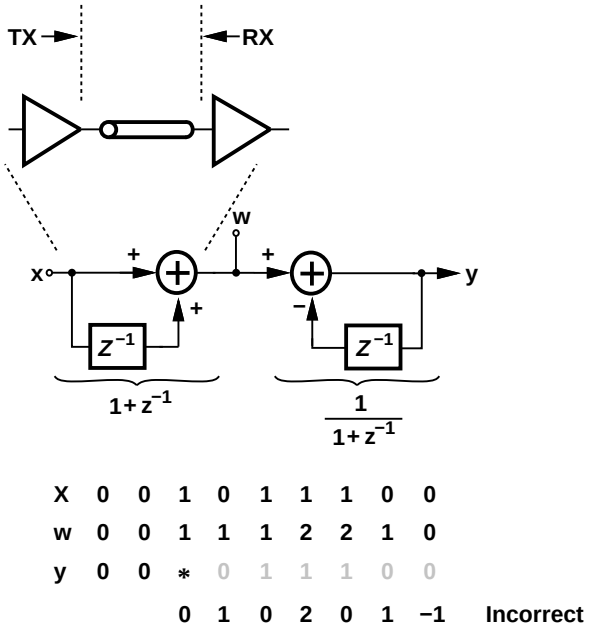


Fig. 11.25 Domino effect of wrong bit if decoder is used in the receiver side.

How do we realize a precoder? Although it looks simple and feasible, the precoder in Fig. 11.24 is difficult to implement, primarily due to the stringent timing requirement in the feedback loop. Using a clock-driven flipflop seems to be the only choice, but it suffers from severe phase requirement. This effect can be clearly explained by Fig. 11.26(a), where the XOR gate and the flipflop experience a delay of T_{XOR} and $T_{D\to Q}$, respectively. To make this precoder work properly, these two delays must comprise an exact bit period T_b:

$$T_{XOR} + T_{D\to Q} = T_b. \tag{11.2}$$

That is, the input clock CK_{in} has very little margin for phase movement in order to produce a proper D-to-Q delay for the flipflop. Such a timing issue becomes aggravated at high speed and requires a complex control scheme.

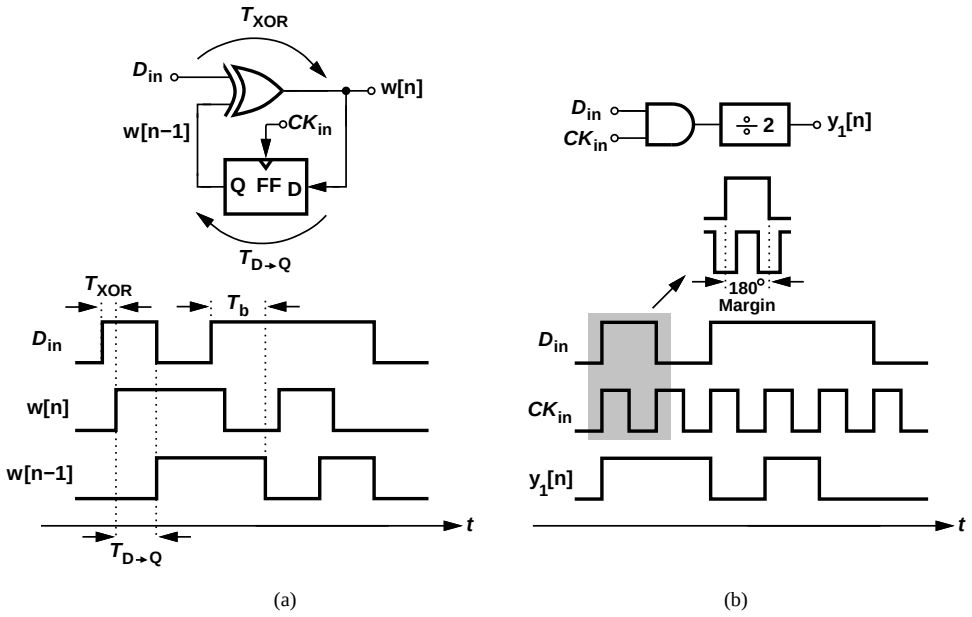


Fig. 11.26 Duobinary precoder design: (a) conventional, (b) open-looped.

To overcome the difficulties, we realize the precoder in an alternative way as illustrated in Fig. 11.26(b), [2]. The input data and clock pass through an AND gate, which is followed by a divided-by-2 circuit. The output thus toggles whenever a data ONE arrives, leading to the following operation:

$$y_1[n] = y_1[n-1] \oplus D_{in}[n]. \tag{11.3}$$

This structure provides advantages over that in Fig. 11.26(a) in breaking the loop and allowing much more relaxed phase relationship between the input clock and data. The clock CK_{in} now reveals a margin as wide as 180° for skews, which is no longer a limiting factor in most designs. Note that the initial state of the divider has no influence on the final result; $y_1[n]$ with opposite polarity still yields the same output after decoding.

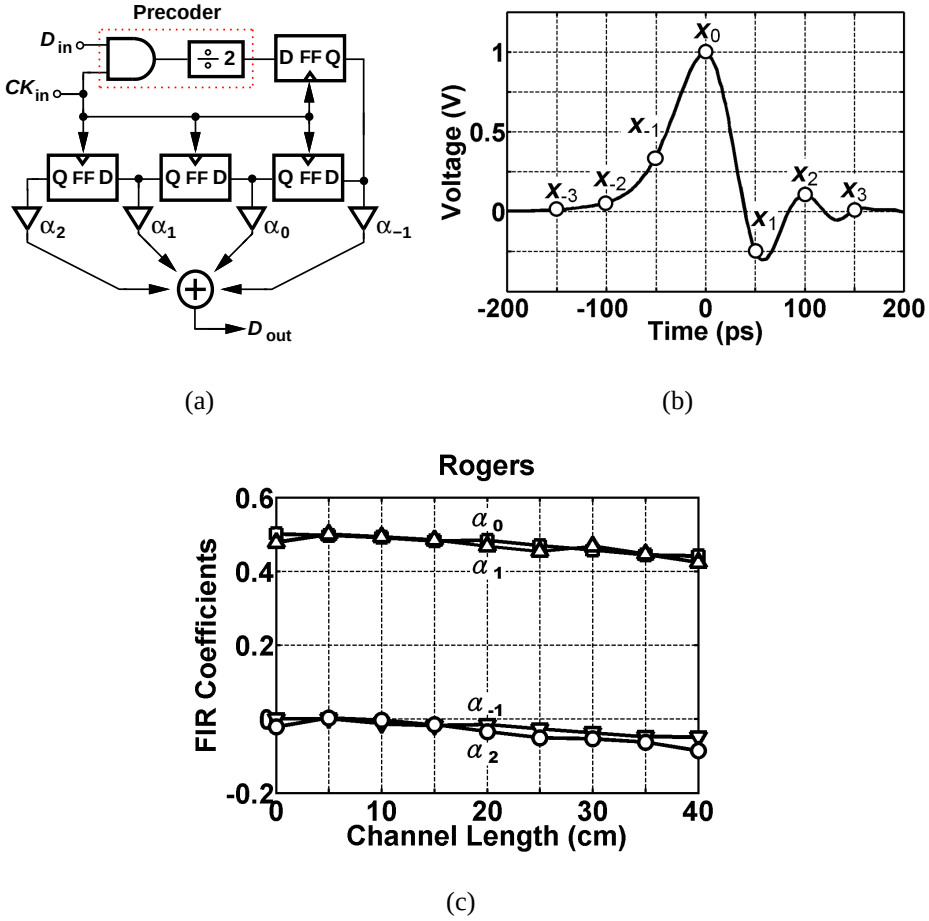


Fig. 11.27 (a) Duobinary TX with FFE, (b) typical 20-Gb/s pulse response, (c) example of coefficient.

The popular feedforward equalizer also proves useful in duobinary systems. Similar to that for NRZ data, FFE for duobinary should limit the number of taps if we are targeting high-speed

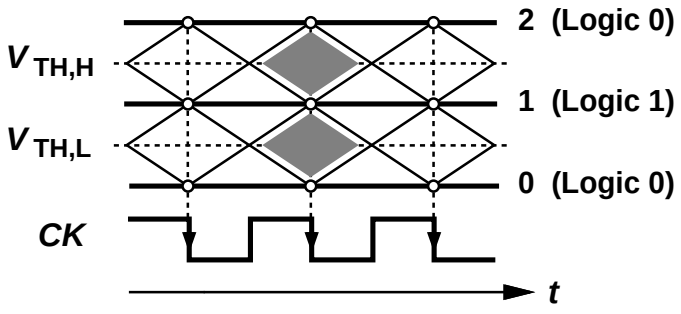


Fig. 11.28 Threshold setup for duobinary signal.

operation. Figure 11.27(a) illustrates an example of simplified duobinary TX with 4-tap FFE as an output driver. All the FIR equalizing methods and techniques that have been extensively used for NRZ data can be applied in duobinary, except that a single pulse ONE (preceded and followed by successive ZEROs) is expected to generate two consecutive bits of $1/2$ at the far end. With a pulse response shown in Fig. 11.27(b),[3] the coefficients α_k are readily available by solving the following equations:

$$\begin{bmatrix} x_0 & x_{-1} & x_{-2} & x_{-3} \\ x_1 & x_0 & x_{-1} & x_{-2} \\ x_2 & x_1 & x_0 & x_{-1} \\ x_3 & x_2 & x_1 & x_0 \end{bmatrix} \begin{bmatrix} \alpha_{-1} \\ \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1/2 \\ 1/2 \\ 0 \end{bmatrix}. \tag{11.4}$$

Figure 11.27(c) reveals the calculated coefficients for it. Note that we have two main cursors now.

