Title

Adaptive Cancellation of Static and Dynamic Mismatch Error in Continuous-Time DACs

Author

Kong, Derui

Publication Date

2019

UNIVERSITY OF CALIFORNIA SAN DIEGO

Adaptive Cancellation of Static and Dynamic Mismatch Error in Continuous-Time DACs

A dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Electrical Engineering (Electronic Circuits and Systems)

by

Derui Kong

Committee in charge:

Professor Ian Andrew Galton, Chair
Professor Peter Michael Asbeck
Professor William S. Hodgkiss
Professor Thomas Tao-Ming Liu
Professor Patrick Philip Mercier

2019

The dissertation of Derui Kong is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2019

DEDICATION

To my wife, my daughter and my parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Ian Galton, for his endless support and guidance throughout my PhD journey. His passion for engineering, his high standard for academic research, and his dedication for his students are admirable. Without his help, I would not even be able to embark on my PhD journey, let alone to reach the end of it.

I would like to thank my lab mates for their valuable advices that improve the quality of my research. Special thanks to Kevin Rivas-Rivera for his help in the layout and the design of the mother board.

I would like to thank my wife, An Zhang, for giving me a wonderful daughter, for supporting my decision of pursuing my PhD degree. I would not have reached this stage of my PhD without her faith in me.

Chapter 1, in full, has been published in the IEEE Transactions on Circuits and Systems I: Regular Papers, volume 65, number 2, pages 421-433, February 2018. D. Kong, I. Galton, 2018. The dissertation author is the primary investigator and author of this paper. Professor Ian Galton supervised the research which forms the basis for this paper.

Chapter 2, in full, has been submitted to the IEEE Journal of Solid-State Circuits. D. Kong, K. Rivas-Rivera, I. Galton, 2018. The dissertation author is the primary investigator and author of this paper. Professor Ian Galton supervised the research which forms the basis for this paper.

Chapter 3, in full, has been submitted to the IEEE Transactions on Circuits and Systems I: Regular Papers. D. Kong, I. Galton, 2019. The dissertation author is the primary investigator and author of this paper. Professor Ian Galton supervised the research which forms the basis for this paper.

Chapter 4, in part, is currently being prepared for submission for publication of the material. D. Kong, I. Galton. The dissertation author is the primary investigator and author of this paper. Professor Ian Galton supervised the research which forms the basis for this paper.

VITA

2019	Doctor of Philosophy in Electrical Engineering (Electronic Circuits and Systems), University of California San Diego
2009	Master of Science in Electrical Engineering, Stanford University
2007	Bachelor of Science in Microelectronics, Fudan University

ABSTRACT OF THE DISSERTATION

Adaptive Cancellation of Static and Dynamic Mismatch Error in Continuous-Time DACs

by

Derui Kong

Doctor of Philosophy in Electrical Engineering (Electronic Circuits and Systems)

University of California San Diego, 2019

Professor Ian A. Galton, Chair

Inadvertent but inevitable mismatches among nominally identical unit element 1-bit DACs within a multi-bit Nyquist-rate DAC cause both static and dynamic error in the DAC’s continuous-time output waveform. Prior calibration techniques are able to suppress static mismatch error, but they have had limited success in suppressing dynamic mismatch error.

This first chapter of the dissertation presents a mismatch noise cancellation (MNC) technique that adaptively measures and cancels both static and dynamic mismatch error over the DAC’s first Nyquist band. The proposed digital calibration technique is capable of either foreground or background operation and is relatively insensitive to non-ideal circuit behavior. The chapter presents a rigorous mean convergence analysis of the technique and demonstrates the results of the paper with both behavioral and transistor-level circuit simulations.

The second chapter of the dissertation presents an integrated circuit DAC which implements the MNC technique of chapter one together with other circuit-level improvement techniques. With MNC enabled, this DAC demonstrates state-of-the-art performance.

This third chapter of the dissertation presents an improved version of MNC that addresses a practical concern. The original MNC technique requires an oversampling ADC clocked at a much higher clock rate than that of the DAC to measure the DAC's mismatch error, while the new technique presented in this chapter overcomes this limitation.

This fourth chapter of the dissertation presents a comprehensive mean-square convergence analysis of MNC proposed in chapter one, it proved that the noise impact on each coefficient in MNC, characterized by a steady-state mean square error metric, is bounded and can be arbitrarily reduced under certain practical conditions. It also established an analytical lower bound of DAC signal-to-noise-ratio (SNR) contributed by noise present in the system during calibration. The results of this paper provide guidance into the design of MNC.

CHAPTER 1

ADAPTIVE CANCELLATION OF STATIC AND DYNAMIC MISMATCH ERROR IN CONTINUOUS-TIME DACS

Abstract—Inadvertent but inevitable mismatches among nominally identical unit element 1-bit DACs within a multi-bit Nyquist-rate DAC cause both static and dynamic error in the DAC's continuous-time output waveform. Prior calibration techniques are able to suppress static mismatch error, but have had limited success in suppressing dynamic mismatch error. This paper presents a digital calibration technique that adaptively measures and cancels both static and dynamic mismatch error over the DAC's first Nyquist band. The technique is capable of either foreground or background operation, and is relatively insensitive to non-ideal circuit behavior. The paper presents a rigorous mathematical analysis of the technique, and demonstrates the results of the paper with both behavioral and transistor-level circuit simulations.

I. INTRODUCTION

High-resolution Nyquist-rate DACs with continuous-time output signals are required in critical applications such as wireless transmitters. Each such DAC interpolates a discrete-time input sequence to create a continuous-time output signal, so it can be viewed as a device that generates an analog output pulse for each input code. Ideally, the output pulse during the *n*th clock interval is scaled by the *n*th input code value, and except for this scale factor all the pulses have the same shape.

Such DACs generally consist of several nominally identical *unit element 1-bit DACs* in parallel. Unfortunately, inadvertent but inevitable fabrication mismatches among the unit element 1-bit DACs often limit performance. The mismatches cause non-ideal deviations of both the scale factor and shape of each overall DAC output pulse. Error in the overall DAC's output waveform from mismatch-induced pulse scale factor deviations is called *static mismatch error* and that from mismatch-induced pulse shape deviations is called *dynamic mismatch error*. Both types of error can significantly limit performance in practice.

Of course, there are many other types of non-ideal circuit behavior that contribute error in addition to static mismatch error and dynamic mismatch error. For example, if any of the unit element 1-bit DAC output waveforms depend on prior DAC input values in addition to the current DAC input value, a type of dynamic error called inter-symbol interference (ISI) is introduced. Nevertheless, these other types of error can be mitigated to a large extent by known circuit and system-level techniques. The same is true of static mismatch error. In contrast, prior techniques have been less successful in mitigating dynamic mismatch error.

Dynamic element matching (DEM) and digital calibration have been applied to address this problem in prior work, but with mixed results. DEM has been shown to prevent both static and dynamic mismatch error from causing nonlinear distortion, but it does so at the expense of degrading signal-to-noise ratio (SNR) [1-4]. Digital calibration techniques have been demonstrated that reduce static mismatch error, but prior calibration techniques do not significantly reduce dynamic mismatch error [5-11].

The difficulty arises from a fundamental property of continuous-time output DACs. Each DAC output pulse has a bandwidth that far exceeds the DAC’s signal bandwidth, because the pulse’s duration is limited to one clock interval. Hence, any technique to cancel dynamic mismatch error must either have a bandwidth that is much wider than the DAC’s signal bandwidth, or must somehow perform frequency selective cancellation over a particular band of interest such as the first Nyquist band. The situation is different in systems that only use sampled versions of DAC output signals, such as switched-capacitor delta-sigma ADCs and pipelined ADCs, and well-known techniques have been developed to cancel or otherwise suppress the effects of component mismatches in such cases [12-14]. Unfortunately, these techniques are not applicable to DACs with continuous-time output signals that are not resampled such as in wireless transmitters.

This paper proposes a mismatch noise cancellation (MNC) technique that addresses this problem. The MNC technique consists of a feedback path around a main DEM DAC. The feedback path adaptively measures and cancels both static and dynamic mismatch error within the DEM DAC’s first Nyquist band. The feedback path consists of an ADC, digital signal processing logic, and a correction DAC. As demonstrated in the paper, the performance requirements of the ADC and correction DAC are modest compared to the overall system performance.

The feedback path forms an estimate of the Nyquist-band portion of the main DEM DAC’s static and dynamic mismatch error by driving the correction DAC with the sum of the outputs of multiple digital filters driven by different pseudo-random digital sequences. The pseudo-random sequences are generated explicitly within the main DAC’s DEM logic so they are known a priori, but the filter coefficients depend on the component mismatches, so they must be estimated by the MNC technique. The feedback path correlates a digitized version of the overall system's analog output waveform by time-shifted versions of each pseudo-random sequence and uses the results to adaptively estimate the filter coefficients. Thus, the feedback path implements several feedback loops that operate in parallel.

The MNC technique functions regardless of the DAC's input sequence, so it can be used in both foreground and background calibration modes. The convergence rate can be maximized in foreground mode, though, so foreground mode can be used to minimize the initial convergence time and background mode can be used to adaptively track out temperature variation effects.

The paper describes the proposed MNC technique in detail, presents a rigorous mathematical convergence-rate analysis, and presents simulation results. Section II presents DEM DAC background information. Section III describes the MNC technique and its analysis in detail. Section IV presents behavioral and transistor-level simulation results that support the theoretical findings of the paper.

II. BACKGROUND INFORMATION

A. Ideal Behavior of a Practical DAC

As illustrated in Fig. 1, a DAC converts a discrete-time digital sequence, $x[n]$, with a sample-rate of f_s, into a continuous-time analog waveform, $y(t)$. The ideal output of a practical DAC is

$$y(t) = \alpha(t)x\left[n_t\right] \quad \text{where} \quad n_t = \lfloor f_s t \rfloor, \tag{1}$$

and $\alpha(t)$ is a periodic *pulse shaping waveform* with period $1/f_s$.[1] It can be verified that the continuous-time Fourier transform of $y(t)$ is

$$Y\left(j\omega\right) = X\left(e^{j\omega T_s}\right) A_p\left(j\omega\right) \tag{2}$$

where $X(e^{j\omega})$ is the discrete-time Fourier transform of $x[n]$, $A_p(j\omega)$ is the continuous-time Fourier transform of

$$\alpha_p(t) = \begin{cases} \alpha(t) & \text{if } 0 \le t \le T_s, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

and $T_s = 1/f_s$ is the sample period of the DAC [15].

Example spectra are shown in Fig. 2. The periodicity of the discrete-time Fourier transform gives rise to multiple *Nyquist bands*, three of which are shown in the figure.[2] A practical DAC is designed to faithfully represent its input sequence over a single Nyquist band, most commonly the first Nyquist band. Strictly speaking, this would require that $A_p(j\omega)$ have a magnitude of unity and a constant group delay over the desired Nyquist band, which is not easy to achieve with practical circuits. However, a digital filter can be inserted between $x[n]$ and the DAC's input to compensate for deviations of $A_p(j\omega)$ from unity magnitude and constant group delay over the desired Nyquist band. Therefore, moderate deviations of $A_p(j\omega)$ from unity magnitude and constant group delay over the desired Nyquist band are not problematic in practice.

[1] By definition, n_t is the largest integer less than or equal to $f_s t$ at time t, so it is a continuous-time waveform. Hence, $x[n_t]$ is a continuous-time waveform even though $x[n]$ is a discrete-time sequence.

[2] The kth Nyquist band for k = 1, 2, …, is defined as the set of frequencies that satisfy $\pi(k-1)f_s<|\omega|<\pi k f_s$.

B. Dynamic Element Matching

Fig. 3 shows the general form of a DEM DAC for an input sequence which takes on values in the range $\{-\frac{1}{2}L\Delta, \Delta-\frac{1}{2}L\Delta, 2\Delta-\frac{1}{2}L\Delta, \ldots, L\Delta-\frac{1}{2}L\Delta\}$, where L is the number of input levels minus one and Δ is the DAC's minimum input step-size [3]. The DEM DAC consists of an all-digital DEM encoder followed by I 1-bit DACs, the outputs of which are summed to form $y(t)$. The output of the ith 1-bit DAC has the form

$$y_i(t) = \left(c_i[n_t] - \tfrac{1}{2}\right) K_i \Delta + e_i(t) \tag{4}$$

where the 1-bit DAC's f_s-rate input bit sequence, $c_i[n]$, takes on values of 1 and 0, K_i is a constant called the 1-bit DAC's weight, and $e_i(t)$ represents all deviations from pure two-level behavior including effects such as intentional pulse-shaping and unintentional error from non-ideal analog circuit behavior.

By design, each K_i is an integer, $K_1 = 1$, and $K_{i-1} \leqslant K_i \leqslant K_1 + K_2 + \cdots + K_{i-1} + 1$ for $i = 2, 3, \ldots, I$ [16]. In practice, 1-bit DAC weights of $K_i > 1$ are implemented by combining multiple unit element 1-bit DACs in parallel. Thus, the ith 1-bit DAC consists of K_i unit element 1-bit DACs in parallel.

The DEM encoder maps each input sample, $x[n]$, to I output bits, $c_i[n]$, for $i = 1, 2, \ldots, I$, under the constraint

$$x[n] = \sum_{i=1}^{I} K_i \left(c_i[n] - \frac{1}{2} \right) \Delta . \tag{5}$$

This constraint is sufficient to ensure that the DEM DAC satisfies (1) with $\alpha(t) = 1$ if $e_i(t) = 0$ for every 1-bit DAC and that the number of input levels, L, is $K_1 + K_2 + \cdots + K_I$ [16].

In practice, $e_i(t)$ in (4) is often well-modeled as

$$e_i(t) = \begin{cases} e_{11i}(t), & \text{if } c_i\left[n_t - 1\right] = 1,\ c_i\left[n_t\right] = 1, \\ e_{01i}(t), & \text{if } c_i\left[n_t - 1\right] = 0,\ c_i\left[n_t\right] = 1, \\ e_{00i}(t), & \text{if } c_i\left[n_t - 1\right] = 0,\ c_i\left[n_t\right] = 0, \\ e_{10i}(t), & \text{if } c_i\left[n_t - 1\right] = 1,\ c_i\left[n_t\right] = 0, \end{cases} \tag{6}$$

where $e_{00i}(t)$, $e_{01i}(t)$, $e_{10i}(t)$, and $e_{11i}(t)$, are T_s-periodic waveforms corresponding to the four different possibilities of the current and previous 1-bit DAC input bit values [15]. During any given T_s clock period, $e_i(t)$ is equal to exactly one of the $e_{00i}(t)$, $e_{01i}(t)$, $e_{10i}(t)$, and $e_{11i}(t)$ waveforms, so $e_i(t)$ is non-periodic and signal-dependent in general.

In DEM DACs, the 1-bit DAC weights by design are such that for most values of $x[n]$ there are multiple distinct sets of DEM encoder output bit values that satisfy (5). During each T_s clock period, the DEM encoder sets its output bits to one of these sets chosen as a function of a pseudo-random variable and, when spectral shaping of the DEM DAC error is required, also as a function of past input samples. This causes

$$y(t) = \alpha(t)x[n_t] + \beta(t) + e_{DAC}(t) \tag{7}$$

where $\alpha(t)$ and $\beta(t)$ are T_s-periodic functions of $e_{00i}(t)$, $e_{01i}(t)$, $e_{10i}(t)$, and $e_{11i}(t)$ that are independent of the type of DEM used, and $e_{DAC}(t)$ is an error waveform, called DAC noise, that depends on the type of DEM used, $x[n]$, and $e_{00i}(t)$, $e_{01i}(t)$, $e_{10i}(t)$, and $e_{11i}(t)$ [15]. The first term on the right side of (7) corresponds to the ideal DAC behavior given by (1). The $\beta(t)$ term is T_s-periodic so it consists only of tones at multiples of f_s. As these tones do not fall within any Nyquist band of the DAC output and do not depend on the DAC input, they do not cause significant problems in most DAC applications. Hence, $e_{DAC}(t)$ is the only significant undesirable component of the DAC output.

It can be shown that (6) implies that $e_{DAC}(t)$ contains two types of error in general, one that depends only on the current DEM DAC input sample, and one that depends on both the prior and current DEM DAC input samples [15]. The first type of error is caused by mismatches among the nominally identical unit element 1-bit DACs, so it is the sum of all static mismatch error and dynamic mismatch error, and is called mismatch noise. The second type of error results from non-ideal memory effects *within* each unit element 1-bit DAC that cause $e_i(t)$ to depend not only on $c_i[n_t]$ but also on $c_i[n_t-1]$. Hence, this latter type of error is ISI error.

DEM causes the mismatch noise to be a pseudo-random noise waveform that is free of nonlinear distortion, and in some cases spectrally shaped so as to minimize the noise within a desired frequency band. DEM causes much of the ISI error to be a pseudo-random waveform too, but even with DEM the ISI error contains a second-order distortion component. If DEM were not used (i.e., if the encoder were to choose only one of the possible sets of output bits for each given input value), (7) would still hold, but $e_{DAC}(t)$ would be a deterministic high-order nonlinear function of $x[n]$.

III. MISMATCH NOISE CANCELLATION TECHNIQUE

A. Problem Statement

DEM DACs achieve high linearity by effectively converting much of what would otherwise be nonlinear distortion into pseudo-random noise. While often preferable to nonlinear distortion, the noise is nevertheless a problem in wideband analog signal generation applications.

In the absence of ISI, if all of the unit element 1-bit DACs were perfectly matched and clocked at exactly the same time, then $e_{DAC}(t)$ would be zero. In this case, the $e_{00i}(t)$, $e_{01i}(t)$, $e_{10i}(t)$, and $e_{11i}(t)$ waveforms would differ from 1-bit DAC to 1-bit DAC only by the ideal 1-bit DAC scale factors, K_i. However, mismatches among the unit element 1-bit DACs including relative skew among their clock signals inevitably result from random errors introduced during fabrication as well as from systematic circuit design and layout constraints. Some of these errors change the scale factors of the $e_{00i}(t)$, $e_{01i}(t)$, $e_{10i}(t)$, and $e_{11i}(t)$ waveforms thereby giving rise to static mismatch error in the DAC's output waveform. Others change the relative shapes of the $e_{00i}(t)$, $e_{01i}(t)$, $e_{10i}(t)$, and $e_{11i}(t)$ waveforms across the 1-bit DACs thereby giving rise to dynamic mismatch error in the DAC's output waveform. As examples, in current steering 1-bit DACs, threshold voltage mismatches among the current source transistors contribute static mismatch error whereas capacitance mismatches and clock skew contribute dynamic mismatch error.

The objective of the proposed MNC technique is to adaptively measure and cancel the entire mismatch noise component of $e_{DAC}(t)$ over the first Nyquist band, which includes both static and dynamic mismatch error. The MNC technique cancels only a portion of the ISI error component of $e_{DAC}(t)$, so it should be applied to DEM DACs in which ISI error is not the dominant type of error. This requires that the rise and fall transients of each unit element 1-bit DAC are sufficiently well matched or else that return-to-zero (RZ) 1-bit DACs are used. RZ 1-bit DACs reset their outputs to a fixed value (usually zero) at the end of each T_s clock period. This causes $e_{00i}(t) = e_{10i}(t)$ and $e_{11i}(t) = e_{01i}(t)$ in (6), so ISI is avoided because $e_i(t)$ does not depend on past values of $c_i[n]$.

B. Proposed Solution

The MNC technique is explained below in the context of a design example that targets an effective number of bits (ENOB) of 13.5 over a 200 MHz first Nyquist band. The purpose of presenting the MNC technique in the context of the design example is to simplify the explanation, but the technique is not restricted to the specific design example details.

Fig. 4 shows a high-level block diagram of the design example system. It consists of a main DAC and a feedback path. The feedback path consists of a VCO-based oversampling ADC of the type described in [17] with an oversampling ratio of $R = 5$, a digital lowpass decimation filter, a bank of digital residue error estimators, and a correction DAC. The details of each block and the overall system's theory of operation are described in the remainder of this section and in Section IV, respectively.

The main DAC is a 14-bit DEM DAC with a DEM encoder of the type described in [3], 36 current-steering RZ 1-bit DACs, and a clock rate of $f_s = 400$ MHz. As shown in [16], it converts the input sequence, $x[n]$, into an analog waveform, $y(t)$, given by (7) with

$$e_{DAC}(t) = \sum_{k=1}^{35} d_k(t) s_k\left[n_t\right] \tag{8}$$

where each $d_k(t)$ is a T_s-periodic linear combination of the 36 sets of $e_{00i}(t)$, $e_{01i}(t)$, $e_{10i}(t)$, and $e_{11i}(t)$ waveforms, and the $s_k[n]$ sequences for $k = 1, 2, \ldots, 35$ are white random sequences that are uncorrelated with $x[n]$, uncorrelated with each other, zero-mean, and restricted to values of −1, 0, and 1. The DEM encoder randomly chooses the sign of $s_k[n]$ independently for all k and n, so all non-zero values of $s_k[n]$ are zero-mean, independent random variables. As the $d_k(t)$ waveforms are functions of component mismatches, they are not known a priori. In contrast, the $s_k[n]$ sequences are generated explicitly within the DEM encoder, so they are known to the system a priori.

Like the main DAC, the correction DAC is based on current-steering 1-bit DACs, and both DACs have differential outputs. The differencing operation in Fig. 4 is implemented at the circuit level by simply connecting the negative and positive outputs of the correction DAC to the positive and negative outputs, respectively, of the main DAC.

Although not shown explicitly in Fig. 4, the output of the bank of error residue estimators is re-quantized to have the same minimum step-size as the correction DAC. This step-size must be small enough that both the quantization error and any additional error introduced by the correction DAC have negligible effects on the performance of the overall system. It was found that a step-size equal to a quarter of that of the main DAC is more than sufficient to meet this objective. The maximum swing of the main DAC's output, $y(t)$, is much greater than that of $e_{DAC}(t)$ in practice, so the maximum swing of the correction DAC need only be a fraction of that of the main DAC. This makes it practical for the correction DAC's resolution to be modest despite its reduced minimum step-size relative to that of the main DAC. Accordingly, in this design example the correction DAC has a resolution of 9-bits and does not incorporate DEM.

The VCO-based ADC and lowpass decimation filter are designed such that the f_s sample-rate output of the decimation filter is equivalent to a digitized version of just the first Nyquist band of the overall output, $v(t)$. Although the design example system has a 200 MHz Nyquist band and an ADC oversampling ratio of $R = 5$, simulation results suggest that fairly high ADC noise and nonlinear distortion can be tolerated. In particular, they indicate that the noise and nonlinear distortion introduced by the VCO-based ADC prototype in [17] would negligibly affect the performance of the feedback loop even without the digital linearization described in [17]. They also indicate that the high input impedance of the ADC would negligibly load the outputs of the DACs.

The $s_k[n]$ residue estimators in Fig. 4 for $k = 1, 2, \ldots, 35$ are digital blocks that together generate the correction DAC's input sequence. Each $s_k[n]$ residue estimator is responsible for adaptively generating an output sequence that contributes a component in the correction DAC's output equal to the portion of the *k*th term in (8) over the first Nyquist band.

The details of the $s_k[n]$ residue estimator for each k are shown in Fig. 5, wherein N, P, Q, and K have values of 9, 3, 15, and $6 \cdot 10^{-5}$, respectively, for the example system. As described in more detail shortly, N represents a tradeoff between cancellation accuracy and digital complexity, P and Q are chosen according to the delay and impulse response spread, respectively, of the MNC feedback path, and K represents a tradeoff between MNC convergence speed and accuracy.

The $s_k[n]$ residue estimator consists of N f_s-rate channels, the inputs of which are the decimation filter output sequence, $r[n]$, and the outputs of which are summed to form the $s_k[n]$ residue estimator's output. The *m*th channel multiplies the decimation filter output by a time-shifted version of the $s_k[n]$ sequence, accumulates the result to generate a sequence $a_{k,m}[n]$, and multiplies $a_{k,m}[n]$ by another time-shifted version of the $s_k[n]$ sequence. As described above, the $s_k[n]$ sequences are restricted to values of −1, 0, and 1 which greatly simplifies the multipliers, and they are known to the system because they are calculated explicitly within the main DAC's DEM encoder. P-sample advanced versions of the $s_k[n]$ sequences are required, but this is not an issue provided $x[n]$ is known P samples in advance.

It can be seen from Fig. 5 that the output of the $s_k[n]$ residue estimator can be written as

$$\sum_{m=0}^{N-1} h_k[m] s_k[n+P-m]. \tag{9}$$

where $h_k[m] = a_{k,m}[n]$ for $m = 0, 1, \ldots, N-1$. It follows that the output of the $s_k[n]$ residue estimator is equivalent to the output of an N-tap FIR filter with input $s_k[n+P]$ and impulse response $h_k[m]$. The filter is not time-invariant because the impulse response evolves over time, n. As proven in Section IV, the feedback system causes the impulse response to adaptively converge such that the correction DAC's output contains a component equal to the portion of the kth term in (8) over the first Nyquist band. Therefore, the bank of $s_k[n]$ residue estimators in Fig. 4 can be viewed as the bank of adaptive FIR filters shown in Fig. 6, where $H_k(z)$ denotes the z-transform of $h_k[m]$.

C. Mismatch Noise Cancellation Principle

Even though the correction DAC does not incorporate DEM, its output has the same form as (7), i.e.,

$$y_c(t) = \alpha_c(t)x_c[n_t] + \beta_c(t) + e_{DAC\text{-}c}(t) \tag{10}$$

where the subscript c is used to distinguish the various terms from their main DAC counterparts, except $e_{DAC\text{-}c}(t)$ is harmonic distortion rather than noise [18]. Analysis as well as transistor-level simulations with realistic mismatches indicate that the correction DAC's minimum step-size is sufficiently small relative to that of the main DAC that $e_{DAC\text{-}c}(t)$ is negligible relative to $e_{DAC}(t)$. Hence, $e_{DAC\text{-}c}(t)$ is neglected in the analysis below. The $\beta_c(t)$ term is also neglected, because it does not have any components within the first Nyquist band, so it does not interfere with the cancelation process.

Therefore, by the same reasoning that led to (2), the continuous-time Fourier transform of the correction DAC output over the first Nyquist band is well-approximated as

$$Y_c\left(j\omega\right) = X_c\left(e^{j\omega T_s}\right)A_{p\text{-}c}\left(j\omega\right) \tag{11}$$

where $X_c(e^{j\omega})$ is the discrete-time Fourier transform of $x_c[n]$ and $A_{p\text{-}c}(j\omega)$ is the continuous-time Fourier transform of the right side of (3) with $\alpha(t)$ replaced by $\alpha_c(t)$. Also by the same reasoning that led to (2), the continuous-time Fourier transform of (8) is

$$E_{DAC}\left(j\omega\right)=\sum_{k=1}^{35}S_k\left(e^{j\omega T_s}\right)D_{p\text{-}k}\left(j\omega\right) \tag{12}$$

where $S_k(e^{j\omega})$ is the discrete-time Fourier transform of $s_k[n]$ and $D_{p\text{-}k}(j\omega)$ is the continuous-time Fourier transform of the right side of (3) with $\alpha(t)$ replaced by $d_k(t)$. To cancel $e_{DAC}(t)$ over the first Nyquist band it is necessary for (11) and (12) to equal each other for all $|\omega| < \pi f_s$. It follows from (11), (12), and Fig. 6 that this is achieved if

$$H_k(e^{j\omega})=e^{-j\omega P}\,\frac{D_{p\text{-}k}(j\omega)}{A_{p\text{-}c}(j\omega)}\quad\text{for }\ |\,\omega\,|\le\pi\, f_s \tag{13}$$

The inverse discrete-time Fourier transform of the right side of (13) is the ideal $H_k(z)$ filter impulse response and it is both infinite-length and two-sided, yet the actual $H_k(z)$ filters only have impulse responses that are nonzero for $n = 0, 1, \ldots, N-1$. Consequently, it is not possible to satisfy (13) perfectly. However, (13) represents a stable system, so the ideal impulse response converges to 0 as $n \rightarrow \pm\infty$. It follows from (13) that P is just a delay term, so increasing P simply shifts the ideal impulse response to the right. Consequently, P can be chosen large enough that the terms of the ideal impulse response are negligible for $n < 0$. Similarly, N can be chosen large enough that the terms of the ideal impulse response are negligible for $n \geqslant N$. So choosing N and P ensures that the error incurred by using length-N $H_k(z)$ filters to approximate (13) is negligible. As demonstrated in Section IV, $N = 9$ and $P = 3$ are sufficient to achieve more than 2.5 bits of both static mismatch error and dynamic mismatch error cancellation in the design example system.

It remains to show that the feedback causes $a_{k,m}[n]$ for $m = 0, 1, \ldots, N-1$ and $k = 1, 2, \ldots, 35$ to converge to values that cause (13) to be well approximated. A rigorous analysis that proves this result is presented next.

D. Convergence Analysis

The decimation filter's output can be written as $r[n]$=$r_{ideal}[n]$+$r_e[n]$+$r_c[n]$, where $r_{ideal}[n]$ is the decimation filter output sequence that would have occurred in the absence of both $e_{DAC}(t)$ and the correction DAC feedback loop, $r_e[n]$ is the additional error caused by $e_{DAC}(t)$ that would have occurred in the absence of the correction DAC feedback loop, and $r_c[n]$ is the additional component introduced by the correction DAC feedback loop. Therefore, the objective of the correction DAC feedback loop is to adjust the $a_{k,m}[n]$ values such that $r_c[n] = -r_e[n]$ for all n.

Each term in the summation on the right side of (8) has the form of with $d_k(t)$ playing the role of $\alpha(t)$ and $s_k[n_t]$ playing the role of $x[n_t]$. Consequently, each term can be viewed as being contributed by a separate ideal DAC with input sequence $s_k[n]$ and pulse shaping function $d_k(t)$. It follows that the relationship between $s_k[n]$ and its contribution to $r_e[n]$ must be that of a causal linear time-invariant (LTI) discrete time system. Denoting the impulse response of this LTI system by $b_k[n]$, it follows that

$$r_e[n] = \sum_{k=1}^{35} b_k[n] * s_k[n] = \sum_{k=1}^{35} \sum_{i=0}^{\infty} b_k[i] s_k[n-i]. \tag{14}$$

To the extent that nonlinearity and aliasing from the ADC can be neglected, similar reasoning implies that the relationship between $x_c[n]$ and $r_c[n]$ must also be that of a causal discrete-time LTI system. Hence,

$$r_c[n] = x_c[n] * \left(-h_c[n]\right) \tag{15}$$

where $-h_c[n]$ is the LTI system's impulse response (the −1 factor in this definition of $h_c[n]$ simplifies notation in the subsequent analysis). Furthermore, $h_c[n] = 0$ for all $n < 0$ for causality and also for $n = 0$ to prevent the feedback loop from being delay-free.

