LOW POWER CIRCUIT TOPOLOGIES FOR DIGITAL-TO-ANALOG CONVERTERS WITH A MM-WAVE SAMPLING CLOCK

by

Yingying Fu

A thesis submitted in conformity with the requirements for the degree of master of applied science
Graduate Department of Electrical and Computer Engineering
University of Toronto

© Copyright 2015 by Yingying Fu
Abstract

Low power circuit topologies for digital-to-analog converters with a mm-wave sampling clock

Yingying Fu
master of applied science
Graduate Department of Electrical and Computer Engineering
University of Toronto
2015

This thesis describes the design of an 8-bit, 75GS/s, full-rate, low-power DAC in 55nm SiGe BiCMOS process. This is the highest sampling frequency (75GHz) broadband DAC to the best of the author’s knowledge. It has an output swing of 1.2Vppd, which is sufficient to directly drive high performance VCSELs for short-reach data center communication links. The low power circuit topologies minimize the power consumption of the DAC to < 0.5W. Most of the cells use a 1.8V supply only.

The analysis and design of the low power, high speed digital library for SiGe BiCMOS technology is presented. The design of the DAC is based on this library. Using the novel MOS-HBT, quasi-CML topology, a test chip of a 108GHz retimer with equalization and a broadband clock amplifier is designed, fabricated and measured.
Acknowledgements

I would like to sincerely thank Professor Sorin Voinigescu for his guidance and advice. His passion for circuit design and technology is very inspirational. It is a great honour to have him as my supervisor. I would also like to acknowledge my examination committee Professor Carusone, Professor Liscidini and Professor Poon.

I would like to thank Andreea Balteanu, Ioannis Sarkas for their patience and help at the beginning of my studies. I would also like to thank my colleagues from BA4192: Ivan Krotnev, Guy Alter, Stefan Shopov, James Hoffman, Konstantinos Vasilakopoulos, Hassan Farooq and James Bateman.

I would also like to thank Dan Case and The’ Linh Nguyen from Finisar Corporation for giving me the opportunity to work with them on the 75GS/s DAC.

The DAC project was supported by Finisar Corporation. The retiming flipflop project was supported by Gennum Corporation and the Ontario Centre of Excellence.
## Contents

1 Introduction ........................................... 1
   1.1 Motivation ........................................ 1
   1.2 Objective ......................................... 4
   1.3 Thesis Overview ................................. 5

2 High speed Digital-to-analog converter architectures .. 6
   2.1 Introduction .................................... 6
   2.2 DAC performance metrics ....................... 7
   2.3 DAC architecture ................................ 8
      2.3.1 Binary-weighted current DAC .............. 8
      2.3.2 R-2R based DAC ............................... 10
      2.3.3 Fully segmented DAC ......................... 12
      2.3.4 Partially segmented DAC ..................... 13
   2.4 Half-rate vs. full-rate Nyquist DAC ............ 14
   2.5 Summary ....................................... 15

3 High speed digital library design ...................... 16
   3.1 Logic cell design ............................... 16
      3.1.1 Current-mode logic ......................... 16
         3.1.1.1 MOS CML vs. Bipolar CML ............. 18
         3.1.1.2 MOS-HBT cascode ....................... 20
         3.1.1.3 Emitter-coupled Logic (ECL) ........... 21
         3.1.1.4 Quasi-CML MOS-HBT cascode .......... 22
         3.1.1.5 Shunt Peaking ............................ 24
   3.2 Common mode oscillation ........................ 25
   3.3 Driving long interconnect ....................... 28

4 108GHz retiming flipflop .......................... 30
# 4. Prior Art

- **4.1 Prior Art** .................................................. 30
- **4.2 Motivation for including an on-chip equalizer** ............... 32
- **4.3 Circuit description** .......................................... 34
- **4.4 Flip-flop design and simulation** .......................... 35
- **4.5 Broadband clock amplifier design and simulation** .......... 37
- **4.6 Eye diagram simulation** .................................. 41
- **4.7 Fabrication technology** .................................... 42
- **4.8 Experimental characterization** ................................
  - **4.8.1 40Gb/s measurements** .................................. 43
  - **4.8.2 75Gb/s Measurements** .................................. 46
  - **4.8.3 Oversampling measurements with 108GHz clock signal** .. 47
  - **4.8.4 Power Consumption** .................................... 49
  - **4.8.5 Comparison with the state-of-art** ....................... 50
  - **4.8.6 Summary** .................................................. 50

## 5. 75GS/s DAC

- **5.1 Technology** .................................................. 51
- **5.2 8-bit 75GS/s DAC design** .................................. 51
  - **5.2.1 DAC specification** .................................... 51
  - **5.2.2 Block diagram** ........................................ 52
  - **5.2.3 Investigation of the optimal bit segmentation split** ...... 53
- **5.3 DAC block design** ........................................... 54
  - **5.3.1 Clock path** .............................................. 54
  - **5.3.2 Flip-flop design** ....................................... 56
  - **5.3.3 Data path** ............................................... 59
  - **5.3.4 Output stage** ............................................ 62
- **5.4 DAC block placement and layout** ........................... 63
  - **5.4.1 DAC block placement** .................................. 63
  - **5.4.2 DAC top level layout** .................................. 68
- **5.5 Simulation after layout extraction** .......................... 69
  - **5.5.1 Flipflop timing margin** .................................. 69
  - **5.5.2 Impact of glitch energy** ................................ 72
  - **5.5.3 Mismatch simulation** .................................... 74
  - **5.5.4 Ramp simulation** ....................................... 77
  - **5.5.5 Dynamic performance** .................................. 80
  - **5.5.6 Comparison with the state-of-art** ....................... 81
## 6 Conclusion

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1 Summary</td>
<td>84</td>
</tr>
<tr>
<td>6.2 Contribution and Publication</td>
<td>85</td>
</tr>
<tr>
<td>6.3 Future Work</td>
<td>85</td>
</tr>
</tbody>
</table>

## Bibliography

88
Chapter 1

Introduction

1.1 Motivation

With the proliferation of cell phones and other mobile devices, there is increasing demand for more bandwidth. To meet this demand, all physical layers of the communication system need to be improved by increasing the bandwidth of the device, the data center and the larger communication backbone network. Each physical layer has its own unique requirements, hence different designs are necessary. Data centers typically require communication links of a few hundred meters in reach. Currently, most data centers use 4x28 Gb/s optical interconnect to reach 100 Gigabit Ethernet speed. Cost and power considerations are the most important factors in choosing a design for data center. On the other hand, commercial long haul fibre-optic communication networks currently operate with 100Gb/s line rate with DWDM (dense wavelength-division multiplexing) and Coherent receiver technology [1]. Up to 3.6Tb/s capability has been achieved in a single bay [1]. Higher data-rate, longer transmission distance and greater energy efficiency per bit per kilometer are more critical than per-unit cost.

The current IEEE standards for fiber optic communications are 40-gigabit Ethernet (40GbE) and 100-gigabit Ethernet (100GbE). For example, 100GE-LR4 can be achieved by multiplexing 4 wavelengths of 25Gb/s optics [2]. The next standard is 400Gb/s and a possible way of reaching this data rate is by using 8x50Gb/s [3]. The bandwidth is limited to 50GHz, because the optical filter has a bandwidth of 50GHz, as a result of the technologies of filters and standard requirements [4].

In the 1990s, the trend in the optical communication industry was to use simple ON-OFF
intensity modulation and push the bit-rate, while using a direct detection receiver architecture \[4\]. This brute force method becomes increasingly difficult and power hungry as the data rate increases. There are two other ways to increase the bit-rate for an optical network: use higher order modulation to increase the number of bits per symbol, and increase the number of carriers \[4\]. These two approaches have been intensively investigated to increase spectral efficiency of the system. In particular, increasing the modulation level is very attractive, because it can be applied to both direct detection receivers and digital coherent receivers. These two types of receivers are both being investigated for the next generation optical transport network. In the last 10 years, digital coherent receivers have become popular, because they allow distortion compensation, such as chromatic dispersion compensation and polarization mode dispersion compensation, in the digital domain \[4\]. On the other hand, direct detection receivers are sometimes preferred over digital coherent receivers because of the relative simplicity in the receiver design.

In order to employ multilevel modulation, a digital-to-analog converter (DAC) is necessary in the transmitter. Research into tens of Giga-sample per second DACs has been going on for many years. There are two distinctively different high speed DAC topologies depending on the output swing required. If greater than 2V peak-to-peak swing and large bandwidth are desired, a distributed topology is needed. Some examples could be found in \[5\], \[6\]. A distributed topology is necessary to have enough gain to generate the required swing, while maintaining the bandwidth of the circuit. Such a large swing DAC can be used to directly drive the optical modulator, such as a Mach-Zehnder modulator, without needing a very linear, large power amplifier following it.

However, most published high speed DACs have between 150mV \[7\] to 650mV \[8\] output voltage swing per side. These DACs do not use the distributed topology, but require a linear post-amplifier to drive the Mach-Zehnder modulator for long haul transmission of over 1440km \[9\]. Another interesting application is to use the DAC to directly drive a vertical cavity surface emitting laser (VCSEL). The VCSEL is a lower cost, low power and high performance alternative to the distributed-feedback and Fabry Perot lasers. It is particularly suitable for short reach (<2km) fibre optic networks, such as those in a data center. Currently, it has been demonstrated that VCSELs can be used for up to 56.1Gb/s optical link \[10\].

Because of the mm-wave sampling clocks, early high-speed DACs have mostly used InP HBT and SiGe BiCMOS technologies, instead of CMOS. One of the early pioneering work, in terms of speed, was published in 2005 by Nortel. They presented a 6-bit 22GS/s
DAC using segmented topology in 0.13um SiGe BiCMOS technology [8]. Following this work, Nagatani et al, from NTT, published three different DACs operating at 32GS/s (half rate), 28GS/s (full rate) and 60GS/s (full rate), in 2009, 2010 and 2011 respectively. Their work was done in InP HBT technology. However, the power consumption for these DACs are high, because of the higher supply voltage needed for bipolar technologies.

As $f_T$ and $f_{MAX}$ increased for CMOS technology, researchers also explored nanoscale CMOS for building high speed DACs. One example is a 6-bit, half-rate, 56GS/s DAC with 4.3 ENOB up to 26.7GHz [11]. The obvious advantage of using CMOS is the lower static power consumption, ease of integration with DSP and scaling with CMOS technology. However, it is also important to note that the limitations of CMOS technology at such high speed also means that more digital signal processing is necessary to compensate for bandwidth limitation, mismatch and signal distortion. The additional digital blocks also consume significant power. Also, it should be noted that CMOS scaling has not resulted in significant improvement in the DAC performance. A recently published DAC in 28nm CMOS [12] showed lower ENOB and higher power than the DAC in 65nm CMOS published by Ciena [11]. On the other hand, the SiGe BiCMOS technology has the unique advantages of having a better metal backend than nano-scale CMOS for making passive components, fast bipolar transistors with high transconductance and fast MOSFETs to enable lower supply, low power circuit topologies. These advantages represent a compelling reason to implement a high speed, low power DAC in SiGe BiCMOS.

As shown in Figure 1.1 [13], with high speed, high resolution DACs and a powerful DSP engine, a programmable optical transceiver becomes feasible. Figure 1.2 shows an example of the constellation created with two 5-bit DACs. The constellation on the right includes distortion compensation to correct for non-linearity in the E/O (electrical to optical transduction) [13]. A high speed DAC is a crucial building block to enable this kind of software-defined optics. The transceiver can be dynamically programmed to provide different modulation schemes and enable digital compensation. This is a very powerful idea with significant challenges, too. On the digital side, the digital streams feeding the DAC should operate at a high data rate and the DSP needs to be powerful enough to process and generate all the bits going to the DAC. On the DAC side, the challenge is to design a high speed DAC with high resolution to enable higher order modulation.
Chapter 1. Introduction

Figure 1.1: Potential architecture for software-defined optical transceiver

Flexible Output Field Using DAC Technology

Figure 1.2: Demonstration of using two 5-bit DACs to create complex modulation

1.2 Objective

The objective of this thesis is to study and design the necessary building blocks to enable 400Gb/s transmitters with low power consumption. A high speed digital library with current mode logic (CML) and BiCMOS (Bipolar and CMOS logic) needs to be built and characterized first. Using this library, an 8-bit, 75GS/s DAC with 600mVpp per side output swing is designed for fiber optic communication. This DAC can be used to drive VCSELs directly for low-cost data center designs. It can also be used with a high power linear amplifier, which can drive optical modulators requiring higher voltage swing. With an 8-bit resolution, this DAC can also be used in instrumentation equipment to generate arbitrary waveforms.
High output bandwidth (> 25GHz), low power (< 0.5W) and good dynamic range (>36dB SFDR up to 36.5GHz) are three key objectives of this project. Since the effective resolution of the DAC decreases with increasing output signal frequency, different modulation complexity can be applied to the DAC depending on the output signal frequency. At lower output frequencies, the DAC has higher effective number of bits (ENOB), so a higher order modulation, such as 64PAM, can be used. At higher output frequencies, the DAC has lower ENOB and a simpler modulation scheme, such as 4PAM and OOK can be used.

A full-rate retiming flip-flop is a critical block for the high speed DAC. It is the most demanding block to design because it is the fastest switching logic block. Hence, a test chip of a high speed flipflop needs to be designed and verified on silicon first to assess the speed limitations of the high speed digital library. To enable testing in the lab, an equalizer and a broadband clock amplifier also need to be designed and included on the test chip to compensate for the loss from the cables in the measurement setup at high frequency.

1.3 Thesis Overview

Chapter 2 covers the background knowledge on high-speed DACs and DAC architectures with an emphasis on DACs with conversion rates greater than 10GS/s. The high speed digital library is described in Chapter 3. The 108GHz retimer with equalization is described in Chapter 4. The design and simulation results of the 75GS/s DAC are presented in Chapter 5.
Chapter 2

High speed Digital-to-analog converter architectures

2.1 Introduction

A digital-to-analog converter (DAC) produces an analog output voltage that is proportional to the digital word input. For a binary encoded digital word, the analog output voltage is defined as

\[ V_{out} = V_{ref}(b_12^{-1} + b_22^{-2} + b_N2^{-N}), \quad (2.1) \]

where \( V_{ref} \) is the full-scale output voltage and \( b_1 - b_N \) represent the binary digital word. \( b_1 \) is the most significant bit and \( b_N \) is the least significant bit \([14]\).