Owing to the limited bandwidth, the duobinary signal seen at the input of RX has long rise-fall times during transition. A typical duobinary eye diagram is shown in Fig. 11.28, where two outer levels represents logic Zero and the middle level logic One. Two diamond-shaped eyes locate on top of each other, and two thresholds ($\mathrm{V}_{TH,H}$ and $\mathrm{V}_{TH,L}$) symmetric with respect to the middle level are needed. Theoretically, three threshold can also be used for clock recovery, given that transitions of duobinary signals are pretty linear. We investigate duobinary RX design in the next section.

[3]The example pulse response shown in Fig. 11.27(a) is obtained from a 20-cm Rogers channel.

11.8 DUOBINARY RX

11.8.1 DFE

As we know, the FFE and CTLE collaborate with each other to reshape the channel response. A DFE can still be helpful in the final step of waveform reconstruction. Similar to that for PAM4 signals, a DFE for duobinary systems must have feedback from both data paths. Illustrated in Fig. 11.29 is an example, where the outputs of both slicers (or comparators) are fed back to the summer with the same coefficients. The rule of thumb here is to make the summer's output (which is a 3-level duobinary signal) as ideal as possible.

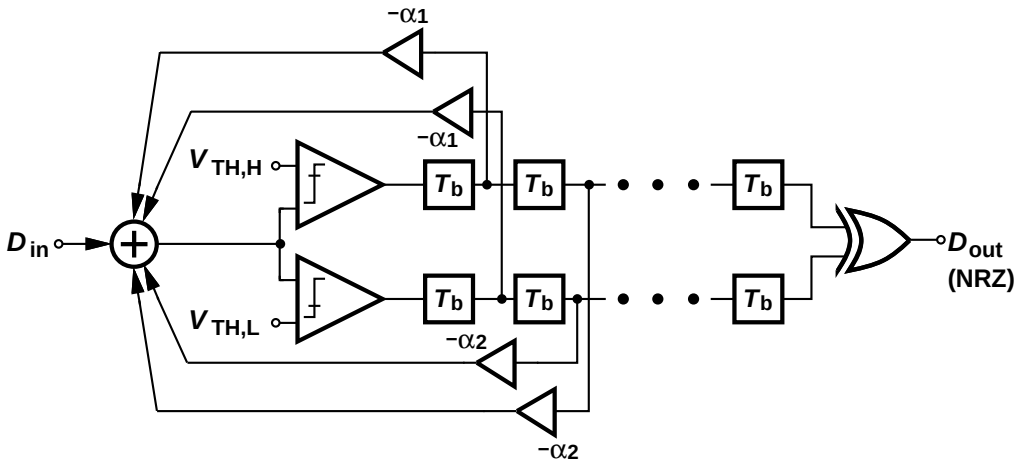


Fig. 11.29 Generic DFE in duobinary RX.

Example 11.4

Consider the setup in Fig. 11.30, where a single pulse $\{\ldots, 0, 0, 1, 0, 0, \ldots\}$ would be reshaped and amplified by the FFE-channel-CTLE combination as $\{\ldots, 0, 0.7, 0.85, 0.25, 0.2, 0, \ldots\}$. As a part of duobinary RX, the DFE tries to clean out the bit sequence with its best effort. Determine the coefficient and draw the waveform at node y if the DFE has 2 taps.

Example 11.4 ***(Continued)***

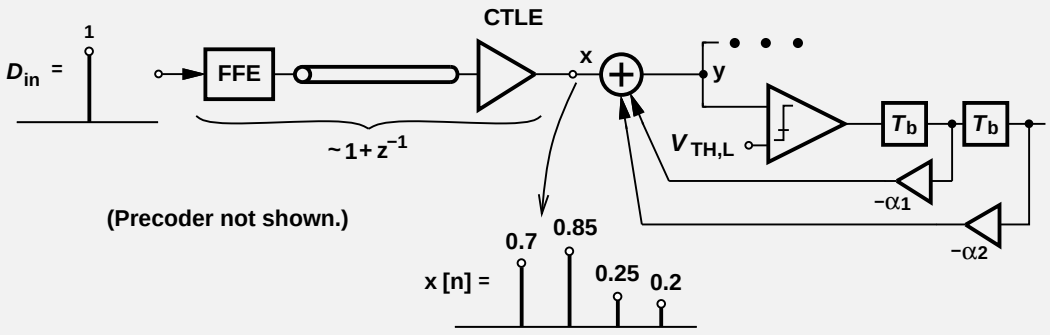


Fig. 11.30 Example of DFE coefficient setting.

Solution:

Only the lower part of the DFE is involved in the case here. Following the same approach as we did for NRZ data, we need $\alpha_1 = 0.15$ to quell the difference between the two One bits. By the same token, we have

$$0.25 - \alpha_1 - \alpha_2 = 0. \tag{11.5}$$

to null the first post cursor. It leads to $\alpha_2 = 0.1$. However, the second post cursor can not be compensated completely, leaving behind a magnitude of 0.1. The subsequent bits are all Zeros. Fig. 11.31 plots the results. Actually we need infinite taps to achieve a complete compensation.

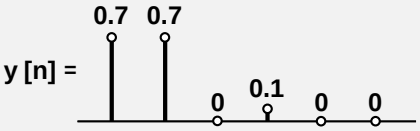


Fig. 11.31 Summer output $y[n]$.

A practical DFE design may resort to level generation technique that we introduced for PAM4. One possible realization is depicted in Fig. 11.32, where a level generator is responsible for creating the first level, i.e., Level 0, $V_{TH,L}$, Level 1, $V_{TH,H}$, Level 2, respectively. With Level 0 and Level 2 fitting to the average lines of the two outer levels, the two thresholds are produced accordingly. Sign-sign LMS algorithm is thus executed to optimize the coefficients dynamically.

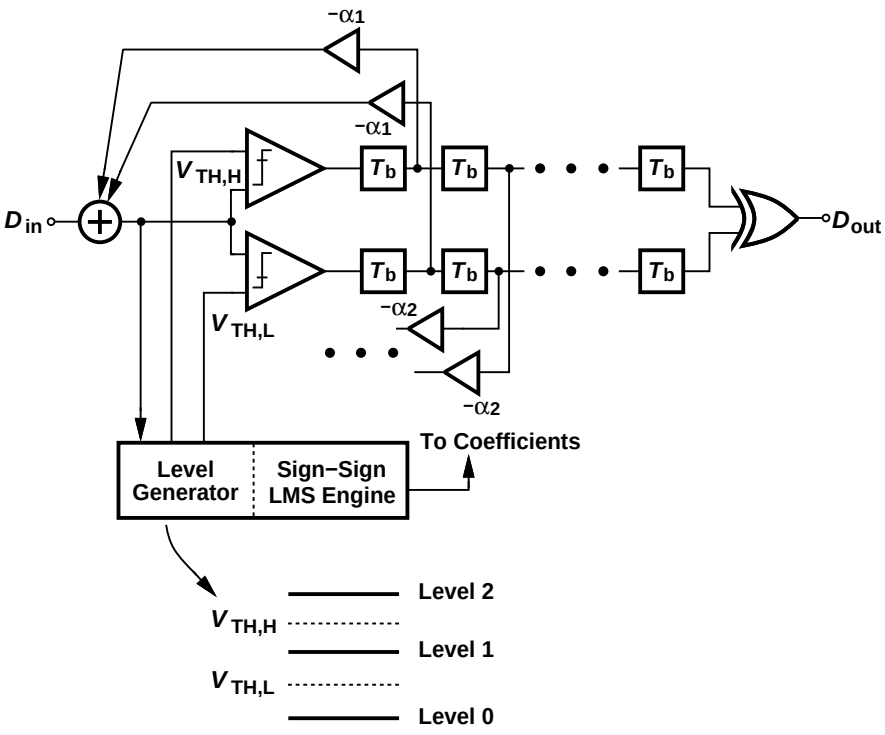


Fig. 11.32 Complete DFE with level generator for duobinary.

Certainly, there are other ways to determine the coefficient of duobinary DFEs. We leave the further exploration to the reader.