These observations and the signal processing operations shown in Fig. 4 and Fig. 5 imply that the input to and the output of the mth accumulator in the $s_k[n]$ residue estimator can be written as

$$u_{k,m}\left[n\right]=s_k\left[n-m+P-Q\right]\Bigg(r_{ideal}\left[n\right]+r_e\left[n\right] \\ -\sum_{l=1}^{35}\sum_{i=-\infty}^{n-1}\sum_{j=0}^{N-1}h_c[n-i]a_{l,j}[i]s_l\left[i+P-j\right]\Bigg), \tag{16}$$

and

$$a_{k,m}[n]=a_{k,m}[n-1]+Ku_{k,m}[n], \tag{17}$$

respectively. It follows that $u_{k,m}[n]$ = 0 at each value of n for which $s_k[n–m+P−Q]$ = 0, so $a_{k,m}[n]$ only changes at values of n for which $s_k[n–m+P−Q] \neq 0$. Given that the only non-zero values of $s_k[n]$ are 1 and −1, this implies that $a_{k,m}[n]$ only changes at values of n for which $s_k^2[n–m+P−Q]$ = 1.

Given that the convergence rate of each $a_{k,m}[n]$ sequence depends on the particular pattern of zeros and ones taken on by $s_k^2[n]$ for all n, the expected values of $u_{k,m}[n]$ and $a_{k,m}[n]$ conditioned on this pattern of zeros and ones are of interest. In the following, these conditional expectations are denoted as $\overline{u}_{k,m}[n]$ and $\overline{a}_{k,m}[n]$, respectively. As described above, all non-zero values of $s_k[n]$ are independent zero-mean random variables that take on values of 1 and −1. Furthermore, (16) and (17) imply that $a_{l,j}[n]$ does not depend on $s_k[n']$ for any $n' \geq n + P$. These properties with (14) and (16) imply that

$$\overline{u}_{k,m}\left[n\right]=s_k^{\,2}\left[n-m+P-Q\right]\Bigg(b_k\left[m-P+Q\right] \\ -\sum_{l=1}^{35}\sum_{i=n-m-Q}^{n-1}\sum_{j=0}^{N-1}h_c[n-i]E_{l,i,j}[n]\Bigg) \tag{18}$$

where $E_{l,i,j}[n]$ is the mean of $a_{l,j}[i]s_l[i+P–j]s_k[n–m+P−Q]$ conditioned on the pattern of zeros and ones taken on by $s_k^2[n]$ for all n.

By definition, $E_{l,i,j}[n] = \overline{a}_{k,j}[n-m-Q+j]s_k^2[n-m+P-Q]$ when $l = k$ and $i - j = n - m - Q$. Given that K is very small (e.g., $K = 6\cdot10^{-5}$ in the design example) it follows from (17) that $a_{l,j}[i]$ is only very weakly correlated with $s_k[n-m+P-Q]$ for all other values of l, i, and j in the triple sum of (18). Hence, any of these terms that are non-zero are very close to zero because all non-zero values of $s_k[n-m+P-Q]$ are independent, zero-mean random variables. Consequently, (18) can be well approximated as

$$\overline{u}_{k,m}\left[n\right] = s_k{}^2\left[n-m+P-Q\right]\Bigg(b_k\left[m-P+Q\right] \\ -\sum_{j=0}^{N-1} h_c\left[Q+m-j\right]\overline{a}_{k,j}\left[n-m-Q+j\right]\Bigg) \tag{19}$$

The expectation operator is linear, so (17) implies

$$\overline{a}_{k,m}[n] = \overline{a}_{k,m}[n-1] + K\overline{u}_{k,m}[n]\,. \tag{20}$$

The set of difference equations given by (20) with $\overline{u}_{k,m}[n]$ given by (19) for $m = 0, 1, \ldots, N-1$ specifies the evolution of the expectation of the coefficients of the kth FIR filter in Fig. 6. However, these difference equations present two analysis complications because of the s_k^2 term in (19). One complication is that the difference equations, while linear, are not time-invariant because the s_k^2 terms are zero for some values of n. The other complication is that the s_k^2 terms across the different equations are not zero for the same values of n.

The latter complication can be solved by replacing n with $n+m$ in each of the difference equations, because, as can be verified from (19), the s_k^2 terms in the expressions for $\overline{u}_{k,m}[n+m]$ are identical for all $m = 0, 1, \ldots, N-1$. The N equations obtained by substituting (19) into (20) for every $m = 0, 1, \ldots, N-1$ and replacing every occurrence of n by $n+m$ can be written in matrix form as

$$\begin{aligned}\mathbf{a}_k[n] = \mathbf{a}_k[n-1] \\ -\begin{cases} \mathbf{0} & \text{if } s_k\left[n+P-Q\right]=0, \\ K\mathbf{H_c}\mathbf{a}_k[n-Q]-K\mathbf{b}_k, & \text{otherwise,} \end{cases}\end{aligned} \tag{21}$$

where

$$\mathbf{a}_k[n] = \begin{bmatrix} \bar{a}_{k,0}[n] \\ \bar{a}_{k,1}[n+1] \\ \vdots \\ \bar{a}_{k,N-1}[n+N-1] \end{bmatrix}, \quad \mathbf{b}_k = \begin{bmatrix} b_k[Q-P] \\ b_k[Q-P+1] \\ \vdots \\ b_k[Q-P+N-1] \end{bmatrix}, \tag{22}$$

and

$$\mathbf{H_c} = \begin{bmatrix} h_c[Q] & h_c[Q-1] & \cdots & h_c[Q-N+1] \\ h_c[Q+1] & h_c[Q] & \cdots & h_c[Q-N+2] \\ \vdots & \vdots & \ddots & \vdots \\ h_c[Q+N-1] & h_c[Q+N-2] & \cdots & h_c[Q] \end{bmatrix}. \tag{23}$$

This is an *N*-dimensional, *Q*th-order, time-varying matrix difference equation. It converges if and only if $\mathbf{a}_k[n]\rightarrow\mathbf{a}_k'$ as $n\rightarrow\infty$ where $\mathbf{a}_k'$ is the constant steady-state solution of (21). Furthermore, if the system converges it follows from taking the limit of (21) as $n\rightarrow\infty$ that

$$\mathbf{H_c}\mathbf{a}_k' = \mathbf{b}_k. \tag{24}$$

Defining $\mathbf{z}_k[n] = \mathbf{a}_k[n] - \mathbf{a}_k'$, (21) and (24) imply that

$$\mathbf{z}_k[n] = \begin{cases} \mathbf{z}_k[n-1], & \text{if } s_k\left[n+P-Q\right]=0, \\ \mathbf{z}_k[n-1]-K\mathbf{H_c}\mathbf{z}_k[n-Q], & \text{otherwise,} \end{cases} \tag{25}$$

and the system converges if and only if $\mathbf{z}_k[n] \rightarrow \mathbf{0}$ as $n\rightarrow\infty$.

If $s_k[n]$ were never zero, then (25) would be a time-invariant as well as linear matrix difference equation. In this case (25) could be rewritten as a *QN*-dimensional, first-order matrix equation and shown to converge provided the eigenvalues of its system matrix all have magnitude less than one. Unfortunately, $s_k[n] = 0$ for some values of *n* as described above, which complicates the analysis. A new analysis is presented in the remainder of the section that addresses this problem. The analysis shows that the system parameters can be chosen such that $\mathbf{z}_k[n] \rightarrow \mathbf{0}$ as $n\rightarrow\infty$ and provides a measure of the convergence rate.

The analysis makes use of the following standard matrix theory definitions and results [19]. For any N-dimensional vector $\mathbf{v} = [v_j]$ and $N\times N$ matrix $\mathbf{A} = [a_{j,k}]$, the *max norm* of $\mathbf{v}$ and the maximum absolute row sum norm of $\mathbf{A}$ are defined as

$$\|\mathbf{v}\| = \max_{1\le m\le N} |v_m| \quad \text{and} \quad \|\mathbf{A}\|_1 = \max_{1\le m\le N} \sum_{n=1}^{N} |a_{m,n}|, \tag{26}$$

respectively, and these definitions imply that

$$\|\mathbf{A}\mathbf{v}\| \le \|\mathbf{A}\|_1 \|\mathbf{v}\|. \tag{27}$$

For any two vectors $\mathbf{v}$ and $\mathbf{w}$ of equal dimension

$$\|\mathbf{v}\| - \|\mathbf{w}\| \le \|\mathbf{v}+\mathbf{w}\| \le \|\mathbf{v}\| + \|\mathbf{w}\|. \tag{28}$$

The following system-related definitions are used by the theorems presented below:

$$r = \frac{1}{h_c[Q]} \sum_{m\ne Q} |h_c[m]|, \tag{29}$$

and

$$g = \frac{\left\|\mathbf{H}_\mathbf{c}^{\,2}\right\|_1 \left[1-\left(1-2h_c[Q]K\right)^{Q-1}\right]}{2h_c^{\,2}[Q](1-r)\left(1-2h_c[Q]K\right)^{2Q-2}}. \tag{30}$$

The following theorem shows that $\mathbf{z}_k[n] \rightarrow \mathbf{0}$ as $n\rightarrow\infty$ for the case where the system is started at time $n = 0$ with all registers initialized to zero. It does so by showing that $\|\mathbf{z}_k[n]\| \rightarrow 0$ as $n\rightarrow\infty$. From the definition of $\mathbf{z}_k[n]$, this initial condition implies that $\mathbf{z}_k[n] = -\mathbf{a}_k'$, for all $n < 0$

Theorem 1: If $0 \le r < 1$, $0 < g < 1$, and $\mathbf{z}_k[n] = -\mathbf{a}_k'$ for all $-Q \le n < 0$, then

$$\|\mathbf{z}_k[J_m]\| \le \|\mathbf{a}_k{}'\| \left(1-K(1-r)(1-g)h_c[Q]\right)^m \tag{31}$$

for all $m \ge 1$, where J_m is the mth largest non-negative integer n for which $s_k[n+P-Q] \ne 0$.

□

As implied by (25), $\mathbf{z}_k[n] = \mathbf{z}_k[n-1]$ when $n \neq J_m$ for any $m = 1, 2, \ldots$, so the theorem implies that $\mathbf{z}_k[n] \rightarrow \mathbf{0}$ at least exponentially with the number of times that $s_k[n+P-Q] \neq 0$ over n provided the theorem's hypothesis is satisfied.

As explained below, the conditions placed on $h_c[n]$ and K by the theorem's hypothesis are easy to meet in a practical design, and the dependence of the convergence on how frequently $s_k[n+P-Q]$ is non-zero does not present a problem in practice.

The theorem also gives insight into the choice of Q. The requirement that $0 \leq r < 1$ implies that $h_c[Q]$ must be positive and that it must be the maximum value of the impulse response.

Proof of Theorem 1:

If $\mathbf{a}_k' = 0$ then (25) implies that $\mathbf{z}_k[n] = \mathbf{0}$ for all $n \geq 0$, so Theorem 1 holds for this case. The remainder of the proof considers the case of $\mathbf{a}_k' \neq 0$.

The proof uses mathematical induction. The *inductive step*, which is proven shortly, is: for any $m = 1, 2, 3, \ldots$, if

$$\frac{\|\mathbf{z}_k[i]\|}{\|\mathbf{z}_k[i-1]\|} \geq 1 - 2h_c[Q]K, \tag{32}$$

for all $-Q+1 \leq i < J_m$,[3] then the conditions of the theorem's hypothesis are sufficient to ensure that (32) holds for $i = J_m$ and

$$\frac{\|\mathbf{z}_k[J_m]\|}{\|\mathbf{z}_k[J_m-1]\|} \leq 1 - K(1-r)(1-g)h_c[Q]. \tag{33}$$

The induction *base step*, i.e., that (32) holds for $-Q+1 \leq i < J_1$, follows directly from (25), the max norm definition in (26), and the condition that $\mathbf{z}_k[n] = -\mathbf{a}_k'$ for all $-Q \leq n < 0$. Therefore, given that $\mathbf{z}_k[n] = \mathbf{z}_k[n-1]$ when $n \geq 0$ and $n \neq J_m$ for any $m = 1, 2, \ldots$, provided

[3] By limiting the amount that $\|\mathbf{z}_k[i]\|$ can decrease over each iteration, (32) prevents the possibility of convergence with ringing, which is necessary for (31) to hold.

the inductive step is true, it follows from induction that (32) and (33) hold for all integers $m \geq 1$. Furthermore, recursively applying (33) when $n = J_m$ and $\mathbf{z}_k[n] = \mathbf{z}_k[n-1]$ when $n \neq J_m$ with $\mathbf{z}_k[J_1-1] = -\mathbf{a}_k'$ yields (31).

Hence, it remains to show that the inductive step is true. This is done in the remainder of the proof.

For any $m = 1, 2, 3, \ldots$, let $n = J_m$ (to simplify the notation). Then (25) reduces to

$$\mathbf{z}_k[n] = \mathbf{z}_k[n-1] - K\mathbf{H_c}\mathbf{z}_k[n-Q] \tag{34}$$

which can be rewritten as

$$\begin{aligned}\mathbf{z}_k[n] = \mathbf{z}_k[n-1] - K\mathbf{H_c}\mathbf{z}_k[n-1] \\ -K\mathbf{H_c}\left(\mathbf{z}_k[n-Q] - \mathbf{z}_k[n-1]\right)\end{aligned} \tag{35}$$

and further rewritten as

$$\begin{aligned}\mathbf{z}_k[n] = \left(\mathbf{I} - K\mathbf{H_c}\right)\mathbf{z}_k[n-1] \\ -\sum_{m=1}^{Q-1} K\mathbf{H_c}\left(\mathbf{z}_k[n-m-1] - \mathbf{z}_k[n-m]\right).\end{aligned} \tag{36}$$

where $\mathbf{I}$ is the $N\times N$ identity matrix. Taking the L_1 norm of (36) and applying (28) multiple times yields

$$\begin{aligned}\left\|\mathbf{z}_k[n]\right\| \leq \left\|\left(\mathbf{I} - K\mathbf{H_c}\right)\mathbf{z}_k[n-1]\right\| \\ +\sum_{m=1}^{Q-1}\left\|K\mathbf{H_c}\left(\mathbf{z}_k[n-m-1] - \mathbf{z}_k[n-m]\right)\right\|\end{aligned} \tag{37}$$

and

$$\begin{aligned}\left\|\mathbf{z}_k[n]\right\| \geq \left\|\left(\mathbf{I} - K\mathbf{H_c}\right)\mathbf{z}_k[n-1]\right\| \\ -\sum_{m=1}^{Q-1}\left\|K\mathbf{H_c}\left(\mathbf{z}_k[n-m-1] - \mathbf{z}_k[n-m]\right)\right\|.\end{aligned} \tag{38}$$

Let $\mathbf{v}$ be any real N-dimensional column vector. Then

$$\left\|\left(\mathbf{I} - K\mathbf{H_c}\right)\mathbf{v}\right\| = \left\|\left(1 - h_c[Q]K\right)\mathbf{v} + K\left(h_c[Q]\mathbf{I} - \mathbf{H_c}\right)\mathbf{v}\right\|. \tag{39}$$

Applying (27) with $\mathbf{A} = h_c[Q]\mathbf{I} - \mathbf{H_c}$ and (28) to (39) gives

$$\left\|\left(\mathbf{I}-K\mathbf{H_c}\right)\mathbf{v}\right\| \le \left(1-h_c\left[Q\right]K\right)\left\|\mathbf{v}\right\| + K\left\|h_c\left[Q\right]\mathbf{I}-\mathbf{H_c}\right\|_1\left\|\mathbf{v}\right\| \tag{40}$$

and

$$\left\|\left(\mathbf{I}-K\mathbf{H_c}\right)\mathbf{v}\right\| \ge \left(1-h_c\left[Q\right]K\right)\left\|\mathbf{v}\right\| - K\left\|h_c\left[Q\right]\mathbf{I}-\mathbf{H_c}\right\|_1\left\|\mathbf{v}\right\|. \tag{41}$$

The requirement that $0 \le r < 1$ and (29) imply that $h_c[Q]$ is positive. This, (23), (26), and (29) imply that $\|h_c[Q]\mathbf{I} - \mathbf{H_c}\|_1 \le h_c[Q]r$, so (40) and (41) imply

$$\left\|\left(\mathbf{I}-K\mathbf{H_c}\right)\mathbf{v}\right\| \le \left(1-h_c\left[Q\right]K\left(1-r\right)\right)\left\|\mathbf{v}\right\| \tag{42}$$

and

$$\left\|\left(\mathbf{I}-K\mathbf{H_c}\right)\mathbf{v}\right\| \ge \left(1-h_c\left[Q\right]K\left(1+r\right)\right)\left\|\mathbf{v}\right\|. \tag{43}$$

Substituting $\mathbf{v} = \mathbf{z}_k[n-1]$ into (42) and (43), and the results into (37) and (38) yields

$$\begin{aligned}\left\|\mathbf{z}_k[n]\right\| \le &\left(1-h_c\left[Q\right]K\left(1-r\right)\right)\left\|\mathbf{z}_k[n-1]\right\| \\ &+\sum_{m=1}^{Q-1}\left\|K\mathbf{H_c}\left(\mathbf{z}_k[n-m-1]-\mathbf{z}_k[n-m]\right)\right\|\end{aligned} \tag{44}$$

and

$$\begin{aligned}\left\|\mathbf{z}_k[n]\right\| \ge &\left(1-h_c\left[Q\right]K\left(1+r\right)\right)\left\|\mathbf{z}_k[n-1]\right\| \\ &-\sum_{m=1}^{Q-1}\left\|K\mathbf{H_c}\left(\mathbf{z}_k[n-m-1]-\mathbf{z}_k[n-m]\right)\right\|\end{aligned} \tag{45}$$

Equation (25) for $n \ge 0$ and the condition $\mathbf{z}_k[n] = -\mathbf{a}_k'$ for $-Q \le n < 0$ imply that each $\mathbf{z}_k[n-m-1] - \mathbf{z}_k[n-m]$ in (44) and (45) is either $K\mathbf{H_c}\mathbf{z}_k[n-m-Q]$ or $\mathbf{0}$. Consequently,

$$\begin{aligned}\left\|\mathbf{z}_k[n]\right\| \le &\left(1-h_c\left[Q\right]K\left(1-r\right)\right)\left\|\mathbf{z}_k[n-1]\right\| \\ &+\sum_{m=1}^{\min\{Q-1,n\}}\left\|K^2\mathbf{H_c}^2\mathbf{z}_k[n-m-Q]\right\|\end{aligned} \tag{46}$$

and

$$\begin{aligned}\left\|\mathbf{z}_k[n]\right\| \ge &\left(1-h_c\left[Q\right]K\left(1+r\right)\right)\left\|\mathbf{z}_k[n-1]\right\| \\ &-\sum_{m=1}^{\min\{Q-1,n\}}\left\|K^2\mathbf{H_c}^2\mathbf{z}_k[n-m-Q]\right\|.\end{aligned} \tag{47}$$

Applying (27) with $\mathbf{A} = K^2\mathbf{H_c}^2$ to (46) and (47) yields

$$\begin{aligned}\left\|\mathbf{z}_k[n]\right\| &\le \left(1-h_c\left[Q\right]K\left(1-r\right)\right)\left\|\mathbf{z}_k[n-1]\right\| \\ &\quad +K^2\left\|\mathbf{H}_{\mathbf{c}}{}^2\right\|_1\sum_{m=1}^{\min\{Q-1,n\}}\left\|\mathbf{z}_k[n-m-Q]\right\|\end{aligned}\tag{48}$$

and

$$\begin{aligned}\left\|\mathbf{z}_k[n]\right\| &\ge \left(1-h_c\left[Q\right]K\left(1+r\right)\right)\left\|\mathbf{z}_k[n-1]\right\| \\ &\quad -K^2\left\|\mathbf{H}_{\mathbf{c}}{}^2\right\|_1\sum_{m=1}^{\min\{Q-1,n\}}\left\|\mathbf{z}_k[n-m-Q]\right\|.\end{aligned}\tag{49}$$

Recursively applying (32) to itself for i = 2, 3, 4, …n+Q, yields

$$\left\|\mathbf{z}_k[n-i]\right\| \le \left\|\mathbf{z}_k[n-1]\right\|\left(1-2h_c\left[Q\right]K\right)^{-i+1}.\tag{50}$$

Hence,

$$\sum_{m=1}^{\min\{Q-1,n\}}\left\|\mathbf{z}_k[n-m-Q]\right\| \le \left\|\mathbf{z}_k[n-1]\right\|\sum_{m=1}^{Q-1}\left(1-2h_c\left[Q\right]K\right)^{-m-Q+1}.\tag{51}$$

The right side of (51) can be expanded via the geometric series formula as

$$\frac{1-\left(1-2h_c\left[Q\right]K\right)^{Q-1}}{2h_c\left[Q\right]K\left(1-2h_c\left[Q\right]K\right)^{2Q-2}}\ .\tag{52}$$

Substituting (52) into (51) and the result into (48) and (49) yields

$$\begin{aligned}\frac{\left\|\mathbf{z}_k[n]\right\|}{\left\|\mathbf{z}_k[n-1]\right\|} &\le 1-h_c\left[Q\right]K\left(1-r\right) \\ &\quad +K\left\|\mathbf{H}_{\mathbf{c}}{}^2\right\|_1\frac{1-\left(1-2h_c\left[Q\right]K\right)^{Q-1}}{2h_c\left[Q\right]\left(1-2h_c\left[Q\right]K\right)^{2Q-2}}\end{aligned}\tag{53}$$

and

$$\begin{aligned}\frac{\left\|\mathbf{z}_k[n]\right\|}{\left\|\mathbf{z}_k[n-1]\right\|} &\ge 1-h_c\left[Q\right]K\left(1+r\right) \\ &\quad -K\left\|\mathbf{H}_{\mathbf{c}}{}^2\right\|_1\frac{1-\left(1-2h_c\left[Q\right]K\right)^{Q-1}}{2h_c\left[Q\right]\left(1-2h_c\left[Q\right]K\right)^{2Q-2}}\end{aligned}\tag{54}$$

Given that n = J_m, substituting (30) into (53) results in (33) and substituting (30) into (54) results in

$$\frac{\|\mathbf{z}_k[J_m]\|}{\|\mathbf{z}_k[J_m-1]\|} \geq 1 - h_c[Q]K(1+r) - h_c[Q]K(1-r)g\ . \tag{55}$$

This finishes the proof because (55) implies that (32) with $i = n$ is satisfied provided $g < 1$.

□

Theorem 2 extends the result of Theorem 1 to cover all possible initial conditions. It shows that while the specific form of $\mathbf{z}_k[n]$ depends on the system's initial conditions, the convergence of $\|\mathbf{z}_k[n]\|$ is still exponential for any K and $h_c[n]$ that satisfy the hypothesis of Theorem 1 regardless of the initial conditions.

Theorem 2: Provided $0 \leq r < 1$ and $0 < g < 1$, $\mathbf{z}_k[n]$ can be written as

$$\mathbf{z}_k[n] = \sum_{j=1}^{Q} \mathbf{z}_{k,j}[n], \tag{56}$$

where for every $J_m \geq Q - j$,

$$\|\mathbf{z}_{k,j}[J_m]\| \leq \left(1 - K(1-r)(1-g)h_c[Q]\right)\|\mathbf{z}_{k,j}[J_m-1]\|, \tag{57}$$

and for all non-negative $n \neq J_m$

$$\|\mathbf{z}_{k,j}[n]\| = \|\mathbf{z}_{k,j}[n-1]\|. \tag{58}$$

□

Proof:

Let $\mathbf{z}_{k,j}[-1]$, $\mathbf{z}_{k,j}[-2]$, …, $\mathbf{z}_{k,j}[-Q]$ for $j = 1, 2, …, Q$, be

$$\mathbf{z}_{k,j}[n] = \begin{cases} \mathbf{z}_k[-j] - \mathbf{z}_k[-j-1], & \text{if } j < Q, -j \leq n < 0, \\ \mathbf{0}, & \text{if } j < Q, -Q \leq n < -j \\ \mathbf{z}_k[-Q], & \text{if } j = Q, -Q \leq n < 0, \end{cases} \tag{59}$$

It can be verified by substituting (59) into (56) that (59) is a solution of (56) for $-Q \leq n < 0$. It follows from (25) that $\mathbf{z}_k[n]$ for all $n \geq 0$ is uniquely determined by (25) and the values of $\mathbf{z}_k[n]$ for $-Q \leq n < 0$. Consequently, $\mathbf{z}_k[n]$ for all $n \geq 0$ is uniquely determined by (25), (56), and (59).

Equation (25) is a linear matrix difference equation, so, as can be seen by substituting (56) into (25), $\mathbf{z}_{k,j}[n]$ for $n \geq 0$ can be defined as

$$\mathbf{z}_{k,j}[n] = \begin{cases} \mathbf{z}_{k,j}[n-1], & \text{if } s_k\left[n+P-Q\right] = 0, \\ \mathbf{z}_{k,j}[n-1] - K\mathbf{H_c}\mathbf{z}_{k,j}[n-Q], & \text{otherwise.} \end{cases} \tag{60}$$

This with (59) completely specifies $\mathbf{z}_{k,j}[n]$ for $n \geq -Q$.

It follows from (60) and the definition of J_m that (58) holds for all non-negative $n \neq J_m$, so it remains to show that (57) holds for all $J_m \geq Q - j$. This is done below by induction.

For all $n \geq 0$, (60) implies that $\mathbf{z}_{k,j}[n] = 0$ if $\mathbf{z}_{k,j}[-1]$, $\mathbf{z}_{k,j}[-2]$, …, $\mathbf{z}_{k,j}[-Q]$ are all zero. In this case, (57) holds for all $J_m \geq Q - j$, and (58) holds for all non-negative $n \neq J_m$. All other cases are considered in the remainder of the proof.

As can be seen from (59), the first j values of $\mathbf{z}_{k,j}[-1]$, $\mathbf{z}_{k,j}[-2]$, …, $\mathbf{z}_{k,j}[-Q]$ are non-zero and equal, and the remaining $Q–j$ values are $\mathbf{0}$. This and (60) imply that all Q values of $\mathbf{z}_{k,j}[n+Q–j]$ for $n = -Q, -Q+1, \ldots, -2, -1$ are non-zero and equal. Therefore, by exactly the same reasoning used for the induction base step in the proof of Theorem 1,

$$\frac{\left\|\mathbf{z}_{k,j}[i]\right\|}{\left\|\mathbf{z}_{k,j}[i-1]\right\|} \geq 1 - 2h_c\left[Q\right]K, \tag{61}$$

for all $-j+1 \leq i < J_p$. where p is the smallest integer for which $J_p \geq Q - j$. This is the induction *base step*.

By exactly the same reasoning used in the proof of Theorem 1, the following inductive step holds for each $\mathbf{z}_{k,j}[n]$: for any $m = p, p+1, p+2, p+3, \ldots$, if (61) holds for all $-j+1 \leq i < J_m$ then the conditions of the theorem's hypothesis are sufficient to ensure that (61) holds for $i = J_m$ and (57) holds.

It follows from induction that (57) holds for all $J_m \geq Q - j$.

□

Theorems 1 and 2 provide conditions for which the convergence of $\|\mathbf{z}_k[n]\|$ is bounded from above by a decaying exponential sequence. The following corollary shows that these same conditions ensure that the convergence of $\|\mathbf{z}_k[n]\|$ is also bounded from below by a decaying exponential sequence.

Corollary: Provided $0 \le r < 1$ and $0 < g < 1$,

$$\left\|\mathbf{z}_{k,j}[J_m]\right\| \ge \left(1-K\left(2-(1-r)(1-g)\right)h_c[Q]\right)\left\|\mathbf{z}_{k,j}[J_m-1]\right\|, \tag{62}$$

for every $J_m \ge Q - j$.

Proof: The proof follows directly from that of Theorem 2.

E. Noise Versus Convergence Rate Tradeoff

As described in Section III-C, the MNC technique causes the impulse responses of the adaptive filters shown in Fig. 6, i.e., $h_k[m] = a_{k,m}[n]$ for $m = 0, 1, \ldots, N-1$ and $k = 1, 2, \ldots, 35$, to converge toward their ideal values as $n \to \infty$. As shown in Section III-D, the $a_{k,m}[n]$ coefficients are well-modelled as random variables with means that converge to their ideal values as $n \to \infty$. Thus, once the convergence transient has died out, each $a_{k,m}[n]$ is equal to its ideal value plus zero-mean noise.

As with most adaptive filter analyses, the analysis of Section III-D does not provide insight into the variance of the noise component in each $a_{k,m}[n]$ sequence. It does not even rule out the possibility that the variance could diverge as $n \to \infty$, which, of course, would be catastrophic for the MNC technique. Fortunately, intuitive reasoning and extensive simulations run by the authors, some of which are presented in Section IV, indicate that the variance of the noise can be made arbitrarily small by reducing the feedback loop gain, K. Specifically, it is reasonable to expect from (17) that reducing K reduces the sample-to-sample variability, and therefore the variance, of the noise component of $a_{k,m}[n]$. Simulation results presented in the next section bear this out. This and the results of Section III-D imply the usual tradeoff between convergence rate and accuracy in adaptive systems. Reducing K reduces the convergence error variance, but it also reduces the convergence rate.