The spectrum of the output of an ideal DAC can be expressed as

\[ X(f) = \frac{\sin(\pi f/F_s)}{\pi f/F_s} \sum_{k=-\infty}^{+\infty} x(f - kF_s) \quad (2.2) \]

where \( k \) consists of integer numbers, \( x(f) \) represents the spectrum of the input data and \( F_s \) is the sampling frequency.

This shows that the spectrum of the ideal DAC is the sum of the shifted \( x(f) \) and the frequency shift is equal to the sampling frequency of the DAC. The additional spectra, as
Chapter 2. High speed Digital-to-analog converter architectures

A result of sampling the input, are called the images. They are usually undesirable and should be filtered out. An example of the ideal DAC output is shown in Figure 2.1.[16] The diagram on the left shows the DAC output in time domain. A stair case waveform is observed because of the zeroth order hold response, where the input changes at the sampling instant and the value is kept for the whole sampling period[15]. The diagram on the right shows the output spectrum of the DAC. Four image frequencies are observed. The amplitude of the image signals are attenuated because of the zeroth-order hold from the DAC, which is modeled by the Sinc function. The undesirable effect of the zeroth-order hold response is that at a frequency close to the Nyquist frequency ($F_s/2$), the inband output signal amplitude is reduced.

![Ideal DAC Output - Time domain](image)

Figure 2.1: Ideal DAC output in time domain and frequency domain.[16]

### 2.2 DAC performance metrics

The performance of a DAC is characterized by its static and dynamic behavior. Static parameters of a DAC include gain error, offset, INL (Integral non-linearity) and DNL (Differential non-linearity). Some of the metrics are listed and briefly discussed below. More details can be found in [14] and [17].

- **Gain error** = \( \frac{V_{out}(1\ldots1)-V_{out}(0\ldots0)}{LSB} - (2^N - 1) \) [14].
- **INL**: difference between the actual output voltage and the value extrapolated from the best-fit straight line at each input word.
- **DNL** = \((V_{out}(n)-V_{out}(n-1))-LSB\) for \(n\) from 1 to \(2^N - 1\).

The dynamic parameters of a DAC include its output bandwidth and dynamic range. The dynamic range can be specified by the SFDR (spurious free dynamic range) and
SNDR (signal to noise and distortion ratio). Both SFDR and SNDR are defined at a given input power.

- SFDR = \(10\log_{10}(P_S/P_H)\), where \(P_S\) is the signal power and \(P_H\) is the power of the largest spur.

- SNDR = \(10\log_{10}(P_S/(P_N + P_D))\), where \(P_S\) is the signal power and \(P_N\) is the noise power and \(P_D\) is the combined power of the harmonic distortions.

- ENOB = \(\frac{SNDR - 1.76dB}{6.02}\) bits \[14\].

Several factors limit the dynamic range of a DAC. For a high-speed DAC, the timing misalignment of the switches at the DAC output causes glitches, which contributes to dynamic non-linearity. The timing misalignment can be caused by clock jitter and delay mismatch between data lanes. Larger timing mismatch results in a larger glitch energy.

### 2.3 DAC architecture

There are two popular DAC architectures: i) resistor divider ladders and ii) current-steering arrays \[17\]. For high speed DACs, current-steering is the preferred choice. DACs based on resistor divider ladders suffer from large RC delay. They also require a high speed and very linear amplifier at the output to drive a 50 ohm load for all codes.

Current-steering DACs can be further divided into four categories: binary weighted, R-2R based, fully segmented and partially segmented architectures. Several papers have summarized the advantages and disadvantages of each architecture \[13\], \[18\]. In this section, all four architectures are reviewed and examples of high speed DACs (> 10GS/s) using these architectures are reviewed.

#### 2.3.1 Binary-weighted current DAC

A simplified circuit diagram of a binary-weighted DAC is shown in Figure \[2.2\]. Input binary codes switch the binary-weighted current sources to a 50\(\Omega\) on-chip load resistor and a 50\(\Omega\) off-chip termination resistor to generate the analog voltage output. For high speed DACs, a 50\(\Omega\) output impedance is necessary to match with the input termination of the measurement equipment, such as an oscilloscope or spectrum analyzer. It is important
to have proper termination for high speed DACs to avoid signal degradation. From the diagram, the obvious advantage of this architecture is that it only requires N current sources for a N-bit DAC. This can result in area and power savings by having fewer data lanes, which reduces the number of buffers necessary to drive them. This is very important for high speed DACs, because smaller area means smaller footprint and less layout parasitic capacitance to drive. As the conversion rate of the DAC increases, the bandwidth of the DAC becomes increasingly limited by the interconnect capacitance. Examples of DACs using binary-weighted current sources include a 6-bit, 20GS/s DAC demonstrated by Baranauskas et al in 2006 [19] and a 6-bit, 20.5GS/s DAC [20] by Khafaji et al.

The area and power savings are somewhat limited because the current switches need to be scaled proportionally larger with the current to ensure the transistors are biased at peak $f_T$ current density. As a result, all the buffers driving the current switches also need to be scaled appropriately. This architecture is also demanding in terms of device matching to achieve high resolution. Sometimes, in order to reduce mismatch of the current sources caused by the large ratio, it is desirable to use unit-sized current sources. This further increases the area of the DAC.

Having a large current ratio also negatively impacts the dynamic performance of the DAC. In Baranauskas’ design, scaled buffers were used to drive the weighted current sources. However, in order to save power, scaling was done such that there was still a fanout difference of 4:1 between the MSB and LSB. As a result, MSB lanes driving a large load will have more delay, which will cause glitches at the DAC output. The authors compensated for this delay by retiming the MSB earlier than the LSB. Also, to further reduce the current ratio, they applied the LSB current to a 25Ω load, whereas the rest of the binary currents go to a 50Ω load.

There are two drawbacks of a binary-weighted DAC that plague all DACs not using thermometer decoding. Firstly, the DAC may not be monotonic. A monotonic DAC is one where the output always increases with increasing input code [14]. Secondly, the DAC may suffer from large glitch energy when the MSB toggles, e.g. when the code changes from 1000 to 0111 for a 4-bit DAC. This problem is especially pronounced for high speed DACs, because the duration of the bit period can be too short for the glitch to settle within one bit period. Hence, it is important to align the digital bits to reduce the glitch energy as much as possible. This issue is discussed further in Chapter 5.

The main advantage of the binary-weighted DAC is its simplicity. It does not require a
thermometer decoder, which can be challenging to design at very high speed with low power. Some designs have taken advantage of this fact while minimizing the impact of the glitches by using binary-weighted currents for the LSBs only [11]. This architecture is known as partial segmentation, which is discussed in more depth in section 2.3.4.

![Binary-weighted current DAC diagram](image)

Figure 2.2: Binary-weighted current DAC.

### 2.3.2 R-2R based DAC

The challenges associated with the large current ratio of a binary-weighted DAC is mitigated with a R-2R resistor ladder where all current sources have identical values. This implies that the current switch for each bit steers the same amount of current. The issue of current source mismatch as a result of the large current ratio is avoided. A block diagram is shown in Figure 2.3. For this topology, the binary output levels are generated with the resistors. The operation of this topology can be understood by examining the Norton equivalent circuit of each stage [17]. For this topology, only a resistor ratio of 2:1 is required, hence further reducing the impact of mismatch. If only unit sized resistors are used, the matching can be even further improved. Resistor process variation imposes the minimum resistor width needed to achieve the desired resolution.

Since only a fraction of the current is passed to the following stage, an obvious disadvantage of the R-2R architecture is that a portion of the current is wasted in the resistor ladder. As the resolution of the DAC increases, the amount of wasted current increases and the efficiency decreases.
This architecture also suffers from strong glitches if the delays of the data signals to the current switches are not well matched. In fact, the timing skew between the data lanes is the main factor that limits the sampling speed of the DAC [21]. In [21], Nagatani et al. showed that for a 6-bit 25GS/s R-2R DAC, a timing skew of 3ps results in a 5.8LSB glitch when the MSB toggles. They designed two R-2R DACs: a 6-bit 32GS/s DAC [22] and a 6-bit 28GS/s DAC [7]. Different methods were applied to address timing skew. The 32GS/s DAC uses only a half rate clock and performs double sampling of the data to relax the timing constraint. The 28GS/s DAC uses a full rate clock and relies on matching the interconnect length to minimize timing skew between the data lanes. For a more recent version of their DAC at 60GS/s, they used the same layout technique and demonstrated a clear four-level eye diagram at 60GS/s [23]. All three DACs used InP HBT technology with $f_T$ and $f_{MAX}$ of 175GHz and 260GHz [22], respectively and much less lossy, semi-insulating substrate than that of the commercially available CMOS processes.

The impact of timing skew between data lanes becomes more serious at higher sampling frequency. With flip-flops, the data can be retimed before reaching the current switch. However, as Naganati pointed out, at higher frequencies, the timing margin of the flip-flop is reduced [21]. As a result, it becomes harder to suppress the timing skew of the input data.

As in the binary-weighted current DAC, the R-2R topology also does not require a
thermometer decoder, which simplifies circuit design. Moreover, since the output current switches have the same sizes, the data buffers, retimers and clock buffers driving the current switches can be identical for all the lanes as well. This simplifies the clock tree design, which is the highest frequency block in the system.

### 2.3.3 Fully segmented DAC

As shown on Figure 2.4, fully segmented N-bit DACs require $2^N - 1$ current sources. A thermometer decoder is used to convert the binary bits to thermometer code. As the digital code increases by one, the thermometer decoded bits switch one more current source to the output. Fully segmented DACs address the two drawbacks of the binary-weighted current and R-2R DACs. First, the device matching requirement is relaxed, because the DAC is essentially made up of $2^N - 1$ identical modules. All the current sources have the same current, so the current switches and the circuit driving the current switches are also identical. 50% matching of the current is enough to guarantee DNL is < 0.5LSB. Second, the glitch energy is reduced. This is because when the digital word increases/decreases by 1, only one current switch is toggled. This also guarantees the monotonicity of the DAC.

![Figure 2.4: Thermometer code based current DAC.](image)

Both the static and dynamic performance is improved with a fully segmented DAC. However, the number of data lanes increases exponentially with the number of bits. This increases the area and, to a lesser extend, the power consumption. With a larger footprint and a large number of transistors connected in parallel, the capacitance at the output of
the DAC can be quite large. This explains why there are no published high speed, high resolution DACs with full segmentation.

### 2.3.4 Partially segmented DAC

To balance the tradeoffs between binary-weighted and fully segmented DACs, a partially segmented architecture is typically employed, where the LSBs are binary-weighted and the MSBs are segmented. The block diagram for this architecture is illustrated in Figure 2.5. This topology is very popular in high speed, high resolution DACs. The thermometer-coded MSBs result in a higher resolution DAC by improving the static and dynamic linearity of the DAC, while suppressing glitch energy. The binary-weighted LSBs save area and power and reduce the thermometer decoder complexity. The partition of the DAC into binary-weighted bits and fully segmented bits is a matter of design tradeoff of power, area, effective resolution, glitch height and decoder complexity.

In 2001, Van den Bosch et al., reported a 1GS/s DAC with 10 effective number of bits at Nyquist frequency [18]. The DAC has 5 binary-weighted LSBs and 5 fully segmented MSBs (i.e. 31 bits in thermometer code). The authors described many techniques used to achieve the high effective resolution. Some of those include:

1. Careful layout to minimize capacitance and ensure symmetry.
2. Properly sizing the current source transistors to achieve good matching between current sources.
3. Making sure the current source has large enough output impedance at Nyquist frequency to satisfy the required SFDR.
4. Use of many dummy cells to ensure good matching.

This shows that high resolution often comes at the cost of increased chip area.

Another example is a partially segmented, 22GS/s, 6-bit DAC published by Schvan et al. in 2005 [8]. The DAC has 3 binary weighted LSBs and 3 segmented MSBs (i.e. 7 bits in thermometer code). This creates 10 data lanes for the DAC. The binary portion was implemented with R-2R resistor ladder. The authors emphasized the importance of reducing layout parasitics and the matching of interconnect length in order to achieve high speed and good linearity. This DAC was implemented in a 0.13um SiGe BiCMOS process where the HBTs have $f_T$ and $f_{MAX}$ of 150GHz [8].
The highest speed DAC implemented in CMOS is a half-rate, 6-bit, 56GS/s DAC published in 2011 [11]. By implementing the majority of the circuits leading up to the current switches in CMOS logic, this DAC has a small area ($0.25\text{um}^2$) and consumes low power (0.75W). Also, the smaller area of CMOS logic cells means that the DAC can have a more aggressive 4:2 split between the thermometer-coded MSBs and binary-weighted LSBs, resulting in 17 data lanes in total. The larger number of segmented MSBs leads to improved linearity. Greater than 4.3 ENOB was achieved at 26.7GHz.

2.4 Half-rate vs. full-rate Nyquist DAC

In order to reach mm-wave sampling frequencies, half-rate clocks are sometimes used. A half-rate clock tree requires less bandwidth than a full-rate clock tree, so it is easier to design and consumes less power. However, there are two main drawbacks of a half-rate DAC. First of all, for a half-rate DAC, full-rate clock is not available to retime the data signals. As a result, jitter is higher for a half-rate DAC. Any duty-cycle-distortion in the output will not be suppressed. Secondly, for a full-rate DAC, the clock signal at $F_s=75\text{GHz}$ can be easily filtered out, since the output bandwidth is at most $F_s/2$ or (37.5GHz). However, for a half-rate DAC, the clock frequency $F_s=37.5\text{GHz}$ is close to the roll-off portion of the output bandwidth and it is not possible to filter out. Hence, the effect of clock feedthrough is more serious for a half-rate DAC.
2.5 Summary

Table 5.2 summarizes published DACs with greater than 10GS/s conversion rate. A few observations can be made. Firstly, with only three exceptions([11],[12],[24]), all of the DACs use some variant of bipolar based technology, such as SiGe BiCMOS, SiGe HBT, InP HBT. These processes offer transistors with higher $f_T$ and $f_{MAX}$ that are necessary for high sample rate converters. Although the $f_T$ and $f_{MAX}$ of the transistors in recent nano-scale CMOS technologies are comparable to that of InP HBTs, the metal backend is usually not as good for making passive components, such as inductors and transmission lines. In particular, the InP HBT ([21],[22],[23]) process also has a much higher resistivity substrate, which is better for making low-loss passive components, minimizing parasitic capacitance and the noise coupling through the substrate.

The most popular architecture is the partially segmented DAC, especially when the technology has a lossy substrate. While designs using InP process can afford to make the resistors very large to reduce mismatch ([21]), designs in processes with lossy substrate cannot do the same. This is because the parasitic capacitance from the large resistors will limit the bandwidth of the DAC. The partially segmented architecture offers more flexibility, which allows the circuit designer additional degrees of freedom to optimize the architecture for the desired speed and resolution.