11.8.2 CDR

Unlike PAM4, duobinary signals always transit between adjacent levels. This feature facilitates the CDR design. As illustrated in Fig. 11.33(a), we can take the outputs of the two slicers[4] to determine the data transitions. If duobinary signal transmits linearly between adjacent levels, the points crossing over $V_{TH,H}$ and $V_{TH,L}$ coincide with the rising edges of clock. In other words, a bang-bang CDR engine could be directly adopted here to accomplish clock recovery. Data retiming and decoding can be included as well. Sub-rate architecture is also possible to achieve.

[4]They are binary signals.

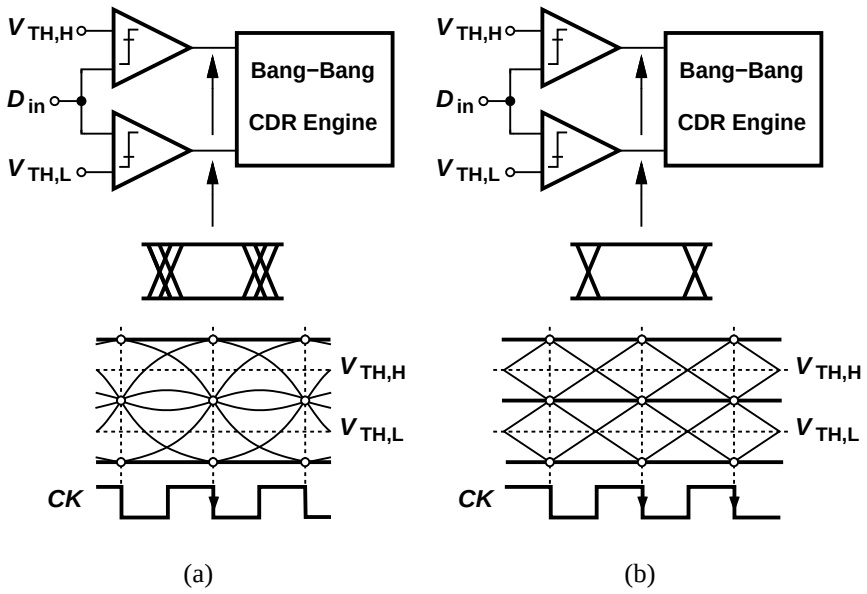


Fig. 11.33 Impact of duobinary waveforms on clock recovery.

It is interesting to note that, for over-compensated channels (i.e., high-frequency part is not sufficiently suppressed), the duobinary waveforms become rounder [Fig. 11.33(b)]. Under such a circumstance, the original thresholds $V_{TH,H}$ and $V_{TH,L}$ may fail to serve as good crossover levels for CDR, as multiple traces would occur. More sophisticated structures must be exploited to further improve the performance.

In some applications, a simplified duobinary RX may be sufficient to achieve reasonable performances with low power consumption. Figure 11.34 shown an example. Here, we have a reference-free comparator and a servo controller to dynamically optimize the output data eye. The comparator compares the input with two threshold levels virtually equivalent to $V_{TH,L}$ and $V_{TH,H}$, generating two outputs V_{out1} and V_{out2}. Amplified to logic level by the subsequent hysteresis buffers [3], V_{out1} and V_{out2} are then XORed to produce the final output D_{out}. The recovered data inevitably bears jitter, since (1) the threshold levels may drift due to mismatches and PVT variations; (2) the threshold-crossing points for the rising and falling would differ intrinsically. Here, the pulsewidth distortion associated with the first issue is corrected by means of a negative feedback loop, which contains a low-pass filter (LPF), and a V/I converter. With the assumption that the input data is purely random, the high loop gain forces the thresholds to stay at the optimal positions

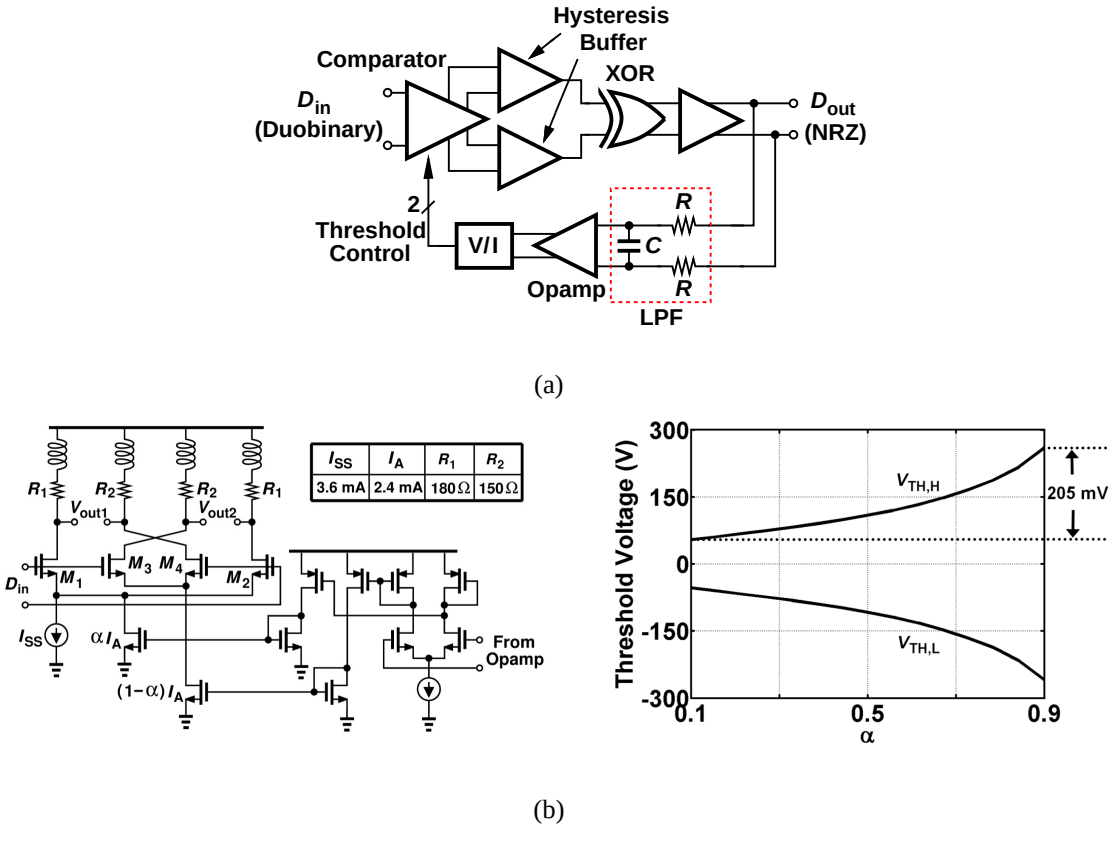


Fig. 11.34 (a) Duobinary RX with dynamic thresholds, (b) comparator and V/I comparator design.

such that the waveform of D_{out} reaches an equal pulsewidth for ZEROs and ONEs. In contrast to the design in [4], this arrangement recovers the data without extracting the clock, providing a compact solution. If necessary, the remaining jitter due to the second issue can be further removed by placing a regular CDR circuit behind it. Note that for simplicity, no receive-side equalization is used in this prototype.

The comparator and V/I converter design is depicted in Fig. 11.34(b), where the input quad $M_1 - M_4$ along with the tail currents and loading resistor form two zero-crossing thresholds for V_{out1} and V_{out2}. Mirrored from the V/I converter, the two variable current αI_A and $(1-\alpha)I_A$ create a threshold tuning range of 205 mV for $\alpha = 0.1 - 0.9$. Fig. 11.34(b) illustrates the variation of threshold levels as a function of α. The key point here is that the threshold adjustment is fully

symmetric with respect to the input common-mode level. It not only eliminates reference offset issue but facilitates the pulsewidth equalization.

REFERENCES

[1] H. Shankar, Duobinary modulation for optical systems, Inphi Corp.[Online]. Available: http://www.inphi-copr.com/products/whitepapers/Duobinary Modulation For Optical Systems.pdf.

[2] J. Lee, A 75-GHz PLL in 90-nm CMOS, Digest of International Solid-State Circuits Conference, pp. 432-433, Feb. 2009.

[3] K. Yamaguchi et al., 12 Gb/s duobinary signaling with $\times$ 2 oversampled edge equalization, Digest of International Solid-State Circuits Conference, pp. 70-71, Feb. 2009.

MEASUREMENT AND LAYOUT TECHNIQUE 12

In the final chapter we look at practical issue regarding layout and testing. Performance of analog circuits is highly related to layout, which is especially true for high-speed SerDes. We present several layout technique proven useful in advanced CMOS processes. Meanwhile, testing of highly-integrated SerDes circuits and system become more and more challenging, as data rates are approaching tens of Gb/s. Measurement techniques are investigated in this chapter as well.

12.1 FUNDAMENTAL MEASUREMENTS

Similar to other analog circuits, wireline chips and system need to be verified by time-domain and frequency-domain testing. The former can be conducted by oscilloscope, bit-error-rate tester (BERT), and other similar equipments, while the latter necessitates spectrum analyzer and network analyzer. Figure 1 illustrated the two categories with main measurements.