At first glance it might also appear that there is a tradeoff between the convergence rate and how frequently non-zero values of each $s_k[n]$ sequence occur. As described in [16], the values of n for which $s_k[n] = 0$ are partly dependent on the DEM DAC's input sequence, so it follows from the results of Section III-D that the convergence rate of $a_{k,m}[n]$ has a dependency on the DEM DAC's input sequence. For example, if the input sequence were such that $s_k[n] = 0$ for all n, then $a_{k,m}[n]$ would remain constant. On the other hand, it can be seen from (8) that the error term in $e_{DAC}[n]$ corresponding to $s_k[n]$ would be zero for this case, so the lack of convergence would not be a problem. More generally, the less frequently non-zero values of each $s_k[n]$ sequence occur, the slower the convergence rate with n but the lower the noise introduced by the corresponding term in $e_{DAC}[n]$. These two effects tend to cancel each other out in practice.

It can be seen from Figures 4 and 5 that a change in the ADC gain is mathematically equivalent to a change in K. Therefore, any variations in the ADC gain simply change the tradeoff between the convergence error variance and the convergence rate. This suggests that the system is not highly sensitive to ADC gain variations such as might be caused by temperature variations during background calibration. Indeed, simulation results performed by the authors during which the ADC gain was varied by up to 50% during background calibration showed negligible effect on MNC accuracy.

F. Clock Jitter and Feedback Path Noise and Nonlinearity

It follows from the analysis in Section III-D that the multiplication of the decimation filter output, $r[n]$, by $s_k[n+P-Q-m]$ in Fig. 5 causes $\overline{u}_{k,m}[n]$ to be the signal of interest and $u_{k,m}[n] - \overline{u}_{k,m}[n]$ to be noise from the perspective of estimating $a_{k,m}[n]$. It can be verified by subtracting (19) from (16) that the signal to noise ratio associated with each $a_{k,m}[n]$ estimation is low even in the absence of any noise from the ADC. This is because $r[n] = r_{ideal}[n] + r_e[n] + r_c[n]$, where $r_{ideal}[n]$ and all but small portions of $r_e[n]$ and $r_c[n]$ contribute only noise terms to $u_{k,m}[n]- \overline{u}_{k,m}[n]$ given that they are uncorrelated with $s_k[n+P-Q-m]$. For example, error introduced anywhere in the system by clock jitter is generally uncorrelated with $s_k[n+P-Q-m]$, so it is simply another noise term in $u_{k,m}[n] - \overline{u}_{k,m}[n]$, and it only needs to be on the order of 6 dB lower than the variance of the other terms in $u_{k,m}[n] - \overline{u}_{k,m}[n]$ to have a negligible effect on the error variance of $a_{k,m}[n]$.

The same is true of ADC noise provided it is uncorrelated with $s_k[n+P-Q-m]$. Consequently, an ADC with a low SNR can be tolerated as demonstrated in the next section. In the design example a VCO-based ADC is used because the noise it introduces is essentially uncorrelated with its input signal, which ensures that it is uncorrelated with $s_k[n+P-Q-m]$. Most other types of $\Delta\Sigma$ ADCs have this property too, so they could be used in place of the VCO-based ADC, although in most such cases a high-impedance input buffer would be necessary to prevent the ADC's input network from disturbing the main DAC's output waveform.

It is also demonstrated in the next section that an ADC with relatively high nonlinearity can be tolerated by the MNC technique. The reasons for this nonlinearity tolerance are explained in the remainder of this section.

As described in Section III, the additive terms in the ADC's input signal which are proportional to $s_k[n]$ for $k = 1, 2, \ldots, 35$ are the terms that the MNC technique measures. In this sense they can be viewed as the *desired terms* from the perspective of the MNC technique's measurement process. Each desired term consists of two additive parts: one that comes from $e_{DAC}[n]$ so it has the form $s_k[n]d_k(t)$, and the other that comes from the correction DAC. The first part is very small relative to the ADC's input range because $d_k(t)$ arises from component mismatches. The second part is similarly small by design because it is intended to cancel the first part over the first Nyquist band.

It follows that nonlinear distortion from the ADC causes the decimation filter output to contain numerous additive terms that are each proportional to the products of multiple values of $(s_i[j])^p$ for different integer values of i, j, and p. From the perspective of estimating $a_{k,m}[n]$, most of these terms contribute noise to $u_{k,m}[n] - \overline{u}_{k,m}[n]$ because they get multiplied by $s_k[n+P-Q-m]$. Only the terms from nonlinear distortion that are proportional to $(s_k[n+P-Q-m])^p$ where p is 1, 3, 5, 7, …, and not also proportional to $s_i[j]$ for any $i \neq k$ or $j \neq n+P-Q-m$ contribute an error bias to the estimate of $a_{k,m}[n]$. Not only are there relatively few such terms, but the terms are much smaller than the corresponding desired terms even when the ADC is fairly nonlinear. Each such error term is proportional to one of the ADC's second-or-higher-order Taylor coefficients, which is much less than unity, as well as one of the desired terms raised to the pth power. For $p = 3, 5, 7, \ldots$, the terms are particularly small because the desired terms are small to begin with.

Furthermore, the estimate of $e_{DAC}[n]$ need not be highly accurate to significantly improve the system's overall SNR. For example, suppose that $e_{DAC}[n]$ degrades the main DEM DAC's peak SNR in the absence of the MNC technique by more than 6 dB. Then, even if the MNC technique were applied for a case where the error terms described above are so severe that they cause the estimate of $e_{DAC}[n]$ to deviate from the actual $e_{DAC}[n]$ by 50%, the MNC technique would still improve the overall SNR by as much as 6 dB.

IV. SIMULATION RESULTS

The system shown in Fig. 4 and described above was simulated in the Cadence Virtuoso environment with the STMicroelectronics FDSOI 28 nm CMOS process design kit. Relevant additional design details and two sets of simulation results are presented in this Section. The first set of simulation results demonstrates the performance of the MNC technique after convergence. The second set demonstrates the convergence behavior of the MNC technique.

Both the main and correction DACs incorporate RZ 1-bit DACs similar to the type described in [20] with an RZ duration of 25% of the clock period. All operate from a 1.8 V supply and their combined differential outputs are loaded with a 15 Ω resistor and 14 pF capacitor to ground on each side. The main DAC has a differential minimum step-size of $\Delta = 2.44$ μA. It has 36 1-bit DACs, 16 of which have a weight of 1024, and 20 of which have respective weights of 1, 1, 2, 2, 4, 4, 8, 8, …, 512, 512. The correction DAC has a differential minimum step-size of $\Delta_c = \Delta/4 = 0.61$ μA. It has 14 1-bit DACs, 7 of which have a weight of 64, 3 of which have a weight of 16, and 4 of which have respective weights of 1, 2, 4, 8.

The static mismatch of each of the smallest 1-bit DACs in the main DAC was chosen as a Gaussian random variable with a standard deviation of 3.2% of the 1-bit DAC's step-size, Δ. That of each larger 1-bit DAC in the main DAC was chosen the same way except with a standard deviation of 3.2% divided by the square root of the 1-bit DAC's weight, e.g., the standard deviation of the largest 1-bit DACs is 0.1% of their 1024Δ step-size. The static mismatches in the correction DAC were chosen in the same fashion except starting from minimum-size 1-bit DACs with a standard deviation of 6.4% of their step-size, Δ_c.

The dynamic mismatches of the 1-bit DACs were implemented in two ways. A random Gaussian time skew with a standard deviation of 1.8 ps was applied to each 1 bit DAC switch driver. Additionally, for the 1-bit DAC of lower weights, the sizes of their current steering switches were not scaled in proportion due to minimal width limitation of technology, which introduces systematic dynamic mismatches.

The VCO-based ADC is similar to that presented in [17] except without the digital calibration circuitry. As in [17], each VCO consists of an open-loop voltage-to-current (V/I) converter followed by a current-controlled ring oscillator (ICRO). The V/I converter is a source degenerated differential pair, and the ICRO is a pseudo-differential ring of current-starved inverters. Accordingly, the VCO, and, thus, the VCO-based ADC, are highly nonlinear. For example, simulations indicate that for a full-scale sinusoidal input signal the ADC's 2nd, 3rd, and 4th harmonics are −26 dBc, −47 dBc, and −64 dBc, respectively. As demonstrated below, and for the reasons described in Section III-F, this nonlinearity does not limit the simulated system's performance.

The decimation filter is implemented as a 33-tap polyphase FIR filter for low hardware complexity [21]. As described in Section III, $h_c[n]$ is defined as the impulse response from the input of the correction DAC to the output of the decimation filter. The values of $h_c[n]$ were extracted from circuit simulation of the correction DAC, ADC, and decimation filter operating together, and the gain of the decimation filter was normalized such that $h_c[Q] = 1$. The extracted values were found to depend only weakly on the behavior of the correction DAC and ADC so they do not change significantly over process and temperature variations. Substituting the extracted values of $h_c[n]$ into (23), (29), and (30) results in g = 0.0018 and r = 0.25, which easily satisfy the hypotheses of the theorems in Section III-D.

In the first set of simulations (shown in Figures 7, 8, and 9) all the 1-bit DAC current sources and switches and the ADC's V/I converters were simulated at the transistor level. The remaining analog circuitry, e.g., the 1-bit DAC switch drivers and the ADC's ICROs, as well as all the digital logic was simulated at the behavioral level using Verilog-AMS to reduce simulation time. The transistor-level portions of the simulations enhance realism, but significantly increase simulation time, so the simulations were run with the MNC technique implemented in foreground mode to minimize convergence time and, therefore, simulation time.

The DEM DAC was driven by a digital sequence that toggles back and fourth between −2389.5Δ and −2388.5Δ at the clock rate. This input sequence was chosen because it is both simple and ensures that each $s_k[n]$ sequence is non-zero at least 30% of the time. Two minor enhancements were applied to reduce convergence time. The first enhancement is the use of a few extra 1-bit DACs to cancel most of the signal component of the main DACs output prior to the ADC. This allowed the loop gain, K, to be increased without a significant noise penalty. The extra 1-bit DACs were simulated at the transistor-level with mismatches chosen as described above. The second enhancement is to use a 4× larger value of K for the first 100 μs of convergence time than used for the remaining convergence time. With these enhancements the total convergence time was 250 μs, which corresponds to approximately three weeks of simulation time.

Representative simulated output spectra are shown in Fig. 7 for a −1 dB full-scale sinusoidal input without and with the MNC technique enabled. In each case the 14-bit input signal was generated by adding a dither sequence that is white and uniformly distributed between $-\Delta/2$ and $\Delta/2$ to a floating point sinusoidal signal and quantizing the result to 14 bits. Output spectra of the main DAC for the ideal case of no unit element mismatches are also shown in Fig. 7 to provide a comparison baseline. The decimation filter's relatively short length resulted in aliasing that limits MNC performance in the top 16% of the first Nyquist band.[4] This was considered a reasonable design tradeoff, so the signal band is taken to range from zero to $0.42\,f_s = 168$ MHz.

The simulation results indicate that the MNC technique increased the signal-to-noise-and-distortion ratio (SNDR) from 66.4 dB (10.8 bits) to 81.9 dB (13.4 bits). Separate simulations suggest that the static mismatch error and dynamic mismatch error for this case contribute roughly equal SNR degradation over the first Nyquist band.

Additional simulated output spectra are shown in Figures 8 and 9 for different input signal amplitudes without and with the MNC technique enabled. In each case the results show the expected SNDR improvement when the MNC technique is enabled. Other simulations that have been run by the authors for many different input signals and random number seeds yield comparable results.

The second set of simulations model the system with the same parameters and non-ideal behavior described above except that K was set to its final value from the start, and all components were simulated at the behavioral level to avoid excessive simulation time. The left plot in Fig. 10 shows the convergence of the elements of $[a_{k,0}[n],\ a_{k,1}[n+1],\ \ldots,\ a_{k,N-1}[n+N-1]]^{\mathrm{T}} - \mathbf{a}_k'$, for a representative value of k and the artificial case of no ADC quantization noise. It also shows the upper and lower bounds of the means of these trajectories predicted

[4] This percentage can be arbitrarily reduced at the costs of greater hardware complexity and power consumption by increasing the decimation filter length.

by Theorem 1, i.e., ±$\|\mathbf{z}_k[n]\|$. As expected, all coefficients converge to their ideal values within the bounds predicted by Theorem 1. The right plot in Fig. 10 shows the corresponding results with ADC quantization noise included. The results suggest that the means of the trajectories are still within the predicted bounds even though the noise causes the instantaneous values to exceed the bounds from time to time. These results as well as those from all of many other such simulations run by the authors are in agreement with the theoretical results of Section III-D.

ACKNOWLEDGEMENTS

Chapter 1, in full, has been published in the IEEE Transactions on Circuits and Systems I: Regular Papers, volume 65, number 2, pages 421-433, February 2018. D. Kong, I. Galton, 2018. The dissertation author is the primary investigator and author of this paper. Professor Ian Galton supervised the research which forms the basis for this paper.

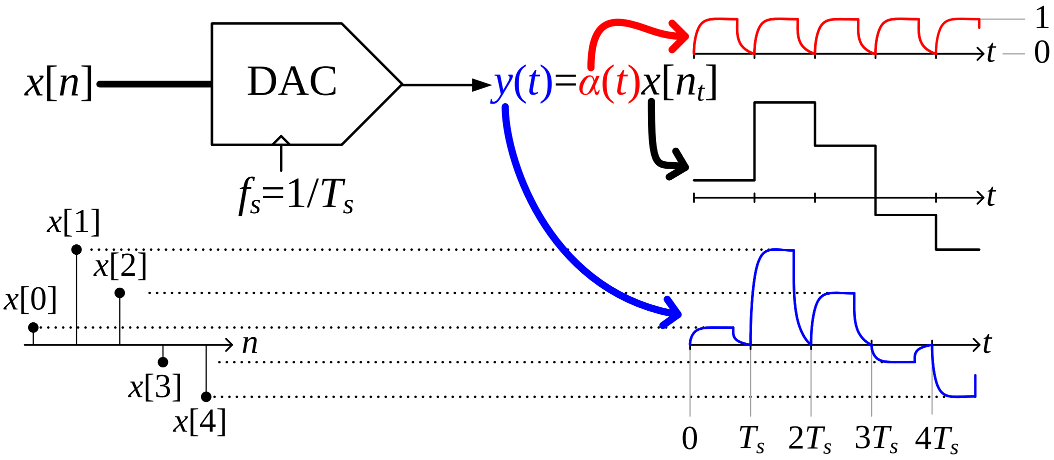


Figure 1 : Desired DAC behavior.

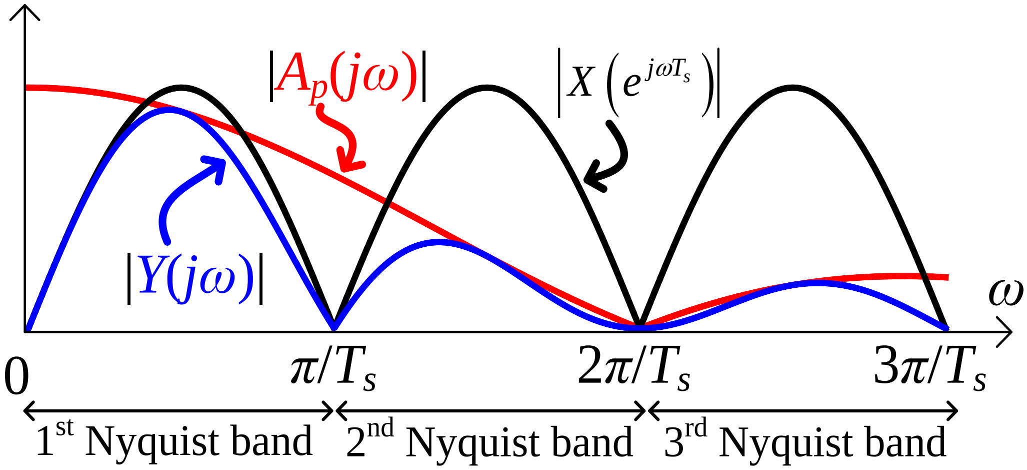


Figure 2 : Example DAC spectra.

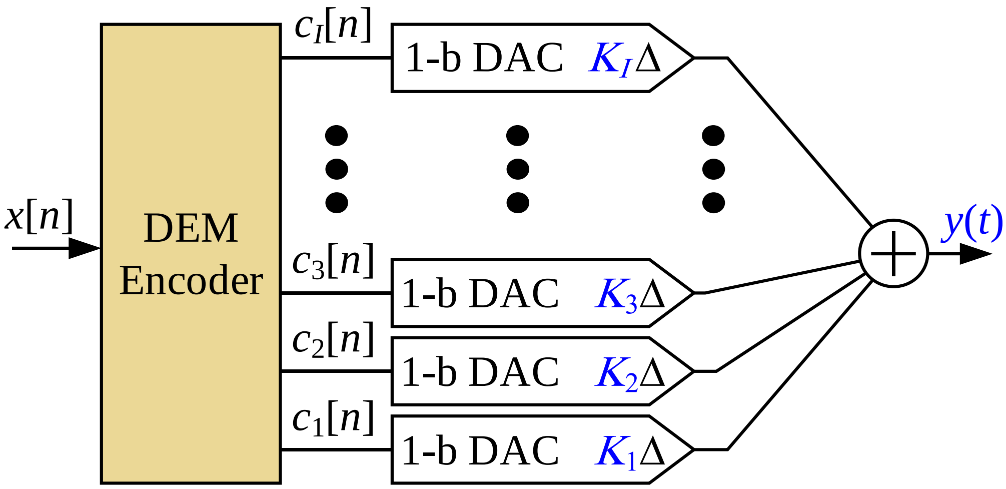


Figure 3: General form of a DEM DAC.

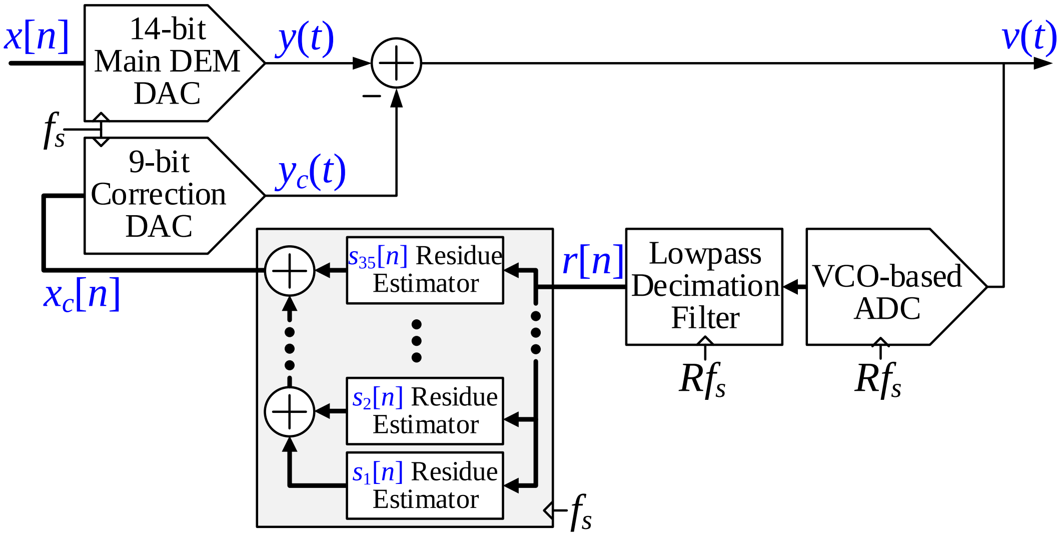


Figure 4: Proposed MNC technique applied to a main DEM DAC.

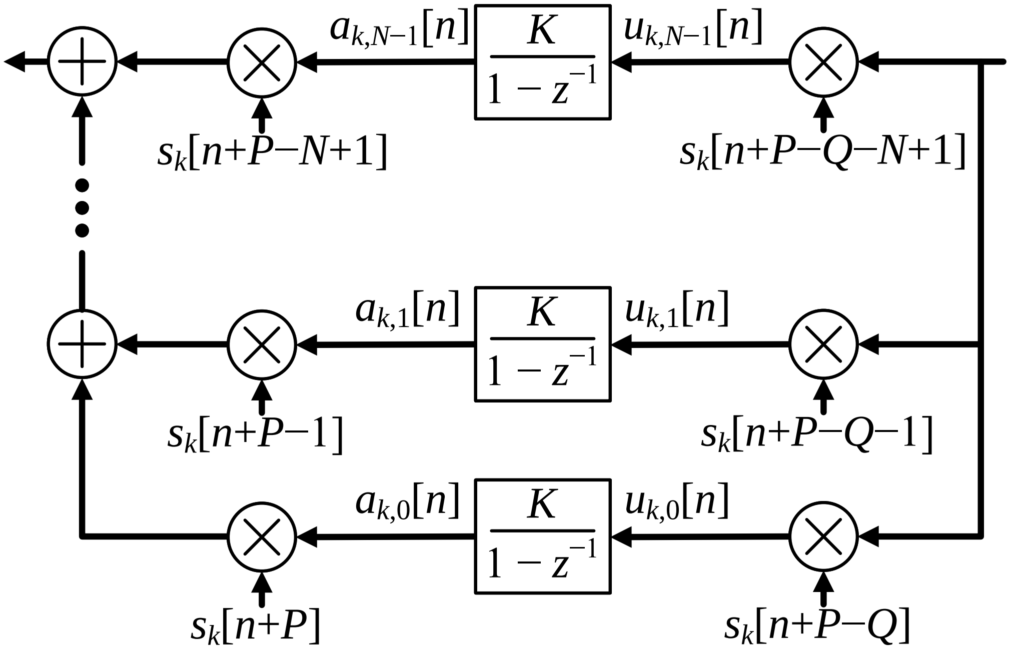


Figure 5: Details of each $s_k[n]$ residue estimator.

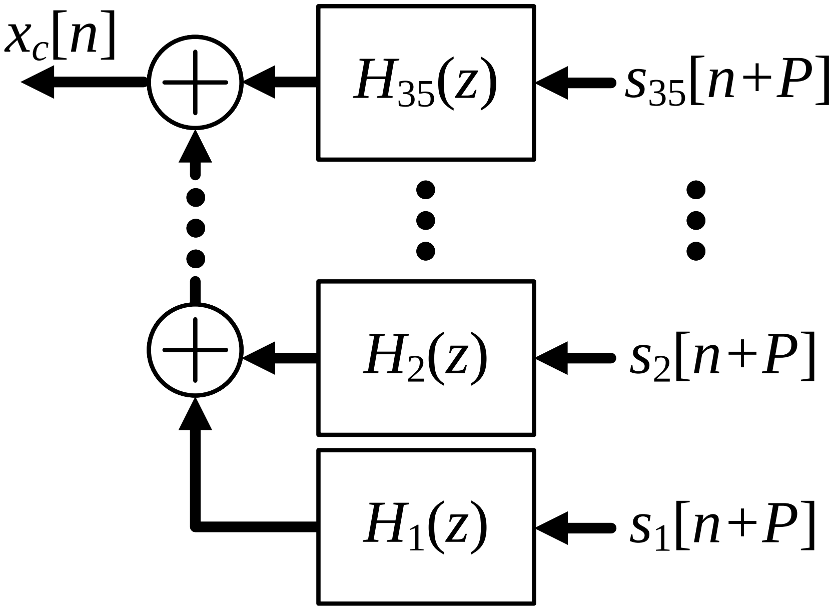


Figure 6: Equivalent behavior of the $s_k[n]$ residue estimator bank.

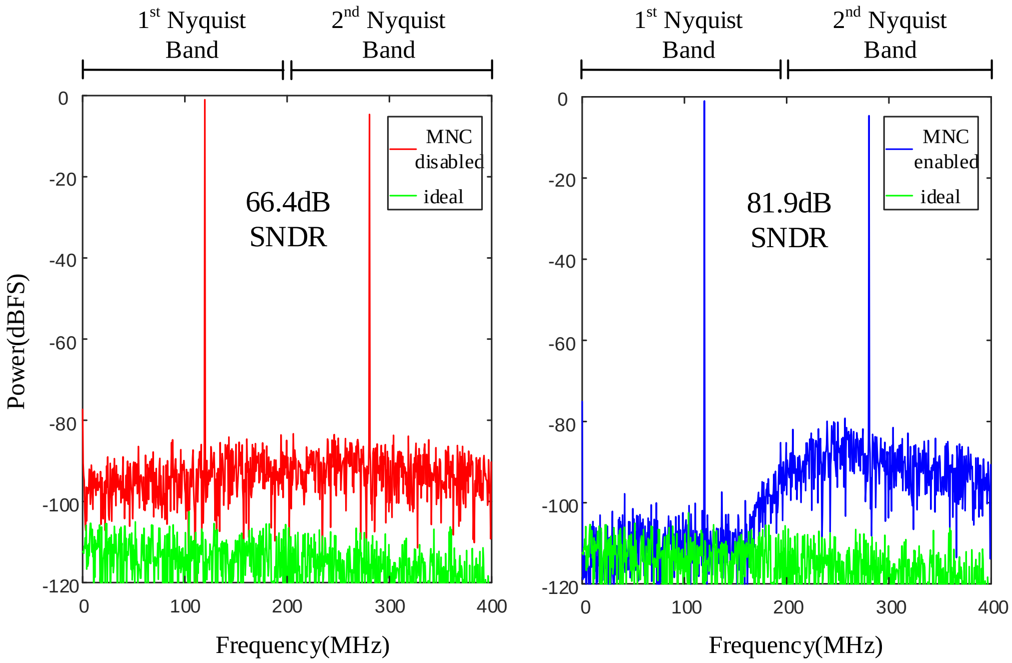


Figure 7: Representative simulated output Spectra without/with MNC for a −1 dB full scale signal. The SNDR bandwidth is 0 to 0.42f_s.

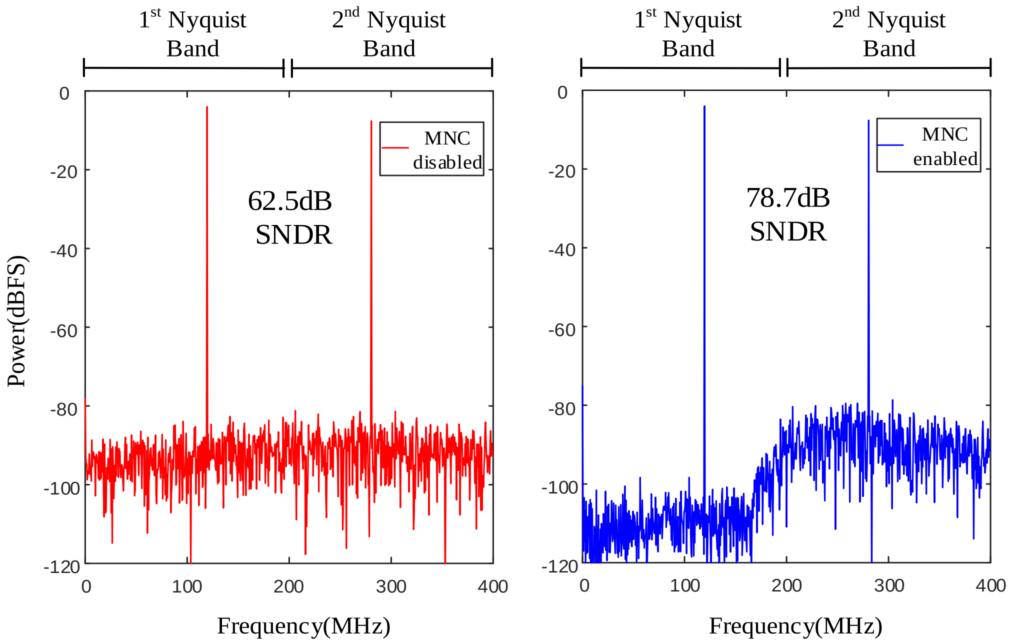


Figure 8: Representative simulated output Spectra without/with MNC with -4dBFS input tone.

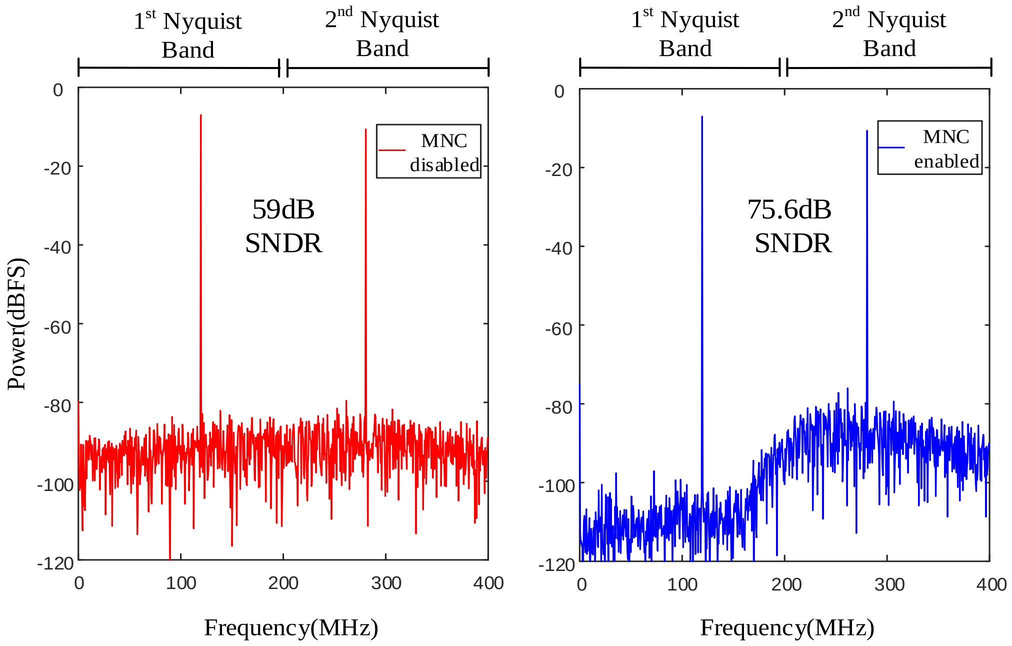


Figure 9: Representative simulated output Spectra without/with MNC with -7dBFS input tone.