Most high speed DACs have 6 bits of resolution. Unfortunately, dynamic performance is not always reported. An ENOB of 3-4 bits is reported at an output frequency close to Nyquist frequency.

At high frequency, it becomes increasingly challenging to distribute the clock signal in phase to a large number of data lanes for retiming the input data. A significant amount of the total DAC power is consumed in the clock distribution network. For example, for the full-rate 60GS/s DAC in [23], only 17% of the total power is consumed by the DAC core. Some designs were implemented at half rate clock to reduce power consumption in the clock tree.

In the end, the DAC architecture choices are often guided by the technology used. The designer needs to take advantage of the features that the technology provides, while working around its limitations.
Chapter 3

High speed digital library design

This section describes the high speed digital logic cells used in the 75GS/s DAC and the 108GHz retimer. Even at today’s most advanced technology nodes, only current-mode logic (CML) topology has greater than 50GHz bandwidth. CMOS has many advantages, such as low static power and small layout area. However, at higher frequency, the transistor gain is lower and there is not enough bandwidth for rail-to-rail switching. Hence, for high speed digital circuits, current-mode logic circuits are necessary.

3.1 Logic cell design

3.1.1 Current-mode logic

There are two main differences between current-mode logic (CML) and CMOS logic families. Examples of inverters implemented in both logic families are shown in Figure 3.1. The operation of CML is based on current switching between the transistors in the differential pair. The output logic swing is \( R_d \times I_d \), where \( R_d \) is the load resistance and \( I_d \) is the current from the tail current source. If the tail current is not completely switched, the output voltage swing is lower. Output voltage swing is one of the main differences between CML and CMOS logic. The output swing of CML is always smaller than that of the CMOS logic. This property allows CML to operate at a higher speed. This makes intuitive sense, because it takes less time to charge or discharge the load capacitor by a smaller voltage difference. The inverter delay can also be studied analytically through open circuit time constant (OCTC) analysis. It has been found that the switching speed
is dominated by RC delay \[25\]. Figure 3.2 and Figure 3.3 show the small signal model used for calculating the open-circuit time constant of the CML inverter and the CMOS inverter, respectively. This analysis is based on work done in \[26\]. When the CML inverter drives a fan-out of \(k\), the inverter delay is:

\[
\tau_{\text{CML}} = (C_{gd} + C_{db})R_L + k\{C_{gs} + (1 - A)C_{gd}\}(R_g/k + R_L) \quad (3.1)
\]

To analyze the CMOS inverter, it is assumed that the CMOS inverter is biased midway between \(V_{dd}\) and \(V_{ss}\) with both transistors ON as in an amplifier. For the CMOS inverter, the time constant is:

\[
\tau_{\text{CMOS}} = (C_{gd} + C_{db})R_{ds} + k\{C_{gs} + (1 - A)C_{gd}\}(R_g/k + R_{ds}) \quad (3.2)
\]

As can be seen from the equations, the inverter delay of a CML inverter is smaller if \(R_L\) is smaller than \(R_{ds}\). This is usually the case. \(R_{ds}\) is in the kilo-ohms range while \(R_L\) is in the hundreds of ohms range depending on the tail current and swing required.

The second difference between CML and CMOS logic is that CML requires a tail current source. This means that CML consumes static power that is data-rate independent. On the other hand, CMOS logic only consumes dynamic power, which is data-rate dependent. This is a drawback of using CML at lower frequency; however, at higher frequencies, the difference in power consumption is reduced, as the dynamic power of the CMOS logic increases linearly with frequency. One advantage of having a tail current source is that the current switching noise on the supply and ground is much smaller for current-mode logic than CMOS logic when data transition occurs.
3.1.1.1 MOS CML vs. Bipolar CML

Examples of MOS CML and Bipolar CML inverters are shown in Figure 3.1(a) and Figure 3.2, respectively. A major difference between the two inverters is that they require different input voltage swing to fully switch. For a bipolar CML circuit, the input swing only needs to be approximately 250mVpp (dependent on the thermal voltage of a pn junction[27]); while the input swing required for MOS CML is technology dependent and data-rate dependent[27]. The input swing required for complete switching is dependent on the bias current density of the MOSFETs[27]. On the other hand, the minimum CML gate delay is when the MOSFET is biased at 0.3mA/um when fully switched, regardless of technology node[25]. When biased at 0.3mA/um, the input swing required for a MOS CML inverter is still larger than that of a bipolar CML inverter. As shown in [28], given the same power consumption, a lower signal swing results in a smaller CML inverter delay. This can be seen by rearranging the delay equations shown in Equation (3.1) into
\[ \tau_{\text{CML}} = R_L \{ (C_{gd} + C_{db}) + k (C_{gs} + (1 - A)C_{gd}) \left( \frac{R_g}{kR_L} + 1 \right) \}. \] (3.3)

The term \( R_L \) can be substituted by \( \Delta V / I_{\text{tail}} \), where \( \Delta V \) is the output logic swing. The delay equation for a MOS CML inverter becomes:

\[ \tau_{\text{MOS CML}} = \frac{\Delta V}{I_{\text{tail}}} \{ (C_{gd} + C_{db}) + k (C_{gs} + (1 - A)C_{gd}) \left( \frac{R_g}{kR_L} + 1 \right) \}. \] (3.4)

For bipolar CML:

\[ \tau_{\text{Bipolar CML}} = \frac{\Delta V}{I_{\text{tail}}} \{ (C_{bc} + C_{cs}) + k (C_{be} + (1 - A)C_{bc}) \left( \frac{R_b}{kR_L} + 1 \right) \}. \] (3.5)

From Equation (3.4), it becomes obvious that with fixed power consumption (i.e. \( I_{\text{tail}} \)), inverter delay is lowered by minimizing voltage swing. Since bipolar CML requires smaller input swing than MOS CML, it leads to smaller inverter delay with the same tail current. In addition, \( C_{bc} \) and \( C_{cs} \) are smaller than \( C_{gd} \) and \( C_{db} \) for the same tail current.

Comparing two 4mA CML buffers implemented in 55um SiGe BiCMOS, the simulated 20% to 80% rise times are 1.62ps and 2.63ps for bipolar CML and MOS CML buffers, respectively.

Comparing delay equations of MOS CML and bipolar CML, three observations can be made[25]. First, the output time constant of the bipolar inverter is lower because \( C_{db} \) is higher than \( C_{cs} \) with the same tail current source. Second, the Miller capacitor of the MOS inverter is generally lower than bipolar inverter because the transconductance and gain of the bipolar transistor is larger. Third, gate resistance of the MOS transistor can be improved through layout without increasing \( C_{gs} \) or \( C_{gd} \). Hence, the input time constant can be improved. However, for the bipolar transistor, the input time constant does not improve with layout. The last two observations lead to the conclusion that MOS inverter can have a smaller input time constant.

A more subtle way that the larger transconductance of a bipolar transistor contributes to its higher switching speed is explained in [29]. The \( f_T \) equations for the two transistors with some parasitic interconnect capacitance \( C_{\text{int}} \) at the gate/base are:

\[ f_{T - \text{MOS}} = \frac{g_{m\text{MOS}}}{2\pi(C_{gs} + C_{gd} + C_{\text{int}})} \] (3.6)
\[ f_{T-Bipolar} = \frac{g_{m_{bip}}}{2\pi(C_{be} + C_{bc} + C_{int})}. \] (3.7)

By writing the equations in terms of the ideal \( f_T \) without the additional parasitic capacitance \( C_{int} \) as \( f'_T \), equation (3.6) and equation (3.7) become:

\[ f_T = \frac{f'_T}{1 + \frac{2\pi(f'_T C_{int})}{g_{m}}}. \] (3.8)

From equation (3.8), it is clear to see that a larger transconductance reduces the impact of layout parasitics. Even as the MOSFET dimensions shrink resulting in higher \( f_T \), the layout parasitics will not reduce much because the interconnect is not scaling as fast. The lower transconductance of a MOSFET will limit the high speed performance of the MOS CML even if its \( f_T \) becomes the same as that of the bipolar transistor.

### 3.1.1.2 MOS-HBT cascode

As mentioned in the previous section, a simple CML inverter suffers from the large Miller capacitor. This capacitance can be reduced by using a cascode. For a BiCMOS process, where both bipolar and MOS transistors are available, there are four possible cascode inverter topologies: MOS-only, HBT-only, MOS-HBT (i.e. cascoded HBT) and HBT-MOS (i.e. cascoded MOSFET). The delay comparison is studied in detail in [25]. The conclusion from the study is that MOS-HBT cascode, shown in Figure 3.4, has superior delay time constant compared to the other three topologies. This topology takes advantage of the smaller input time constant of the MOSFET and the lower output time constant of the bipolar transistor, as discussed in the previous section.

An additional time constant at node A in Figure 3.4(a), the drain of the MOSFET, also needs to be included, which is

\[ \frac{(C_{gd} + C_{db} + C_{be})}{g_{m_{Q2}}}. \] (3.9)

The time constant at node A is higher for MOS-only and HBT-MOS cascode topologies because \( g_{m_{MOS}} \) is lower than \( g_{m_{HBT}} \). Although the HBT-only cascode has a lower time constant at node A than the MOS-HBT due to lower capacitance at the collector node, it also has a larger input time constant.
The time constant of the MOS-HBT cascode is:

\[
\tau_{\text{BiCMOS}} = \frac{\Delta V}{I_{\text{tail}}} \left\{ (C_{bc}+C_{cs}) + k \left( C_{bc} + \left(1 + \frac{g_{m,HBT}}{g_{m,MOS}} \right) C_{be} \right) \left( R_b \frac{1}{kR_L} + 1 \right) + \frac{(C_{gd} + C_{db} + C_{be})}{g_{m,HBT}} \right\} \tag{3.10}
\]

The results in [25] were based on 130nm SiGe BiCMOS technology. Another technology node is available since then and it was used for some of the work in this thesis. The 55nm SiGe BiCMOS technology features MOSFETs with 55nm gate length and faster HBTs. The \( f_T \) and \( f_{MAX} \) are improved for both the MOSFETs and HBTs in the 55nm SiGe BiCMOS node. To demonstrate that the MOS-HBT cascode still has the highest bandwidth, AC simulation was performed for all four cascode topologies. The result in Figure 3.5 shows that the 3dB bandwidth of the HBT-MOS, HBT-only and MOS-only cascode topologies are 41GHz, 58GHz, 84GHz, respectively; while the MOS-HBT cascode has a 3dB bandwidth of 114GHz. The gain of the MOS-HBT cascode remains above 0dB up to 126GHz.

Due to its low time constant, the MOS-HBT cascode is a critical component in the high speed digital library. It is used in the high speed latch, buffer and the output stage of the DAC.

![MOS-HBT cascode inverter: (a)single-ended; (b)Differential.](image)

Figure 3.4: MOS-HBT cascode inverter: (a)single-ended; (b)Differential.

### 3.1.1.3 Emmitter-coupled Logic (ECL)

The bandwidth of the circuit can be increased by adding an emitter-follower between the inverters. This technique results in another logic family known as the emitter-coupled
Figure 3.5: AC gain comparison between the four cascode inverters in 55nm SiGe BiCMOS process.

logic (ECL) when one emitter follower is used, or $E^2CL$ when two emitter followers are in cascade. An example of a flipflop implemented in ECL is shown in Figure 3.6. The emitter follower stage has less than unity voltage gain, but has current gain and higher bandwidth than a CML buffer. It can be used to buffer the large load capacitance from the previous stage.

The disadvantage of ECL circuits is higher power consumption because of the additional current drawn by the emitter follower stage and the higher supply voltage needed. It becomes harder to use the emitter-follower for buffering as the supply voltage reduces, because the EF reduces the DC voltage on the input of the following stage. For example, with a 1.7V worst-case supply, it is not possible to use an EF to drive a bipolar CML, because it will leave only 50mV at the output of the tail current source.

3.1.1.4 Quasi-CML MOS-HBT cascode

The emitter-coupled MOS-HBT flipflop shown in Figure 3.6 is also an example of quasi-CML topology first introduced in [26]. The quasi-CML topology is different from the standard CML because it doesn’t have a tail current source. This saves an additional 200-300mV of DC voltage drop, which is what allows this circuit to operate with only
a 1.7V supply in the worst corner. This is a useful topology for implementing high speed logic with only a 1.8V supply. With this scheme, the MOSFET has a $V_{DS}$ of 650mV, while the HBT has a $V_{CE}$ of 900mV. Large $V_{DS}$ and $V_{CE}$ are necessary for high speed operation. With a 1.7V supply, this leaves a 150mV DC voltage drop on the load resistor, which provides an adequate logic swing of 300mVpp to switch the HBTs in the following stage. The transistors should be biased at the peak $f_T$ current density when fully switched. The penalty for removing the tail current source is that this topology has no common mode rejection.

The clock input can be AC-coupled or DC-coupled. Having a DC-coupled clock path is desirable because it can be used with a very low frequency clock. Having a broadband clock path that extends down to DC is especially useful for test chips. It enables testing at different clock frequencies. However, the DC-coupled flipflop is more sensitive to PVT variation and has higher power consumption. As shown in Figure 3.6, the current through the latch is set by the gate voltage, which is set by the base voltage of the emitter follower stage before. If the emitter-follower is driven by a resistor-loaded buffer referenced to the supply, as shown in Figure 3.8, the bias voltage at the input of the emitter can vary significantly over PVT. Since $V_{be}$ does not change significantly over PVT, the change in bias voltage at the base will translate to a change in the MOSFET gate voltage, which changes the current in the latch. Hence, the DC-coupled flipflop in Figure 3.6 is more sensitive to PVT variation. In situations where it is important that the circuit operation is maintained over a large variation in PVT, it is better to bias the gates of the MOSFETs with a current mirror and AC-couple the clock signals, as shown in Figure 3.7. In addition, AC-coupled pseudo CML flipflops consume less power
because the emitter followers are eliminated. The four resistors between the gate and the bias mirror node must have small layout footprint and be $> 2$-3KΩ. They provide a large impedance looking into the diode-connected current mirror and do not affect the switching behavior.