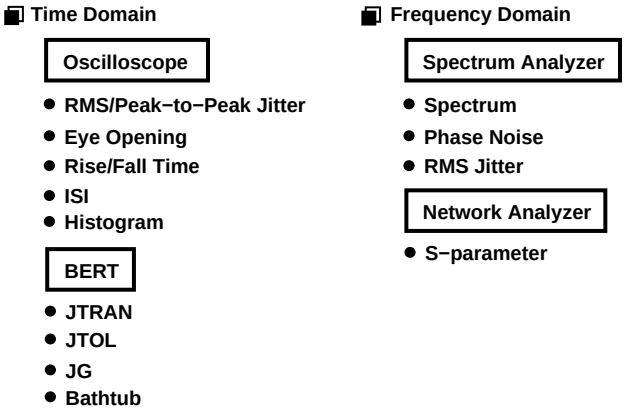
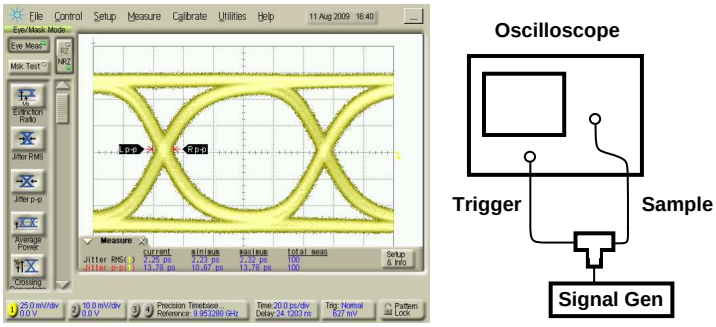


Fig. 12.1 Testing categorization.

The easiest way to check data eye quality is to use an oscilloscope [Fig. 2(a)]. Modern digital scopes can perform tens of statistic functions on the waveforms of the device under test (DUT), including rms and peak-to-peak jitter, rise/fall time, eye opening and so on. At high data rates (i.e., $>$10 Gb/s), precise triggering becomes mandatory. For sensitive measurements or very high-speed signals, the intrinsic jitter associated with the oscilloscope itself may have significant influence on the measurement accuracy. To de-embed the equipment jitter, one can estimate the intrinsic scope jitter ΔT_{scope} using the setup shown in Fig. 2(b). Here, a clock at the frequency of interest is power splitted, forming a self-trigger signal shown on the scope. The measured rms jitter is fully composed of rms jitter from oscilloscope, ΔT_{scope}. Now if we go back to normal testing, the actual rms jitter of the DUT (ΔT_{DUT}) is therefore given by



(a) (b)

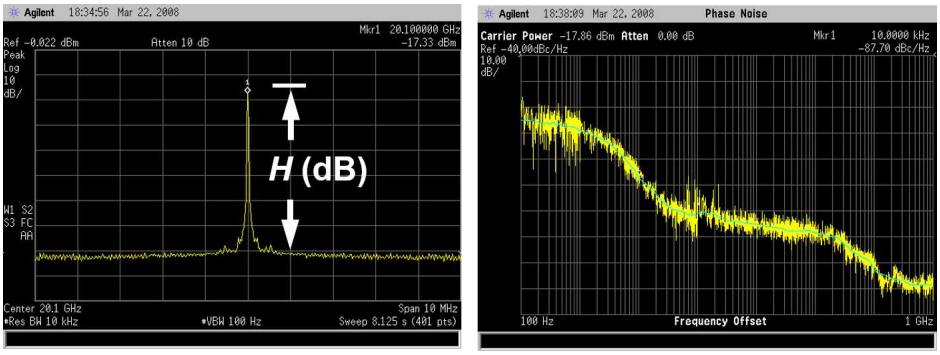
Fig. 12.2 (a) Jitter measurement on oscilloscope (Captured from Keysight DCA-X 86100D), (b) de-embedding equipment jitter .

$$\Delta T_{DUT} = \sqrt{\Delta T_{tot}^2 - \Delta T_{scope}^2}, \tag{12.1}$$

where ΔT_{tot} denotes the raw data (rms jitter) directly captured on the scope. Here we assume the reasons to cause jitters are uncorrelated, which is true in most cases. Typical ΔT_{scope} is on the order of tens to hundreds of femto-seconds. Peak-to-peak jitter would need a large number of samples (e.g., 10,000) to be meaningful.

In addition to scope, a versatile BERT is essential for SerDes testing. It basically provides PRBS with different length for the DUT, and or conducts error checking on the data returned to

BERT. Jitter tolerance and other jitter testing can be covered as well if the BERT is sophisticated enough. Frequency-domain testing is of great important as well. Figure 3(a) illustrates typical spectrum of a recovered clock. As we know the spectrum is plotted based on the boise power accumulated over a certain bandwidth interval (i.e., resolution bandwidth). The phase noise could be directly obtained from the spectrum. For example, if the noise power level at 1-MHz away from the carrier is -65 dBc (with respect to the carrier) and the resolution bandwidth is 10 KHz, we conclude phase noise is -105 dBc/Hz at 1-MHz offset. A more convenient way to observe the phase noise is to plot it in log scale. As shown in Fig. 3(b), such a phase noise plot clearly show the transition behavior of spectrum. Many spectral analyzers provide integration over the spectrum, revealing rms jitter or jitter generation directly. Network analyzer is critical for high-speed signals. It basically measures the power of traveling waves at each port of the DUT, including the incident and reflected waves. By solving the matrix, we arrive at S-parameters (Fig. 4). Careful calibration must be taken to eliminate environmental effects which may possibly cause inaccuracy.



(a) (b)

Fig. 12.3 Spectrum and phase noise plot (in log scale).

12.2 TESTING TECHNIQUES

Let us consider the testing setup for SerDes and its building blocks. Figure 5(a) illustrates a setup for TIAs and LAs. Small signal behavior (e.g., gain, impedance matching, etc) can be checked by using network analyzers. Large signal testing can be accomplished by applying a data stream

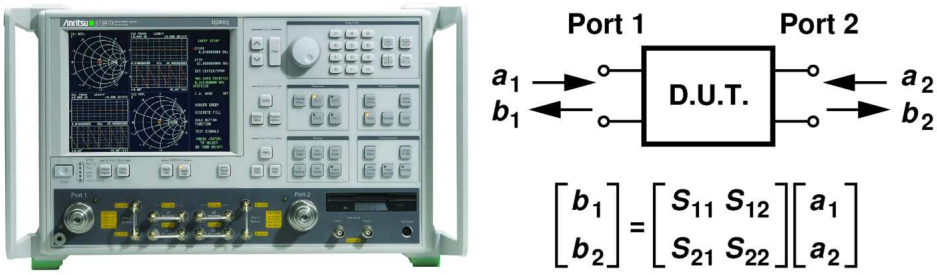


Fig. 12.4 Network analyzer.

with small magnitude and observing the amplified eye diagram on scope. Optical signals can be captured as well if the scope contains an optical sampling head.

Testing of closed-loop blocks is more complicated. Figure 5 (b) depicts a possible arrangement for testing PLLs, which requires a clean source as a reference. A similar for CDRs can be found in Fig. 5(c), where the input is now the data stream from pattern generator or BERT. To do BER test or JTOL, the recovered clock must be fed back to the error detector (ED) of the BERT. PLLs and CDRs are synchronized blocks, whose outputs can be easily recovered on scope with proper trigger.

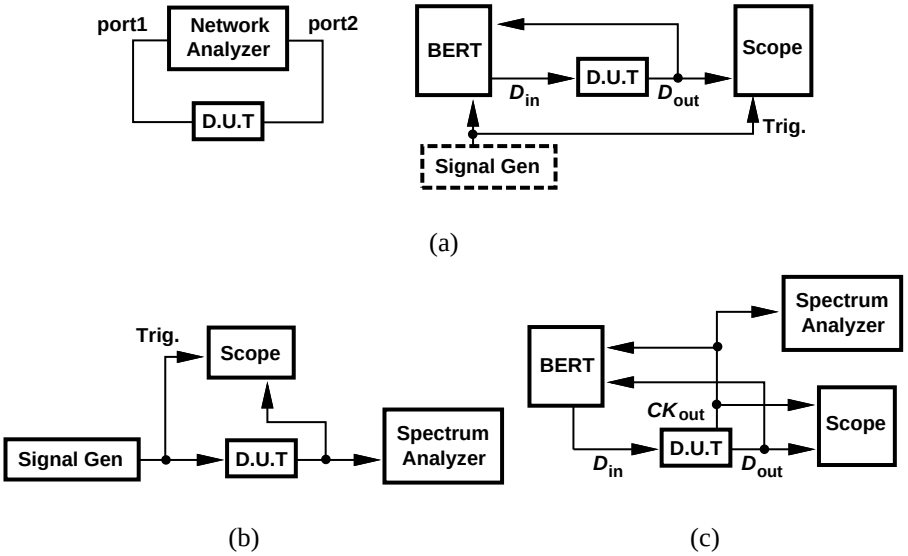
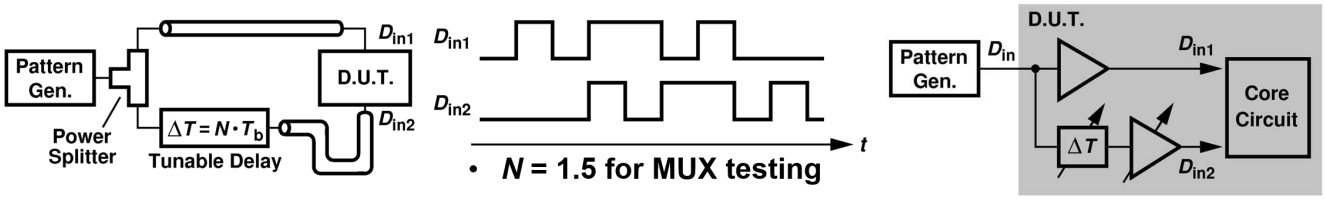


Fig. 12.5 Testing setup for (a) TIAs and LAs, (b) PLLs, (c) CDRs.