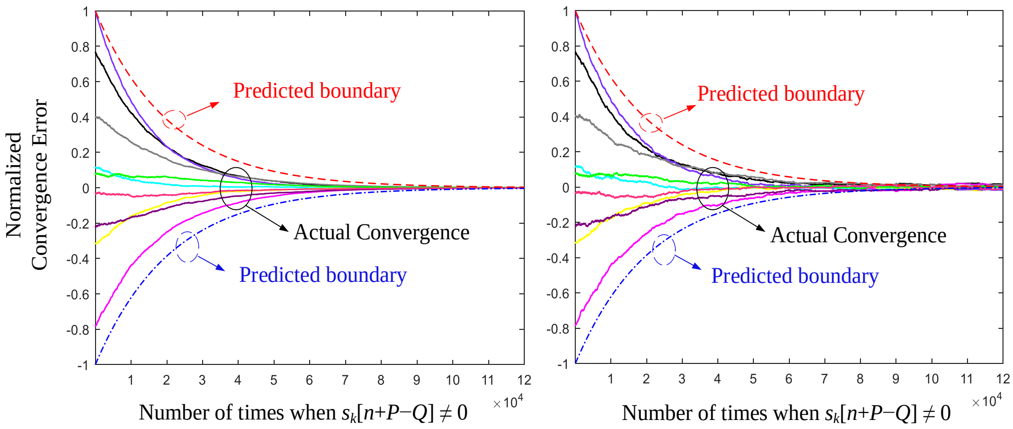


Figure 10: Simulated coefficient convergence without ADC noise (left plot) and with ADC noise (right plot).

REFERENCES

1. B. Jewett, J. Liu, and K. Poulton, "A 1.2 Gs/s 15b DAC for Precision Signal Generation," in *IEEE ISSCC Digest Technical Papers*, 2005, pp. 110-111.

2. K. O. Sullivan, C. Gorman, M. Hennessy, and V. Callaghan, "A 12-bit 320-MSample/s Current-Steering CMOS D/A Converter in 0.44 mm^2," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 7, pp. 1064-1072, July 2004.

3. K. L. Chan, J. Zhu, and I. Galton, "Dynamic Element Matching to Prevent Nonlinear Distortion from Pulse-Shape Mismatches in High-Resolution DACs," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 9, pp. 2067-2078, September 2008.

4. W.-T. Lin, H.-Y. Huang, and T.-H. Kuo, "A 12-bit 40 nm DAC Achieving SFDR > 70 dB at 1.6 GS/s and IMD < -61dB at 2.8 GS/s With DEMDRZ Technique," *IEEE Journal of Solid-State Circuits*, Vol. 49, no. 3, pp. 708-717, March 2014.

5. W. Schofield, D. Mercer, L. St. Onge, "A 16b 400MS/s DAC with <80dBc IMD to 300MHz and 160dBm/Hz noise Power Spectral Density," *IEEE International Solid State Circuits Conference*, February 2003.

6. Q. Huang, P. A. Francese, C. Martelli, and J. Nielsen,"A 200Ms/s 14b 97 mW DAC in 0.18μm CMOS," *IEEE International Solid State Circuits Conference*, February 2004.

7. H.-H. Chen, J. Lee, J. Weiner, Y.-K. Chen, and J.-T. Chen, "A 14-bit 150 MS/s CMOS DAC with Digital Background Calibration," *Symposium on VLSI Circuits*, pp. 51-52, June 2006.

8. M. Clara, W. Klatzer, B. Seger, A. Di Giandomenico, and L. Gori, "A 1.5V 200MS/s 13b 25mW DAC with Randomized Nested Background Calibration in 0.13 μm CMOS," *IEEE International Solid State Circuits Conference*, February 2007.

9. M. Clara, W. Klatzer, D. Gruber, A. Marak, B. Seger, and W. Pribyl, "A 1.5 V 13 bit 130-300 MS/s Self-calibrated DAC with Active Output Stage and 50 MHz Signal Bandwidth in 0.13μm CMOS," *European Solid-State Circuits Conference*, pp. 262-265, September, 2008.

10. Y. Tang, J. Briaire, K. Doris, R. van Veldhoven, P. van Beek, H. Hegt, and A.van Roermund, "A 14 bit 200 MS/s DAC With SFDR >78 dBc, IM3 < −83 dBc and NSD < −163 dBm/Hz Across the Whole Nyquist Band Enabled by Dynamic-Mismatch Mapping," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 6, pp. 1371-1381, June 2011.

11. B. Catteau, P. Rombouts, J. Raman, and L. Weyten, "An on-line calibra- tion technique for mismatch errors in high-speed DACs," *IEEE Trans. Circuits Syst.–I, Reg. Papers* , vol. 55, no. 7, pp. 1873–1883, Aug. 2008

12. R. Schreier, G. C. Temes, *Understanding Delta-Sigma Data Converters*, John Wiley and Sons, 2005.

13. I. Galton, "Digital Cancellation of D/A Converter Noise in Pipelined A/D Converters," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 47 no. 3, pp. 185-196, March 2000.

14. E. Siragusa, I. Galton, "A Digitally Enhanced 1.8V 15b 40MS/s CMOS Pipelined ADC," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 12, pp. 2126-2138, December 2004.

15. J. Remple, I. Galton, "The Effects of Inter-Symbol Interference in Dynamic Element Matching DACs," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 1, pp. 14-23, January 1017.

16. K. L. Chan, N. Rakuljic, I. Galton, "Segmented Dynamic Element Matching for High-Resolution Digital-to-Analog Conversion," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 55, no. 11, pp. 3383-3392, December 2008.

17. G. Taylor, I. Galton, "A Mostly-Digital Delta-Sigma ADC With a Worst-Case FOM of 160 dB," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 4, pp. 983-995, April 2013.

18. I. Galton, "Why Dynamic Element Matching DACs Work," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 57, no. 2, pp. 69-74, February 2010.

19. R. A. Horn, C. R Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

20. S. Kim, J. Kang, M. Lee, "A 12 bit 250 MS/s 28 mW +70 dB SFDR DAC in 0.11 µm CMOS Using Controllable RZ Window for Wireless SoC Integration," *IEEE Asian Solid-State Circuits Conference*, pp. 93-96, November 2014.

21. P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, 1993.

CHAPTER 2

A 600 MS/S DAC WITH OVER 87DB SFDR AND 77DB PEAK SNDR ENABLED BY ADAPTIVE CANCELLATION OF STATIC AND DYNAMIC MISMATCH ERROR

Abstract— This paper presents a Nyquist-rate current-steering DAC that achieves a peak SFDR better than 87 dB and a peak SNDR better than 77 dB over a 265 MHz signal band. It is enabled by a fully integrated digital calibration technique that measures and cancels both static and dynamic mismatch error over the first Nyquist band, and various circuit-level techniques that mitigate the effects of jitter and ISI.

I. INTRODUCTION

Nyquist-rate DACs with continuous-time output waveforms are widely used in moderate-to-high-bandwidth applications such as wireless base stations. Such DACs generate a continuous-time analog output pulse once every clock period. Ideally, the amplitude of each pulse is scaled by the value of the DAC’s input sequence during its clock period, but otherwise the pulses have identical shapes.

Unfortunately, non-ideal circuit behavior causes input-dependent deviations of both the amplitude and shape of each output pulse, which introduce nonlinear error in the DAC’s output waveform. The portion of the error from pulse amplitude deviations is called static error, and that from pulse shape deviations is called dynamic error. Both types of error significantly limit DAC performance in practice. In many Nyquist-rate DACs, clock skew and mismatches among nominally identical DAC components are the dominant causes of these errors. Clock skew causes dynamic error and component mismatches cause both static and dynamic error.

Several previously published DACs incorporate methods to mitigate error from clock skew and component mismatches. These methods include randomization techniques such as dynamic element matching (DEM) and digital random return-to-zero (DRRZ), dynamic mismatch mapping (DMM), and various mixed-signal calibration techniques [1-15]. Randomization techniques cause static error and, in some cases, dynamic error to be wideband noise instead of harmonic distortion. Hence, they improve DAC linearity, but at the expense of significantly reduced signal-to-noise ratio (SNR). DMM is a foreground calibration technique that reorders the usage pattern of nominally identical components to reduce integral nonlinearity (INL). While beneficial, it does not improve differential nonlinearity (DNL) and it tends to be limited by compromises made between improving static and dynamic error. Previously published on-chip mixed-signal calibration techniques have been demonstrated that suppress static error, but not dynamic error.

A mixed-signal calibration technique called mismatch noise cancellation (MNC) was recently proposed in [16] that adaptively measures and cancels both static and dynamic error from clock skew and component mismatches over the DAC's signal band. This paper presents the first DAC IC implemented with MNC. With MNC enabled, the DAC's measured spurious-free dynamic range (SFDR) is better than 87 dB and its peak signal-to-noise-and-distortion ratio (SNDR) is better than 77 dB over a 265 MHz signal band. With MNC disabled, the SFDR and SNDR drop by more than 24 dB and 20 dB, respectively.

Additional measured results further demonstrate that MNC cancels dynamic error as well as static error, as predicted by theory. As [16] presents a theoretical analysis of the MNC technique, this paper focusses on its practical implementation details and presents several circuit-level techniques incorporated in the DAC to reduce jitter and inter-symbol interference (ISI).

II. SIGNAL PROCESSING OVERVIEW

As shown in Fig. 11, the prototype IC consists of a 600 MHz 14-bit main DAC, and an MNC feedback path that measures and cancels the main DAC's signal band error from clock skew and component mismatches. The feedback path consists of a 3 GHz VCO-based ADC, a lowpass decimation filter, a digital error estimator block, and a 600 MHz 9-bit correction DAC.

The sampling theorem implies that no matter what error is introduced by the main DAC, there must exist a correction DAC input sequence, $x_c[n]$, that would result in a correction DAC output waveform, $y_c(t)$, which would cancel the error over the first Nyquist band up to the accuracy of the correction DAC. The correction DAC's minimum step-size must be small enough that error from the quantization of $x_c[n]$ is well below the post-cancellation target noise and distortion floor of the main DAC, and the correction DAC's output range must be large enough to cancel the main DAC's error components. As explained in Section III-B, the correction DAC's resolution of 9-bits and step-size equal to a quarter of that of the main DAC are sufficient for this purpose.

The main DAC's static and dynamic output error from clock skew and component mismatches has the form

$$e_{DAC}(t) = \sum_{k=1}^{35} d_k(t) s_k\left[n_t\right] \tag{63}$$

where each $d_k(t)$ is a 600 MHz periodic waveform that depends on the main DAC's clock skew and component mismatches but not on the DAC's input sequence, and the $s_k[n]$ sequences are generated explicitly within the DEM encoder so they are known to the system a priori [6,17]. The $s_k[n]$ sequences each take on values of −1, 0, and 1, and when DEM is enabled they are zero-mean, white pseudo-random sequences that are uncorrelated with the main DAC's input sequence and each other.

The objective of the MNC feedback loop is to make $y_c(t)$ well-approximate $e_{DAC}(t)$ over the signal band. To do this, the MNC feedback loop must measure $e_{DAC}(t)$ over the first Nyquist band, which requires a digitized version of the main DAC's output waveform that has been filtered to include only the first Nyquist band. The oversampling VCO-based ADC and decimation filter in Fig. 11 perform this operation, so $r[n]$ contains a component equal to the portion of $e_{DAC}(t)$ restricted to the first Nyquist band that is left over from imperfect MNC cancellation.

Ideally, once the MNC feedback loop converges, $r[n]$ becomes free of $e_{DAC}(t)$, in which case it is uncorrelated with all of the $s_k[n]$ sequences. Otherwise, $r[n]$ contains a residual component of $e_{DAC}(t)$ restricted to the first Nyquist band, so it is correlated with at least some of the $s_k[n]$ sequences. Furthermore, the Nyquist-band filtering has an impulse response that is many 600 MHz samples long, so prior to full MNC convergence $r[n]$ is correlated with multiple time-shifted versions of the $s_k[n]$ sequences.

The MNC technique exploits these properties of $r[n]$. As shown in Fig. 12, the digital error estimator block in the MNC feedback loop consists of 35 $s_k[n]$ residue estimators, each of which correlates $r[n]$ with 9 shifted versions of one of the $s_k[n]$ sequences. Given that each $s_k[n]$ sequence is restricted to values of −1, 0, and 1, each correlation is performed by multiplying $r[n]$ by a −1, 0, or 1 during the nth 600 MHz clock cycle. The result is multiplied by a small loop gain constant, $K = 6{\cdot}10^{-7}$, and accumulated. As proven in [16], the feedback loop causes the accumulator outputs to increase or decrease as necessary for $y_c(t)$ to well-approximate $e_{DAC}(t)$ over the first Nyquist band.

Even though $e_{DAC}(t)$ is a broadband waveform which depends on the main DAC's input sequence, $x[n]$, the $d_k(t)$ waveforms in (63) are periodic and independent of $x[n]$ [17]. They depend only on component mismatches and clock skew within the main DAC, so they do not change significantly over time. The MNC feedback loop causes the 315 accumulator outputs in the digital error estimator to converge to coefficients which depend only on the $d_k(t)$ waveforms. Thus, like the $d_k(t)$ waveforms these coefficients depend only on the main DAC's component mismatches and clock skew.

In principle, the MNC technique can perform foreground or background calibration with only minor differences, but in the prototype IC it was limited to foreground calibration to simplify the project. During foreground calibration, the MNC feedback loop measures 315 coefficients described above. During normal DAC operation, the coefficients that were measured during foreground calibration continue to be used to generate $y_c(t)$, and thereby continue to cancel the error components.

III. CIRCUIT IMPLEMENTATION DETAILS

The IC was implemented in the GlobalFoundries 22 nm FDSOI process. In addition to the blocks shown in Fig. 11 and Fig. 12, the IC contains a direct digital synthesizer (DDS), a serial peripheral interface (SPI), full ESD protection circuitry, and miscellaneous control and test circuitry. As shown in Fig. 13, an off-chip 1:1 balun converts the IC's differential output signal to a non-differential signal which is used during testing to drive the 50 Ω input of a laboratory signal analyzer.

A. Main DAC

The 14-bit main DAC consists of the DEM encoder and the subsequent 36 current-steering 1-bit DACs shown in Fig. 13. The DEM encoder [5] converts the 14-bit $x[n]$ sequence into 36 1-bit sequences, each of which drives a 1-bit DAC with weight K_i. For i = 1, 2, …, 20 the values of K_i are 1, 1, 2, 2, 4, 4, …, 512, 512, respectively, and for i = 21, 22, …, 36, each K_i has a value of 1024. The main DAC's minimum current step averaged over a 600 MHz clock interval, Δ, is 1.56 μA.

Non-return-to-zero (NRZ) 1-bit DACs are a common design choice for current-steering DACs. In the context of the main DAC, an ideal ith NRZ 1-bit DAC would steer $K_i\Delta$ amperes of current to either its top output or its bottom output during the nth clock period depending on whether its input bit during that clock period is high or low, respectively. Unfortunately, inevitable asymmetries within any practical NRZ 1-bit DAC in conjunction with parasitic capacitances cause the 1-bit DAC's output waveform to depend nonlinearly on its input bit during at least one prior clock period in addition to that of the current clock period. The resulting ISI causes even DACs that incorporate DEM to introduce harmonic distortion [17].

Simulations suggest that ISI from NRZ 1-bit DACs would reduce the achievable Nyquist band SFDR below the project's target of better than 85 dB.[5] To circumvent this problem, the IC incorporates return-to-zero (RZ) 1-bit DACs, each of which is reset to a data-independent state every clock period to make its output waveform independent of its input sequence during prior clock periods. In principle, this eliminates ISI provided all analog and digital circuit blocks that make up the 1-bit DACs and that control them are fully reset each period. However, conventional RZ 1-bit DACs are more sensitive to clock jitter than their NRZ counterparts. As explained in the remainder of this subsection, various new circuit techniques are implemented in the IC to mitigate this issue while ensuring that ISI does not limit performance.

Fig. 14 shows the high-level structure and timing of each 1-bit DAC and its interface to the automatically placed and routed (P/R) digital block. Each 1-bit DAC is implemented as a parallel combination of two current-steering RZ 1-bit sub-DACs to mitigate the effect of clock jitter as explained shortly. The sub-DACs each operate on the same input bit sequence, and their outputs are connected so their output currents add. Each is reset for 20% of the clock period and generates output current for 80% of the clock period.[6] The only difference between the two sub-DACs is that they are reset at different times: sub-DAC 1 is reset during the first 20% of each clock period and sub-DAC 2 is reset during the second 20% of each clock period as shown in Fig. 14. Error from mismatches and clock skew between the sub-DACs are cancelled by the MNC technique so they are not a significant issue in this design.

The average output current magnitude from the *i*th 1-bit DAC is $I_i = K_i\Delta$, so each of the two RZ sub-DACs has an output current magnitude of $0.625I_i$ during its 80% data phase. A single RZ 1-bit DAC with an 80% data phase output current magnitude of $1.25I_i$ or a

[5] IC measurements further support this conclusion: when the 1-bit DACs are configured to run in NRZ mode (via a debug feature of the IC) the measured SFDR drops by 10 dB.

[6] The commonly used alternative of interlacing 50% RZ sub-DACs was not used here because of its high current consumption [18].

single NRZ 1-bit DAC with an output current magnitude of I_i are each comparable alternative 1-bit DACs in that they too have average output current magnitudes of I_i. Of these comparable alternatives, the single RZ 1-bit DAC is significantly more sensitive to clock jitter than the single NRZ 1-bit DAC, because the former has two output current transitions each clock period whereas the latter has at most one output current transition each clock period and no transition if the input bit remains unchanged.

The timing of the 1-bit DAC in Fig. 14 is such that each rising edge of output current from sub-DAC 1 aligns with a falling edge of output current from sub-DAC 2. The sub-DACs share most of their timing circuitry, so their jitter is highly correlated. Hence, most of the error from clock jitter at the aligned edges cancels when the present and prior 1-bit DAC input bits are equal, and add in amplitude otherwise. It follows that the error from jitter introduced by the pairs of aligned edges has the same form as the total error from jitter of the comparable single NRZ 1-bit DAC, but with |20log(0.625)| = 4 dB less power. Each of the non-aligned edges has half the transition magnitude of each edge from the single comparable RZ 1-bit DAC, so the combined error from clock jitter introduced by the two non-aligned sub-DAC edges has the same form as that of the single comparable RZ 1-bit DAC, but with 6 dB less power. Hence, the error from jitter of the 1-bit DAC of Fig. 14 contains a component similar to that from an NRZ 1-bit DAC and a component similar to that from an RZ 1-bit DAC. As the former can be much smaller than the latter for broadband input sequences, it follows that the total error from jitter is up to 6 dB lower than that of the comparable single RZ 1-bit DAC.

As shown in Fig. 15, each sub-DAC’s current-steering cell consists of two cascode current sources, each of which is steered to one of the two sub-DAC outputs by a differential pair controlled by the switch driver. During the data phase, the differential pairs steer both currents to the I+ output if the input bit, $c_i[n]$, is high and to the I− output if $c_i[n]$ is low. During the RZ phase, they steer the currents to opposite outputs so the differential output current is zero. In contrast to a conventional RZ 1-bit DAC, which steers a single cascode current source to one of the two outputs during the data phase and to a dummy load during the RZ phase, the common mode output current does not change during the RZ phase so unwanted large output slewing transients are avoided. Simulations indicate that the largest-weight 1-bit DACs have a minimum output impedance of 30 kΩ across the DAC's signal band, which is sufficient to prevent input code dependent impedance variations from limiting performance.

In addition to controlling the current-steering cell as described above, the switch driver converts from the input sequence's 0.8 V power supply domain to the current-steering cell's 1.8 V power supply domain, and its design ensures data-independent switching current. As shown in Fig. 16, the switch driver circuit consists of separate signal paths for the complementary input bit sequences $c_i[n]$ and $\overline{c}_i[n]$, each of which consists of first and second latch stages that operate from 0.8 V and 1.8 V power supplies, respectively. The latches in both stages are briefly reset each DAC clock period, so the two-path design ensures that the same numbers of positive-going and negative-going logic transitions occur each DAC clock period regardless of $c_i[n]$. This ensures that the current drawn by the switch driver is data-independent, thereby preventing data-dependent supply modulation which would be a source of ISI and nonlinear distortion.

The interface circuitry shown in Fig. 14 generates retimed complementary versions of the DEM encoder's *i*th output bit and includes an additional ISI-mitigation technique. It resets the complementary outputs to zero just prior to updating them with their next data values so as to mitigate ISI that would otherwise result from data-dependent coupling from the digital to analog supply domains.

B. Correction DAC

The quantization step-size of the correction DAC, Δ_C, is a quarter that of the main DAC, i.e., 0.39 μA, because behavioral simulations during the design phase suggested that error from quantizing the correction DAC's input sequence to this step-size is well below the post-cancellation target noise and distortion floor of the main DAC. Circuit simulations further suggested: 1) that 9-bits of resolution is sufficient, because the main DAC's error spans only small fraction of its total output range, and 2) that these step-size and resolution values are sufficiently small, even without DEM, calibration, or the sub-DAC interleaving technique, that the correction DAC's error is well below the post-cancellation target noise and distortion floor.

Thus, the correction DAC consists of the non-DEM encoder and the subsequent 14 current-steering 1-bit DACs shown in Fig. 13. The non-DEM encoder converts the correction DAC's 9-bit digital input sequence into 14 1-bit sequences, each of which drives a 1-bit DAC with weight L_i. For i = 1, 2, 3, and 4, the values of L_i are 1, 2, 4, and 8, respectively, for i = 5, 6, and 7, each L_i has a value of 16, and for i = 8, 9, …, 14, each L_i has a value of 64. The 1-bit DACs are identical to those of the main DAC, except without the bottom sub-DAC shown in Fig. 14.

C. VCO-based ADC

It was explained heuristically in [16], but not proven, that the MNC technique is highly insensitive to ADC nonlinearity and noise. An objective of this project is to provide experimental support of this claim. The implemented ADC does not include calibration or special linearization techniques, so it is quite nonlinear: circuit simulations indicate that its 2nd, 3rd, and 4th output harmonics are −26 dBc, −47 dBc, and −64 dBc, respectively, for a full-scale sinusoidal input. Furthermore, it has only first-order quantization noise shaping and an oversampling ratio of only 5, so its noise floor is high. Nevertheless, the experimental results presented in Section IV suggest that the ADC's error negligibly affects the MNC coefficients.

The ADC requirements are even further relaxed during foreground calibration, because in this case the ADC's input range need only be a fraction of the main DAC's full-scale output range. Specifically, the main DAC's input during foreground calibration is toggled back and forth between -2389.5Δ and -2388.5Δ. In principle, any other input sequence could have been used, but this choice has the benefit of a very small dynamic rage and it ensures rapid MNC loop convergence because it results in $s_k[n]$ sequences with a low percentage of zero values. With this choice, the ADC's differential input range of only 20 mV is sufficient to accommodate the maximum expected error from component mismatches and clock skew.

A strict requirement, however, is that the digital error estimator input must contain negligible aliased power from outside the DAC's first Nyquist band. This is why an oversampling ADC is required, which is the MNC technique's primary downside. A VCO-based ADC is used in the IC because its inherent lowpass sinc filtering helps suppress the input signal above the DAC's first Nyquist band [19], which made it possible to use the relatively low oversampling ratio of 5. Also its design is particularly simple given that the MNC technique's insensitivity to nonlinearity makes ADC calibration or other linearization techniques unnecessary.

As shown in Fig. 17, the VCO-based ADC includes a differential voltage-to-current (V/I) converter, each output of which is followed by a 15-element pseudo-differential current controlled ring oscillator (ICRO), a ring sampler, a phase decoder, and a $1 – z^{-1}$ digital differentiator block. The high level structure is similar to that presented in [19], except it consists of one instead of two signal paths, and it does not include dither or digital calibration because of the relaxed linearity and noise requirements.

The V/I converter (Fig. 18) generates currents I_{ICRO+} and $I_{ICRO–}$ that drive the ICROs. As in [19], each ICRO consists of two pseudo-differential rings, each made up of 15 current-starved inverters. The V/I converter's input common-mode voltage is that of the IC's output signal, i.e., 1.8 V, and its common-mode output current is 1 mA. The two PMOS cascode bias voltages, V_{bp1} and V_{bp2}, are generated separately to reduce kick-back from the second stage to the first stage, and V_{bp2} is set to the 1.8 V during start-up while the other V/I converter nodes settle to protect the ICRO's thin-oxide devices from start-up transients. The V/I converter's DC gain is programmable but was set to its nominal value of 40 mS during testing. Its −3 dB bandwidth is slightly above the Nyquist frequency of the DAC.

The ring sampler, phase decoder, $1 - z^{-1}$ block, are similar to those described in [19]. They are not implemented as part of the P/R digital block because their data rate is 3 GHz and the P/R digital block is clocked at 600 MHz. The decimation filter is implemented in the P/R digital block, so its input data must have a 600 MHz sample-rate. This is achieved by a digital interface circuit close to the ADC which parallelizes the six 3 Gb/s ADC output lines to thirty 600 Mb/s lines.

D. Clock Generator

The IC is externally clocked by a single 3 GHz differential clock signal, from which the on-chip clock generator derives all the IC's internal clocks. The clock generator consists of the three-stage differential to single-ended amplifier and the 1.2 V to 0.8 V level shifter shown in Fig. 19, followed by a clock divider that generates several 600 MHz clock signals including those shown in Fig. 14. The amplifier operates from a 1.2 V supply and generates a nearly rail-to-rail squared-up version of the 3 GHz clock. The third stage is a transimpedance amplifier which provides a signal-dependent load to the second stage that limits the second-stage's swing sufficiently to prevent its transistors from entering triode operation. The level shifter generates the ADC's 3 GHz clock signal which is also the input to the clock divider.

To achieve the post-cancellation DAC noise performance target, the main DAC's critical 600 MHz clock paths must have RMS jitter values of less than 80 fs. The clock generator was designed such that the simulated RMS jitter values of these clock paths are below 50 fs to leave margin. The noise performance of the three-stage amplifier is the most critical component of these clock paths. Accordingly, the amplifier dissipates approximately 80% of the clock generator's total power dissipation.

E. P/R Digital Block

The P/R digital block contains the main DAC's DEM encoder, the correction DAC's non-DEM encoder, the lowpass decimation filter, the MNC digital error estimator, the DDS, the SPI, a pseudo-random sequence generator block, and miscellaneous control and test logic. It consists of approximately 170,000 standard logic cells, occupies an area of 700 μm

× 250 µm, and operates from a 0.8 V digital power supply. All registers except those in the SPI are clocked at 600 MHz.

The DDS provides the main DAC's 14-bit input sequence. It is capable of generating one-tone and two-tone test signals with frequencies at arbitrary integer multiples of 600/512 MHz, and amplitudes of 0, −6, and −12 dBFS. The DDS internally generates an 18-bit version of the desired sequence and performs dithered requantization to obtain the final 14-bit sequence to suppress spurious tones in the quantization error.

The lowpass decimation filter is implemented as a 33-tap digital poly-phase finite impulse response (FIR) filter [20]. Dithered requantization is subsequently performed to reduce its 16-bit output sequence to a 4-bit sequence prior to the digital error estimator to save area.

F. Mixed-Signal Isolation and Process-Specific Details

The FDSOI process provides good isolation of the IC's transistors from substrate noise. Additionally, various measures were applied to reduce coupling of digital noise into sensitive analog circuitry. The P/R digital block is surrounded by a 2 nF ring of on-chip MOS power supply decoupling capacitors and substrate connections, and is separated from the analog circuit blocks by a 250 µm BFMOAT isolation region with reduced substrate doping. All analog transistors reside in triple N-wells, and the analog power supplies are each decoupled with 200 pF on-chip MOS capacitors. On-chip ground planes are used to shield critical clock signals, and the clock generator is placed as far as possible from the P/R digital block. Multiple parallel package bond wires are used to reduce the inductance of critical power supplies.

Several blocks within the IC take advantage of the FDSOI IC technology. The back gates of all 0.8 V pMOS transistors in the P/R digital logic, 1-bit DACs, ADC, and interface circuitry are tied to ground to reduce threshold voltages and increased speed. The back gates of the pull-down nMOS transistors in the second latch stages of the 1-bit DAC switch drivers are tied to 1.8 V to increase pull-down strength.

IV. MEASUREMENT RESULTS

Fig. 20 shows an annotated IC die photograph. The die dimensions are 2.5 mm by 2 mm, and the IC's active area is 1.15 mm^2. The IC is packaged in a 36-pin QFN package with an exposed paddle to which all the IC's ground pads are down-bonded. The package is mounted to a printed circuit board (PCB) via an Ironwood GHz elastomer QFN socket.

The PCB includes clock input and DAC output signal conditioning circuitry, low-noise LDO regulators, and a microcontroller for SPI communication. A Rohde & Schwarz SMA100A signal generator was used to provide a single-ended 3 GHz clock signal which was passively bandpass filtered to suppress noise and harmonics prior to the PCB. The clock signal is converted to differential form by a PCB balun, the outputs of which are AC coupled to 50 Ω impedance-controlled PCB traces. Series 5 Ω resistors between the clock traces and the IC's input clock pins mitigate clock ringing associated with the package bond wire inductance. A PCB balun (Fig. 13) provides a non-differential version of the DAC output, which was measured with a Keysight N9030B PXA signal analyzer.

To fully characterize continuous-time DAC performance, it is necessary to measure both noise and nonlinear distortion over the signal band relative to the signal power. Yet many DAC publications report limited or no noise measurements, and most report measurements of SFDR—the dB power difference between the DAC output's fundamental tone and its largest spurious tone for a full-scale sinusoidal input signal—as the sole means of quantifying nonlinear distortion. Unfortunately, SFDR can be misleading because the number of spurious tones changes with input frequency, and as this number increases the SFDR tends to decrease even when the total distortion power remains relatively constant. To avoid these limitations, the IC was extensively tested to measure several values of not only SFDR, but also SNDR, noise spectral density (NSD), and noise and distortion spectral density (NDSD) as described below.

Each measurement was taken over a signal band that extends from 1 MHz to 265 MHz. The DC to 1 MHz band was excluded because it is suppressed by the output balun, and the upper 35 MHz of the first Nyquist band was excluded because aliasing from the decimation filter's transition band reduces MNC accuracy over this band. This latter exclusion band represents a design tradeoff. It can be reduced by increasing the digital filter's complexity and, therefore, power consumption. Alternatively, the filter complexity can be kept relatively low, e.g., in the current design it is just a 33-tap FIR filter, but the DAC's sample-rate can be increased slightly to compensate for the exclusion band. In lieu of other constrains, the best choice in practice is that which minimizes power dissipation for a given process.