![Figure 3.7: Pseudo CML flip-flop with AC-coupled clock](image)

### 3.1.1.5 Shunt Peaking

In addition to the emitter-follower buffering technique mentioned in the previous section, another method to increase the circuit bandwidth without consuming more current is by adding an inductor in series with the output load resistor as shown in Figure 3.8. Bandwidth increase is due to the zero introduced by the inductor. With the inductor, the output impedance becomes

$$Z_L = (\omega L + R_L) \parallel \frac{1}{j\omega C_L}. \quad (3.11)$$

The voltage transfer function of the circuit becomes

$$A = g_m \left( (\omega L + R_L) \parallel \frac{1}{j\omega C_L} \right). \quad (3.12)$$

When the equation is recast into pole-zero form as shown [27], the gain becomes

$$A = g_m R_L \frac{1 + s \frac{L}{R_L}}{1 + s R_L C_L + s^2 C_L L}. \quad (3.13)$$
It is obvious to see that there is a zero added at \( \frac{R_L}{2\pi L} \).

The penalty of having inductors is increased area. However, since these inductors are in series with resistors, the Q-factor of the inductors does not need to be high. This means that a compact spiral inductor with several metal layers stacked vertically can be used to achieve the highest inductance per unit area. Minimum width metal can be used as well. If lower metal layers are used, it is also necessary to make sure that the self-resonant frequency is at least two times larger than the highest frequency of operation.

![Figure 3.8: An example of a resistively loaded bipolar buffer with inductive peaking.](image)

The inductor should be sized as

\[
L = \frac{R_L^2 C_L}{3.1}
\]  

(3.14)

to achieve a bandwidth improvement of 60% and a maximally flat group delay response \cite{27}.

### 3.2 Common mode oscillation

A sometimes neglected issue in high speed circuit design is common mode oscillation. Figure 3.9 shows the the half circuit of a differential pair with current source. The impedance looking into the current source at node X is

\[
Z_{cs} = 2R_o \parallel \frac{2}{j\omega C_{cs}}.
\]  

(3.15)
At high frequency, \( C_{cs} \) will lower the total impedance looking into the current source. This has two negative effects. First, the common mode rejection of this differential circuit becomes worse at higher frequency. Second, this capacitor can lead to common mode oscillation.

To demonstrate, the common mode impedance looking into the base of the bipolar transistor in a CML inverter with current source is

\[
Z_{icm} = \frac{R_b + R_e}{2} - \frac{w_{Teff}}{w^2 C_{cs}} + \frac{1}{jw C_{cs}} + \frac{w_{Teff}}{2jw g_{meff}} \tag{3.16}
\]

where \( R_b \) is the base resistance, \( R_e \) is the emitter resistance, \( w_{Teff} \) is the effective \( f_T \) and \( g_{meff} \) is the effective \( g_m \) which include the effect of \( R_e \) [27]. The negative term in the equation shows that the output capacitance of the current source \( (C_{cs}) \) can result in negative resistance looking into the base of the transistor. Increasing \( C_{cs} \) reduces negative resistance. This implies that since MOS current source has larger drain-bulk capacitance than the collector-substrate capacitance of a bipolar transistor, it causes less negative resistance at the input of the differential pair. In addition, it is important to minimize the interconnect capacitance from the output of the current source to the differential pair by using minimum distance as allowed by design rules.

Sometimes, an inductor is added in series with the current source to increase the common mode rejection of the circuit at high frequency. It is important to make sure that the inductor has a small footprint and a large self-resonant frequency so as not to add too much additional capacitance to the common node. A series resistors can also be added to further reduce the Q of the inductor to avoid oscillation.
Equation (3.16) ignores the effect of Ro. At lower frequencies, a smaller output resistance actually reduces negative resistance. When Ro is included, the common mode impedance becomes

$$Z_{\text{im}} = \frac{R_b + R_e}{2} + \frac{W_{\text{Teff}}}{2j\omega g_{\text{meff}}} + (R_0 \parallel \frac{1}{j\omega C_{\text{cs}}}) \times \left( \frac{f_{\text{Teff}}}{j\omega} + 1 \right).$$

(3.17)

The equation above can be rearranged as

$$Z_{\text{im}} = \frac{R_b + R_e}{2} + \frac{W_{\text{Teff}}}{2j\omega g_{\text{meff}}} - \frac{f_T}{2\pi f^2 C_{\text{cs}}} \frac{f}{R_0} + \frac{jf}{R_0} - \frac{2\pi f^2 C_{\text{cs}}}{2\pi f^2 C_{\text{cs}}}. $$

(3.18)

The third and fourth terms in the equation can be further rearranged to reveal the real part of the equation by multiplying the numerator and denominator by denominator’s complex conjugate. The real part of the two terms becomes

$$- \frac{f_T \times (2\pi f^2 C_{\text{cs}})}{(2\pi f^2 C_{\text{cs}})^2 + (\frac{f}{R_0})^2} + \frac{f}{2\pi f R_0 C_{\text{cs}}^2}.$$

(3.19)

This equation shows that a smaller Ro will reduce the negative resistance. A smaller Ro reduces the negative term and increases the positive term. Once again, a simple MOS current source results in a lower negative resistance at the common-mode input node than a bipolar current source because it has lower output resistance. Cascode MOS current source also suffers from the same high output impedance problem as bipolar current source.

A simple current source model with an ideal current source in parallel with Ro and Ccs is used to demonstrate the impact of Ro and Ccs at low frequency, the value of Ro dominates the change in Zim. As can be seen from Figure 3.10, the negative resistance at the input increases from 237 ohms to 400 ohms when R0 increases by 10 times. However, at high frequency, the value of Ccs has a bigger influence. Figure 3.11 shows that as Ccs increases, the negative resistance decreases.

Having a larger negative common-mode resistance is only a problem when there is a low-loss LCR parallel network connected to the input, which can lead to oscillation. This can
happen when a long interconnect, with a large inductance and capacitance, is connected to the input. If the load resistor at the output of the driving buffer is small and the resistance of the wire is small, a large negative resistance can cause oscillation. However, in general, when the interconnect between buffers is short and common-mode oscillation is not an issue, then a bipolar current source or cascode current source with higher output impedance is desired to increase common-mode rejection across frequency.

![Simulation results showing change in Zim as the value of Ro is swept while Ccs is kept constant.](image)

**3.3 Driving long interconnect**

As transistor speed increases, the bandwidth of a circuit becomes more interconnect limited as opposed to being limited by the intrinsic capacitance of the transistor. In addition, with the increase of the circuit’s signal frequency, the length of a long on-chip interconnect becomes a significant fraction of the wavelength. For example, at 75GHz, the wavelength in silicon dioxide is approximately 2.7mm \[30\]. 10% of the wavelength corresponds to 270um. Hence, interconnect longer than 270um is electrically large and should be modeled as transmission lines. Proper termination may be necessary to prevent the circuit from ringing as a result of reflection.
To improve the bandwidth of the circuit, one can use an emitter follower after the long interconnect. This reduces the load capacitance of the driving stage. Sometimes it is necessary to reduce the voltage swing of the stage to increase bandwidth when it is driving a large capacitive load. Follow this stage with an inverter with gain to get the necessary voltage swing. Also, the designer has the flexibility to break up long interconnect and put inverters in between where necessary.

Figure 3.11: Simulation results showing change in Zim as the value of Ccs is swept while Ro is kept constant.
Chapter 4

108GHz retiming flipflop

As discussed in the previous chapters, the high speed flip-flop is a critical building block of any DAC with a mm-wave sampling clock. As the data conversion rate increases, the timing alignment between the data lanes plays an important role in determining the dynamic linearity of the DAC. As explained in Chapter 2, this is especially critical for binary-weighted DACs, because the switching of the MSB can result in a large glitch in the output signal if the data bits are not well aligned due to data lane mismatch or random jitter in the data. By including a retiming flip-flop, the input data phase can be re-aligned first. In addition, the regenerating stage of the flip-flop can also increase the eye height if the input data signal is small. In this chapter, the design and measurement of a 108GHz retiming flipflop are presented. The attenuation through cables can be very high at mm-wave frequency range (> 20dB). As a result, in order to fully characterize the flip-flop in this frequency range, an input equalizer is also included on the test chip to compensate for the frequency dependent attenuation of the cable connecting the data source (i.e PRBS generator) and the test chip. A DC to 110GHz clock amplifier and a broadband 50Ω output driver are also designed and included in the test chip.

A paper based on this circuit was presented at IEEE CSICS in October 2013.

4.1 Prior Art

Examples of the highest clock frequency flipflops fabricated in several different technologies are discussed below. As shown in Figure 4.1, an 80Gb/s flipflop was demonstrated in InP HEMT technology in [31]. This circuit operated from a -5.7V supply and the power
consumption of the entire test chip was 1.2W. Inductive peaking was used to extend the bandwidth of the flipflop. Emitter followers were used between the two latches to reduce the loading of the slave latch. However, this topology is not possible with a 1.8V supply. A similar topology was used to implement a 50GHz SiGe MOS-HBT latch in a low-pass delta sigma ADC in [32] as shown in Figure 4.2. An additional emitter-follower was added on the clock path to reduce the loading on the clock path and maximize the bandwidth of the clock amplifier. This topology required a 2.5V supply. Another example is an 81Gb/s flipflop implemented in standard 65nm CMOS process, shown in Figure 4.3 [33]. This circuit operates from a 1.2V supply. The MOSFETs used in this flipflop have higher $f_T$ and $f_{MAX}$ than the MOSFETs used in the flipflop presented in this section of the thesis (130nm MOSFETs). Since only MOSFETs are available, the latches were designed with 400mVpp swing.

In a recent paper, a minimum sized quasi-CML, MOS-HBT flipflop with 3mA current and without inductive peaking was demonstrated to operate with clock frequency of
Chapter 4. 108GHz retiming flipflop

Figure 4.3: 81Gb/s retiming flipflop in 65nm CMOS [33]

Figure 4.4: 12GHz flipflop in 130nm SiGe BiCMOS process [34]

12GHz [34]. The schematic of the flipflop is shown in Figure 4.4. By eliminating the tail current source in the latch, this topology only required a 1.8V supply. The high speed operation of this topology with minimum sized transistors shows the promise of extending the frequency of operation to a higher rate.

4.2 Motivation for including an on-chip equalizer

The goal of this project was to design a test chip that could demonstrate > 75Gb/s, full-rate retiming after equalization with only 1.8V supply topologies in 130nm SiGe BiCMOS process with a minimum gate length of 130nm and a minimum emitter width of 130nm.

As explained in the introduction, as the frequency increases, the attenuation of the coaxial cable also increases. The measured insertion loss of a 4.9 meter long cable is shown in Figure 4.5. For 40Gb/s operation, the cable loss at the fundamental frequency is 17.85dB. For 80Gb/s operation, the cable loss is approximately 28dB at the fundamental frequency.
of 40GHz. To demonstrate its impact on the eye diagram, Figure 4.6 shows the measured 40Gb/s input eye diagram to the test chip after a 6 meter long cable. The attenuation of the cable connecting the PRBS generator and the test chip caused the eye to be completely closed at the input to the test chip. Hence, in order for the flipflop to be fully characterized at speed in the lab, additional circuitry needs to be included in the circuit breakout.

Figure 4.5: Measured loss of a 5 meter long high quality coaxial cable with 2.5mm connectors.

Figure 4.6: Eye diagram at the input of the chip after a 6 meter long cable at 40Gb/s.
4.3 Circuit description

The fully differential retimer, whose block diagram is shown in Figure 4.7, consists of a low-noise broadband equalizer, a retiming flip-flop, a broadband clock amplifier, and an output driver. The data come from off-chip and are first equalized by the equalizer. The equalizer helps to compensate for the frequency-dependent loss of the cable and probe connected between the PRBS generator and the circuit. It ensures that the signal swing is large enough at the input of the flipflop. The equalizer is based on a modified version of the equalizer previously published in [35]. The clock signal is also generated off-chip and is amplified by the on-chip clock amplifier. The clock amplifier is based on a modified version of the equalizer with gain > 0dB up to 110GHz. The flip-flop retimes the data. The role of the output driver is to provide the interface between the flipflop and the external loads, such as the oscilloscope and spectrum analyzer.

![Figure 4.7: Block diagram.](image)

To minimize power consumption, only a 1.8V supply is used. To ensure the bandwidth is not sacrificed, all circuit blocks are realized with low-voltage HBT-only or MOS-HBT topologies. Operation from 1.8V supply is ensured by vertically stacking no more than one MOSFET and one HBT between the power supply and ground. Furthermore, inductive shunt-peaking and the emitter-follower + MOS-HBT cascode topology are also used extensively to maximize the bandwidth. The design of each block is explained in the following sections.
4.4 Flip-flop design and simulation

The design of the flipflop was inspired by the 3mA flipflop published in [34]. To operate with a 1.8V supply, the same emitter coupled quasi-CML MOS-HBT topology was used on the clock path. However, in order to increase the bandwidth from 12GHz to 108GHz in the same technology, the bias current needed to be increased to obtain higher transconductance and reduce the load resistance. As a result, transistor sizes were also increased so that they were biased with the same current density. Two latches and emitter followers with 4mA and 8mA bias current were designed. They are shown in Figure 4.8. The two latches were designed to produce 250mVpp logic swing. Given the tail current of 8mA, the load resistor is chosen as

\[
R_L = \frac{\Delta V}{I_{\text{tail}}} = 31.25\Omega.
\]  

(4.1)

A load resistance of 30Ω was used to account for the additional parasitic resistance in the peaking inductor. The transistor sizes were chosen based on scaling the 1mA latch used in [34]. Given the large current of this latch, the MOSFETs were biased at 0.125mA/um, instead of 0.5x\(J_{\text{p}ft}\) of 0.15mA/um. When the current is completely switched to one side, the MOSFET has a current density of 0.25mA/um. This wider transistor increases the transconductance of the MOSFET differential pair. The HBTs were biased at 0.75x\(J_{\text{p}ft}\). With \(J_{\text{p}ft}\) of 8.5mA/um² and a minimum emitter width of 0.13um, the HBT emitter length was sized at 4.8um for the 8mA latch. The CBEBC layout was used for the HBT because of the large emitter length. This layout improves the transistor’s \(f_T\) and \(f_{MAX}\) by reducing \(R_B\) and \(R_C\).