How do we test some high-speed blocks which need multiple inputs of data streams, e.g., a MUX? The testing of PAM4 TX would encounter the same difficulty. Advanced BERTs with more than one output data are usually expensive at high speed. Synchronizing two pattern generators is one way to do it, but costly equipments are not always available. A quick way to create two random bit sequences from one pattern generator is to duplicate it with a proper delay. As depicted in Fig. 6(a), a data sequence could be split into two. If the two channels are set apart from each other by 2 or bits, two data stream with reasonable randomness are produced. For the case of MUX testing, delays in unit of 0.5 bit period are suggested as they provide intrinsic shifting for serialization. The reader can prove that two new data stream are correlated and the multiplexed output would no longer be a PRBS. The delay could be put on chip to minimize uncertainty [Fig. 6(b)]. Nonetheless, placing a built-in PRBS engine with multiple output channels is a thorough solution, which inevitably requires more effort on design and layout.



(a) (b)

Fig. 12.6 Creating two data streams for testing.

Jitter related testing necessitates sinusoidally-modulated data outputs for the RX to react. In many cases, however, pattern generators or BERTs can only provide limited range of output modulation for jitter testing. For example, some BERTs only allow sinusoidal jitter up to 10 MHz. Besides, the jitter magnitude is quite moderate. That precludes the user to perform JTOL at very high and very low offsets. It is because the internal clocks inside the equipments have restricted capability of modulation. We need solutions to modulate the data externally.

A direct way to do high-speed modulation is to put a broadband delay after the pattern generator, which is driven by a clock with fixed frequency (i.e., in CW mode). The broadband delay is

governed by an arbitrary waveform generator (AWG), as shown in Fig. 7. Depending on the linearity and tuning range, the delay element modulates the data phase directly. For 25-Gb/s data, some broadband delay elements can provide 1∼2 UI of tuning range and very high-speed modulation (∼ GHz). It is quite useful in testing jitter performance at high offsets.

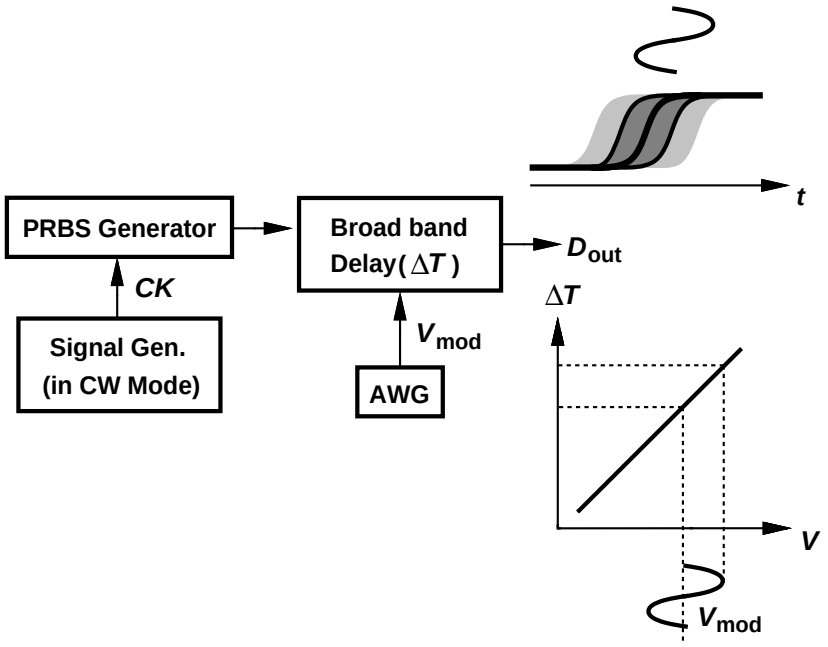


Fig. 12.7 Modulating data stream by broadband delay.

Example 12.1

Consider the data phase modulation setup in Fig. 7. The output data eye can be monitored on scope as the driving clock CK serves as a trigger. Determine the shape of histogram curve if the modulation V_{mod} is in (a) sinusoidal, (b) triangular shapes.

Solution:

(a) Given that the delay unit is purely linear, we realize the excess phase of data is also sinusoidal. Assume the position $x = \sin t$ for simplicity [Fig. 8(a)]. We have

$$\frac{\Delta x}{\Delta t} = \frac{dx}{dt} = \cos t = \sqrt{1 - x^2}. \tag{12.2}$$

Example 12.1 ***(Continued)***

Since the histogram bar Δy for a given position x is proportional to the time slot at which the phase stays in it, we have

$$\Delta y \propto \Delta t = \frac{1}{\sqrt{1-x^2}} \cdot \Delta x. \tag{12.3}$$

That is, the histogram of data phase presents a curve of $(1-x^2)^{-\frac{1}{2}}$ between boundaries (± 1).

(b)With the same approach, we assume $x = at$ for the first quarter of a triangular waveform. It can be easily shown that

$$\Delta y \propto \frac{1}{a} \cdot \Delta x, \tag{12.4}$$

which reveals a uniformly distributed histogram as shown in Fig. 8(b).

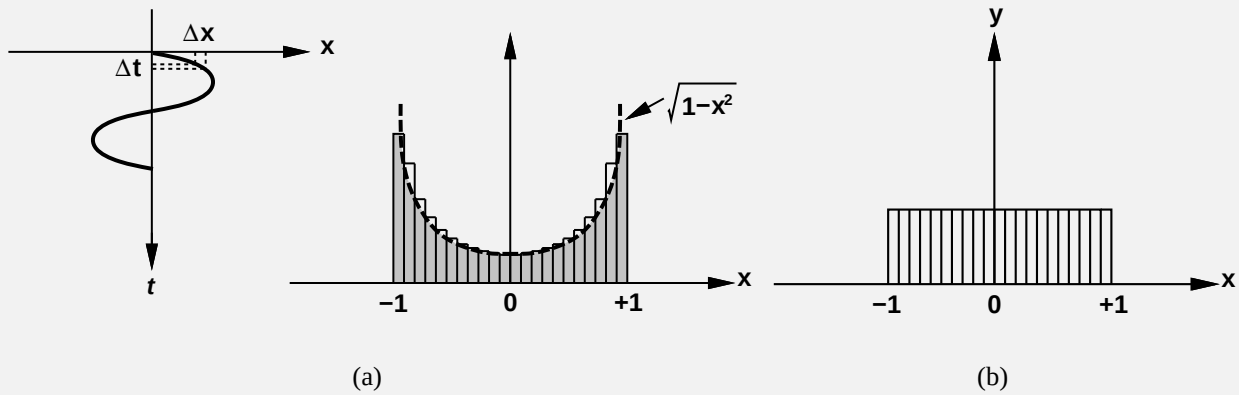


Fig. 12.8 Calculating histogram shape for (a) sinusoidally-modulated, (b) triangularly-modulated data phases.