Each measurement was made with and without DEM enabled during normal DAC operation. The main DAC's error waveform is given by (63) even when DEM is disabled, but in this case the $s_k[n]$ sequences are nonlinear deterministic functions of $x[n]$. For correct MNC coefficient convergence, the $s_k[n]$ sequences must be uncorrelated with each other and with $x[n]$, so DEM is required during foreground calibration. However, once the coefficients have been measured, DEM is optional; error cancellation works regardless of whether DEM is enabled or disabled. With MNC enabled, DEM offers a tradeoff during normal DAC operation: it slightly increases the signal-band noise floor and overall power dissipation, but it slightly reduces harmonic distortion over the signal band and greatly reduces it outside of the signal band.

Fig. 21 shows representative measured output power spectra over the first two Nyquist bands for a full-scale 249.6 MHz single-tone DAC input sequence. The data were measured with a signal analyzer resolution bandwidth of 100 Hz, exported to files, and plotted via software for improved readability. Without MNC and DEM, the signal-band SFDR is 63.7 dB, with MNC but without DEM the signal-band SFDR improves to 86.4 dB, and with MNC and DEM the signal-band SFDR slightly improves further to 87.6 dB. As shown in Fig. 22, these SFDR results are representative of those measured for full-scale single-tone and two-tone input signals throughout the signal band.

The post-cancellation noise floor of the DAC is below that of the signal analyzer, so to measure the DAC's noise floor it was necessary to use the signal analyzer's internal preamplifier. To avoid being limited by the preamplifier's nonlinearity it was further necessary to use passive notch and lowpass filters prior to the signal analyzer to suppress the signal component of DAC's output waveform and limit the spectrum to the first Nyquist band.

Fig. 23 shows representative DAC output power spectra measured with the passive filters and preamplification described above for a 116 MHz full-scale sinusoidal input signal. The data were measured with the signal analyzer's resolution bandwidth set to 30 kHz, the number of frequency trace points set to 1001, and the RMS average detector enabled. As indicated in the figure, enabling MNC reduced the noise by over 20 dB across the 265 MHz signal band.

Table 1 presents values of SNDR, NDSD, and NSD calculated from measured power spectra for full-scale single-tone input signals with frequencies of 50.4 MHz, 116 MHz, and 179.3 MHz. The values were calculated from power spectrum plots like those shown in Fig. 23. Each of the three input frequencies was chosen such that the corresponding notch filter did not hide significant spurious tones, and for each measurement the DAC noise over the notch filter’s 30 MHz stop-band was estimated by extrapolation. For the SNDR and NDSD measurements, the total noise and distortion was calculated by integrating the measured power spectrum from 1 MHz to 265 MHz and then adding the extrapolated noise over the 30 MHz notch filter stop-band. Each NDSD value is this noise and distortion value divided by the integration bandwidth. Each NSD value is equal to the corresponding NDSD value minus the measured power of each non-negligible signal-band spurious tone.

Extensive noise measurements performed by the authors suggest that the DAC’s noise floor is nearly independent of the input signal frequency. The slight drop in SNDR with frequency evident in Table 1 occurs mainly because of the sinc roll-off imposed on the input signal by the 1-bit DAC hold operations. These slight drops with frequency also occur for the NDSD and NSD values in Table 1 because the values are specified in units of dBc/Hz.

Measurements were also performed to assess the IC’s ISI mitigation techniques. By enabling and disabling partial-interleaving and the interface and switch driver ISI mitigation techniques for various test conditions, it was determined that the techniques together prevent an SNDR degradation of about 1.7 dB, with the partial interleaving technique contributing roughly half of this benefit. Configuring the 1-bit DACs to operate in NRZ mode with DEM and MNC enabled reduced the measured SFDR by about 10 dB.

The IC includes a test feature that can be enabled to intentionally delay the clock signals that drive just two of the main DAC’s 256-weight 1-bit DACs by approximately 25 ps. Letting MNC converge in foreground with this feature disabled, and then enabling it during normal DAC operation with a 0 dBFS 116 MHz input signal caused the measured SNDR to degrade from 77.3 dB to 63.0 dB. Given that enabling the feature only introduces clock skew, this 14.3 dB of degradation must be entirely from dynamic mismatch error. Rerunning foreground calibration with the clock delays in place and applying the same 0 dBFS 116 MHz input signal during normal DAC operation caused the measured SNDR to improve to 76.8 dB. This provides experimental confirmation of the theoretical result presented in [16] that the MNC technique effectively cancels dynamic mismatch error.

Fig. 24 shows a representative subset of the 315 MNC coefficient values versus time measured during foreground calibration. The values were obtained by periodically freezing MNC and reading the coefficients from the $s_{11}[n]$ residue estimator (Fig. 12) via the SPI during foreground calibration. The observed coefficient convergence rate is consistent with that predicted by the analysis presented in [16]. Increasing the MNC loop gain, K, reduces the convergence time at the expense of accuracy. For the measurements reported in this paper, the loop gain was set conservatively small. Additional measurements performed by the authors indicate that increasing K by a factor of 16 reduced the convergence time to 2.5 ms while degrading the SNDR by less than 0.5 dB.

All of the measurements presented above were made from a single randomly-selected copy of the IC. Fig. 25 shows representative SFDR and SNDR values measured from this and five other randomly-selected copies of the IC with MNC enabled. As expected, with MNC enabled the performance differences among the ICs are small: less than a dB for SNDR and less than 2 dB for SFDR.

Table 2 summarizes the measured performance described above along with the available corresponding performance of previously published state-of-the-art DACs. Excluding the DAC presented in [15], the DAC reported in this paper achieves at least 7 dB better SFDR than the other DACs, it achieves at least 12 dB better NSD than the other DACs that incorporate randomization to scramble mismatches (i.e., DEM and DRRZ) without calibration, and it achieves at least 3 dB better NSD than those of the remaining DACs. However, it does not outperform the DAC presented in [15]. As this DAC uses NRZ 1-bit DACs, DEM, and calibration that only addresses static error from component mismatches, its astonishingly good performance suggests that special circuit design and layout techniques not described in [15] must have been utilized to reduce ISI and dynamic mismatch error.

ACKNOWLEDGEMENTS

Chapter 2, in full, has been submitted to the IEEE Journal of Solid-State Circuits. D. Kong, K. Rivas-Rivera, I. Galton, 2018. The dissertation author is the primary investigator and author of this paper. Professor Ian Galton supervised the research which forms the basis for this paper.

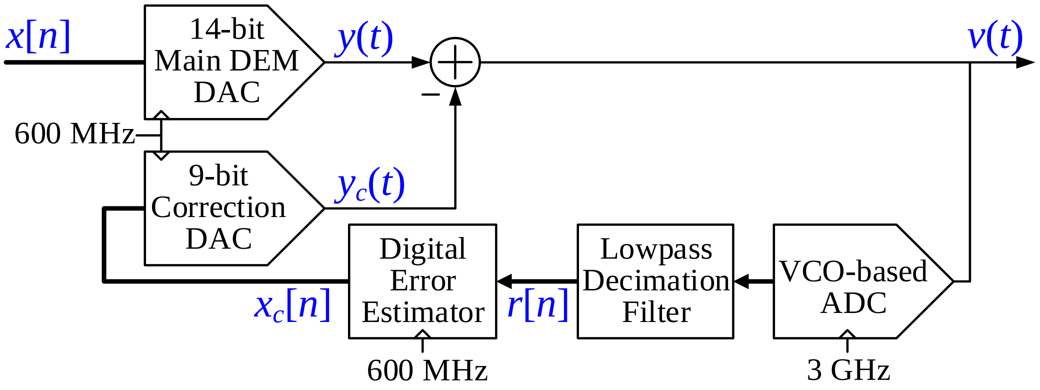


Figure 11: High-level signal processing block diagram of the prototype IC.

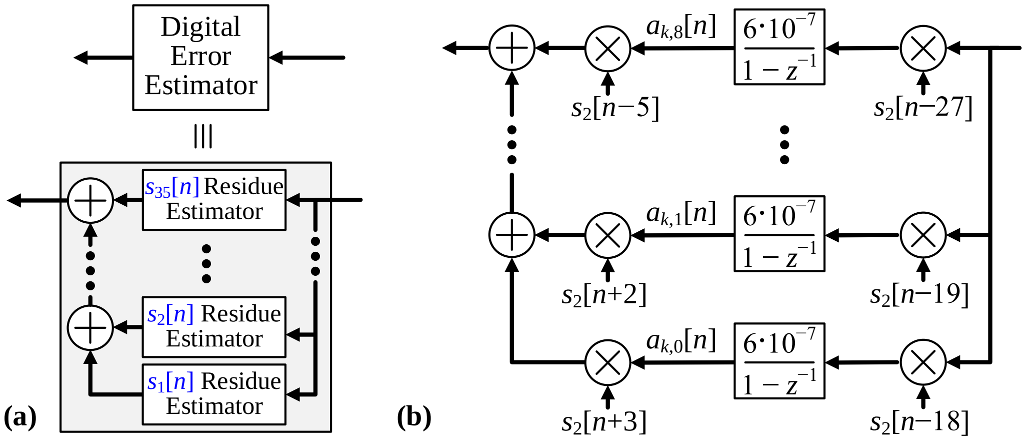

Digital Error Estimator
|||
$s_{35}[n]$ Residue Estimator
$s_2[n]$ Residue Estimator
$s_1[n]$ Residue Estimator
(a)
(b)
$a_{k,8}[n]$
$\frac{6\cdot10^{-7}}{1-z^{-1}}$
$s_2[n-5]$
$s_2[n-27]$
$a_{k,1}[n]$
$\frac{6\cdot10^{-7}}{1-z^{-1}}$
$s_2[n+2]$
$s_2[n-19]$
$a_{k,0}[n]$
$\frac{6\cdot10^{-7}}{1-z^{-1}}$
$s_2[n+3]$
$s_2[n-18]$

Figure 12: a) High-level structure of the digital error estimator, and b) signal processing details of each $s_k[n]$ error estimator.

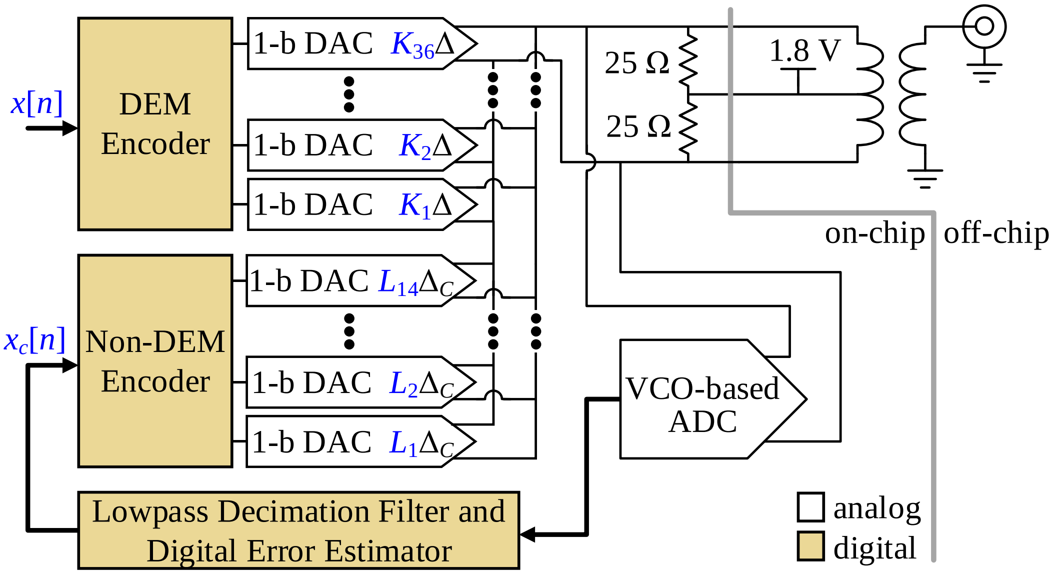


Figure 13: Circuit-level block diagram of the prototype IC.

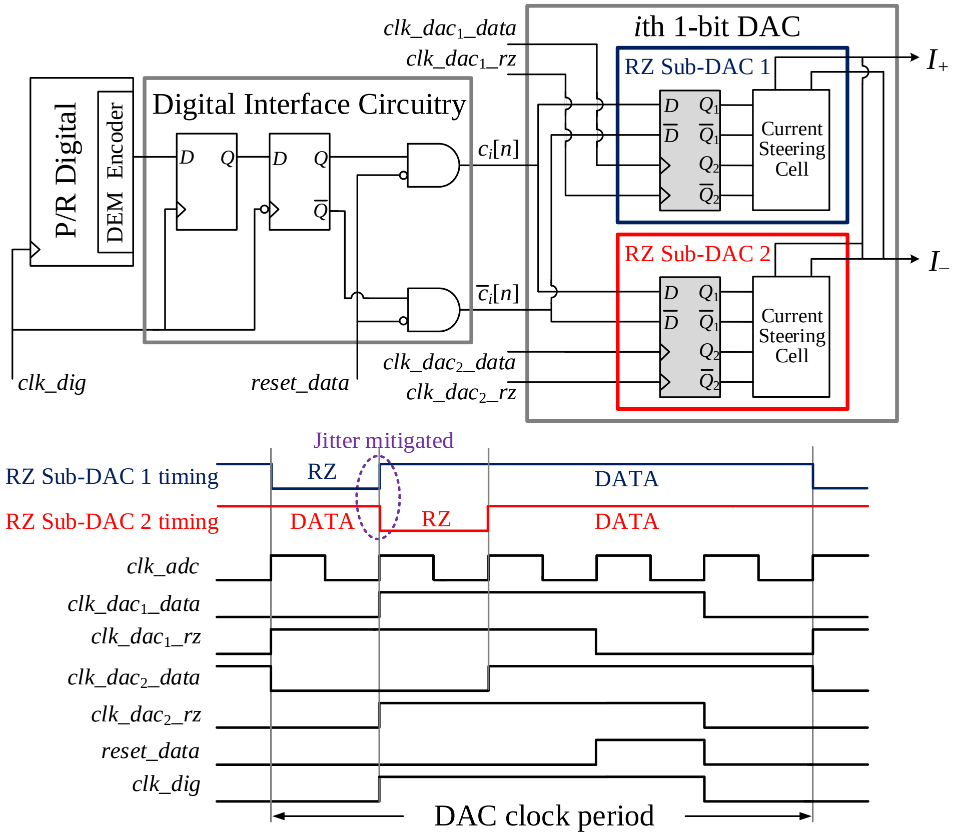


Figure 14: High-level diagram and timing of the ith 1-bit DAC and digital interface.

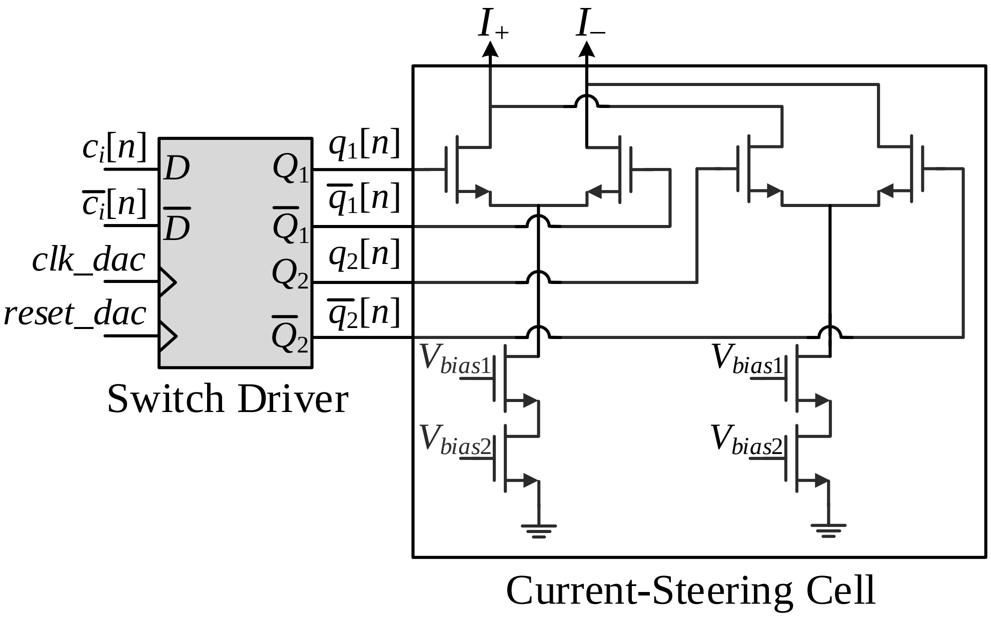


Figure 15: Circuit diagram of the *i*th RZ 1-bit sub-DAC.

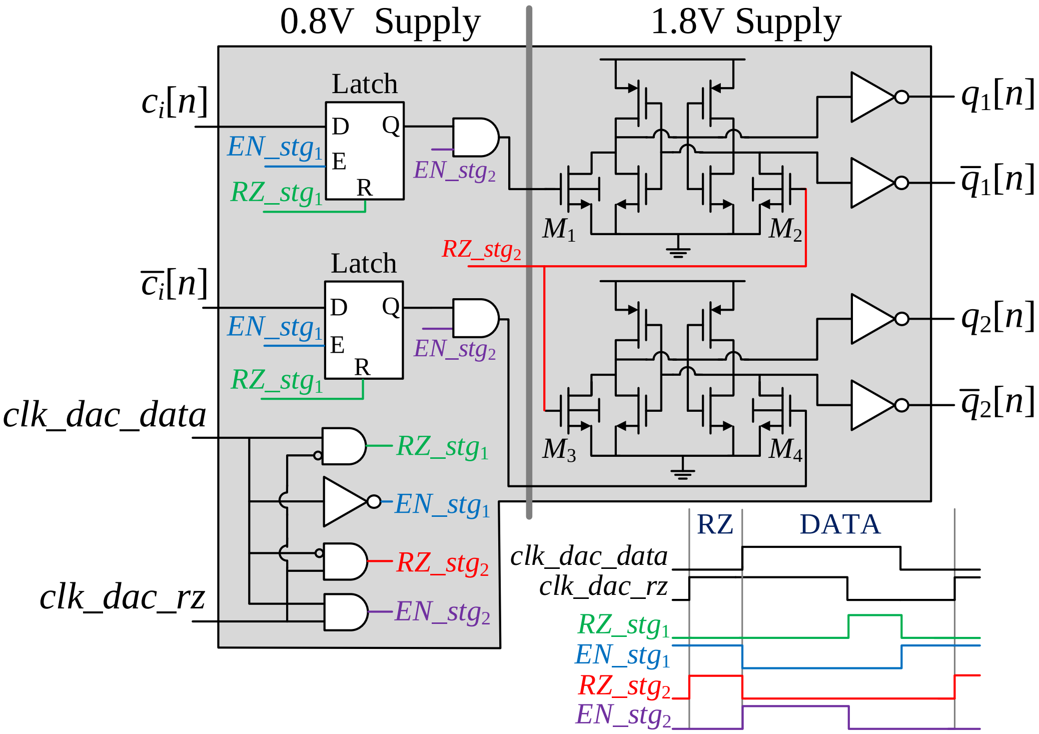


Figure 16: Circuit diagram of the ith 1-bit DAC’s switch driver.

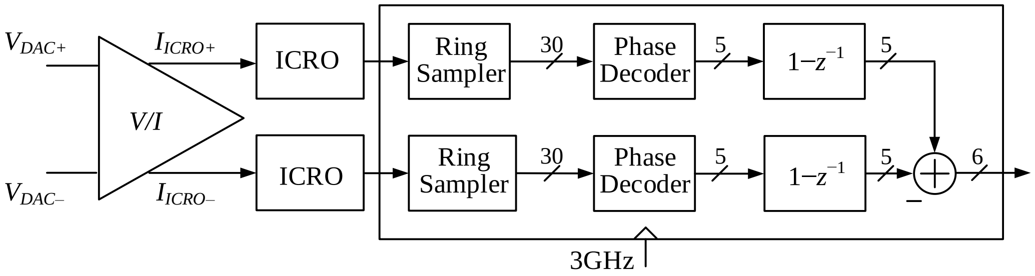


Figure 17: Block diagram of the VCO-based ADC.

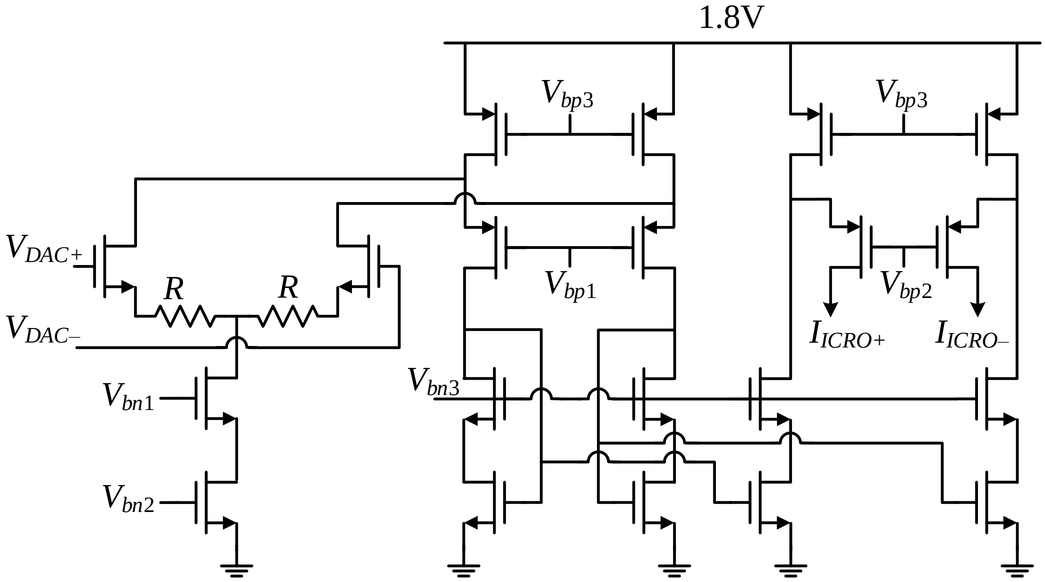


Figure 18: Circuit diagram of the V/I converter.

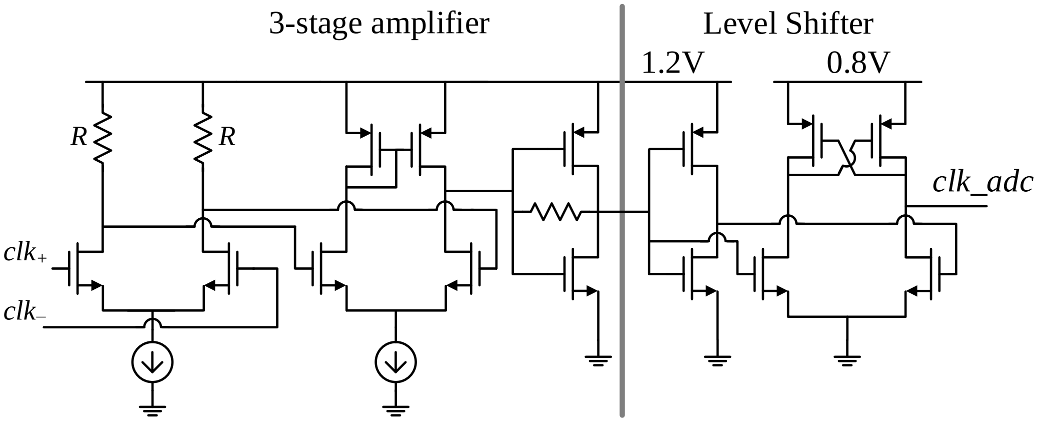


Figure 19: Simplified diagram of the 3 GHz portion of the clock generator.

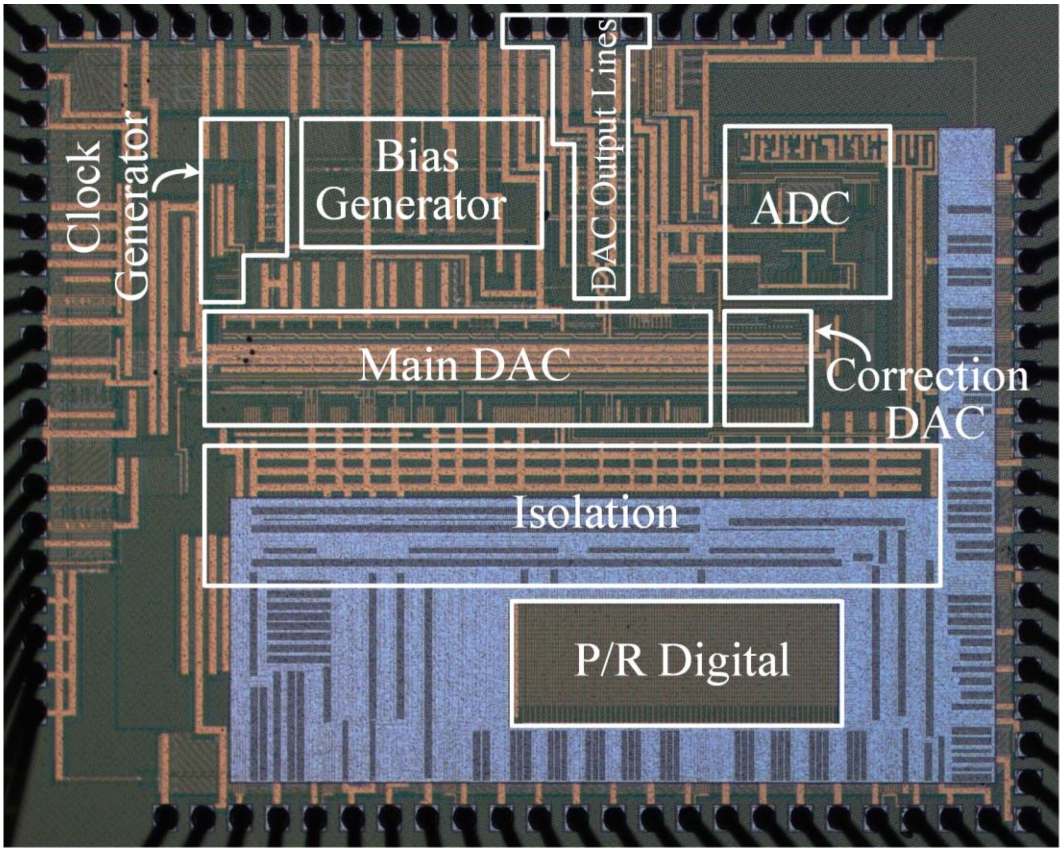


Figure 20: Die photograph.

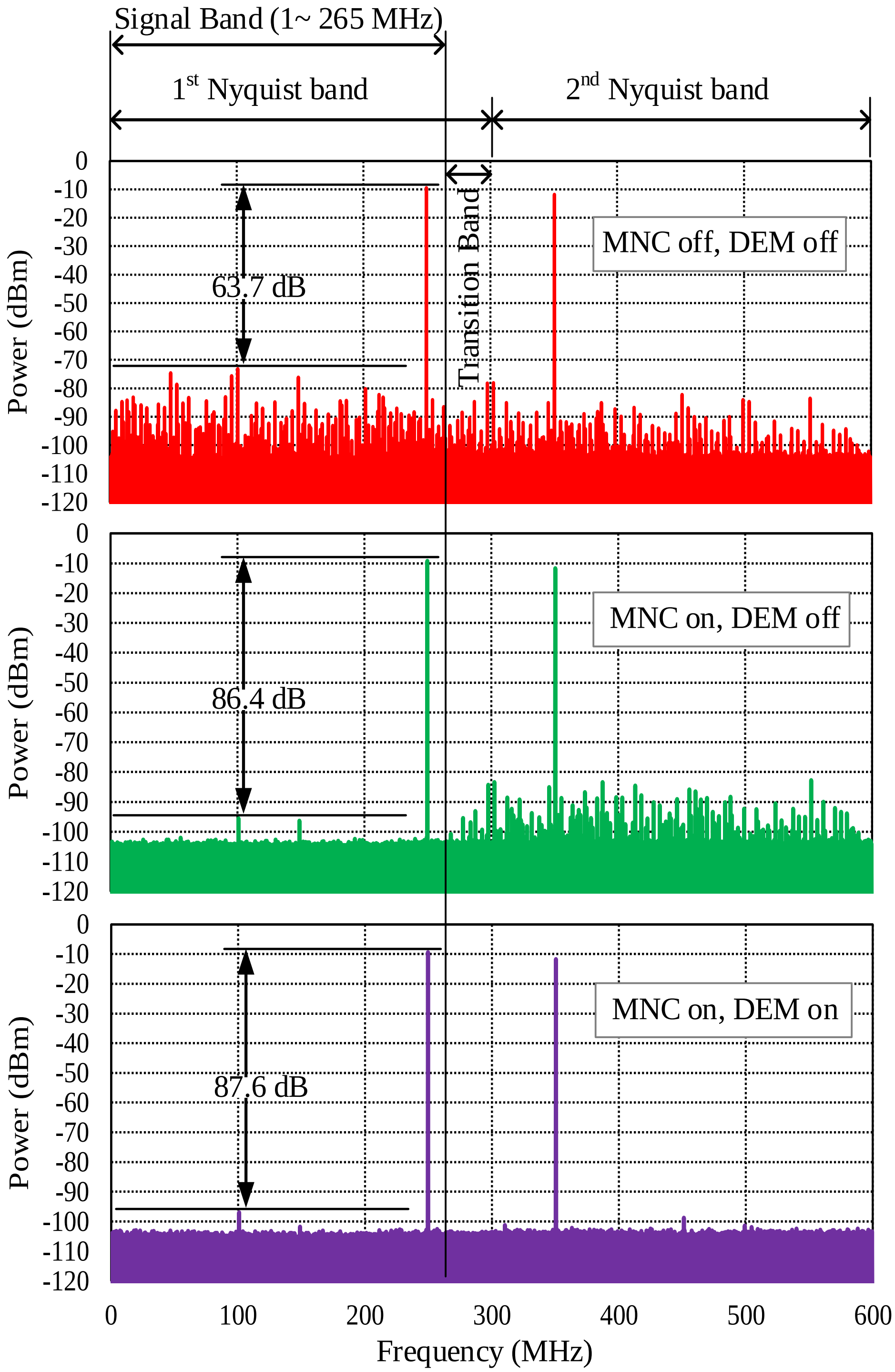


Figure 21: Measured output spectra for a full-scale 249.6 MHz input signal.