Figure 4.8: Emitter-coupled MOS-HBT latches: (a) 4mA latch and 2x4mA emitter followers; (b) 8mA latch and 2x8mA emitter followers.
After layout extraction, the bandwidth of the latches on the clock path is shown in Figure 4.9. For the clock path bandwidth simulation, an AC signal is applied to the clock input at the gate of the MOSFET; while the HBT differential pair is switched. The latch is loaded with a fan-out of 1. Without inductive peaking, the 8mA latch has a 3dB bandwidth of 79GHz. With a load resistance of 30Ω, the equivalent total capacitance is

\[
C_{\text{Total}} = \frac{1}{2\pi f_{3dB} R_L} = \frac{1}{2\pi \times 79 \times 30} = 67fF.
\]  

(4.2)

The bandwidth of the circuit can be improved by adding inductive peaking. To achieve 1.6x increase in bandwidth, the peaking inductor is

\[
L_p = \frac{C_{\text{Total}} R_L^2}{3.1} = 21pH.
\]

(4.3)

This is a small inductance. A few simulation iterations with slightly higher inductance were performed. A final inductance value of 60pH was selected to extend the 3dB bandwidth of the circuit to 111GHz as shown in Figure 4.9. It is also worth noting that even with a 120pH peaking inductor, the 3dB bandwidth of the 4mA latch on the clock path only reaches 71GHz as shown in the Figure. This shows that an 8mA latch with inductive peaking is necessary to reach 110GHz bandwidth on the clock path.

Figure 4.9: AC bandwidth of the latch from clock input to latch output in the transparent phase.
The emitter followers (EF) on the clock path serve two purposes. First, they reduce loading on the driving stage. Second, they shift the DC voltage down so that the clock signal can be DC coupled to the MOSFET input in the MOS-HBT cascode. Since the emitter voltage in the EF stage is fixed (by $V_{BE}$), the resistor connected to the emitter is sized for the desired current in the EF. For example, with an emitter voltage of 520mV and the required current of 8mA, the resistor is equal to 65Ω. Note that the bias current of the latch is set by $V_E$ of the EF stage. Hence, with the MOSFET transistor sizing fixed, the only way to change the bias current is by adjusting the base voltage of the EF, which is set by the previous stage on the clock path.

The schematic for the final flipflop is shown in Figure 4.10. The complete layout of the emitter-coupled flipflop is shown in Figure 4.11. The clock signal comes from the south side, while the data input is on the west side. This placement ensures the symmetry of the clock input to the two latches and also minimizes the crossing between data signal and clock signal. Where possible, the interconnect metals are routed on top of the deep trench isolation (DTI) layer to minimize parasitic capacitance. The substrate ring around each HBT is removed in order to keep the distance between the transistors to the minimum. A single substrate ring surrounds the 8 HBTs in the flipflop.

![Figure 4.10: Proposed 108GHz flipflop in 130nm SiGe BiCMOS process](image36)

4.5 Broadband clock amplifier design and simulation

There are several challenges associated with the high frequency clock amplifier. First, a broadband, DC-100GHz amplifier had not been achieved in silicon at the time of this work. All of the reported high speed retimers use tuned clock path and have lower maximum frequency [31], [33]. A tuned clock path is easier to realize because it can be
designed to have gain at a higher center frequency than can be achieved with a broadband amplifier. Unlike the amplifiers on the data path, the clock amplifier does not need to have flat gain response and group delay variation is not a problem since the clock signal is sinusoidal. However, the disadvantage of a tuned clock path is that it cannot be used to retime data at lower frequencies outside of its bandwidth. Second, a differential signal is difficult to provide from off-chip at > 40GHz. The differential clock signal is either generated with an on-chip VCO or supplied from off-chip signal source single-endedly. In the later case, the circuit must suppress common-mode feedthrough. Common-mode rejection is very difficult to achieve from DC-100GHz.

The clock amplifier used in this chip is a modified version of the equalizer first reported as a stand-alone circuit in [35]. The original equalizer has gain up to at least 80GHz. It was modified to achieve gain from DC to 110GHz. The clock-path amplifier schematic is shown in Figure 4.12.

The first stage of the clock amplifier is a TIA stage. One input of the TIA is connected to the external clock source. This stage is used for broadband amplification, input matching and common-mode rejection. To achieve input matching to 50Ω, the feedback resistor is sized by
This resistor sizing assumes the gain of the TIA is 3. The small signal bandwidth of the TIA is shown in Figure 4.13. To achieve common-mode rejection, two techniques were applied in the TIA and the following stage. A cascode current source was used to improve common mode rejection at low frequency. The current source transistors were sized very large so that the $V_{DS_{-sat}}$ is low and the impact on the voltage head room is small. A common-mode inductor was added to improve common mode rejection at high frequency. The impedance at the common node reduces at higher frequency because of the capacitance at the output of the cascode tail current source. The series inductor improves the impedance at the common node at high frequency. Finally, the common-mode resistor was added to reduce the Q the inductor to avoid common-mode oscillation.
The second stage (b) is a bipolar differential amplifier stage. It provides 9dB of gain and further improves the common-mode rejection. To demonstrate the impact of the common-mode inductor and resistor, an AC simulation was performed to show the common-mode gain. Figure 4.14 and Figure 4.15 show the common-mode gain without and with the common-mode inductor and resistor, respectively. The first two stages provide more than 20dB of attenuation up to 10GHz without the common-mode inductor and resistor. However, there is common-mode gain from 40GHz to 70GHz. On the other hand, with the additional common-mode inductor and resistor, Figure 4.15 shows the common-mode gain is below 10 dB from DC-120GHz.

The bipolar amplifier stage is followed by a chain of MOS-HBT cascode inverter and bipolar differential inverter stages. The third stage (c) is an emitter-follower coupled MOS-HBT cascode. The purpose of this stage is to buffer the next stage (d) with large gain. It does not have much gain, but its 3dB bandwidth is >100GHz. The fourth stage (d) was designed to have a peaking frequency of 50GHz. Gain peaking is necessary to compensate for the frequency dependent cable and probe loss. This is achieved by using inductive peaking and capacitive degeneration. The voltage gain of a common-emitter stage with degeneration is

\[ A = \frac{R_L}{\frac{R_{deg}}{\sqrt{C_{deg}}} \cdot \frac{1 + j\omega R_{deg} C_{deg}}{R_{deg}}} \]  

(4.5)

This equation shows that the DC gain of this stage is \( R_L/R_{deg} \) and it increases with
frequency. With a degeneration capacitor and resistor of 20fF and 200Ω, respectively, and the 200Ω ON resistance of the nMOS transistor, the zero frequency is

\[ F_{\text{zero}} = \frac{1}{2 \pi R_{\text{deg}} C_{\text{deg}}} = \frac{1}{2 \pi \times 50 \times 40 f} = 80 \text{GHz}. \] (4.6)

Stages (c) and (d) are repeated to provide the gain and bandwidth necessary for the clock signal. The last EF-INV (g) stage is the buffer for the clock input to the flipflop.

As illustrated in Figure 4.16, the simulated clock path response has positive gain from DC to 110 GHz.

### 4.6 Eye diagram simulation

Figure 4.17 shows the simulated output eye diagram at the output of the flipflop at 40Gb/s. The peak-to-peak jitter is 0.4ps and the output eye has 6ps rise time (20% - 80%) and 8ps fall time. There are some ripples on the waveform due to clock feedthrough. Figure 4.19 shows the simulated output eye diagram at the output of the flipflop at 75Gb/s.

The timing margin of the flipflop was also investigated. In the testbench, the phase of the data signal was swept while the clock signal was kept constant. Figure 4.18 shows...
the output eyes with the data phases varying over 20ps overlaid on the same plot. This plot shows that with an input data phase change of ± 10ps, there is only a maximum phase shift of ± 0.8ps. At 75Gb/s, the timing margin of the flipflop is 8ps.

4.7 Fabrication technology

The circuit was manufactured in STMicroelectronics’ production 130-nm SiGe BiCMOS process featuring 130-nm Si MOSFETs and high-speed SiGe HBTs with $f_T$ and $f_{MAX}$
of 230 GHz and 280 GHz, respectively [37]. The die microphotograph is shown in Figure 4.20. The die area is 1.2 mm x 1.5 mm.

4.8 Experimental characterization

4.8.1 40Gb/s measurements

The highest data rate commercial PRBS generator available to the group is a 40Gb/s Centallex PRBS generator. Figure 4.21 shows the setup for measuring the output eye diagram with different input cable lengths. Two signal sources were used. One provides the clock signal for the PRBS generator and another provides the clock signal to the chip and locks the time-base trigger of the sampling oscilloscope. At the output of the PRBS generator, the data pass through cables of different lengths before reaching the
Chapter 4. 108GHz retiming flipflop

Figure 4.20: Die photo.

The output eye diagrams were measured for different input cable lengths. Figure 4.6 shows the input eye diagram after the data pass through a 6m long cable with approximately 18dB of loss at 20GHz. The eye diagram is completely closed at the input of the circuit. Figure 4.22 shows the measured eye diagram at the output of the chip after equalization and retiming. The measured 20% to 80% rise time is less than 6ps and RMS jitter is less than 400fs, including output probe and cable loss.

Although the two sources were locked with the 10MHz trigger, the drift between the two clock sources made it impossible to capture the BER with the build-in bathtub function of the oscilloscope. The bit patterns were captured through the pattern lock function with a lower frequency signal source, while the precision time base trigger was turned off. With this method, there is inevitably more jitter at the output. Also, this could only be done with a $2^7 -1$ sequence length due to the limited memory of the oscilloscope.
Figure 4.22: Output eye diagram after equalization and retiming at 40Gb/s.

Figure 4.23 shows the captured bit pattern. The blue line at the top in Figure 4.23 is the ideal pattern. This pattern was obtained by capturing the output of the PRBS generator directly. The transient waveform was converted to ideal binary bit sequence for easier comparison with the output from the circuit. This means that the impact of the finite rise/fall time of the signal from the PRBS generator was removed. The bottom plot shows the captured output of the chip. The pattern sequence was checked bit by bit with a Matlab script to make sure that they are correct.

The benefit of having a retimer is very clear by comparing the measured eye diagram of the equalizer chip (without the retimer) 35 with that of the retimer chip in Figure 4.24 and Figure 4.22, respectively. Both eye diagrams were taken after the input data passed through a 6 meter long cable. The jitter is reduced and the eye opening is greatly increased when the equalizer output is retimed.

Figure 4.23: Pattern checking for $2^7 - 1$ sequence length, 40Gb/s data.
Figure 4.24: Measured equalizer output eye diagram at 40Gb/s after a 6 meter long cable [35].

4.8.2 75Gb/s Measurements

To demonstrate the circuit operation at data rates above 50Gb/s, a custom PRBS chip, designed in 2005 [38] was used. The PRBS generator can provide a 75Gb/s output data stream. Because the PRBS chip was not packaged, it was mounted on a different probe station and the outputs were probed. Hence, two probe stations were necessary; one for the PRBS chip and another for the chip-under-test. The setup block diagram is shown in Figure 4.25 and a photograph of the setup is reproduced in Figure 4.26. The blue arrow indicates the location of the PRBS chip while the pink arrow points to the retimer chip. They were connected with a 3m long cable. In order to generate the high frequency clock, a multiplier was used. To align the clock and data signals, a phase shifter was used on the clock path. Although the two signal sources were synchronized with a 10MHz reference clock, this does not prevent them from drifting relative to each other. The frequency drift would result in additional measured jitter simply from the setup. As a result, it was very difficult to conduct a measurement over a very long time period. Ideally a single signal source should provide the clock signal to the PRBS chip, the retimer chip and the external timebase of the oscilloscope for triggering. However, this was not possible due to the lack of sufficient signal source output power at 40GHz and above.

Figure 4.27 shows the full rate retiming experiment at 75Gb/s. The blue trace at the bottom represents the output of the PRBS generator after a short cable. The effect of attenuation through the cable is small. The pink eye diagram at the top shows the equalized and retimed eye at 75Gb/s. Keep in mind that the data input to the chip needs to first go through a 3m long cable with approximately 16dB of loss at 40GHz. The on-
chip equalizer is a critical block that helped to open up the input eyes and increase the voltage swing before reaching the retimer. Both the jitter and the eye height are improved with the equalized retimer.

Figure 4.25: Measurement setup diagram with 75Gb/s PRBS generator[38].

Figure 4.26: Measurement setup picture with 75Gb/s PRBS generator[38].

### 4.8.3 Oversampling measurements with 108GHz clock signal

To verify the highest clock frequency at which the flipflop operates, the multiplier was used again to obtain clock frequencies above 100GHz. The data generated by the commercial PRBS generator was oversampled. Figure 4.28 shows the measurement setup. In
this setup, a 108GHz clock was used to oversample the 36Gb/s data by a factor of three. This oversampling ratio was used because the highest data rate from the PRBS generator is 40Gb/s and the largest clock frequency (< 110GHz) that is an integer multiple of the data rate is 108GHz.

Figure 4.29(a) shows the output signal at 36Gb/s without retiming by turning off the clock signal. Figure 4.29(b) shows the measured output eye diagram with data retimed with a 3x oversampling clock of 108GHz. The improvement in the jitter is clear after retiming, indicating that the flip-flop operates at 108GHz.

The captured bit pattern is shown in Figure 4.30. The red line at the top represents the measured output data stream of the chip. The blue line is the ideal pattern. The same Matlab script used for the 40Gb/s retiming experiment was used here to verify the correctness of the output data stream.
Figure 4.29: Measured 36Gb/s output eye diagram with and without retiming by an 108GHz clock.

Figure 4.30: Pattern checking for $2^7 - 1$ sequence length, 36Gb/s data, retimed with 108GHz clock.

### 4.8.4 Power Consumption

The entire retimer with equalization consumes 540 mW, with 220 mW attributed to the equalizer, 186 mW to the clock amplifier, 86 mW to the flip-flop, and 48 mW to the output driver.
4.8.5 Comparison with the state-of-art

Table 4.1 shows the comparison with the state-of-art.

The retimer presented in this work operates with the highest clock frequency. Because of the broadband clock amplifier, it can be used with clock frequencies from 40GHz to 108GHz. The integration of the equalizer and retimer is critical, because this enables the characterization of the flipflop at higher frequencies. This test chip demonstrates the highest data rate cable equalizer with retiming.