Jitter testing at low to moderate offset frequencies requires much wider modulation range. The most popular way to create such a data stream is to put the driving clock in FM mode (Fig. 9). Assuming the AWG provides a sinusoidal waveform (in voltage) as $V_{amp}\cos(2\pi f_M t)$, where V_{amp} denotes the amplitude and f_M the modulation range. With an FM gain of K_{FM}, the signal generator's output CK_{mod} presents a sinusoidal modulation in frequency. The amplitude ΔF is therefore equal to

$$\Delta F = K_{FM} \cdot V_{amp}. \tag{12.5}$$

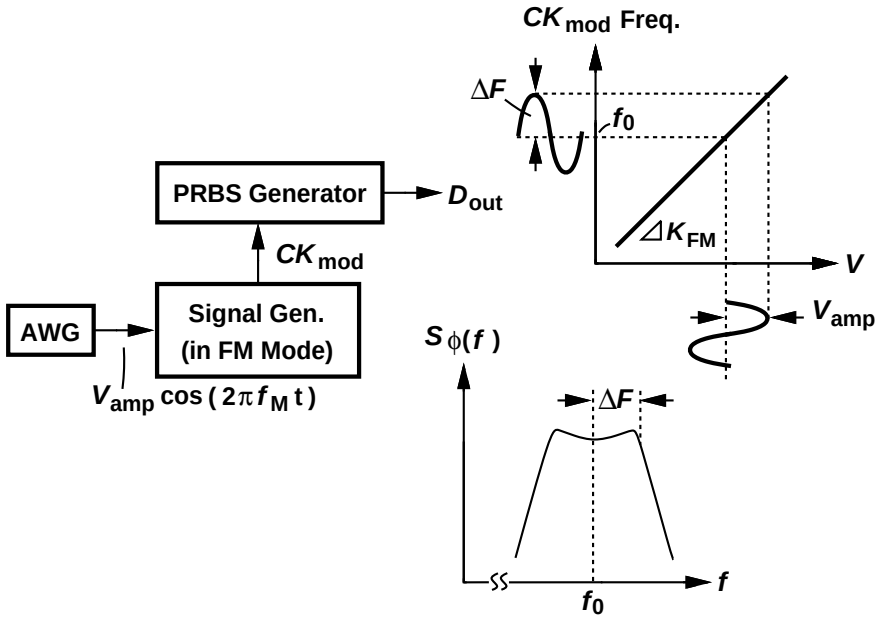


Fig. 12.9 Modulating data stream by FM clock.

Meanwhile, we realize that frequency in rad/sec is given by

$$\omega = 2\pi[f_0 + K_{FM}V_{amp}\cos(2\pi f_M t)]. \tag{12.6}$$

The excess phase is therefore equal to

$$\Delta\phi = \int 2\pi K_{FM}V_{amp}\cos(2\pi f_M t)dt = \frac{K_{FM}V_{amp}}{f_M}\sin(2\pi f_M t). \tag{12.7}$$

Defining UI_{PP} as the peak-to-peak range of sinusoidal phase modulation, we arrive at

$$\frac{UI_{pp}}{2} \times 2\pi = \frac{K_{FM}V_{amp}}{f_M}. \tag{12.8}$$

It follows that

$$UI_{pp} = \frac{K_{FM}V_{amp}}{\pi f_M} = \frac{\Delta F}{\pi f_M}. \tag{12.9}$$

In theory, we are capable to determine the phase modulation UI_{PP} by setting K_{FM}, V_{amp} for a given modulation rate f_M. However, most signal generators have FM operation in analog domain, leaving significant inaccuracy. Observing ΔF in spectrum would be subject to error, as

FM spectrum are discrete spectral lines apart from each other by f_M [4]. We resort to the following methods to calibrate UI_{PP}.

(a) For small modulation ($UI_{pp} < 2$), the accuracy could be checked by comparing the line magnitude of J_0 (at carrier or center frequency) and J_1 (one f_M offset away from J_0). Zooming in the spectrum, we will see a plot like the one shown in Fig. 10(a). It can be proven that

$$\frac{J_1}{J_0} = 1.2 \qquad corresponds \quad to \qquad UI_{pp} = 0.5, \tag{12.10}$$

which provides a useful checking points.

(b) For intermediate modulation ($UI_{pp} < 20$), it is helpful to check to "null" of J_0. Again, it can be shown that J_0 vanishes as

$$UI_{pp} \cong N + 0.75, \qquad N = 0, 1, 2, \cdots \tag{12.11}$$

known as "Bessel Nulling Method", this property also provides quick check in FM accuracy [Fig. 10(b)]. It is because a FM (or PM) signal's spectral lines is actually governed by Bessel function. With all nulls identified, it is not difficult to calibrate desired UI_{pp} by interpolation.

(c) For large modulation ($UI_{pp} > 20$), nulling method becomes improper. The most convenient way to check the modulation accuracy is to lock at spectrum itself. As shown in Fig. 10(c), ΔF is defined as the distance between carrier (center) and the -3-dB point. With careful investigation, one can still achieve accuracy of around 0.1%.

The modulated data is thus ready for different kinds of jitter testing.

12.3 LAYOUT TECHNIQUE

Now we look at layout techniques for high-speed circuits. Like other analog circuits, the performance of SerDes and the associated building blocks highly depends on layout. We summarize general layout rule as follows:

(i) Minimize any possible parasitics, including capacitances, inductance, and resistance. It could be done by sharing diffusion area of active devices, shortening interconnects, and so on.

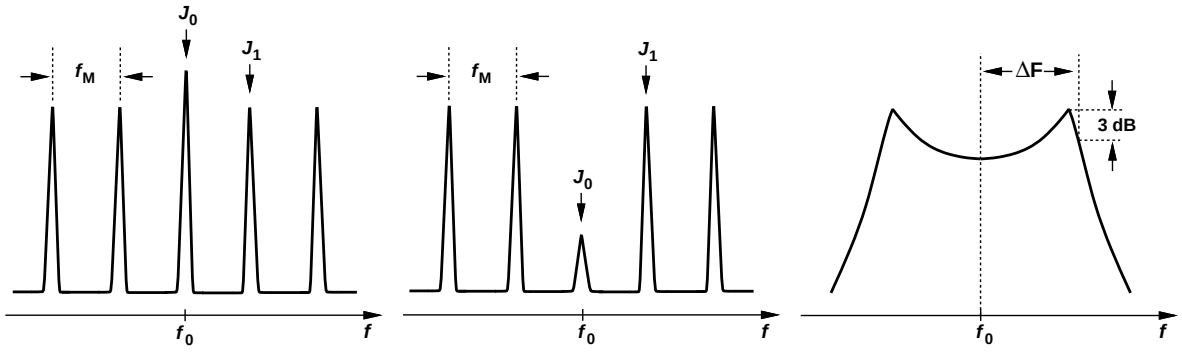


Fig. 12.10 Close look at FM spectrum around the carrier frequency: (a) typical situation, (b) nulling, (c) zoom out for large modulation.

(ii) Make layout of differential circuits as symmetric as possible.

(iii) Add dummy in marginal area if possible. It allows your main circuits facing constant environments. Edge devices are subject to deviations.

(iv) Place substrate contacts all around the layout. Substrate potential needs to be defined at least every tens of μm otherwise the devices threshold voltage vary.

(v) Add bypass capacitors at all important dc nodes, including power lines and bias lines. Use suitable capacitors to optimize bypassing. For example, it is meaningless to bypass a 0.3-V voltage by using MOS capacitor whose V_{TH} is 0.5 V. Capacitance would be developed only after the channel is established.

(vi) Fundamental layout skills (e.g., common-centroid arrangement) always apply to sensitive circuits. Separate analog and digital parts.

(vii) Guard rings and other isolation techniques can be applied to important circuits.

The above guideline are general principles. Let us look at some practical layouts. Shown in Fig. 11(a) is one example of MOS capacitors, where source and drain are shorted together. Channel length should not be too long otherwise the channel cannot be formed evenly. Varactors

(i.e., NMOS in n-well) have similar structure. Figure 11(b) shows a layout of poly resistor. Note that single-row contacts are required in most processes. A normal device can be found in Fig. 12(a), where polyclinics gates are connected by metal at both ends to reduce the resistance effect of polysilicon. Do not make the metal as a ring. Substrate contacts are placed aside. Such a multi-finger structure is popular in analog and mixed-signal circuits, as diffusion region of source and drain are shared. It is important to keep each finger short (no longer than $1 \sim 2\ \mu m$). The junction sharing technique can be further extended to cascode devices. Depicted in Fig. 12(b) is a layout example of it, in which a round-table arrangement is used to further minimize devices internal parasitics.