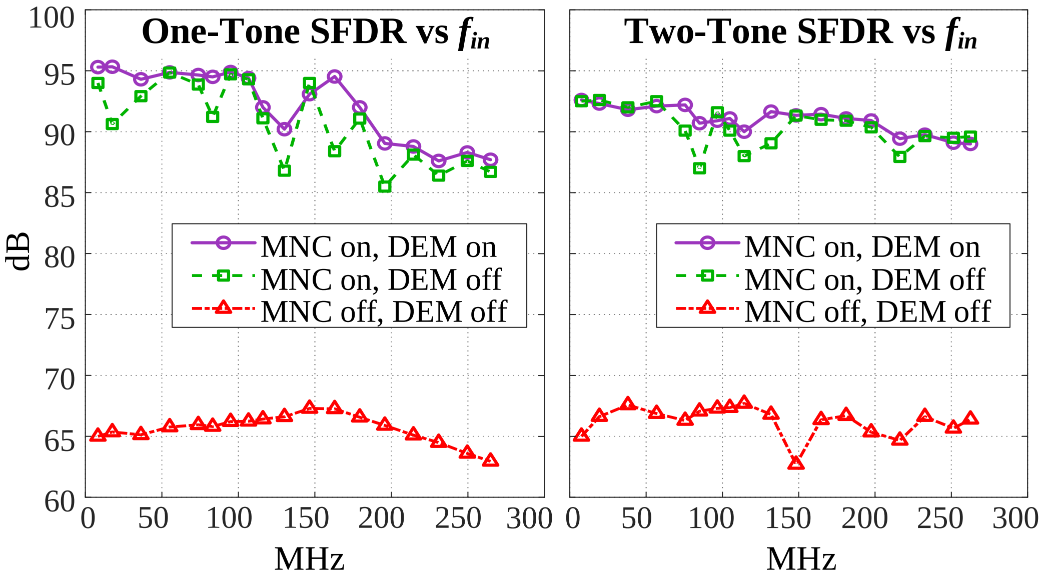


Figure 22: Measured SFDR versus frequency for one-tone input signals and two-tone input signals separated by 3.52 MHz.

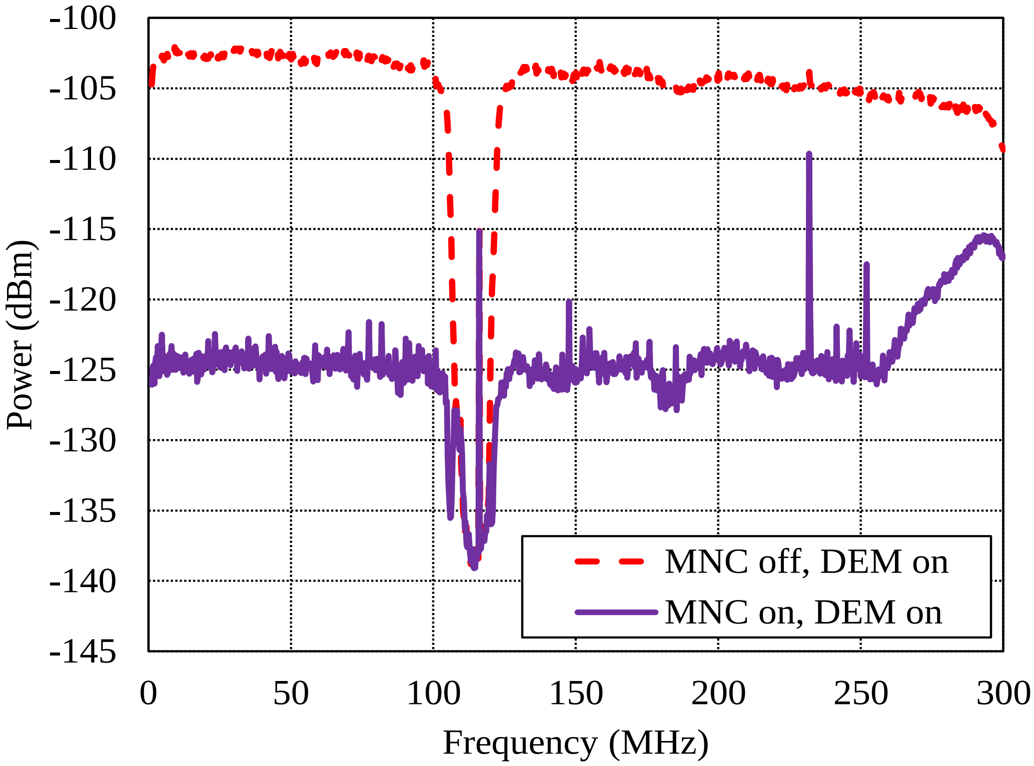


Figure 23: Measured output noise and distortion spectra.

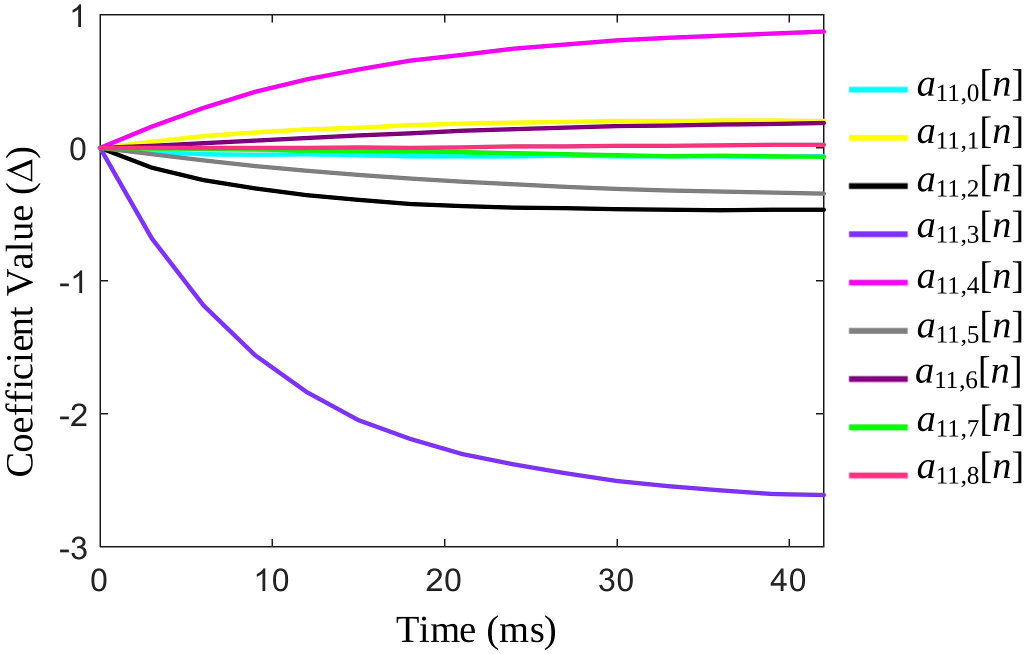


Figure 24: Representative plot of measured coefficient values versus time.

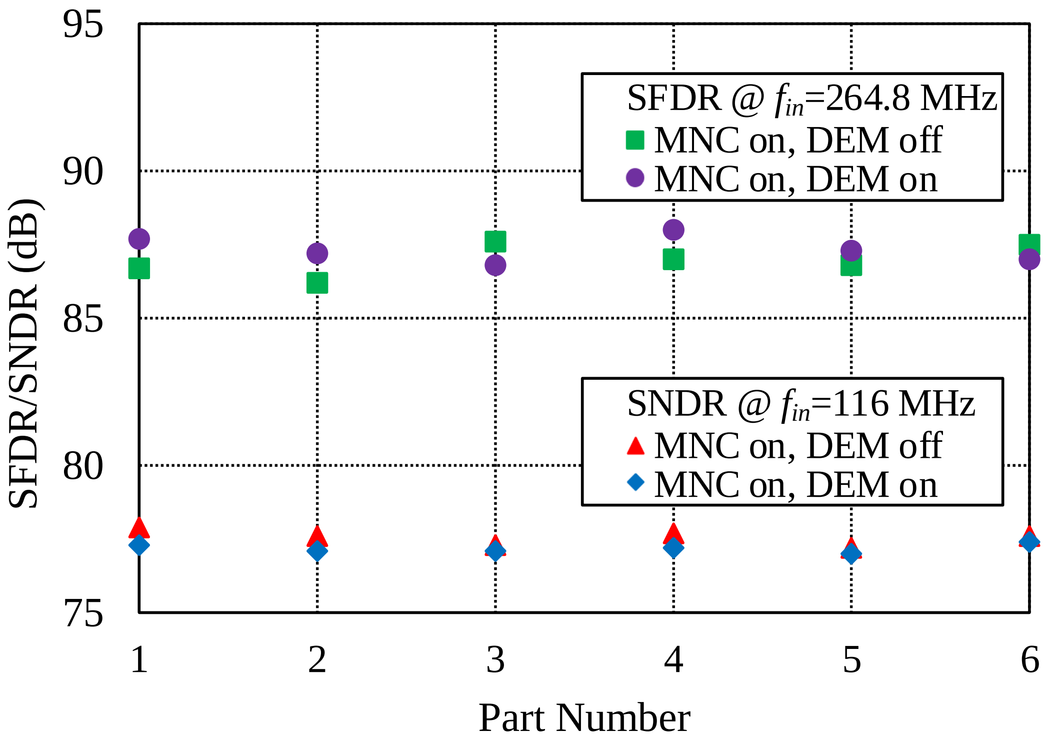
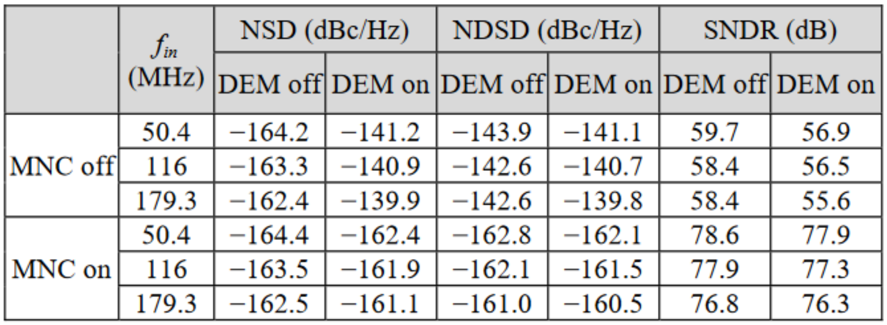


Figure 25: Measured SFDR and SNDR values across 6 parts.

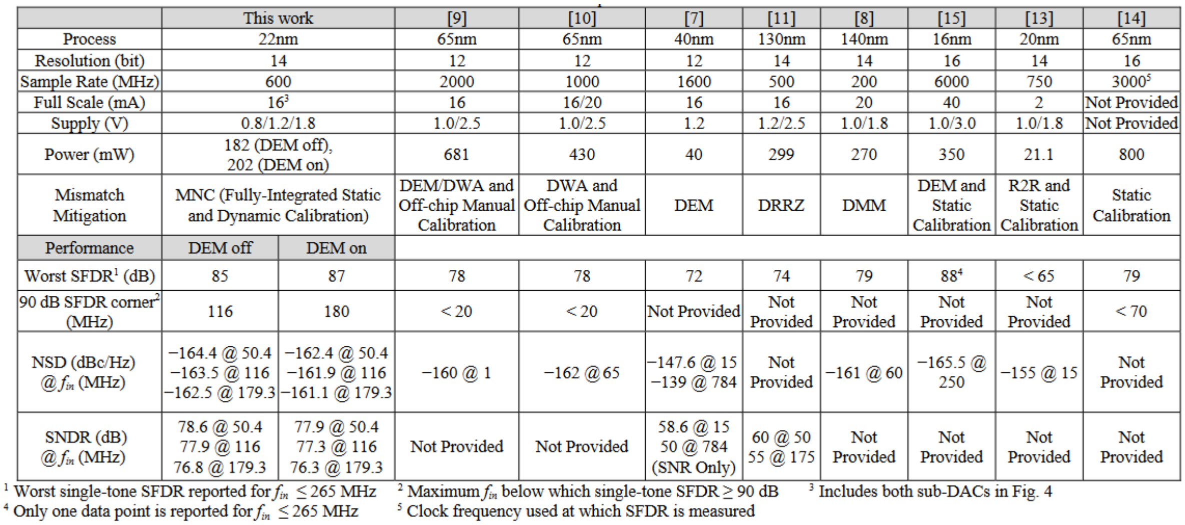
TABLES

Table 1: NSD/NDSD/SNDR Measurement Results



	f_{in} (MHz)	NSD (dBc/Hz)		NDSD (dBc/Hz)		SNDR (dB)	
		DEM off	DEM on	DEM off	DEM on	DEM off	DEM on
MNC off	50.4	−164.2	−141.2	−143.9	−141.1	59.7	56.9
	116	−163.3	−140.9	−142.6	−140.7	58.4	56.5
	179.3	−162.4	−139.9	−142.6	−139.8	58.4	55.6
MNC on	50.4	−164.4	−162.4	−162.8	−162.1	78.6	77.9
	116	−163.5	−161.9	−162.1	−161.5	77.9	77.3
	179.3	−162.5	−161.1	−161.0	−160.5	76.8	76.3

Table 2: Performance Table and Comparison to Prior State-of-the-Art DACs



	This work		[9]	[10]	[7]	[11]	[8]	[15]	[13]	[14]
Process	22nm		65nm	65nm	40nm	130nm	140nm	16nm	20nm	65nm
Resolution (bit)	14		12	12	12	14	14	16	14	16
Sample Rate (MHz)	600		2000	1000	1600	500	200	6000	750	3000[5]
Full Scale (mA)	16[3]		16	16/20	16	16	20	40	2	Not Provided
Supply (V)	0.8/1.2/1.8		1.0/2.5	1.0/2.5	1.2	1.2/2.5	1.0/1.8	1.0/3.0	1.0/1.8	Not Provided
Power (mW)	182 (DEM off), 202 (DEM on)		681	430	40	299	270	350	21.1	800
Mismatch Mitigation	MNC (Fully-Integrated Static and Dynamic Calibration)		DEM/DWA and Off-chip Manual Calibration	DWA and Off-chip Manual Calibration	DEM	DRRZ	DMM	DEM and Static Calibration	R2R and Static Calibration	Static Calibration
Performance	DEM off	DEM on								
Worst SFDR[1] (dB)	85	87	78	78	72	74	79	88[4]	< 65	79
90 dB SFDR corner[2] (MHz)	116	180	< 20	< 20	Not Provided	Not Provided	Not Provided	Not Provided	Not Provided	< 70
NSD (dBc/Hz) @ f_{in} (MHz)	−164.4 @ 50.4 −163.5 @ 116 −162.5 @ 179.3	−162.4 @ 50.4 −161.9 @ 116 −161.1 @ 179.3	−160 @ 1	−162 @ 65	−147.6 @ 15 −139 @ 784	Not Provided	−161 @ 60	−165.5 @ 250	−155 @ 15	Not Provided
SNDR (dB) @ f_{in} (MHz)	78.6 @ 50.4 77.9 @ 116 76.8 @ 179.3	77.9 @ 50.4 77.3 @ 116 76.3 @ 179.3	Not Provided	Not Provided	58.6 @ 15 50 @ 784 (SNR Only)	60 @ 50 55 @ 175	Not Provided	Not Provided	Not Provided	Not Provided

[1] Worst single-tone SFDR reported for $f_{in} \leq 265$ MHz
[2] Maximum f_{in} below which single-tone SFDR ≥ 90 dB
[3] Includes both sub-DACs in Fig. 4
[4] Only one data point is reported for $f_{in} \leq 265$ MHz
[5] Clock frequency used at which SFDR is measured

REFERENCES

1. Y. Cong and R. Geiger, “A 1.5 V 14-bit 100 MSPS self-calibrated DAC,” *IEEE Journal of Solid-State Circuits*, vol. 38, no.12, pp. 2051–2060, Dec. 2003.

2. M. Clara, W. Klatzer, B. Seger, A. Di Giandomenico, and L. Gori, “A 1.5V 200MS/s 13b 25mW DAC with Randomized Nested Background Calibration in 0.13 µm CMOS*,” IEEE International Solid State Circuits Conference,* February 2007.

3. Q. Huang, P. A. Francese, C. Martelli, and J. Nielsen,“A 200Ms/s 14b 97 mW DAC in 0.18µm CMOS,” *IEEE International Solid State Circuits Conference,* February 2004.

4. B. Catteau, P. Rombouts, J. Raman, and L. Weyten, “An on-line calibra- tion technique for mismatch errors in high-speed DACs,” *IEEE Trans. Circuits Syst.–I, Reg. Papers,* vol. 55, no. 7, pp. 1873–1883, Aug. 2008.

5. K. L. Chan, J. Zhu, and I. Galton, “Dynamic Element Matching to Prevent Nonlinear Distortion From Pulse-Shape Mismatches in High-Resolution DACs,” *IEEE Journal of Solid-State Circuits,* vol. 43, no. 9, pp. 2067-2078, September 2008.

6. K. L. Chan, N. Rakuljic, I. Galton, “Segmented Dynamic Element Matching for High-Resolution Digital-to-Analog Conversion,” *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 55, no. 11, pp. 3383-3392, December 2008.

7. W.-T. Lin, H.-Y. Huang, and T.-H. Kuo, “A 12-bit 40 nm DAC achieving SFDR > 70 dB at 1.6 GS/s and IMD < -61 dB at 2.8 GS/s with DEMDRZ technique,” *IEEE Journal of Solid-State Circuits,* vol. 49, no. 3, pp. 708–717, March 2014.

8. Y. Tang, J. Briaire, K. Doris, R. van Veldhoven, P. van Beek, H. Hegt, and A.van Roermund, “A 14 bit 200 MS/s DAC With SFDR >78 dBc, IM3 < −83 dBc and NSD < −163 dBm/Hz Across the Whole Nyquist Band Enabled by Dynamic-Mismatch Mapping,” IEEE Journal of Solid-State Circuits, vol. 46, no. 6, pp. 1371-1381, June 2011.

9. S. Su and M. S.-W. Chen, “A 12-bit 2 GS/s dual-rate hybrid DAC with pulse-error pre-distortion and in-band noise cancellation achieving > 74 dBc SFDR and < −80 dBc IM3 up to 1 GHz in 65 nm CMOS,” *IEEE Journal of Solid-State Circuits,* vol. 51, no. 12, pp. 2963–2978, December 2016.

10. S. Su, T.-I. Tsai, P. K. Sharma, and M. S.-W. Chen, “A 12 bit 1 GS/s dual-rate hybrid DAC with an 8 GS/s unrolled pipeline delta-sigma modulator achieving 75 dB SFDR over the Nyquist band,” *IEEE Journal of Solid-State Circuits,* vol. 50, no. 4, pp. 896–907, April 2015.

11. X. Li, Q. Wei, Z. Xu, J. Liu, H. Wang, and H. Yang, “A 14 bit 500 MS/s CMOS DAC using complementary switched current sources and time- relaxed interleaving

DRRZ," *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 61, no. 8, pp. 2337–2347, August 2014.

12. W.-H. Tseng, C.-W. Fan, and J.-T. Wu, "A 12-Bit 1.25-GS/s DAC in 90 nm CMOS with 70 dB SFDR up to 500 MHz," *IEEE Journal of Solid-State Circuits, vol. 46, no. 12,* pp. 2845–2856, December 2011.

13. S. M. Lee et al., "A 14 b 750 MS/s DAC in 20 nm CMOS with < −168 dBm/Hz noise floor beyond Nyquist and 79 dBc SFDR utilizing a low glitch-noise hybrid R-2R architecture," in *Symp. VLSI Circuits Dig.,* June 2015.

14. Engel, M. Clara, H. Zhu, and P. Wilkins, "A 16-bit 10 Gsps currentsteering RF DAC in 65 nm CMOS achieving 65 dBc ACLR multi-carrier performance at 4.5 GHz Fout," in *Symp. VLSI Circuits Dig.,* June 2015.

15. C.-H. Linet al., "A 16b 6 GS/S Nyquist DAC with IMD < −90 dBc up to 1.9 GHz in 16 nm CMOS," *IEEE International Solid State Circuits Conference,* February 2018.

16. D. Kong, I. Galton, "Adaptive Cancellation of Static and Dynamic Mismatch Error in Continuous-Time DACs," *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 65, no. 2, pp. 421–433, February 2018

17. J. Remple, I. Galton, "The Effects of Inter-Symbol Interference in Dynamic Element Matching DACs," *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 64, no. 1, pp. 14-23, January 1017.

18. R. Adams, K. Q. Nguyen, K. Sweetland, "A 113-dB SNR Oversampling DAC with Segmented Noise-Shaped Scrambling," *IEEE Journal of Solid State Circuits,* vol. 33, no. 12, pp. 1871-1878, December 1998.

19. G. Taylor and I. Galton, "A Mostly-Digital Variable-Rate Continuous-Time Delta-Sigma Modulator ADC," *IEEE Journal of Solid-State Circuits,* vol. 45, no. 12, pp. 2634–2646, December 2010.

20. P. P. Vaidyanathan, *Multirate Systems and Filter Banks,* Prentice Hall, 1993.

CHAPTER 3

SUBSAMPLING MISMATCH NOISE CANCELLATION FOR HIGH-SPEED CONTINUOUS-TIME DACS

Abstract— Clock skew and component mismatches in continuous-time DACs introduce two types of error: static error and dynamic error. Both types of error typically limit the performance of practical, high-resolution, continuous-time DACs, but most prior calibration techniques primarily reduce only static error. An exception is a recently published mismatch noise cancellation (MNC) technique that adaptively measures and cancels both types of error over the DAC's first Nyquist band. However, a disadvantage of the technique is that it requires an oversampling ADC that operates at several times the DAC's Nyquist rate to prevent convergence error that would otherwise be caused by aliasing. This paper presents a sub-sampling version of the MNC technique that avoids this limitation at the expense of a lower calibration convergence rate. As proven in the paper, the subsampling MNC technique allows aliasing to occur, but in such a way that convergence error is avoided.

I. INTRODUCTION

A continuous-time DAC generates an analog output pulse for each digital input code. Ideally, the output pulse during each clock interval is scaled by the DAC's input code value during that clock interval, and except for this scale-factor it has the same shape as all the other pulses. Unfortunately, non-ideal circuit behavior causes input-dependent deviations of both the scale-factor and shape of each output pulse. Error in a DAC's output waveform from pulse scale-factor deviations is called *static error* and that from pulse shape deviations is called *dynamic error*.

The most significant types of static and dynamic error in practical high-resolution continuous-time DACs are caused by 1) inadvertent but inevitable *clock skew* and *component mismatches*, 2) *inter-symbol interference* (ISI), and 3) *signal-dependent output impedance* [1-14]. For DACs implemented in present-day CMOS technology that target signal-to-noise-and-distortion ratios (SNDRs) of greater than about 65 dB, error from clock skew and component mismatches is the most significant limitation. Unlike the other types of error, analog circuit design and layout techniques to reduce error from clock skew and component mismatches below this level are not known.

Yet continuous-time DACs with SNDRs of greater than 65 dB are increasingly necessary in high-performance applications such as 4G and 5G cellular base station transmitters. In such cases, calibration techniques are necessary to suppress error from clock skew and component mismatches. Unfortunately, most prior digital calibration techniques primarily reduce only static error, which leaves dynamic error as a major limitation in high-performance continuous-time DACs [1-14].

The difficulty in suppressing dynamic error arises from a property inherent to continuous-time DACs. Each DAC output pulse has a bandwidth that far exceeds the DAC's sample-rate, because its duration is time-limited to one clock period. Therefore, a technique that cancels dynamic error must either have a bandwidth that is wider than the DAC's signal bandwidth, or must perform frequency selective cancellation over a single Nyquist band.

Recently, a mismatch noise cancellation (MNC) technique was developed that addresses this difficulty [15, 16]. It incorporates a feedback loop that measures and cancels both static and dynamic error caused by clock skew and component mismatches over the DAC's first Nyquist band. While the MNC technique solves the dynamic error problem, it requires an oversampling ADC that operates at many times the DAC's Nyquist rate. This ultimately limits the maximum achievable signal bandwidth for a given power consumption.

This paper presents a subsampling version of the MNC technique that avoids the oversampling requirement. The original version of the MNC technique requires oversampling to avoid aliasing that would otherwise cause convergence error in the technique's error cancellation feedback loop. The modified version does not prevent aliasing, but is designed such that the aliasing does not cause convergence error. By avoiding oversampling, the modified MNC technique removes the potential signal bandwidth limitation of the original version at the expense of a modest reduction in the feedback loop's convergence rate. The paper presents a rigorous mathematical analysis of the proposed technique, and demonstrates the results via computer simulations.

II. BACKGROUND INFORMATION: OVERSAMPLING MNC

Fig. 26 shows a high-level diagram of the IC presented in [16]. It consists of a 14-bit main DAC enclosed in an oversampling MNC feedback loop that adaptively measures and cancels static and dynamic error caused by clock skew and component mismatches within the main DAC over the first Nyquist band. The MNC feedback loop consists of an oversampling ADC, a lowpass decimation filter, a digital error estimator and a correction DAC.

The main DAC incorporates dynamic element matching (DEM) of the type presented in [17]. Its static and dynamic error resulting from clock skew and component mismatches, collectively referred to as mismatch noise in the remainder of this paper, has the form

$$e_{DAC}(t) = \sum_{k=1}^{35} d_k(t) s_k\left[n_t\right] \tag{64}$$

where n_t is the largest integer less than or equal to $f_s t$ with f_s = 600 MHz, each $d_k(t)$ is a 600 MHz periodic waveform that depends on clock skew and component mismatches within the main DAC, and each $s_k[n]$ sequence is generated by digital logic within the main DAC's DEM encoder [18]. Specifically, the $s_k[n]$ sequences are pseudo-random 600 MHz sample-rate sequences that take on values of −1, 0 and 1 and are uncorrelated with each other and with the main DAC's input sequence, $x[n]$. Consequently, $e_{DAC}(t)$ is wideband noise that is uncorrelated with $x[n]$ and free of harmonic distortion.

Without DEM, $e_{DAC}(t)$ would still be given by (64), but the $s_k[n]$ sequences would be deterministic, nonlinear functions of $x[n]$, so $e_{DAC}(t)$ would be entirely nonlinear distortion. Hence, DEM eliminates nonlinear distortion that would otherwise be caused by clock skew and component mismatches. However, it does so by converting the nonlinearity into noise, which severely degrades the DAC's signal-to-noise ratio (SNR). The purpose of the MNC feedback loop is to cancel this noise so as to keep the benefit of DEM without the SNR penalty.

The sampling theorem implies that for any $e_{DAC}(t)$ there must exist a correction DAC input sequence, $x_c[n]$, that would cause the correction DAC output waveform, $y_c(t)$, to cancel $e_{DAC}(t)$ over the first Nyquist band up to the accuracy of the correction DAC. As the dynamic range of $e_{DAC}(t)$ is much smaller than that of the main DAC, the resolution and step-size of the correction DAC, and, therefore, the error it introduces, are considerably smaller than those of the main DAC. Consequently, a 9-bit correction DAC with a step-size equal to a quarter that of the main DAC and no DEM or calibration was found to be sufficient in [16] to achieve more than 24 dB of error cancellation.

To make $y_c(t)$ well-approximate $e_{DAC}(t)$ over the first Nyquist band, the MNC feedback loop must measure $e_{DAC}(t)$ over the first Nyquist band. This requires a digitized version of the main DAC's output waveform that has been filtered to include only the first Nyquist band. The oversampling ADC and decimation filter in Fig. 26 perform this operation, so $r[n]$ contains a residual portion of $e_{DAC}(t)$ restricted to the first Nyquist band that is left over from imperfect MNC cancellation. Given that $e_{DAC}(t)$ is correlated with the $s_k[n]$ sequences as indicated by (64) and the decimation filter's impulse response is many 600 MHz samples long, it follows that the residual portion of $e_{DAC}(t)$ in $r[n]$ must be correlated with multiple time-shifted versions of the $s_k[n]$ sequences.

The MNC feedback loop measures the residual portion of $e_{DAC}(t)$ by correlating $r[n]$ with time-shifted versions of the 35 $s_k[n]$ sequences, and uses the measurement results to generate the correction DAC input sequence. Each of the 35 $s_k[n]$ residue estimators in the digital error estimator consists of a coefficient calculator block and an FIR filter with input $s_k[n+P]$ as shown in Fig. 26c.[7] The coefficient calculator correlates $r[n]$ with $N = 9$ time-shifted versions of $s_k[n]$. Each correlation is performed by multiplying $r[n]$ by a time-shifted version of $s_k[n]$ (which is −1, 0, or 1 during each 600 MHz clock period), and the result is scaled by $K = 8\cdot10^{-6}$ and accumulated. The accumulator outputs, $\alpha_{k,0}[n]$, $\alpha_{k,0}[n]$, …, $\alpha_{k,8}[n]$, form the impulse response of the FIR filter, so each $s_k[n]$ residue estimator operates as an adaptive FIR filter. The 35 adaptive filters converge as necessary for $y_c(t)$ to well-approximate $e_{DAC}(t)$ over the first Nyquist band as proven in [15].

The MNC technique can operate either as a foreground or background calibration technique. While $e_{DAC}(t)$ is a broadband $x[n]$-dependent waveform, the $d_k(t)$ waveforms and the digital error estimator's target FIR filter coefficients depend primarily on component mismatches, clock skew, and other parameters that do not change significantly over time. Hence, the IC in [16] runs the MNC feedback loop during foreground calibration, and subsequently freezes the FIR filter coefficients and disables the ADC during normal DAC operation.