<table>
<thead>
<tr>
<th>Reference</th>
<th>31</th>
<th>25</th>
<th>39</th>
<th>33</th>
<th>36</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function</td>
<td>Flip-flop</td>
<td>Retimer</td>
<td>Mux</td>
<td>Retimer</td>
<td>Equalizer + retimer</td>
</tr>
<tr>
<td>Technology</td>
<td>InP HEMT</td>
<td>130nm SiGe</td>
<td>130nm SiGe</td>
<td>65nm CMOS</td>
<td>130nm SiGe</td>
</tr>
<tr>
<td>Clock (GHz)</td>
<td>80</td>
<td>45</td>
<td>66</td>
<td>81</td>
<td>108</td>
</tr>
<tr>
<td>Rise time (ps)</td>
<td>7.5</td>
<td>&lt; 7</td>
<td>5.67</td>
<td>5.78</td>
<td></td>
</tr>
<tr>
<td>Supply (V)</td>
<td>-5.7</td>
<td>2.5</td>
<td>-3.6</td>
<td>1.2</td>
<td>1.8</td>
</tr>
<tr>
<td>$P_{flipflop}$ (mW)</td>
<td>58</td>
<td>19.2</td>
<td>86.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficiency (uW/GHz)</td>
<td>1290</td>
<td>237</td>
<td>800</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: Comparison with the state-of-art

4.8.6 Summary

A 1.8V quasi-ECL MOS-HBT SiGe BiCMOS full-rate retiming flipflop was designed, fabricated and demonstrated to work up to 75Gb/s. Because of the broadband clock amplifiers, retiming with clock frequencies as high as 108GHz was also shown by over-sampling the input data by a factor of 2 and 3.
Chapter 5

75GS/s DAC

5.1 Technology

The technology used for this project is a 55nm SiGe BiCMOS process from STMicro-electronic. It features 55nm MOSFETs and SiGe HBTs with 0.1um emitter width. The n-MOSFET has measured $f_T/f_{MAX}$ of 200GHz/250GHz and the SiGe HBT has measured $f_T/f_{MAX}$ of 320GHz/380GHz. The combination of fast NMOS and HBTs in the MOS-HBT cascode is ideal for maximum switching speed. When maximum speed is not necessary, lower current designs can be used to trade-off speed for lower power consumption. This technology features one thick backend metal suitable for transmission lines, inductors and transformers.

5.2 8-bit 75GS/s DAC design

This section discusses the system level design of the DAC.

5.2.1 DAC specification

The goal of this project is to design an 8-bit, 75GS/s digital-to-analog converter. The desired output voltage swing is 1.2V peak-to-peak differential (i.e. 1.2Vppd). The output of the DAC should be matched to 50 Ω per side. The power consumption should be less than 0.5W, so low power topologies should be used whenever possible. The sampling
rate must be larger than 64GS/s over all process and supply corners and the output 3dB bandwidth must be larger than 25GHz. Hence, the footprint of the DAC should be minimized to maximize bandwidth.

This DAC is part of a larger system-on-chip. The high-speed inputs to the DAC include: a 75GHz differential clock signal with 300mVpp swing per side and twelve pairs of differential input data lanes (explained in the next section) at 75Gb/s with 300mVpp swing per side. Two voltage supplies of 1.8V and 2.5V are used. The 2.5V supply is only used in the DAC output stage and the two buffers driving the DAC output stage in order to provide the large output voltage swing. The full-rate, 75GHz, differential clock signal is generated with an on-chip VCO. The decoded 12 data lanes are also generated on chip.

5.2.2 Block diagram

This DAC consists of four main blocks: clock path, retimer, data path and the DAC output stage, as shown in Figure 5.1. The clock path distributes the clock signal to the flip-flops, where the input data is aligned. Only 1.8V logic family is used in the clock path and the flip-flops. The data path amplifies the retimed data and removes the effect of clock feedthrough from the flip-flops. In addition, it provides voltage level shifting in order to drive the 2.5V DAC output stage. The DAC output stage consists of current steering switches with associated current sources and the differential R-2R ladder.

The diagram in Figure 5.1 also shows the bit rate or frequency of each stage. Since full-rate retiming is used, the highest frequency block is the clock tree operating up to 75GHz. Twelve data lanes, each at up to 75Gb/s, drive the flip-flops. The number of data lanes is the result of the decision to segment this 8-bit DAC with 5 binary-weighted LSBs and 3 segmented MSBs (converted to 7 thermometer-coded bits). This is explained in more depth in the following section. The DAC output stage switches up to 24mA to the R-2R

![Figure 5.1: DAC block diagram.](image)
ladder and the 50 Ω off-chip termination resistors, which provide 600mV peak-to-peak voltage swing per side at the DAC output.

5.2.3 Investigation of the optimal bit segmentation split

The segmented architecture was chosen because of its advantages mentioned in Chapter 2. The R-2R ladder is used for the binary LSBs, while the MSBs are fully segmented. The resistance value chosen ensures that the output resistance of the ladder is 50Ω. As a result, no additional matching network or output driver is necessary.

The segmentation split between the number of R-2R bits and segmented bits was based on a study comparing the area, power consumption and glitch height. Table 5.1 summarizes the different options. The area and power are calculated based on each data lane having a 2mA retiming flipflop (35umx30um), a 1mA data buffer (25umx35um) and an output current steering switch (25umx25um). In reality, the power and area for DACs with a larger number of segmented bits will be higher, because the number of clock buffers also needs to increase. The tail current sources of the output current steering switches are adjusted so that the required 600mVpp swing per side is maintained. The glitch energy is simulated based on having a 1ps timing skew between the MSB and the LSBs. In reality, the glitch height can be lower after including all parasitic capacitance at the output node. The timing skew may increase depending on the delay mismatch between the data lanes. It is possible to match the delay between the clock lanes and the delay between the data lanes, but this will inevitably increase the size of the DAC.

<table>
<thead>
<tr>
<th>Num. of binary bits</th>
<th>Num. of segmented bits</th>
<th>Num. of data lanes</th>
<th>Area (mm2)</th>
<th>Current (mA)</th>
<th>Max glitch height (mV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8</td>
<td>256</td>
<td>0.653</td>
<td>793.5</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>128</td>
<td>0.326</td>
<td>409.5</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>65</td>
<td>0.166</td>
<td>220.6</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>34</td>
<td>0.087</td>
<td>128.3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>19</td>
<td>0.048</td>
<td>84.2</td>
<td>18.8</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>12</td>
<td>0.031</td>
<td>68.3</td>
<td>38</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>9</td>
<td>0.023</td>
<td>70.8</td>
<td>53</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>8</td>
<td>0.02</td>
<td>100.5</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: Segmentation split comparison table.

The first five options, with most of the bits fully segmented, consume large area and power. However, the glitch energy is lower. On the other hand, the last two options,
with most of the bits binary-weighted, have large glitch energy. A split of 5 binary bits and 3 thermometer bits was chosen because it represents a better balance of all the tradeoffs. Also, it results in an even number of data lanes, which simplifies clock distribution. The block diagram of the singled-ended version of the DAC output stage is shown in Figure 5.2. The actual DAC has differential output stage.

Figure 5.2: DAC output stage architecture single ended.

5.3 DAC block design

5.3.1 Clock path

The clock path is the circuit block that operates at the highest frequency. For this DAC, the clock signal is distributed to 12 flipflops which synchronize data in all lanes. Since the MOS-HBT cascode topology is used for the flipflops, the clock signal needs to be amplified so that the swing at the gate terminals of the MOSFET pair is at least 400mVpp per side across all corners to fully switch. In addition, the clock distribution network also needs to be able to drive longer interconnects because of layout placement.

The block diagram for the complete clock path is shown in Figure 5.3. There are five stages in the clock distribution network. The bandwidth suffers greatly when the buffer needs to drive both a long interconnect with large inductance and parasitic capacitance, as well as a large fanout. Because of this, the clock distribution has several cascaded stages to reduce the large loading. As shown in Figure 5.3, clock buffer 1 drives a short
interconnect and has a fanout of two, where the clock signal is split to two and routed to the top and bottom halves of the DAC. Clock buffer 2 drives a 330um long interconnect and has a fanout of one. Clock buffer 3 is a composite buffer consisting of an emitter-follower stage and an inverter (EF-INV). By using an emitter follower stage between the two inverters, the loading on the driving stage is reduced. Clock buffer 3 has a fanout of three. In order to maintain the bandwidth, Clock buffer 4 is also an EF-INV buffer. The purpose of clock buffer 5 is to increase the swing from 250mVpp to 400mVpp in order to fully switch the MOSFETs in the flipflops.

![Figure 5.3: DAC complete clock path](image)

It is important to note that, during the initial design, the emitter follower stage was not used in clock buffer 3. Clock buffer 2 directly drives a long interconnect and a large CML bipolar buffer with a bipolar current source. Common-mode oscillation was observed on the clock tree as explained in Chapter 3. This problem was mitigated by using a MOS current source instead.

The schematic of one branch of the clock path is shown in Figure 5.4. Please note that the fanout of the clock buffers is labelled on the schematic but not explicitly drawn.
As a result of the large loading and bandwidth requirement, the clock distribution network consumes 90mA from the 1.8V supply, equivalent to 162mW. If the flipflops are included, the clock path consumes 114mA or 205.2mW. The clock network alone consumes 53 percent of the total power.

5.3.2 Flip-flop design

As can be seen in the measurement results in Figure 5.5, the HBTs in the 55nm SiGe BiCMOS process has a 1.5x higher $f_T$ and a 1.43x higher $f_{MAX}$ than those in the 130nm SiGe BiCMOS. While the nMOSFETs in the 55nm SiGe BiCMOS process has a 2x higher $f_T$ and a 1.66x higher $f_{MAX}$ than those in the 130nm SiGe BiCMOS. The faster transistors imply that it is possible to implement the same MOS-HBT cascode flipflop topology presented in Chapter 4, but use less current to reach the same speed. In addition, based on simulation of a 4mA MOS CML inverter, the required swing to switch the current to 95%-5% in a MOSFET differential pair is 540mVpp in 130nm SiGe BiCMOS, while only 400mVpp is necessary in the 55nm SiGe BiCMOS process. As explained in Chapter 3, the reduction in voltage swing results in a lower delay time constant and hence faster switching speed. The reduction in the required swing for MOS CML is expected with smaller gate length transistors.

Since low power is an important goal in this project and there are 12 flipflops in the DAC, the design started with the minimum sized latch. The flipflops need to work up to 75Gb/s. The minimum emitter length HBT is 0.45um in this technology. With a peak $f_T$ current density of 1.5mA/um, this means that the minimum sized latch requires a tail current of 1mA for maximum switching speed. The HBT reaches 1.5x peak $f_T$ current density when the latch is fully switched. The MOSFETs are biased at peak $f_T$ of 0.15mA/um and reaches 0.3mA/um when fully switched. The flipflop schematic is shown in Figure 5.6. The load resistor is sized to provide 260mVpp logic swing. The
Chapter 5. 75GS/s DAC

Figure 5.5: Measured $f_T$ and $f_{MAX}$ of HBTs in 55nm SiGe BiCMOS, 130nm MW SiGe BiCMOS, 130nm SiGe BiCMOS and 250nm SiGe BiCMOS processes [40].

AC-coupled flipflop topology was chosen to eliminate the emitter follower and reduce current consumption. With a minimum sized transistor, the parasitics from the metal interconnect play an important role in determining the bandwidth of the latch.

Figure 5.6: Schematic of the 2mA AC-coupled MOS-HBT quasi-CML flipflop.

AC simulation was performed to evaluate the bandwidth for the clock and data paths of the flipflop. For clock path bandwidth simulation, an AC signal is applied to the clock input at the gate of the MOSFET; while the HBT differential pair is switched. For the data path bandwidth simulation, the AC signal is applied to the data input at the base of the bipolar differential pair; while the clock path is switched. The latch is loaded
with a fanout of one. In schematic level simulation, the minimum sized latch has 80GHz bandwidth on the data path.

The first layout attempt is shown in Figure 5.7. Only the transistors are shown. The inductors and load resistors are placed above the HBTs. To minimize interconnect parasitics, the HBTs are placed at minimum distance with the guard ring of the transistors overlapping. The collectors are connected in metal 8 (dark purple) to enable routing to the load inductors. Most of the lower level metals are placed above the deep trench to minimize parasitics. With this layout, the data path bandwidth of the latch reduced to 21GHz from 80GHz after layout parasitics extraction as can be seen in Figure 5.9. A careful examination of the extracted parasitic capacitance revealed that the parasitic capacitance between the two differential sides is the largest capacitance extracted in the output node. It was approximately 3fF. Looking at the layout, one can observe the crossing of the two differential signals (outp in metal 7 in light purple and outn in metal 8 in dark purple). In addition, the two outputs are in parallel for 10.5um length and spaced at 0.6um apart. Given that metal 8 is 3um thick, the sidewall capacitance is not negligible. These are the causes for the 3fF capacitance extracted.

![Figure 5.7: Bad layout](image)

Taking the above into consideration, a second version of the layout was done. Three major changes were made. First, where the crossing of the two outputs occur, the two differential output lines are placed on metal 6 (orange) and metal 8 (dark purple), instead of metal 7 and metal 8. This reduces the coupling capacitance between them. Second, the spacing of the two differential output lines is increased. Third, the interconnect across the HBT latching quad is further reduced by 1) rotating the bipolar transistors such that the narrower side of HBTs abut, 2) 45 degree interconnects are used in corners. The new layout is shown in Figure 5.8. With these layout changes, the bandwidth increased to 43GHz as can be seen in Figure 5.9. This shows that iteration in layout and extraction can
be tremendously beneficial. Improvement in layout results in higher bandwidth without increasing power consumption.

![Figure 5.8: Better layout](image)

![Figure 5.9: Data path bandwidth of the latch](image)

### 5.3.3 Data path

After the data signals are retimed, several data buffers are necessary to amplify the data signals and attenuate clock feedthrough. The block diagram and schematic for one lane of the data path block are shown in Figure 5.10 and Figure 5.11, respectively. There are 6 data lanes in the top half of the DAC and 6 lanes in the bottom half. The 1.8V buffer
Chapter 5. 75GS/s DAC

attenuates the clock feedthrough and amplifies the data such that the voltage swing is 240mVpp per side. Inductive peaking is added to improve the bandwidth of this stage. The 2.5V buffer has over 6dB of gain to increase the voltage swing from 240mVpp to 400mVpp. Figure 5.12 shows the simulated bandwidth of the data buffer stages driving the output stage. The simulated transfer function has a 3dB bandwidth of 49GHz and the gain is above 6dB from DC to 80GHz.

![Figure 5.10: One of six data lanes](image1)

![Figure 5.11: Data path showing DC biasing adjustment](image2)

The data buffers also help to attenuate clock feedthrough from the flipflop. To illustrate, Figure 5.13(a) shows the eye diagram at the output of the flipflop at 75Gb/s. The maximum ripple due to clock feedthrough is 140mV; while Figure 5.13(b) shows the output eye diagram after an additional data buffer. The clock feedthrough is reduced to 40mV.