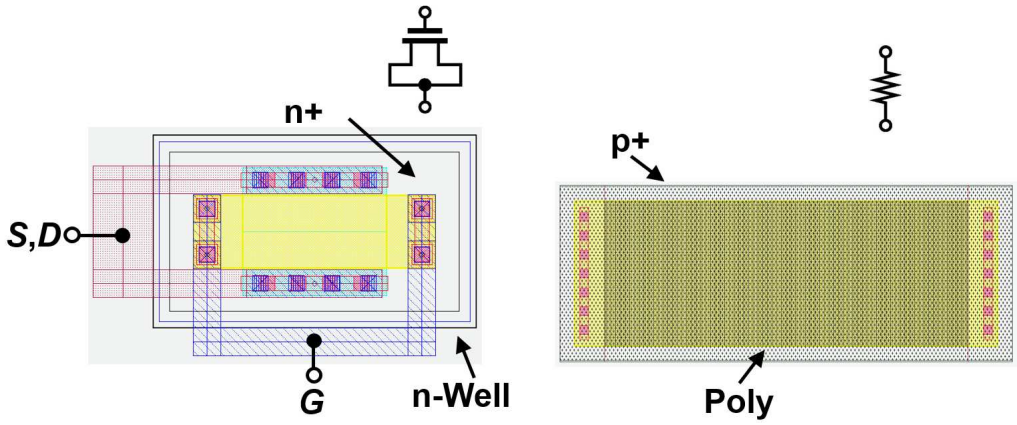
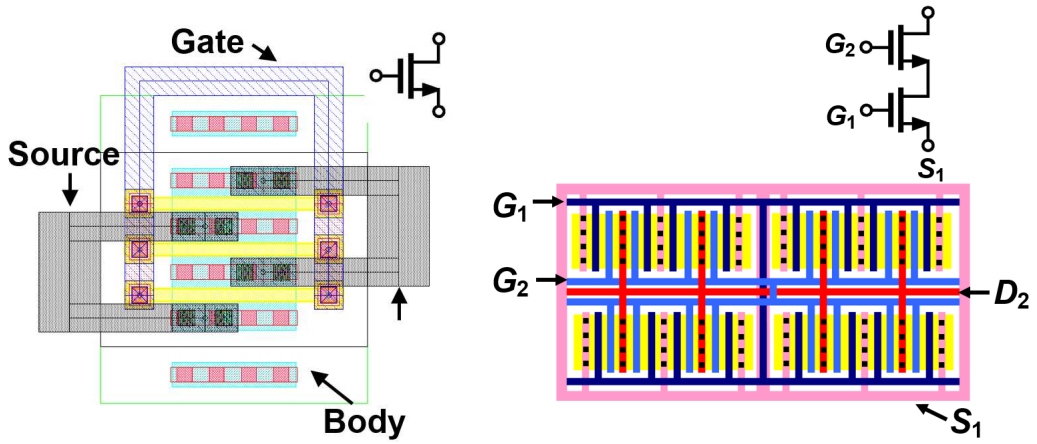


Fig. 12.11 Passive device layout of (a) varactor/MOS cap, (b) poly resistor.

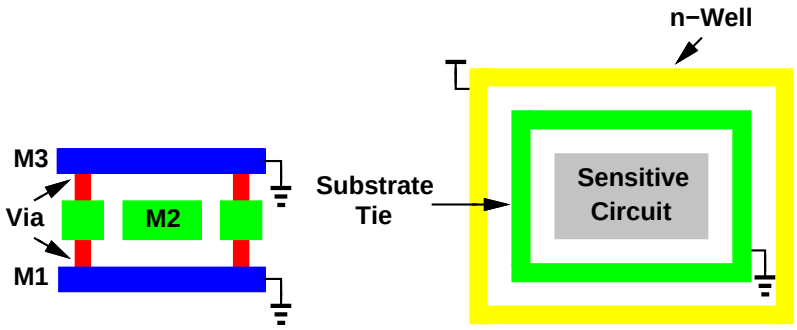
For very sensitive devices or components, proper shielding or guarding is mandatory. For example, the control lines of loop filter in PLLs and CDRs would experience long routing, as the loop capacitors may occupy quick a large area. Perturbation and undesired coupling may cause significant ripples on it if we do not protect the line properly. A good way to shield such important lines is to wrap it with upper and lower metals, as illustrated in Fig. 13(a). Connecting the covers together with vias and shorting them to ground, we achieve a fully isolated signal line here. Similarly, guard rings can be placed around important and sensitive circuits (such as VCO)



(a) (b)

Fig. 12.12 Active device layout of (a) single MOS, (b) cascode MOS.

to increase isolation [Fig. 13(b)]. Substrate tie and n-well are connected to guard and V_{DD}, respectively. "Walls" (made of all layers of metal and polysilicon) can be built around the guard ring to further reduce noise coupling. Routing would be another important issue. It is well known that

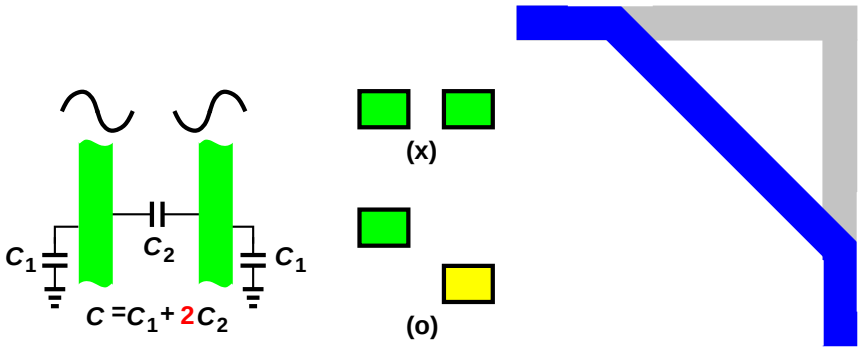


(a) (b)

Fig. 12.13 Layout technique of (a) shielding, (b) guard ring.

for differential signals, the mutual capacitance between the two signals lines are actually doubled because if Miller effect. As shown in Fig. 14(a), each line forces a total parasitic of $C_1 + 2C_2$, where C_1 and C_2 denote the self and mutual capacitance. To de-couple the differential signals, it is preferable to place two lines in different layers of metal. By doing so, mutual capacitance would be minimized to fringe capacitance. Signal lines can swap their metal layers in the midway of routing

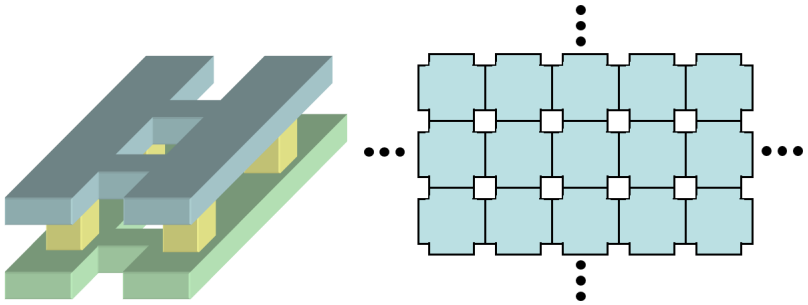
to balance the self capacitance. Diagonal routes are commonly used in analog layouts to reduce approximately 30% of parasitics [Fig. 14(b)]. Other than routing, power lines are of great concern



(a) (b)

Fig. 12.14 Routing skills: (a) separate differential signals, (b) diagonal route.

as well. In order to minimize IR drop, it is possible to realize power lines with multiple metal layers. Modern CMOS processes (especially with copper interconnect) would require power planes to open slots all over the place. Possible realizations are shown in Fig. 15. With fundamental layout



(a) (b)

Fig. 12.15 Power line placement: (a) multi-layer with metal slot, (b) ground plane.

skills understood, we study the higher-level arrangement in the next section.

12.4 LAYOUT PLACEMENT FOR BUILDING BLOCKS

It is quite important to arrange the layout of differential circuits symmetrically. Figure 16 illustrates one popular approach, where a CML flipflop is implemented. The tail current M_7 is placed

underneath the ground line, and the clock and data paths are evenly distributed on both sides. Since the whole circuit is split evenly into two parts, it can be easily connected to other differential circuits with the same structure. As we described in chapter 6, multi-layer inductors are suitable for peaking due to their compact size. Figure 17 reveals another example for Miller divider, which is a differential circuit with class-AB biasing. LC-tank VCOs must be taken care of in layout. Unlike

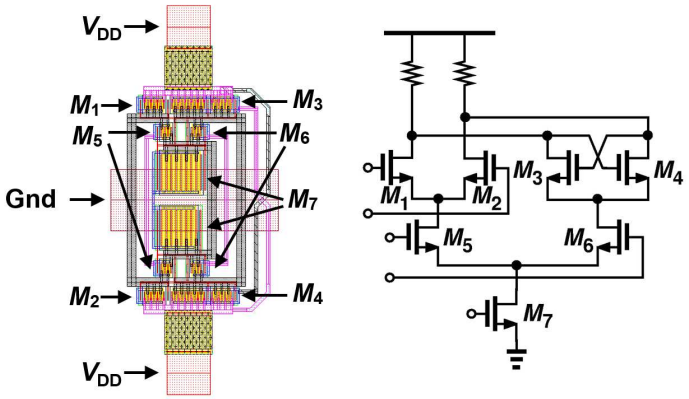


Fig. 12.16 Layout example of a CML latch.