III. SUBSAMPLING MNC

As explained in [15], the accuracy required of the oversampling MNC technique's ADC is modest, e.g., in the IC presented in [16] the ADC's SNDR is less than 30 dB while the post-calibration signal-band SNDR of the DAC is over 77 dB. Yet the oversampling requirement poses a practical problem for DAC samples-rates above a few GHz. For instance, modifying the IC presented in [16] to have a DAC sample-rate of 6 GHz, would require an ADC with a sample-rate of about 30 GHz. While low-SNDR ADCs at such high sample-rates are not impossible, a modified MNC technique that allows for an ADC sample-rate closer to that of the DAC would be preferable in terms of reducing power consumption, all other things being the same.

[7] In the IC presented in [16] P, Q, and N are set to 3, 21, and 9, respectively.

A. MNC Convergence Accuracy in the Presence of Aliasing

If the oversampling ADC and decimation filter in Fig. 26 were replaced by a Nyquist-rate ADC sampled at the same rate as the main DAC, the ADC output would contain all of the content of the main DAC's Nyquist bands aliased down onto its signal band. As each of the main DAC's Nyquist bands contains components correlated to the $s_k[n]$ sequences, the digital error estimator would adaptively cancel the sum of the error from all the aliased bands simultaneously, but it would fail to cancel error in any one of the Nyquist bands individually. This problem could be solved by inserting an anti-aliasing filter prior to the ADC, but this is not a practical option given the wide bandwidth and narrow transition band required of the filter.

Although it is necessary to avoid aliasing in the oversampling version of the MNC technique to measure the necessary MNC FIR filter coefficients, the following line of reasoning implies that it is at least mathematically possible to measure the necessary MNC FIR filter coefficients in the presence of aliasing. The output of the correction DAC in Fig. 1a has the form $y_c(t) = \alpha_c(t)x_c[n_t]$ where $\alpha_c(t)$ is a 600 MHz periodic waveform [18]. As shown in [15], the MNC feedback loop causes the impulse response of the kth FIR filter in Fig. 26c to converge such that the filter's transfer function well-approximates

$$H_k\left(e^{j\omega T_s}\right) = e^{-j\omega PT_s}\frac{D_{p\text{-}k}(j\omega)}{A_{p\text{-}c}(j\omega)} \text{ for } |\,\omega| \leq \pi f_s \tag{65}$$

where f_s = 600 MHz, $T_s = 1/f_s$, and $D_{p\text{-}k}(j\omega)$ and $A_{p\text{-}c}(j\omega)$ are the are the continuous-time Fourier transforms of one period of the T_s-periodic waveforms $d_k(t)$ and $\alpha_c(t)$, respectively.

It follows from (65) that the FIR filter coefficients could be calculated directly from one period of $\alpha_c(t)$ and one period of each $d_k(t)$ for k = 1, 2, …, 35, and they could be calculated approximately from sampled versions of these 35 one-period waveforms.

Moreover, the samples could be measured directly from the main DAC and correction DAC outputs. For example, to measure five samples of $\alpha_c(t)$ over one f_s-rate clock period the input to the correction DAC could be set to a non-zero constant value, and the five samples could be measured at its output over one clock period. Although more complicated, each of the $d_k(t)$ waveforms could be isolated by appropriately manipulating the DEM encoder and then similarly sampled.

This procedure would still require oversampling, but it can be further modified to avoid oversampling by recognizing that the measurements described above could be spread over five clock periods rather than over a single clock period. As depicted in Fig. 27, the f_s-rate periodicity of the $\alpha_c(t)$ and $d_k(t)$ waveforms ensures that an ADC sampled at a rate of $5f_s/6$ would collect the same information over a duration of $6T_s$ as an ADC sampled at a rate of $5f_s$ would collect over a duration of T_s, where $T_s = 1/f_s$. Hence, oversampling can be avoided at the expense of a longer data collection duration.

The argument above is the outline of a proof-by-construction that subsampling MNC is mathematically possible. However, the constructed procedure would only work as a foreground calibration technique, whereas the oversampling MNC technique works as either a foreground or background calibration technique, and it would be computationally expensive.

B. The Subsampling MNC Technique

A more practical way of exploiting the effect described above is the proposed subsampling MNC (SMNC) technique shown in Fig. 28. It differs from the oversampling MNC technique in three ways: an $Rf_s/(R+1)$-rate subsampling ADC is used in place of the Rf_s-rate oversampling ADC, where R is an integer greater than 1 (Fig. 26 is drawn for the specific case of R = 5), a fractional decimation filter is used in place of the lowpass decimation filter, and a bank of latches updated at times n = 0, (R+1), 2(R+1), … separate each coefficient calculator and FIR filter. The fractional decimation filter is equivalent to the cascade of an R+1-fold up-sampler, a digital filter with impulse response $g[m]$, and an R-fold down-sampler, but it can be implemented as the polyphase structure shown in Fig. 29 such that all its components are clocked at a rate of f_s [19]. Therefore, the highest clock-rate in the system is f_s.

The ADC sample-rate is slightly lower than f_s whereas the DAC output spectra are non-zero over several $f_s/2$-wide Nyquist bands. Therefore, the ADC output, $w[q]$, contains significant aliasing. However, as explained shortly, the subsampling effect depicted in Fig. 27 (for the specific case of R = 5) prevents the aliasing from causing MNC convergence error. In particular, as proven in the remainder of the paper the subsampling MNC technique converges to the same set of FIR filter coefficients as the original oversampling MNC technique, but with a lower convergence rate.

To show that the SMNC technique converges to the same FIR filter coefficients as the oversampling MNC technique, it is helpful to first redraw Fig. 28a in an equivalent form that is easier to compare to Fig. 26a. Theorem 1 presented below provides this equivalent form.

Theorem 1: The system shown in Fig. 30 with

$$g^{(l)}[m] = \begin{cases} g[m], & \text{if } \left(l-m\right) \bmod (R+1) = 0, \\ 0, & \text{otherwise}, \end{cases} \tag{66}$$

and

$$t[n] = r^{(l)}[n] \quad \text{where} \quad l = (-n) \bmod (R+1), \tag{67}$$

(i.e., $t[n]$ is the output of the R+1 to 1 multiplexer) generates the same $t[n]$, $x_c[n]$, $y_c(t)$, $y(t)$, and $v(t)$ as that shown in Fig. 28a if both systems start with the same initial conditions and have the same input sequence, $x[n]$.

Proof: It follows from the definition of an up-sampler that the output of the (R+1)-fold up-sampler in Fig. 28a can be written as $d[m]p[m]$ where $d[m]$ is the output of an Rf_s sample-rate ADC in Fig. 30 and

$$p[m] = \begin{cases} 1, & \text{if } m \bmod (R+1) = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{68}$$

This and the signal processing shown in Fig. 28a imply that

$$t[n] = \sum_{m=-\infty}^{Rn} d[m]p[m]g[Rn-m] \tag{69}$$

in Fig. 29. The signal processing shown in Fig. 30 and (67) imply that

$$t[n] = \sum_{m=-\infty}^{Rn} d[m]g^{((-n) \bmod (R+1))}[Rn-m] \tag{70}$$

in Fig. 30. Therefore, it is enough to show that the right sides of (69) and (70) are equal, which is equivalent to showing that

$$g^{((-n) \bmod (R+1))}[Rn-m] = p[m]g[Rn-m]. \tag{71}$$

Given that [($-n$) mod (R+1)] – m] mod (R+1) = (–n – m) mod (R + 1), (66) implies

$$g^{((-n) \bmod (R+1))}[m] = \begin{cases} g[m], & \text{if } (-n-m) \bmod (R+1) = 0, \\ 0, & \text{otherwise,} \end{cases} \tag{72}$$

Given that [$-n$ − (Rn – m)] mod (R + 1) = m mod (R + 1), replacing m with Rn − m in (72) results in

$$\begin{aligned} & g^{((-n) \bmod (R+1))}[Rn-m] \\ & \qquad = \begin{cases} g[Rn-m], & \text{if } m \bmod (R+1) = 0, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{73}$$

Substituting (68) into the right side of (71) results in the right side of (73), which shows that (71) holds.

□

The SMNC equivalent system of Fig. 30 is a useful analysis tool because it can be related to the original oversampling MNC technique as follows. Equation (66) implies that

$$\sum_{l=0}^{R} g^{(l)}[m] = g[m] \tag{74}$$

and Fig. 30 implies that

$$r^{(l)}[n] = \sum_{m=-\infty}^{Rn} d[m]g^{(l)}[Rn-m], \tag{75}$$

so

$$\sum_{l=0}^{R} r^{(l)}[m] = \sum_{m=-\infty}^{Rn} d[m]g[Rn-m]. \tag{76}$$

The right side of (76) is equal to $r[n]$ in the oversampling MNC technique shown in Fig. 26 (generalized with 600 MHz replaced by f_s and 3 GHz replaced by Rf_s). Therefore, the output of the oversampling MNC technique's decimation filter can be written as

$$r[n] = \sum_{l=0}^{R} r^{(l)}[m], \tag{77}$$

with $r^{(l)}[n]$ given by (75).

It follows that $t[n]$ in Fig. 28a (which is identical to that in Fig. 30 as implied by Theorem 1) is different from $r[n]$ in Fig. 26, even when the $v(t)$ waveforms in the two systems are equal. In particular, for equal $v(t)$ waveforms in the two systems, $t[n]$ in Fig. 28a for each n is equal to one of the $r^{(l)}[n]$ sequences whereas $r[n]$ in Fig. 26a is equal to the sum of the $r^{(l)}[n]$ sequences. This difference between $t[n]$ and $r[n]$ is the result of aliasing caused by the SMNC technique's subsampling. As explained in Section III-A, the oversampling ADC is required in Fig. 26a to prevent aliasing that would cause convergence error. However, as proven in the next section, the SMNC technique converges correctly despite the aliasing caused by subsampling.

A qualitative explanation of this paradox is as follows. During foreground calibration, $x[n]$ is chosen such that the statistics of the $s_k[n]$ sequences do not change over time. The latches following each coefficient calculator in Fig. 28c ensure that R+1 samples of $t[n]$ are correlated against the shifted versions of the $s_k[n]$ sequences before the FIR filter coefficients are updated, and it follows from (66) that each of the $g^{(l)}[n]$ impulse responses have only one non-zero value for each set of R+1 samples. These observations imply that the average change of each coefficient calculator's accumulator during each set of R+1 samples is the same as it would be if $t[n]$ were replaced by $r[n]$ as given by (77) and the corresponding coefficient calculator were updated on just the first of every R+1 samples. Thus, instead of performing correlations on all R+1 of the $r^{(l)}[n]$ sequences simultaneously at each sample time, n, as done by the oversampling MNC technique, the SMNC technique equivalently performs correlations on all R+1 of the $r^{(l)}[n]$ sequences sequentially over successive sets of R+1 sample times.

C. Extension to Background Operation

With the $s_k[n]$ residue estimators implemented as shown in Fig. 28c, it is necessary for the statistics of the $s_k[n]$ sequences to be time-invariant as described above. This is easy to achieve during foreground calibration by ensuring that the statistics of $x[n]$ do not change over time. During background calibration, though, $x[n]$ is arbitrary, so it cannot be assumed that its statistics are time-invariant.

This problem can be solved by modifying the $s_k[n]$ residue estimators during background calibration as follows. The main DAC's DEM encoder ensures that the probability distribution of each $s_k[n]$ conditioned on $s_k[n] \neq 0$ is constant and independent of $x[n]$ [17]. Therefore, the problem can be solved by applying two changes to Fig. 28c during background calibration. The first change is to only update the bank of latches once every accumulator has been clocked R+1 times since the last time the bank of latches was clocked. The second change is to only clock the mth accumulator when $s_k[n+P-Q-m] \neq 0$ and n mod (R+1) is distinct from n' mod (R+1) for every prior time index n' of $s_k[n'+P-Q-m] \neq 0$ since the last time the bank of latches was updated. These modifications ensure that each accumulator in the kth coefficient calculator is updated with $r^{(l)}[n]$ information once for each value of l = 0, 1, …, R prior to each time the bank of latches is clocked and that the probability distribution of each $s_k[n]$ when the accumulators are updated is time-invariant.

IV. CONVERGENCE ANALYSIS

Each $r^{(l)}[n]$ sequence in Fig. 30 can be written as

$$r^{(l)}[n] = r^{(l)}_{ideal}[n] + r^{(l)}_e[n] + r^{(l)}_c[n] \tag{78}$$

where $r^{(l)}_{ideal}[n]$ is what $r^{(l)}[n]$ would have been without the main DAC’s mismatch noise and without the SMNC feedback loop, $r^{(l)}_e[n]$ represents error that would have been caused by the main DAC’s mismatch noise without the SMNC feedback loop, and $r^{(l)}_c[n]$ represents the effect of the SMNC feedback loop. The correction DAC’s error can be neglected, because it is much smaller than that of the main DAC as explained in Section II. Consequently, the relationship between $x_c[n]$ and $r^{(l)}_c[n]$ well-approximates that of a linear time-invariant (LTI) discrete-time system with impulse response $-h^{(l)}_c[n]$ (the negative sign simplifies the subsequent analysis). The system is causal and at least one clock delay is introduced by the ADC, so $h^{(l)}_c[n] = 0$ for all $n < 1$. Therefore,

$$r^{(l)}_c[n] = \sum_{i=1}^{\infty} x_c[n-i]\left(-h^{(l)}_c[i]\right), \tag{79}$$

where, as can be seen from Fig. 28,

$$x_c[n] = \sum_{k=1}^{M}\sum_{m=0}^{N-1} a_{k,m}[n] s_k[n+P-m]. \tag{80}$$

The *k*th portion of the main DAC's mismatch noise, $d_k(t)s_k[n_t]$ in (64), has the same form as the output of a DAC with input sequence $s_k[n]$ and T_s-periodic pulse shaping waveform, $d_k(t)$. Thus, the relationship between $s_k[n]$ and its contribution to $r_e^{(l)}[n]$ must also be that of a causal LTI discrete-time system with at least one clock delay. Denoting the LTI system's impulse response as $b_k^{(l)}[n]$, it follows from (64) that

$$r_e^{(l)}[n] = \sum_{k=1}^{M}\sum_{i=1}^{\infty} s_k[n-i] b_k^{(l)}[i]. \tag{81}$$

It follows from (67) that

$$t[n-l] = r^{(l)}[n-l] \text{ if } n \bmod (R+1) = 0. \tag{82}$$

As indicated in Fig. 28c, each FIR filter coefficient, $\alpha_{k,m}[n]$, only changes at times n = 0, R+1, 2(R+1), …, i.e., when n mod (R+1) = 0. Therefore, Fig. 28c and (82) imply that for each of these values of n and for each m = 0, 1, …, N−1,

$$a_{k,m}\left[n\right] = a_{k,m}\left[n-1\right] + K\sum_{l=0}^{R} s_k\left[n-l+P-Q-m\right] r^{(l)}[n-l]. \tag{83}$$

For all other values of n, $\alpha_{k,m}[n]$ = $\alpha_{k,m}[n-1]$. Substituting (79)-(81) into (78), and substituting the result into (83), implies that

$$\begin{aligned} a_{k,m}\left[n\right] = a_{k,m}\left[n-1\right] + \sum_{l=0}^{R}\Bigg\{ K\cdot s_k{}^2\left[n-l+P-Q-m\right]\Bigg(b_k^{(l)}\left[Q-P+m\right] \\ -\sum_{q=0}^{N-1} a_{k,q}\left[n-l-Q-m+q\right] h_c^{(l)}\left[Q+m-q\right]\Bigg) + Ke_{k,m}^{(l)}\left[n-l\right]\Bigg\} \end{aligned} \tag{84}$$

for each n that satisfies n mod (R+1) = 0 and m = 0, 1, …, N−1, where

$$
\begin{aligned}
e_{k,m}^{(l)}\left[n\right] = s_k\left[n+P-Q-m\right]\Bigg\{\sum_{i=1}^{\infty}\sum_{\substack{j=1\\ j\neq k}}^{35}\Bigg(s_j\left[n-i\right]b_j^{(l)}\left[i\right] \\
-\sum_{q=0}^{N-1} a_{j,q}\left[n-i\right]s_j\left[n+P-i-q\right]h_c^{(l)}\left[i\right]\Bigg)+\Bigg(\sum_{\substack{i=1\\ i\neq Q-P+m}}^{\infty} s_k\left[n-i\right]b_k^{(l)}\left[i\right] \qquad (85) \\
-\sum_{\substack{i=1\\ i\neq Q+m-q}}^{\infty}\ \sum_{q=0}^{N-1} a_{k,q}\left[n-i\right]s_k\left[n+P-i-q\right]h_c^{(l)}\left[i\right]\Bigg)+r_{ideal}^{(l)}\left[n\right]\Bigg\}.
\end{aligned}
$$

Equations (84), for m = 0, 1, …, N−1 and each n that satisfies n mod (R+1) = 0, can be written in matrix form as

$$
\begin{aligned}
\mathbf{a}_k\left[n\right] = \mathbf{a}_k\left[n-1\right]+K\sum_{m=0}^{N-1}\sum_{l=0}^{R} s_k{}^2\left[n-l+P-Q-m\right]\Bigg(\mathbf{b}_{k,m}^{(l)} \\
-\sum_{q=0}^{N-1}\mathbf{H}_{m,q}^{(l)}\mathbf{a}_k\left[n-l-Q-m+q\right]\Bigg)+K\mathbf{e}_k\left[n\right] \qquad (86)
\end{aligned}
$$

where

$$\mathbf{a}_k[n] = \left[a_{k,0}[n], a_{k,1}[n], \cdots,\ a_{k,N-1}[n]\right]^T, \qquad (87)$$

$\mathbf{H}_{m,q}^{(l)}$ is an N×N matrix given by

$$\mathbf{H}_{m,q}^{(l)} = \left[\ h_{j,k} = \begin{cases} h_c^{(l)}[Q+j-k], & \text{if } j=m, k=q, \\ 0, & \text{otherwise,} \end{cases}\right], \qquad (88)$$

$\mathbf{b}_{k,m}^{(l)}$ is an N×1 vector given by

$$\mathbf{b}_{k,m}^{(l)} = \left[\ b_j = \begin{cases} b_k^{(l)}\left[Q-P+j\right], & \text{if } j=m, \\ 0, & \text{otherwise,} \end{cases}\right], \qquad (89)$$

and $\mathbf{e}_k[n]$ is an N×1 vector given by

$$\mathbf{e}_k[n] = \sum_{l=0}^{R}\left[e_{k,0}^{(l)}[n-l], e_{k,1}^{(l)}[n-l],\ \ldots, e_{k,N-1}^{(l)}[n-l]\right]^T. \qquad (90)$$

The $\mathbf{a}_k[n]$ vector represents the kth adaptive FIR filter’s coefficients at time n. The loop gain, K, is small by design to ensure that the coefficients converge to values with low variances, so (86) implies that $\mathbf{a}_k[n]$ depends only very weakly on any one of the time-shifted $s_k[n]$ sequences. Furthermore, all of the time-shifted $s_k[n]$ sequences are statistically independent. Consequently, $\mathbf{a}_k[n]$ is well-approximated as being statistically independent of each time-shifted $s_k[n]$ sequence. This type of *independence assumption* is widely used in the analysis of adaptive filters wherein slowly updated adaptive filter coefficients are assumed to be approximately independent from the data processed by the system [20-22].

Expanding the right side of (85) results in a sum of several products. Of these, $s_k[n+P-Q-m]s_j[n+P-i-q]\alpha_{j,q}[n-i]h_c^{(l)}[i]$ and $s_k[n+P-Q-m]s_k[n+P-i-q]\alpha_{k,q}[n-i]h_c^{(l)}[i]$ are the only products whose means are not exactly zero. However, their means are nearly zero by the independence assumption because $s_k[n+P-Q-m]s_j[n+P-i-q]$ and $s_k[n+P-Q-m]$ $s_k[n+P-i-q]$ are zero mean. This implies that the mean of $\mathbf{e}_k[n]$ is well-approximated as zero, i.e.,

$$\overline{\mathbf{e}}_k[n] = \mathbf{0}\,. \tag{91}$$

Given that $s_k[n]$ is restricted to values of -1, 0, and 1, and its statistics are time-invariant, the mean of $s_k^2[n]$ is a constant, c_k, between 0 and 1, i.e.,

$$\overline{s_k{}^2}[n] = c_k \quad \text{for all } n \quad \text{where } \; 0 < c_k \leq 1. \tag{92}$$

Taking the expectation of (86), and applying (91), (92), and the independence assumption yields

$$\begin{aligned}\overline{\mathbf{a}}_k\left[n\right] = {} & \overline{\mathbf{a}}_k\left[n-1\right] \\ & + c_k K \sum_{m=0}^{N-1}\sum_{l=0}^{R}\left(\mathbf{b}_{k,m}^{(l)} - \sum_{q=0}^{N-1}\mathbf{H}_{m,q}^{(l)}\overline{\mathbf{a}}_k\left[n-l-Q-m+q\right]\right)\end{aligned} \tag{93}$$

where $\overline{\mathbf{a}}_k[n]$ is the mean $\mathbf{a}_k[n]$ for each n that satisfies n mod (R+1) = 0. This can be rewritten as

$$\begin{aligned}\bar{\mathbf{a}}_k\left[n\right] &= \bar{\mathbf{a}}_k\left[n-1\right]\\ &\quad -c_k K\sum_{m=0}^{N-1}\sum_{l=0}^{R}\sum_{q=0}^{N-1}\mathbf{H}_{m,q}^{(l)}\bar{\mathbf{a}}_k\left[n-l-Q-m+q\right]+c_k K\mathbf{b}_k,\end{aligned} \tag{94}$$

where

$$\mathbf{b}_k = \sum_{m=0}^{N-1}\sum_{l=0}^{R}\mathbf{b}_{k,m}^{(l)} . \tag{95}$$

A simplification can be made by defining $\mathbf{H}_{\mathbf{c}}^{(J)}$ to be the sum of all $\mathbf{H}_{m,q}^{(l)}$ over l = 0, 1, …, R, m = 0, 1, …, N−1, and q = 0, 1, …, N−1, restricted to values of m, l, and q that satisfy l+Q+m–q = J, such that

$$\sum_{m=0}^{N-1}\sum_{l=0}^{R}\sum_{q=0}^{N-1}\mathbf{H}_{m,q}^{(l)} = \sum_{J=1}^{R+Q+N-1}\mathbf{H}_{\mathbf{c}}^{(J)} . \tag{96}$$

The lower limit of J is 1 because (88) implies that $\mathbf{H}_{m,q}^{(l)} = \mathbf{0}$ for m–$q \le -Q$ given that $h_c^{(l)}[n] = 0$ for all $n \le 0$. Applying (96) to rearrange the triple sum in (94) and applying $\bar{\mathbf{a}}_k[n] = \bar{\mathbf{a}}_k[n-1]$ for values of n that satisfy n mod (R+1) ≠ 0 gives

$$\begin{aligned}&\bar{\mathbf{a}}_k[n] = \bar{\mathbf{a}}_k[n-1]\\ &-\begin{cases} c_k K\displaystyle\sum_{J=1}^{R+Q+N-1}\mathbf{H}_{\mathbf{c}}^{(J)}\bar{\mathbf{a}}_k[n-J]+c_k K\mathbf{b}_k, & \text{if } n \bmod (R+1)=0,\\ \mathbf{0}, & \text{otherwise,}\end{cases}\end{aligned} \tag{97}$$

for each integer, n.

Equation (97) is an N-dimensional matrix difference equation that converges if and only if $\bar{\mathbf{a}}_k[n] \to \mathbf{a}_k'$ as $n \to \infty$, where $\mathbf{a}_k'$ is a constant vector. Taking the limit of (97) as $n \to \infty$ implies that if the system converges then

$$\mathbf{a}_k' = \mathbf{a}_k' - c_k K\mathbf{H}_{\mathbf{c}}\mathbf{a}_k' + c_k K\mathbf{b}_k \tag{98}$$

where

$$\mathbf{H}_{\mathbf{c}} = \sum_{J=1}^{R+Q+N-1}\mathbf{H}_{\mathbf{c}}^{(J)} . \tag{99}$$

It follows from (98) that if the system converges, then

$$\mathbf{a}_k' = \mathbf{H}_{\mathbf{c}}^{-1}\mathbf{b}_k . \tag{100}$$

Equations (88), (96) and (99) imply that

$$\mathbf{H_c} = \left[h_{j,k} = h_c[Q + j - k] \right], \quad \text{where} \quad h_c[n] = \sum_{l=0}^{R} h_c^{(l)}[n]. \tag{101}$$

Given that $-h_c^{(l)}[n]$ is the impulse response of the transfer function between $x_c[n]$ and $r_c^{(l)}[n]$, it follows from (77) and Theorem 1 in Section III that $h_c[n]$ is the impulse response of the transfer function between $x_c[n]$ and $r[n]$ in the oversampling version of the MNC technique shown in Fig. 26. As proven in [15], the FIR filter coefficients in the oversampling MNC technique converge to values that satisfy (100) with $\mathbf{H_c}$ = [$h_{j,k}$ = $h_c[Q+j-k]$]. Therefore, provided the FIR filter coefficients in the subsampling version of the MNC technique converge, they must converge to the same values as those of the oversampling version of the MNC technique.

It remains to show that the subsampling MNC technique's coefficients converge, i.e., that $\bar{\mathbf{a}}_k[n]$ always converges to $\mathbf{a}_k'$ as $n \rightarrow \infty$ for each k. This is done by showing that $\mathbf{z}_k[n]$ converges to $\mathbf{0}$ as $n \rightarrow \infty$, where

$$\mathbf{z}_k[n] = \bar{\mathbf{a}}_k[n] - \mathbf{a}_k\,', \tag{102}$$

and, as implied by (97) and (98),

$$\mathbf{z}_k[n] = \mathbf{z}_k[n-1] - \begin{cases} c_k K \displaystyle\sum_{J=1}^{R+Q+N-1} \mathbf{H}_\mathbf{c}^{(J)} \mathbf{z}_k[n-J] & \text{if } n \bmod (R+1) = 0 \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{103}$$

The analysis makes use of vector and matrix norms. For any N-dimensional vector $\mathbf{v}$ = [v_j] and N×N matrix $\mathbf{H}$ = [$h_{j,k}$], the *vector norm* of $\mathbf{v}$ and the *matrix norm* of $\mathbf{H}$ are defined as

$$\left\| \mathbf{v} \right\| = \max_{1 \le m \le N} \left| v_m \right| \quad \text{and} \quad \left\| \mathbf{H} \right\|_1 = \max_{1 \le m \le N} \sum_{n=1}^{N} \left| h_{m,n} \right|. \tag{104}$$

Theorem 2 presented below, and proven in the appendix, shows that $\mathbf{z}_k[n]$ converges to $\mathbf{0}$ as $n \to \infty$ for each k provided that $h_c[n]$, Q, and K satisfy certain conditions. It does so by showing that $\|\mathbf{z}_k[n]\| \to 0$ as $n \to \infty$. To simplify the notation, the system's initial conditions are taken to be zero, i.e., $\overline{\mathbf{a}}_k[n] = \mathbf{0}$ for all $n < 0$, so (102) implies that $\mathbf{z}_k[n] = -\mathbf{a}_k'$ for all $n < 0$.

Theorem 2: Suppose $0 \leq r < 1$, $0 < g < 1$, $0 < 2Kh_c[Q] < 1$, and $\mathbf{z}_k[n] = -\mathbf{a}_k'$ for all $n < 0$, where

$$r = \frac{1}{h_c[Q]} \sum_{\substack{m=Q-(N-1) \\ m \neq Q}}^{Q+(N-1)} \left| h_c[m] \right| \tag{105}$$

$$g = \sum_{J_1=1}^{R+Q+N-1} \sum_{J_2=1}^{R+Q+N-1} \frac{\left\| \mathbf{H}_\mathbf{c}^{(J_1)} \mathbf{H}_\mathbf{c}^{(J_2)} \right\|_1 \left(1 - \left(1 - 2Kh_c[Q]\right)^{J_1 - 1}\right)}{2h_c^{\,2}[Q]\left(1-r\right)\left(1 - 2Kh_c[Q]\right)^{J_1 + J_2 - 2}} . \tag{106}$$

Then

$$\left\| \mathbf{z}_k[n] \right\| \leq \left\| \mathbf{a}_k' \right\| \left(1 - c_k K \left(1-r\right)\left(1-g\right) h_c\left[Q\right]\right)^{\lfloor n/(R+1) \rfloor + 1} \tag{107}$$

for all $n \geq 0$, where $\lfloor n/(R+1) \rfloor$ is the largest integer less or equal to $n/(R+1)$.

□

Inequality (107) implies that $\|\mathbf{z}_k[n]\|$ converges to 0 following an exponential-like trajectory for each k. This and (102) imply $\overline{\mathbf{a}}_k[n] \to \mathbf{a}_k'$ for each k. Therefore, the conditions in the hypothesis of the theorem are sufficient to guarantee the convergence of SMNC.

The theorem's hypothesis places certain requirements on the values of $h_c[n]$, Q, and K. The $0 \leq r < 1$ requirement and the definition of r in (105) imply that the $h_c[Q]$ must be positive and larger than the sum of multiple adjacent samples of the impulse response. As explained in [15], $0 \leq r < 1$ is also a necessary condition for the convergence of the oversampling version of the MNC technique and can be easily satisfied in practice. The requirement that $0 \leq g < 1$ and $0 < 2Kh_c[Q] < 1$ sets an upper bound on K.

Theorem 2 also provides insight into the convergence rate. It indicates that increasing K increases the convergence rate. It also implies that reducing the probability of $s_k[n] = 0$ over time, which increases the value of c_k in (92), leads to faster convergence.

While Theorem 2 predicts how the expected value of each filter coefficient evolves over time, but it does not provide insight into the variance of the noise component of each filter coefficient. Intuitive reasoning similar to that in [15] and extensive simulations indicate that the noise variance can be made arbitrarily small by reducing K. Therefore, K represents a tradeoff between convergence accuracy and convergence speed.