The emitter follower provides the appropriate DC level for the output current steering stage. The DC biasing of the current switch is very important. To illustrate, the 2.5V buffer, emitter follower and the output stage are shown in Figure 5.11. If the DC level on the gates of the MOSFETs, Q2 and Q3, is too low, the collector voltage of the bipolar current source, Q1, decreases. The bipolar transistor, Q1, can go into saturation and the output impedance is degraded. If the DC level on the MOSFET transistors, Q2
and Q3, is too high, it will increase the source terminal voltage. Since the drain voltage is fixed by transistors Q4 and Q5’s DC bias at the base and $V_{BE}$, the increase of the source voltage reduces the voltage headroom on the MOSFET current switch, Q2 and Q3, and their switching speed. It is possible to adjust the DC level on the MOSFET transistors by adding a common-mode resistor, $R_{bias,adj}$ in the 2.5V buffer as shown in Figure 5.11.

Figure 5.12: DAC data lane small signal bandwidth

Figure 5.13: Output eye diagram showing clock feedthrough: (a) at output of the flipflop; (b) after an additional buffer.
The 12 data lanes consume 141.6mW, which is 30% of the total DAC power consumption.

### 5.3.4 Output stage

The DAC core consists of current sources and current steering switches. The bipolar current source with degeneration is used to ensure high output impedance. As explained in [17], the static INL depends on the output impedance of the current source. For a fully segmented architecture, the maximum INL is equal to $IR_L^2N^2/4r_{ocs}$, where $I$ is tail current in one segment, $R_L$ is the load resistor, $N$ is the number of bits and $r_{ocs}$ is the output impedance of the current source. Hence, it is important to make sure that the static linearity is not limited by the finite output impedance of the current source. For this DAC, the output impedance is very large with the bipolar current source with degeneration because $V_A$ is $>100V$.

The output impedance of the current switches in the output stage also impacts the dynamic linearity as explained in [11]. The capacitance at the output of the current source creates frequency dependent output impedance with a pole at $\frac{1}{2\pi r_{ocs}c_{ocs}}$ and a zero at $\frac{gm}{2\pi c_{ocs}}$ for a simple current source. Figure 5.14 (a) shows the parasitic capacitance $c_{ocs}$. At higher frequency, the output impedance reduces because of the pole at $\frac{1}{2\pi r_{ocs}c_{ocs}}$. The frequency where the output impedance falls below the required impedance determines the bandwidth of the output stage for a specified dynamic performance. For this DAC, a MOS-HBT cascode is used as shown in Figure 5.14 (b). The low frequency output resistance is $gm_{Q3}gm_{Q1}r_{oQ1}r_{oQ3}r_{oQ5}$, which is larger than the simple current source with MOS current switches in Figure 5.14 (a). The cascode configuration produces another pole and zero pair from $C_{cas}$. The pole from $C_{cas}$ is small because the resistance at that node is approximately $\frac{1}{gm_{Q5}}$. As a result, $c_{ocs}$ still creates the dominant pole and the increased output impedance of the cascode results in increased bandwidth.

The MOS-HBT cascode is used in order to have more voltage headroom for higher switching speed. Not only do the common-base HBTs help to improve output impedance as explained previously, they also reduce the Miller effect of the clock MOSFET pair, which reduces the input time constant. The output current goes to the R-2R ladder and the off-chip termination resistor.

The complete DAC output stage is shown in Figure 5.15. The DAC output stage consumes 36.3mA from a 2.5V supply, which results in a power consumption of 80.75mW.
5.4 DAC block placement and layout

5.4.1 DAC block placement

The placement of the data and clock distribution networks strongly affects the bandwidth and dynamic linearity of the DAC. For example, a binary tree-like topology, such
as the one shown in Figure 5.16 has better delay matching between the clock signals used for retiming. When data signals are retimed with clock signals of the same phase, better linearity is achieved. However, by using only a fanout of two, this topology often ends up consuming more area, adding more parasitic capacitance and consuming more power.

![Binary tree clock distribution network](image)

**Figure 5.16: Binary tree clock distribution network.**

Two placement options were considered for this project. They are shown in Figure 5.17 and Figure 5.18. For option #1, the clock signal comes from the west side and is distributed to the retimer block in tree-like fashion. A fanout of three is used because only 6 clock signals are required. One can ensure that the interconnect (i.e. interconnect a-f in Figure 5.17) are matched in length to minimize clock signal delay mismatch. Once the data is retimed, one can adjust the interconnect length to the current steering stage (i.e. interconnect g-l in Figure 5.17) such that they are matched for all the data lanes. From delay matching perspective, this placement is good if the delay of the data lanes are carefully matched. However, as can be seen in Figure 5.17, this placement requires the data lanes to be staggered, which increases the area and layout parasitics of the DAC.

The second option is shown in Figure 5.18 where the data lanes are split into two halves. Half of the data bits come from the north side and propagate south, while the other
half come from the south side and propagate north. The clock signal comes from the west side, same as in option #1. The clock distribution has a tree-like topology as well. Compared to option #1, the advantage of this placement is that there is no need to compensate for delay mismatch between data lanes after the retiming block. All of the retimed data lanes already have equal delay to the current steering stage at the output. As a result, this placement is more compact than option #1. Also, since the DAC core is split in half, the distance between the outer-most current steering stages and the current summing node in the middle is reduced by half. This reduces the output interconnect length, hence minimizes the capacitive and inductive loading of the DAC core.

One disadvantage of option #2 is that the first clock buffer on the west side now has to drive longer interconnect lines. For this DAC, the length of the interconnect is 330um. The large capacitive and inductive load limits the bandwidth of this stage. To minimize bandwidth degradation, two additional buffers are added such that they each drive one of the long interconnect lines.
Another consideration is the data input distribution. For option #2, the most obvious solution is for the data bits to come from the north and south sides of the DAC, as mentioned before. This topology minimizes the coupling between the data lanes and has fewer crossings between data and clock paths. The situation is different for option #1. The data bits can come from the west side, the same as the clock, or from the north and south sides of the DAC. As shown in Figure 5.19, if all of the data lanes come from the west, there will be a lot of crossings, hence coupling between data and clock. This is undesirable. On the other hand, shown in Figure 5.20, if the data lanes come from the north and south side, the two adjacent data lanes will be very close together, unless the distance between the clock buffer and flip-flop is increased. The coupling between data lanes is undesirable as it will lead to cross-talk. The crowding of data lanes is actually worse than the diagram presented in the figure, because in the actual circuit, there are 12 pairs of differential input data lanes, instead of 6 single lanes as shown in the figure.

In summary, in order to minimize layout footprint, the crossing between clock and data lanes, and the coupling between data lanes, placement option #2 in Figure 5.18 was selected.
Figure 5.19: (a) Data lanes routing from west side of the chip.

Figure 5.20: (b) Data lanes routing from north and south sides of the chip.
5.4.2 DAC top level layout

The complete layout of the DAC is shown in Figure 5.21.
5.5 Simulation after layout extraction

5.5.1 Flipflop timing margin

Since the main application of the retiming flipflop is to re-align the incoming data bits from another part of the chip, it is important to characterize in detail the timing margin of the flipflop. A testbench was set up where the data signal delay is swept, while the clock phase is kept constant. As shown in Figure 5.22, when the flipflop has sufficient setup and hold time, the output data, Q, has the same phase.

![Image of flipflop simulation result](image_url)

Figure 5.22: The simulation result shows that the flipflop output Q has a constant phase as the input data phase (not shown) is swept. The clock signal has a constant phase. This plot shows the flipflop output when it has sufficient setup and hold time.

For the 2mA flipflop used in this DAC, Figure 5.23 shows the range of input data delays where the flipflop fails to retime the data. The simulation was carried out in the typical corner at 75GHz. Note that this is a positive edge triggered flipflop. Ideally, the data should be stable when the differential clock is above 0V. However, if the input data switch in that 3ps window indicated in the figure, the flipflop is unable to realign the data and results in timing failure. In this case, the output data phase changes with the phase of the input data as shown in Figure 5.24. Since the clock period is 13.3ps, this equates to 77.7% timing margin. Similar simulation was performed in the slow corner at 125C, where the flipflop failure time window increases to 7ps. Hence, the timing margin
is reduced to 47%. This shows that in the worst case corner, the retiming ability of the flipflop is reduced significantly.

From Figure 5.22, the clock-to-Q delay can be calculated by taking the difference in the zero-crossing time of the clock signal and output data (i.e., Q). Retiming failure is defined by a change in clock-to-Q by ± 0.5ps. This definition is also used to determine the timing margin in the previous paragraph. With this testbench setup, it is very convenient to calculate the clock-to-Q delay. Figure 5.25 shows the calculated clock-to-Q delay for the TT corner. The simulated clock-to-Q delays are 10ps and 11.3ps for the TT and SS corners, respectively.

Figure 5.23: The simulation result shows the input data phase changing while the clock signal phase remains constant. Only the input data phases that cause timing failure is shown. During the 3ps window indicated on the figure, input data switches while the clock is at the end of the transparent phase. This results in insufficient time for the flipflop to sample the input data.
Figure 5.24: The simulation result shows that the flipflop output $Q$ has a different phase as the input data phase (not shown) is swept. The clock signal has a constant phase. This plot shows the flipflop output when it does not have sufficient setup and hold time.

Figure 5.25: Simulated clock-to-$Q$ delay in typical corner at 75GHz.
5.5.2 Impact of glitch energy

As the data conversion frequency increases, the timing alignment of the data lanes at the input of the current switches becomes more critical. The switching of the MSB can result in a large glitch if data lanes are not aligned. For example, if data transitions from 01111 to 10000, the delay between the MSB turning ON and the LSBs turning OFF can result in a large glitch in the output waveform. This effect is more serious if the the MSBs switch more current than the LSBs, such as the case of a binary-weighted current DAC. Due to the difference in current, the MSBs and LSBs will have different rise and fall time as well. This problem can also be observed with a partially segmented DAC with different currents in the binary section and the segmented section, as it is the case for the DAC designed in this project.

A simulation testbench is setup to find out the impact of timing misalignment. A simplified 5-bit DAC is used where the binary section is implemented with an R-2R ladder and the MSB switches twice the current as each bit in the binary section. The topology is similar to the one shown in Figure 5.2. The purpose of this testbench is to observe the glitch height caused by switching from 01111 to 10000. A variable delay is added between the switching time of the MSB and the LSBs. The output of the DAC is loaded with a 30fF capacitor on each side. Transient simulation results showing the synthesized voltage ramp at the DAC output at 70GS/s are plotted in Figure 5.26. As can be seen in the plot, when there is no delay between data bits, there is no glitch. With 0.5ps of delay, there's a 0.5LSB glitch. With 0.75ps of delay, the glitch height grows to 1LSB. The result agrees with simulations reported in [42]. They showed that a skew smaller than 1ps is necessary to ensure the glitch height is less than 1.5LSB at 25GS/s conversion rate. Their simulation result is reproduced in Figure 5.27.

As the skew becomes even larger, the output takes longer to settle. This can potentially result in a missed code. As the signal frequency increases, the data bit period is shorter; hence the output has less time to settle. This can be observed in Figure 5.26 where a 2ps delay results in a glitch height of 4.5LSB at 70GS/s. The output is not settled by the end of the 14.3ps period. This results in a missed code. On the other hand, at a lower data conversion rate of 25GS/s as in Figure 5.27 the bit period is 40ps. The glitch in the output ramp has 40ps to settle, which is sufficient even with a 5.8LSB of peak glitch height.

The timing misalignment is minimized by ensuring the clock paths to all the flip-flops have less than 50um length mismatch. In the technology used, a 50um mismatch corresponds
to 0.3ps delay mismatch. With the placement choice in Figure 5.18, the biggest delay difference comes from the clock distribution section where the clock signal is split into three branches from clock buffer 3 to clock buffer 4. The path length difference is approximately 50um, which equates to 0.3ps of delay. After the data is retimed, it is important to ensure that the data paths also have the same delay to the DAC output stage.

Figure 5.26: Glitch energy at 70GS/s.

Figure 5.27: Glitch energy at 25GS/s [42].
5.5.3 Mismatch simulation

Transistor and resistor mismatch reduce the effective resolution of the DAC. There are four mismatch contributors for this DAC: the resistors in the R-2R ladder, the current switch transistors, the current source transistors, and the degeneration resistors. The mismatch current going to the output termination resistor should be less than $\pm 1/2$ LSB. For example, for a peak-to-peak current swing of 12mA, the mismatch current has to be less than $\pm 23\mu A$ to achieve 8-bit resolution.

Without laser trimming, the resistors need to be sized large enough so that the output mismatch current is less than $\pm 23\mu A$. In the case of the R-2R ladder, increasing the physical size of the resistors limits the bandwidth of the output node. It is also important to note that the resistors also need to be wide enough to carry the current. In the case of this DAC, the resistor’s current-carrying capability is defined by the amount of current that causes less than 20°C temperature rise. After consulting the design rule manual for this technology, it was determined that the current density requirement sets the minimum resistor size for the resistor carrying the most current in the R-2R ladder rather than the mismatch requirement. Although the amount of current going through the resistor reduces by half for the lower bits, all resistors have the same dimension to achieve good matching. With the resistor size chosen, the simulated output mismatch current contribution from the R-2R ladder alone is $\pm 9.3\mu A$.

The current source mismatch also affects the static linearity of the DAC. As explained in [43], when the voltage drop on the emitter degeneration resistor $R_E$ is smaller than the thermal voltage, the collector mismatch current is dominated by $I_s$, the saturation current. However, for this DAC, the emitter degeneration is 150mV, approximately 6 times greater than thermal voltage. In this case, the mismatch is determined by $R_E$ and bipolar transistor mismatch. Because the current through $R_E$ is small, the width of $R_E$ is determined by mismatch requirement. The mismatch current contribution from the degeneration resistors alone is $1.58\mu A$.

While the mismatch contributions of the R-2R ladder and current source are evaluated at DC, the current switch mismatch needs to be considered at the highest switching frequency. When the current switch is fully switched, the mismatch between current switches contributes very little mismatch current. However, at the highest switching frequency, the current may not be completely switched due to bandwidth limitation. In this case, the mismatch of the current switches will have a bigger impact.
A test bench with one data lane with a flip-flop, data buffers and the output stage was setup to evaluate mismatch current at the highest switching frequency. Only one data lane was included to reduce the simulation time. An ideal schematic with no mismatch was simulated first. This ideal differential output current waveform was saved. Then Monte Carlo simulation with 200 mismatch samples was performed. The traces in Figure 5.28 show the differential output current with mismatch after subtracting the ideal current waveform. The spikes in the waveform is the result of different rise and fall times when the current switches, which is not useful information. The flat part of the waveform indicates the output current after it has settled. Figure 5.28 shows up to \( \pm 40\mu\text{A} \) of mismatch current.