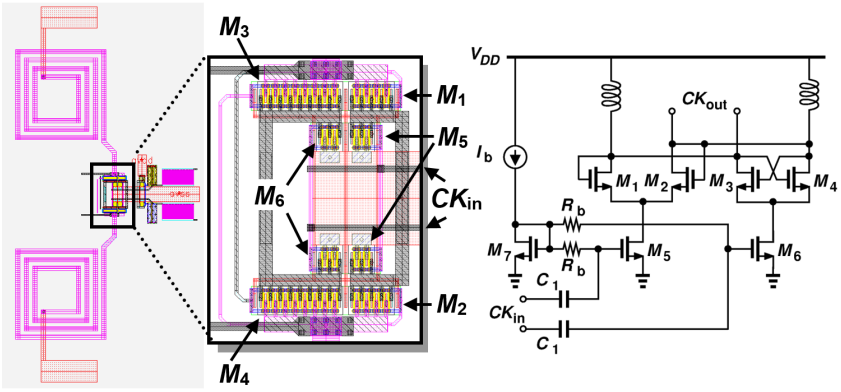


Fig. 12.17 Layout example of a Miller divider.

peaking inductors, the inductors are meant to achieve Q as high as possible. Shown in Fig. 18 is one example. Differential structure is preserved with fully optimized inductors. Ground shielding is placed underneath the spirals. Cross-coupled pair should be allocated in the central part to keep balance, and current source is recommended to step aside to prevent long routing. Other building

blocks with CML structures can be obtained with the same token. Figure 19 demonstrates the case of boosting stage of a CTLE. The degeneration devices $M_3 - M_5$ and R_s are placed in center with proper routing. Figure 20, 21, and 22 provide layout examples for PLL, TX, and RX, respectively.

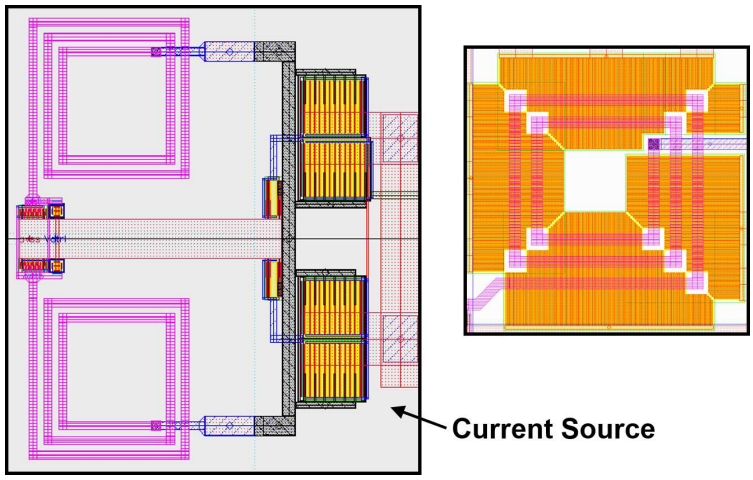


Fig. 12.18 Layout example of a LC-Tank VCO.

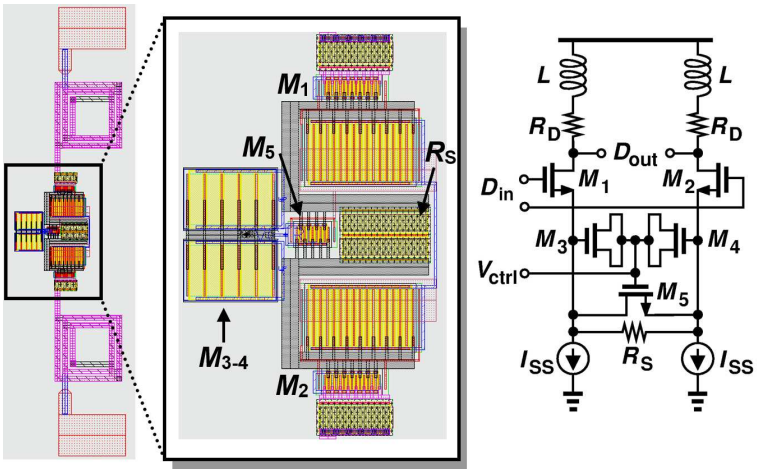


Fig. 12.19 Layout example of an equalizer with RC-Degeneration.

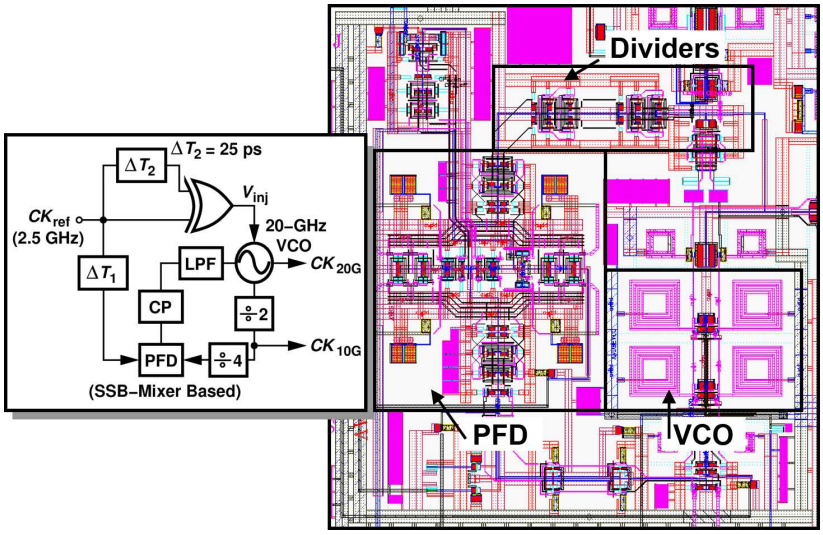


Fig. 12.20 Layout example of a 20-GHz Injection-Locked PLL.

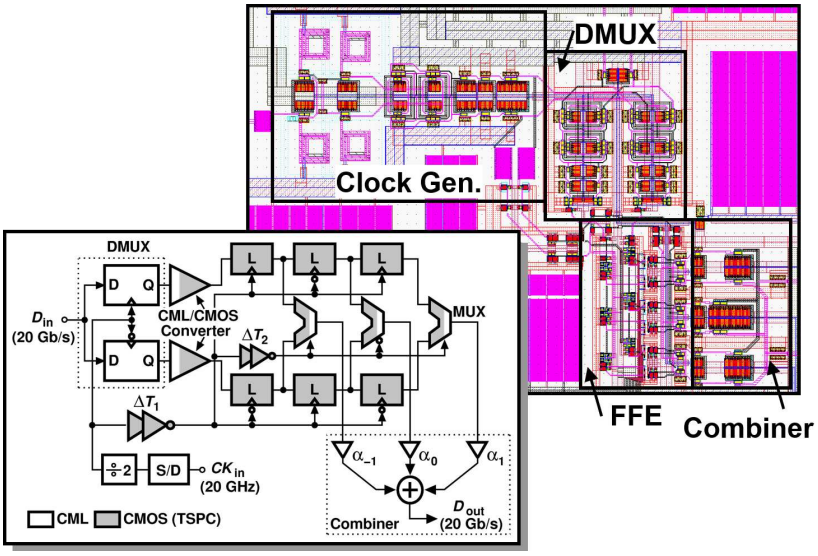


Fig. 12.21 Layout example of a 20-GB/s transmitter.

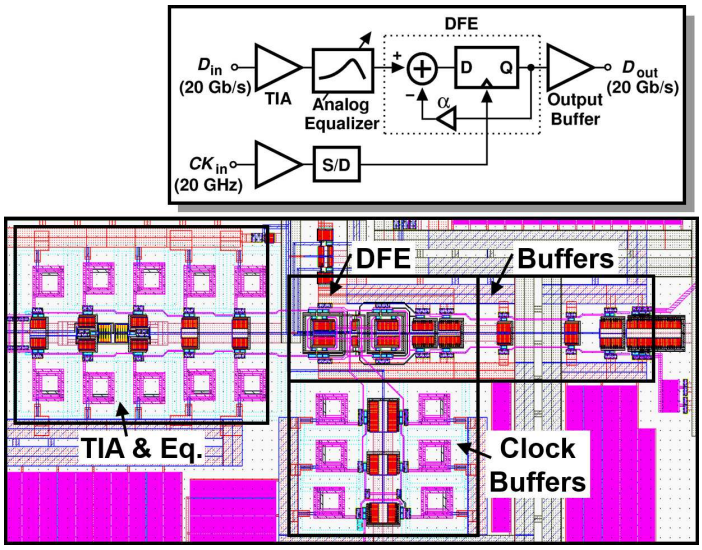


Fig. 12.22 Layout example of a 20-GB/s receiver.

REFERENCES

[1] Jri Lee et al., "A 75-GHz Phase-Locked Loop in 90-nm CMOS Technique," IEEE Journal of Solid-State Circuits, vol. 43, pp. 1414-1426, June 2008.

[2] Jri Lee and H. Wang, "Study of Subharmonically Injection-Locked PLLs," IEEE Journal of Solid-State Circuits, vol. 44, pp. 1539-1553, May 2009.

[3] H. Wang et al., "A 21-Gb/s 87-mW Transceiver with FFE/DFE/Linear Equalizer in 65-nm CMOS Technology," Digest of Symposium on VLSI Circuits, pp. 50-51, June 2009.

[4] Agilent Technologies, "Jitter Fundamentals: Jitter Tolerance Testing with Agilent 81250 ParBERT".