By looking at the traces individually, it was found that the mismatch current consists of current scaling error and current offset. Figure 5.29 shows the two effects separately and combined. Please note the waveforms shown in the third column are idealized examples of the waveforms in Figure 5.28. The current scaling error is defined as the change in the tail current caused by current source mismatch. In particular, it is mostly caused by the bipolar transistor mismatch, since the degeneration resistor is already sized to meet matching requirements. Instead of further increasing the size of the transistor, another option is to add an additional tunable current source in parallel. This way, the current in each data lane can be individually tuned.

Figure 5.28: Monte Carlos simulation result showing the difference in DAC output current as a result of mismatch.
Chapter 5. 75GS/s DAC

Figure 5.29: Waveforms showing the effect of a) current scaling error from current source mismatch, b) current offset between the differential sides from DC voltage offset at the gates of the current switches, c) both effects combined. The first column shows the output current on each side. The second column shows the ideal differential output current in blue and the differential output with mismatch in red. The third column shows the difference between the ideal differential output current and the differential current with mismatch. Note that the waves in the third column represent an example of the mismatch waveforms shown in Figure 5.28.

Current offset is defined as the DC current difference between the two differential outputs. This is caused by the mismatch between the load resistors in the flipflop and the differential CML buffers in the data path. This mismatch causes DC voltage offset between the two differential sides at the gates of the current switches. Combining the DC offset with the limited switching bandwidth also causes the output current to vary. Fortunately, the impact of the DC offset for this DAC is relatively small (up to 22.6mV). If the offset is large and the circuit has sufficient bandwidth for complete switching, duty-cycle distortion will appear at the output. If the offset needs to be corrected, two additional current sources can also be added at the output of the previous stage on each side to adjust the DC level. However, this will increase layout parasitics.

In addition, to achieve even smaller mismatch, dummy cells are often placed around the DAC core to avoid edge effects[18]. This was only done for the R-2R ladder for the DAC in this thesis in order to minimize the added area. Also, the segmented MSBs bits can be calibrated and programmed to minimize random mismatch and achieve better resolution [44]. For example, the switching sequence of the thermometer lanes can be randomized so that gradient errors do not accumulate. Also an optimization algorithm can be adopted to find the best switching sequence for the least static linearity error.
5.5.4 Ramp simulation

The simulation in this section was performed with a 75GHz sinusoidal clock waveform. The output voltage ramp is generated by switching the data signals at 75Gb/s, 37.5Gb/s, 18.75Gb/s, etc. The data inputs are generated with vpwl model with 20% rise and fall time. Twelve additional buffers were used after the ideal signal sources to capture the bandwidth limitation at the data inputs. A Veriloga block was used to generate the binary-to-thermometer code.

All the long interconnects were modeled in EMX simulator, which generates an Nport S-parameter file. The output of the DAC is loaded with a 50 Ω termination resistor and a 30fF load capacitor per side.

Clock path differential voltage waveforms are shown in Figure 5.30. The waveforms below are at the input nodes of clock buffers 2, 3, 4, 5 and at the flipflop input in Figure 5.3. At 75GHz, the input clock swing to the flipflop is 350mVpp per side.

![Waveform Image]

Figure 5.30: Voltage swing at input nodes of all clock buffers at 75GHz in TT corner. The last waveform is the clock swing at the flipflop’s gate nodes.
Figure 5.31 shows the data path differential voltage waveforms at the input nodes of the 1.8V data buffer, 2.5V data buffer, 2.5V emitter follower, output current switch. The block diagram of one data lane is shown in Figure 5.10. The data rate is 75Gb/s. The input voltage swing to the DAC current switches is 450mVpp per side.

![Figure 5.31: Data path voltage swing at 75Gb/s in TT corner.](image)

The single-ended output voltage ramp is shown in Figure 5.32. Seven glitches can be seen. They are the result of timing-misalignment at the data transition when the thermometer-coded bits switch ON, while the binary bits switch OFF. The DAC output swing is 671mVpp per side, or 1.342V differentially.

![Figure 5.32: Single-ended output voltage ramp at 75Gb/s in TT corner.](image)
The differential output voltage ramp is shown in Figure 5.33. The largest glitch height is 15.87mV, which is equivalent to 3LSB.

INL was obtained by sampling the differential ramp above when the signal is stable. This is slightly different from standard static INL measurement, because the data switch at half of the clock frequency, as opposed to DC. As a result, the impact of settling and data skew is also included in the result. Figure 5.34a shows the calculated INL at 75GS/s in TT corner. The LSB was scaled based on the peak-to-peak output voltage. The last few codes have large distortion because they occur when the output transitions from maximum voltage to minimum voltage at full rate. The spikes in the INL are caused by glitches from data skew between the bits. If the glitches are ignored, the INL is within ± 0.7LSB for all other codes.
DNL was simulated in the same test bench as INL. Figure 5.34b shows the calculated DNL at 75GS/s in TT corner. Similar to the INL results, the spikes are from data skew between the bits. If the glitches are ignored, the DNL is within $\pm 0.8\text{LSB}$ for all other codes.

### 5.5.5 Dynamic performance

Ideal sine waves at 1GHz, 10GHz and 36.5GHz were used as inputs to an ideal 8-bit verilog ADC to generate the digital input codes for the DAC. The transient simulation was run for 110ns. The input codes toggled from 0 to 255; hence the DAC output voltage was full scale in this simulation. 8192 points from the differential DAC output were used for DFT. The following plots were generated with the DFT function in Cadence.

DFT of a 1GHz output sine wave with 75GHz sampling clock in TT corner, is shown in Figure 5.35. An SFDR of 53.9dB was achieved.

Figure 5.35: DFT of a 1GHz output sine wave with 75GHz sampling clock in TT corner.

The DFT of a 10GHz output sine wave with 75GHz sampling clock in TT corner, is shown in Figure 5.36. An SFDR of 44.3dB was achieved.
Chapter 5. 75GS/s DAC

Figure 5.36: DFT of a 10GHz output sine wave with 75GHz sampling clock in TT corner.

DFT of a 36.5GHz output sine wave with 75GHz sampling clock in TT corner, is shown in Figure 5.37. An SFDR of 38.7dB was achieved. The tone at 34.5GHz is the intermodulation product of the signal tone at 36.5GHz and the image tone at 38.5GHz.

Figure 5.37: DFT of a 36.5GHz output sine wave with 75GHz sampling clock in TT corner.

5.5.6 Comparison with the state-of-art
<table>
<thead>
<tr>
<th>Reference</th>
<th>Technology</th>
<th>$f_{clk}$ (GHz)</th>
<th>$f_{sample}$ (GSa/s)</th>
<th>bits</th>
<th>Area (mm$^2$)</th>
<th>Power (W)</th>
<th>Architecture</th>
<th>SFDR/ENOB</th>
<th>Output swing (mVpp)</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>[45]</td>
<td>0.12um SiGe HBT</td>
<td>40</td>
<td>40</td>
<td>3</td>
<td>0.66</td>
<td></td>
<td></td>
<td>10dBc</td>
<td>@20GHz</td>
<td>2004</td>
</tr>
<tr>
<td>[46]</td>
<td>0.13um SiGe BiCMOS</td>
<td>22</td>
<td>22</td>
<td>6</td>
<td></td>
<td></td>
<td>Partially segmented 3T:3B</td>
<td>650</td>
<td>2005</td>
<td></td>
</tr>
<tr>
<td>[19]</td>
<td>0.18um SiGe</td>
<td>10</td>
<td>20</td>
<td>6</td>
<td>0.031</td>
<td>0.36</td>
<td>Binary weighted current</td>
<td>50dBc</td>
<td>2006</td>
<td></td>
</tr>
<tr>
<td>[47]</td>
<td>0.18um SiGe BiCMOS</td>
<td>10</td>
<td>10</td>
<td>6</td>
<td>0.12</td>
<td></td>
<td>Partially segmented 2T:3B</td>
<td>2007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[48]</td>
<td>0.25um SiGe BiCMOS</td>
<td>30</td>
<td>30</td>
<td>6</td>
<td>0.7</td>
<td>0.455</td>
<td>Binary weighted resistor</td>
<td>2007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[49]</td>
<td>0.25um SiGe BiCMOS</td>
<td>20.48</td>
<td>20.48</td>
<td>6</td>
<td>0.88</td>
<td>1</td>
<td>Binary weighted current</td>
<td>28.2dBc</td>
<td>500 2010</td>
<td></td>
</tr>
<tr>
<td>[50]</td>
<td>0.7um InP DHBT</td>
<td>21.4</td>
<td>42.8</td>
<td>6</td>
<td>2</td>
<td></td>
<td></td>
<td>36dBc</td>
<td>&gt;13.5GS/s</td>
<td>2011</td>
</tr>
<tr>
<td>[22]</td>
<td>1um InP HBT</td>
<td>16</td>
<td>32</td>
<td>6</td>
<td>1.4</td>
<td>R-2R</td>
<td></td>
<td>29.2dB</td>
<td>@6.66GHz</td>
<td>2009</td>
</tr>
<tr>
<td>[21]</td>
<td>1um InP HBT</td>
<td>28</td>
<td>28</td>
<td>6</td>
<td>0.95</td>
<td>R-2R</td>
<td></td>
<td>&gt;36dB</td>
<td>@13.5GS/s</td>
<td>2010</td>
</tr>
<tr>
<td>[51]</td>
<td>1um InP HBT</td>
<td>60</td>
<td>60</td>
<td>6</td>
<td>1.8</td>
<td>R-2R</td>
<td></td>
<td>6.5b enob</td>
<td>500 2011</td>
<td></td>
</tr>
<tr>
<td>[24]</td>
<td>40nm CMOS</td>
<td>16.25</td>
<td>55-65</td>
<td>8</td>
<td>0.75</td>
<td></td>
<td></td>
<td>30dBc</td>
<td>@8GHz</td>
<td>2009</td>
</tr>
<tr>
<td>[11]</td>
<td>65nm CMOS</td>
<td>28</td>
<td>56</td>
<td>6</td>
<td>0.24</td>
<td>0.75</td>
<td>Partially segmented 4T:2B</td>
<td>30dBc</td>
<td>200 2011</td>
<td></td>
</tr>
<tr>
<td>References</td>
<td>Technology</td>
<td>Resolution</td>
<td>Power Supply (V)</td>
<td>Current (mA)</td>
<td>Overdrive (dB)</td>
<td>Output Bandwidth (GHz)</td>
<td>Device Year</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>-----------------</td>
<td>-------------</td>
<td>---------------</td>
<td>------------------------</td>
<td>--------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[12]</td>
<td>28nm CMOS</td>
<td>25</td>
<td>100</td>
<td>8</td>
<td>1.5</td>
<td>2.5</td>
<td>2014</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This work</td>
<td>55nm SiGe BiCMOS</td>
<td>75</td>
<td>75</td>
<td>8</td>
<td>0.168</td>
<td>0.47</td>
<td>2014</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2: Comparison with the State-of-Art DACs
Chapter 6

Conclusion

6.1 Summary

The analysis and design of a low power, high speed digital library for SiGe BiCMOS technology was presented. Using the MOS-HBT, quasi-CML topology, a test chip of an 108GHz retimer with equalization and a broadband clock amplifier was designed, fabricated and measured. This circuit was implemented in a 0.13um SiGe BiCMOS process. This is the fastest retimed equalizer to-date based on the author’s knowledge. Full-rate equalization and retiming at up to 75Gb/s was measured; while 108GHz operation was demonstrated by oversampling the input data stream.

The design, layout and simulation of an 8-bit, 75GS/s, full-rate, low-power DAC in a 55nm SiGe BiCMOS was presented. This is the highest sampling frequency (75GHz) broadband DAC to the best of the author’s knowledge. It has an output swing of 1.2Vppd, which is sufficient to directly drive high performance VCSELs for short-reach data center communication links. The low power circuit topologies minimize the power consumption of the DAC to < 0.5W. Most of the cells use a 1.8V supply only. A simulated SFDR of 44.7dBC was achieved at 10GHz output frequency; while a 38.7dBC SFDR was achieved at 36.5GHz. The DAC was designed using cells from the high speed digital library. The main challenges associated with designing a full-rate, mm-wave sampling DAC were discussed. Potential solutions were compared and the trade-offs were explained. Based on simulation after layout parasitic extraction, this is the lowest power, highest resolution, full-rate DAC in any process technology.
6.2 Contribution and Publication

The 108GHz retimer chip was published in the IEEE CSICS (Compound Semiconductor Symposium) in 2013 [36].

A record speed and low power 75GS/s DAC was designed, simulated and submitted to industry partner for fabrication and inclusion in a fiber-optic transmitter system.

6.3 Future Work

This DAC was developed as a compact, low power IP block that can be incorporated in larger systems for fiber optic communication. It can also be used as a building block in a higher sampling rate DAC using either time interleaving or analog mixing.

A possible future work is to use time interleaving to extend the conversion rate of this 75GS/s DAC. To satisfy Nyquist’ theorem, the input data bandwidth needs to be limited to below Nyquist to avoid alising and signal loss. As the output signal frequency approaches the Nyquist frequency, the signal frequency and the image frequency becomes closer. This is demonstrated in Figure [6.1]. The design of the reconstruction filter is very challenging because it needs to have a very high order to suppress the image. In fact, most DACs used in an arbitrary waveform generator (AWG) specifies the maximum output frequency as Fs/2.5, instead of the Nyquist frequency [15]. One way to remove the image is by using time interleaving. Figure [6.2] shows the time-interleaving DAC architecture used by Tektronix in an AWG. By using two sub-DACs, such as the one designed in this thesis, and sampling the input with 180 degree phase shifted clocks. One sub-DAC samples the odd data bits; while the second sub-DAC samples the even data bits. By summing the output of the two sub-DACs, the odd images can theoretically be completely cancelled. This results in the doubling of the Nyquist frequency to Fs. By time-interleaving two DACs, the overall conversion rate of the DAC also doubles. This technique relies heavily on the matching of the two DACs.

Another possible extension to this work is to use two of the 75GS/s DAC and up-convert the output of one of the DACs with a mixer as shown in Figure [6.3] to increase the DAC conversion rate.
Figure 6.1: DAC output spectrum with input at close to Nyquist frequency.

Figure 6.2: Example of a time-interleaving DAC by Tektrnoix [15].
Figure 6.3: Example of a DAC architecture that doubles the DAC sample rate from Ciena. [13].
Bibliography


[34] V. Kozlov, P. Park, and S.P. Voinigescu. 4 x 24 gb/s $2^7$-1 prbs generator using 1.8 v mos-hbt quasi-cml logic. In *2012 IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, pages 1–4, September 2012.


[40] S.P. Voinigescu. The design of sige bicmos digital circuits for 100 gbaud wireline and fiberoptics communication systems. In *2014 IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM) short course*, September 2014.