V. SIMULATION RESULTS

Three sets of computer simulation results are presented in this section. The first set demonstrates that oversampling is indeed required for the original version of the MNC technique presented in [15] to work properly. The second set demonstrates the effectiveness and of the SMNC technique. The third set demonstrates the transient convergence behavior of the SMNC technique and compares it to that predicted by Theorem 2 presented in the previous section.

All simulations implement the same main DAC and correction DAC architectures, the same DAC clock-rate of $f_s = 3$ GHz, and the same MNC design parameters P, Q, N and K of 3, 21, 9 and $8 \cdot 10^{-6}$, respectively. As in [16], the main DAC consists of the DEM encoder presented in [17] followed by 36 1-bit DACs. The DEM encoder converts the 14-bit main DAC input sequence, $x[n]$, into 36 1-bit sequences, each of which drives a 1-bit DAC with weight K_i. For i = 1, 2, …, 20 the values of K_i are 1, 1, 2, 2, 4, 4, …, 512, 512, respectively, and for i = 21, 22, …, 36, each K_i has a value of 1024. Each 1-bit DAC implements a 25% return-to-zero (RZ) phase to avoid ISI. Also as in [16], the correction DAC is implemented without DEM or calibration and its minimum step-size is Δ/4, where Δ is the main DAC's minimum step-size.

The same set of mismatch noise parameters was used for each simulation. Dynamic mismatch noise was simulated by inserting a random Gaussian delay with a standard deviation of 0.6 ps on each 1-bit DAC clock time. Static mismatch error was simulated by introducing 1-bit DAC step-size errors. The step-size error for each of the 1024-weight 1-bit DACs was chose as a Gaussian random variable with a standard deviation of 0.15% of the 1-bit DAC's step size, 1024Δ. That of each of the other 1-bit DACs, including those in the correction DAC, were chosen similarly, except that the standard deviation was divided by the square root of the 1-bit DAC's step-size divided by 1024Δ.

Each simulation includes a 5-bit VCO-based ADC of the type implemented in the IC presented in [16]. Aside from its noise and distortion, the VCO-based ADC is equivalent to a sinc lowpass filter followed by a first-order ΔΣ modulator ADC with 5-bit quantization [23]. No ADC calibration was applied, so the ADC's nonlinearity is high: with a full-scale sinusoidal input waveform, the second and third harmonic distortion terms are −26 dBc and −47 dBc, respectively.

Fig. 31 shows simulated output spectra from the system with the original version of the MNC technique and a −1 dBFS sinusoidal input signal, with and without oversampling the ADC. Fig. 31a shows the output spectrum with MNC disabled and Fig. 31b shows the output spectrum with MNC enabled for an oversampling ratio of R = 5. This oversampling ratio in conjunction with the sinc lowpass filtering inherent to the VCO-based ADC is

sufficiently high for the aliasing error to be negligible over the DAC's 0 to $0.42f_s$ signal band.[8] In this case, MNC improves the SNDR by 18 dB over the DAC's signal band. Fig. 31c shows the output spectrum for MNC enabled but without oversampling, i.e., with the ADC sampled at f_s. Some SNDR improvement still occurs in this case relative to the case with MNC disabled, because aliasing does not prevent MNC from canceling a low-frequency portion of the mismatch noise. However, the aliasing prevents cancellation of higher-frequency mismatch noise and, therefore, prevents significant SNDR improvement.
Fig. 32 shows the simulated output spectrum from the system with the SMNC technique and a −1 dBFS sinusoidal input signal for an ADC sample-rate of $5f_s/6$, i.e., $R = 5$. Compared to the case without MNC shown in Fig. 31a, the SMNC technique improves the SNDR by 18 dB. This result supports the paper's assertion that the SMNC technique provides roughly the same SNDR improvement as the original MNC technique despite aliasing from not oversampling.

In the simulations described above, the adaptive FIR filter coefficients were obtained during foreground calibration mode and then frozen for use during normal DAC mode. During foreground calibration, $x[n]$ was chosen to toggle randomly between -2389.5Δ and -2388.5Δ. In principle, any $x[n]$ with time-invariant statistics as required by the foreground mode version of the SMNC technique would work, but this choice of $x[n]$ is attractive because of its small dynamic range, which simplifies the ADC, and it results in $s_k[n]$ sequences with a low percentage of zero values, which is beneficial for rapid convergence.

The SMNC technique's foreground calibration convergence time for the simulation results shown in Fig. 32 was about 3 ms. This is approximately $R = 5$ times longer than that

[8] The decimation filter's non-ideal transition bandwidth causes aliasing at frequencies between $0.42f_s$ and $0.5f_s$, which limits MNC accuracy over this band. As explained in [16],

of the original MNC technique, as expected. Much as in the case of the original MNC technique as explained in [15], the convergence time of the SMNC technique can be decreased by increasing K, but this comes at the expense of increased noise variance of each adaptive FIR filter's coefficients. A practical way to reduce the convergence time without a noise penalty is to use a relatively large value of K during an initial portion of foreground calibration mode so the conversion rate is relatively high while the adaptive FIR filter coefficients get close to their final values, and then reduce K during the final portion of foreground calibration mode to reduce the coefficient variances.

Fig. 33 shows the transient convergence behavior of the SMNC technique's adaptive FIR filter coefficients for a representative value of k and a constant value of K, i.e., $K = 8\cdot 10^{-6}$. The solid curves represent the differences between the instantaneous values of the coefficients, $\alpha_{k,m}[n]$, and their ideal values for $m = 0, 1, \ldots, N-1$ and a representative value of k. The definition of $\mathbf{z}_k[n]$ in (102) implies that the mean of each curve must be bounded by $-\|\mathbf{z}_k[n]\|$ and $\|\mathbf{z}_k[n]\|$. These upper and lower bounds, as predicted in Theorem 1, are plotted as dashed curves in the figure. The simulation results show that although the noise in the system causes the filter coefficients to fluctuate around their mean values, they are still mostly within the predicted upper and lower mean bounds.

APPENDIX

The proof uses the following well-known matrix theory results [24]. For any $N\times 1$ vectors $\mathbf{v}$ and $\mathbf{w}$, and any $N\times N$ matrix $\mathbf{H}$, the vector and matrix norms defined in (104) are such that

this exclusion band can be reduced by increasing the digital filter's complexity.

$$\|\mathbf{Hv}\| \le \|\mathbf{H}\|_1 \|\mathbf{v}\| \tag{108}$$

and

$$\|\mathbf{v}\| - \|\mathbf{w}\| \le \|\mathbf{v}+\mathbf{w}\| \le \|\mathbf{v}\| + \|\mathbf{w}\|. \tag{109}$$

Proof of Theorem 2:

If $\mathbf{a}_k' = 0$, then (103) and the initial condition of $\mathbf{z}_k[n] = -\mathbf{a}_k'$ for all $n < 0$ imply that $\mathbf{z}_k[n] = \mathbf{0}$ for all $n \ge 0$ and (107) holds. The rest of the proof considers the case of $\mathbf{a}_k' \ne 0$.

The proof applies mathematical induction. The *inductive step*, which is proven shortly, is: for any integer $n \ge 0$, if

$$\frac{\|\mathbf{z}_k[i]\|}{\|\mathbf{z}_k[i-1]\|} \ge 1 - 2c_k K h_c[Q], \tag{110}$$

for all $i < n$, then the theorem's hypothesis ensures that (110) also holds for $i = n$ and

$$\frac{\|\mathbf{z}_k[n]\|}{\|\mathbf{z}_k[n-1]\|} \le \begin{cases} 1 - c_k K(1-r)(1-g)h_c[Q], & \text{if } n \bmod (R+1) = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{111}$$

The induction *base step* requires that (110) hold for all $i < 0$. The proof of the base step follows from the initial condition of $\mathbf{z}_k[n] = -\mathbf{a}_k'$ for all $n < 0$ and (104). Hence, if the inductive step is true, it follows from induction that (110) and (111) must hold for all $n \ge 0$. In addition, applying (111) for $n \ge 0$ with the initial condition of $\mathbf{z}_k[-1] = -\mathbf{a}_k'$ leads to (107).

It remains to show that the inductive step is true. This is shown in the remainder of the proof.

If $n \bmod (R+1) \ne 0$, it follows from (103) that $\mathbf{z}_k[n] = \mathbf{z}_k[n-1]$, thus (110) holds for $i = n$ and (111) holds. The rest of analysis considers the case when $n \bmod (R+1) = 0$. In this case, (103) reduces to

$$\mathbf{z}_k[n] = \mathbf{z}_k[n-1] - c_k K \sum_{J=1}^{R+Q+N-1} \mathbf{H}_{\mathbf{c}}^{(J)} \mathbf{z}_k[n-J]. \tag{112}$$

It follows from (99) that (112) can be rewritten as

$$\begin{aligned}\mathbf{z}_k[n] &= \mathbf{z}_k[n-1] - c_k K \mathbf{H_c} \mathbf{z}_k[n-1] \\ &\quad - c_k K \sum_{J=1}^{R+Q+N-1} \mathbf{H_c}^{(J)} \left(\mathbf{z}_k[n-J] - \mathbf{z}_k[n-1] \right)\end{aligned} \tag{113}$$

and further rewritten as

$$\begin{aligned}\mathbf{z}_k[n] &= \left(\mathbf{I} - c_k K \mathbf{H_c} \right) \mathbf{z}_k[n-1] \\ &\quad - c_k K \sum_{J=1}^{R+Q+N-1} \sum_{m=1}^{J-1} \mathbf{H_c}^{(J)} \left(\mathbf{z}_k[n-m-1] - \mathbf{z}_k[n-m] \right)\end{aligned} \tag{114}$$

where **I** is an *N*×*N* identity matrix. Taking the vector norm on both sides of (114) and applying (109) yields

$$\begin{aligned}\left\| \mathbf{z}_k[n] \right\| &\le \left\| \left(\mathbf{I} - c_k K \mathbf{H_c} \right) \mathbf{z}_k[n-1] \right\| \\ &\quad + \sum_{J=1}^{R+Q+N-1} \sum_{m=1}^{J-1} \left\| c_k K \mathbf{H_c}^{(J)} \left(\mathbf{z}_k[n-m-1] - \mathbf{z}_k[n-m] \right) \right\|\end{aligned} \tag{115}$$

and

$$\begin{aligned}\left\| \mathbf{z}_k[n] \right\| &\ge \left\| \left(\mathbf{I} - c_k K \mathbf{H_c} \right) \mathbf{z}_k[n-1] \right\| \\ &\quad - \sum_{J=1}^{R+Q+N-1} \sum_{m=1}^{J-1} \left\| c_k K \mathbf{H_c}^{(J)} \left(\mathbf{z}_k[n-m-1] - \mathbf{z}_k[n-m] \right) \right\|.\end{aligned} \tag{116}$$

The definition of *r* in (105) and the condition 0 ≤ *r* < 1 in Theorem 2 imply that $h_c[Q]$ is positive. Therefore, it follows from the definition of $\mathbf{H_c}$ in (101) and the definition of the matrix norm in (104) that

$$\left\| h_c[Q] \mathbf{I} - \mathbf{H_c} \right\| \le h_c[Q] r \, . \tag{117}$$

For any real *N*-dimensional column vector **v**, the vector norm of $(\mathbf{I} - cK\mathbf{H_c})\mathbf{v}$ can be written as

$$\left\| \left(\mathbf{I} - c_k K \mathbf{H_c} \right) \mathbf{v} \right\| = \left\| \left(1 - c_k K h_c[Q] \right) \mathbf{v} + c_k K \left(h_c[Q] \mathbf{I} - \mathbf{H_c} \right) \mathbf{v} \right\|. \tag{118}$$

Applying (108) and (109) yields

$$\left\| \left(\mathbf{I} - c_k K \mathbf{H_c} \right) \mathbf{v} \right\| \le \left(1 - c_k K h_c[Q] \right) \left\| \mathbf{v} \right\| + c_k K \left\| h_c[Q] \mathbf{I} - \mathbf{H_c} \right\|_1 \left\| \mathbf{v} \right\| \tag{119}$$

and

$$\left\|\left(\mathbf{I}-c_kK\mathbf{H}_\mathbf{c}\right)\mathbf{v}\right\| \ge \left(1-c_kKh_c\left[Q\right]\right)\left\|\mathbf{v}\right\| - c_kK\left\|h_c\left[Q\right]\mathbf{I}-\mathbf{H}_\mathbf{c}\right\|_1\left\|\mathbf{v}\right\|. \tag{120}$$

Applying (117) to (119) and (120) yields

$$\left\|\left(\mathbf{I}-c_kK\mathbf{H}_\mathbf{c}\right)\mathbf{v}\right\| \le \left(1-c_kK\left(1-r\right)h_c\left[Q\right]\right)\left\|\mathbf{v}\right\| \tag{121}$$

and

$$\left\|\left(\mathbf{I}-c_kK\mathbf{H}_\mathbf{c}\right)\mathbf{v}\right\| \ge \left(1-c_kK\left(1+r\right)h_c\left[Q\right]\right)\left\|\mathbf{v}\right\|. \tag{122}$$

Replacing **v** by $\mathbf{z}_k[n-1]$ in (121) and (122), and substituting the results into (115) and (116) gives

$$\begin{aligned}\left\|\mathbf{z}_k[n]\right\| &\le \left(1-c_kK\left(1-r\right)h_c\left[Q\right]\right)\left\|\mathbf{z}_k[n-1]\right\| \\ &\quad + \sum_{J=1}^{R+Q+N-1}\sum_{m=1}^{J-1}\left\|c_kK\mathbf{H}_\mathbf{c}^{(J)}\left(\mathbf{z}_k[n-m-1]-\mathbf{z}_k[n-m]\right)\right\|\end{aligned} \tag{123}$$

and

$$\begin{aligned}\left\|\mathbf{z}_k[n]\right\| &\ge \left(1-c_kK\left(1+r\right)h_c\left[Q\right]\right)\left\|\mathbf{z}_k[n-1]\right\| \\ &\quad - \sum_{J=1}^{R+Q+N-1}\sum_{m=1}^{J-1}\left\|c_kK\mathbf{H}_\mathbf{c}^{(J)}\left(\mathbf{z}_k[n-m-1]-\mathbf{z}_k[n-m]\right)\right\|.\end{aligned} \tag{124}$$

Equation (103) with the initial condition $\mathbf{z}_k[n] = -\mathbf{a}_k'$ for $n < 0$ implies that each $\mathbf{z}_k[n-m-1] - \mathbf{z}_k[n-m]$ in (123) and (124) is either

$$c_kK\sum_{J=1}^{R+Q+N-1}\mathbf{H}_\mathbf{c}^{(J)}\mathbf{z}_k[n-m-J] \quad \text{or} \quad \mathbf{0}. \tag{125}$$

This observation applied to (123) and (124) results in

$$\begin{aligned}\left\|\mathbf{z}_k[n]\right\| &\le \left(1-c_kK\left(1-r\right)h_c\left[Q\right]\right)\left\|\mathbf{z}_k[n-1]\right\| \\ &\quad + \sum_{J_1=1}^{R+Q+N-1}\sum_{m=1}^{J_1-1}\left\|{c_k}^2K^2\sum_{J_2=1}^{R+Q+N-1}\mathbf{H}_\mathbf{c}^{(J_1)}\mathbf{H}_\mathbf{c}^{(J_2)}\mathbf{z}_k[n-m-J_2]\right\|\end{aligned} \tag{126}$$

and

$$\begin{aligned}\left\|\mathbf{z}_k[n]\right\| &\ge \left(1-c_kK\left(1+r\right)h_c\left[Q\right]\right)\left\|\mathbf{z}_k[n-1]\right\| \\ &\quad - \sum_{J_1=1}^{R+Q+N-1}\sum_{m=1}^{J_1-1}\left\|{c_k}^2K^2\sum_{J_2=1}^{R+Q+N-1}\mathbf{H}_\mathbf{c}^{(J_1)}\mathbf{H}_\mathbf{c}^{(J_2)}\mathbf{z}_k[n-m-J_2]\right\|.\end{aligned} \tag{127}$$

Applying (108) with $\mathbf{H}$ replaced by $c_k{}^2 K^2 \mathbf{H}_{\mathbf{c}}^{(J_1)} \mathbf{H}_{\mathbf{c}}^{(J_2)}$, substituting the result into (126) and (127), then applying (109) yields

$$\begin{aligned}\left\|\mathbf{z}_k[n]\right\| \leq &\left(1-c_k K\left(1-r\right) h_c\left[Q\right]\right)\left\|\mathbf{z}_k[n-1]\right\| \\ &+c_k{}^2 K^2 \sum_{J_1=1}^{R+Q+N-1} \sum_{J_2=1}^{R+Q+N-1} \left\|\mathbf{H}_{\mathbf{c}}^{(J_1)} \mathbf{H}_{\mathbf{c}}^{(J_2)}\right\|_1 \sum_{m=1}^{J_1-1} \left\|\mathbf{z}_k[n-m-J_2]\right\|\end{aligned} \tag{128}$$

and

$$\begin{aligned}\left\|\mathbf{z}_k[n]\right\| \geq &\left(1-c_k K\left(1+r\right) h_c\left[Q\right]\right)\left\|\mathbf{z}_k[n-1]\right\| \\ &-c_k{}^2 K^2 \sum_{J_1=1}^{R+Q+N-1} \sum_{J_2=1}^{R+Q+N-1} \left\|\mathbf{H}_{\mathbf{c}}^{(J_1)} \mathbf{H}_{\mathbf{c}}^{(J_2)}\right\|_1 \sum_{m=1}^{J_1-1} \left\|\mathbf{z}_k[n-m-J_2]\right\|.\end{aligned} \tag{129}$$

It follows from (110), $0 < c_k \leq 1$ in (92), and Theorem 2’s hypothesis of $0 < 2Kh_c[Q] < 1$ that

$$\left\|\mathbf{z}_k[n-i]\right\| \leq \left\|\mathbf{z}_k[n-1]\right\| \left(1-2c_k K h_c\left[Q\right]\right)^{-i+1} \tag{130}$$

holds for $i = 2, 3, 4, \ldots$. Therefore,

$$\sum_{m=1}^{J_1-1} \left\|\mathbf{z}_k[n-m-J_2]\right\| \leq \left\|\mathbf{z}_k[n-1]\right\| \sum_{m=1}^{J_1-1} \left(1-2c_k K h_c\left[Q\right]\right)^{-m-J_2+1}. \tag{131}$$

The sum in the right side of (131) can be expanded via the geometric series formula as

$$\sum_{m=1}^{J_1-1} \left(1-2c_k K h_c\left[Q\right]\right)^{-m-J_2+1} = \frac{1-\left(1-2c_k K h_c\left[Q\right]\right)^{J_1-1}}{2c_k K h_c\left[Q\right]\left(1-2c_k K h_c\left[Q\right]\right)^{J_1+J_2-2}}. \tag{132}$$

It follows from (92) that

$$\sum_{m=1}^{J_1-1} \left(1-2c_k K h_c\left[Q\right]\right)^{-m-J_2+1} \leq \frac{1-\left(1-2K h_c\left[Q\right]\right)^{J_1-1}}{2c_k K h_c\left[Q\right]\left(1-2K h_c\left[Q\right]\right)^{J_1+J_2-2}}. \tag{133}$$

Substituting (133) into (131) and substituting the result into (128) and (129) yields

$$\begin{aligned}\frac{\left\|\mathbf{z}_k[n]\right\|}{\left\|\mathbf{z}_k[n-1]\right\|} \leq &1-c_k K\left(1-r\right) h_c\left[Q\right] \\ &+\sum_{J_1=1}^{R+Q+N-1} \sum_{J_2=1}^{R+Q+N-1} \frac{c_k K \left\|\mathbf{H}_{\mathbf{c}}^{(J_1)} \mathbf{H}_{\mathbf{c}}^{(J_2)}\right\|_1 \left(1-\left(1-2K h_c\left[Q\right]\right)^{J_1-1}\right)}{2h_c\left[Q\right]\left(1-2K h_c\left[Q\right]\right)^{J_1+J_2-2}}\end{aligned} \tag{134}$$

and

$$
\begin{aligned}
\frac{\left\|\mathbf{z}_k[n]\right\|}{\left\|\mathbf{z}_k[n-1]\right\|} \geq\, & 1-c_k K\left(1+r\right)h_c\left[Q\right] \\
& -\sum_{J_1=1}^{R+Q+N-1}\sum_{J_2=1}^{R+Q+N-1}\frac{c_k K\left\|\mathbf{H}_{\mathbf{c}}^{(J_1)}\mathbf{H}_{\mathbf{c}}^{(J_2)}\right\|_1\left(1-\left(1-2Kh_c\left[Q\right]\right)^{J_1-1}\right)}{2h_c\left[Q\right]\left(1-2Kh_c\left[Q\right]\right)^{J_1+J_2-2}}.
\end{aligned}
\tag{135}
$$

Substituting (106) into (134) yields (111) for *n* mod (*R*+1) = 0, and substituting (106) into (135) yields

$$
\begin{aligned}
\frac{\left\|\mathbf{z}_k[n]\right\|}{\left\|\mathbf{z}_k[n-1]\right\|} & \geq 1-c_k K\left(1+r\right)h_c\left[Q\right]-c_k Kg\left(1-r\right)h_c\left[Q\right] \\
& =1-c_k K\left(2-\left(1-r\right)\left(1-g\right)\right)h_c\left[Q\right].
\end{aligned}
\tag{136}
$$

This implies that (110) holds for *i* = *n* for any values of *r* and *g* that satisfy $0 \leq r < 1$ and $0 < g < 1$.

□

ACKNOWLEDGEMENTS

Chapter 3, in full, has been submitted to the IEEE Transactions on Circuits and Systems I: Regular Papers. D. Kong, I. Galton, 2019. The dissertation author is the primary investigator and author of this paper. Professor Ian Galton supervised the research which forms the basis for this paper.

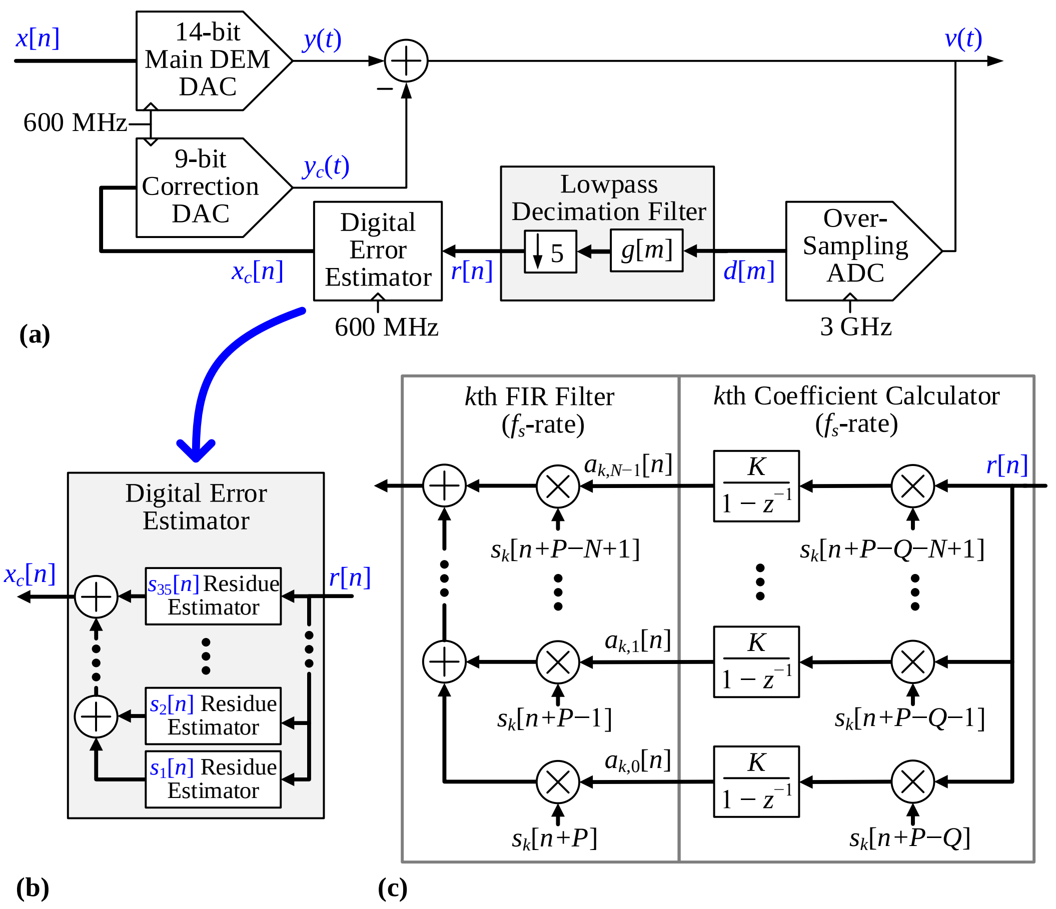


Figure 26: a) High-level structure of the IC presented in [16], b) high-level structure of the digital error estimator, and c) details of each $s_k[n]$ residue estimator.

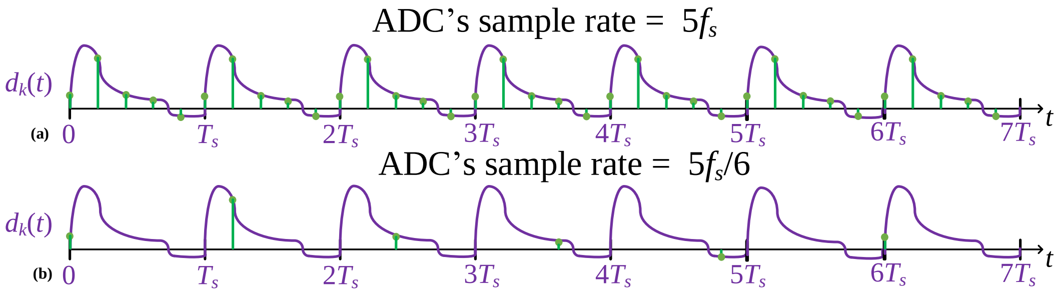


Figure 27: a) Oversampling $d_k(t)$, and b) subsampling $d_k(t)$.

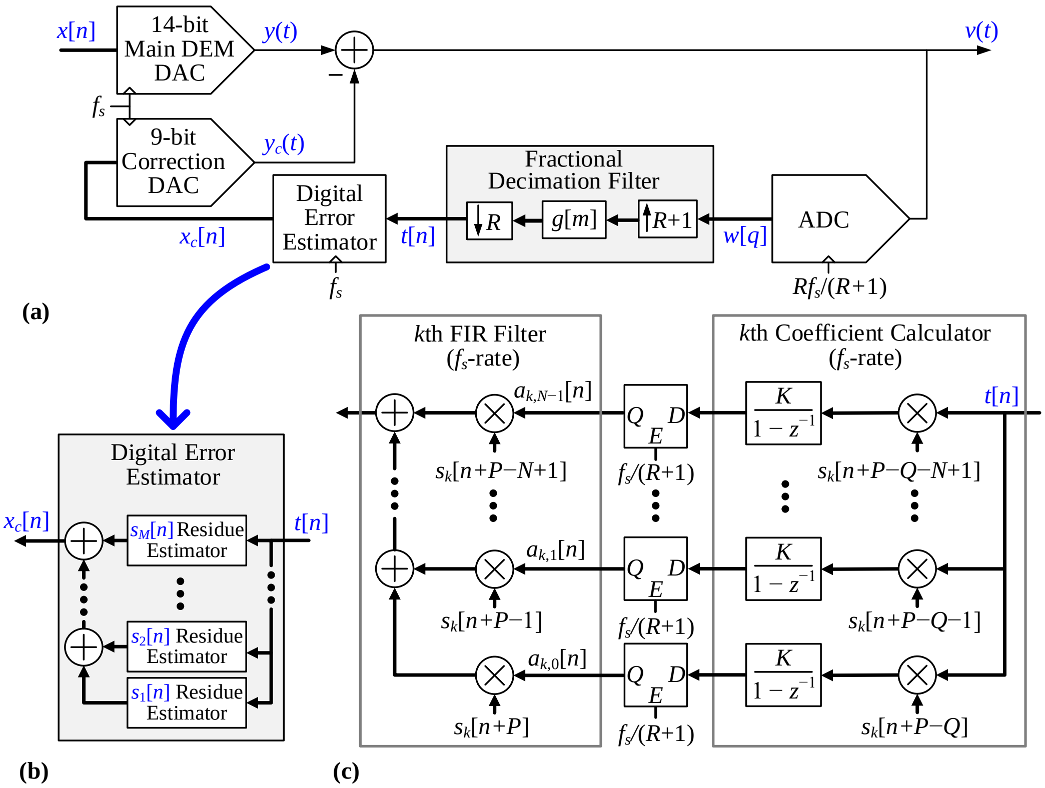


Figure 28: a) High-level structure of the subsampling MNC technique, b) high-level structure of the digital error estimator, and c) details of each $s_k[n]$ residue estimator.

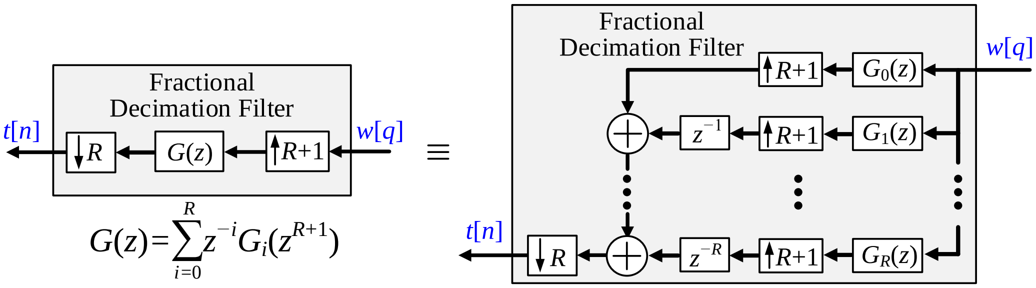

Fractional Decimation Filter
t[n]
↓R
G(z)
↑R+1
w[q]
≡
$G(z)=\sum_{i=0}^{R} z^{-i} G_i(z^{R+1})$
Fractional Decimation Filter
w[q]
↑R+1
G_0(z)
z^{−1}
G_1(z)
t[n]
↓R
z^{−R}
G_R(z)

Figure 29: Polyphase structure for fractional decimation filter.

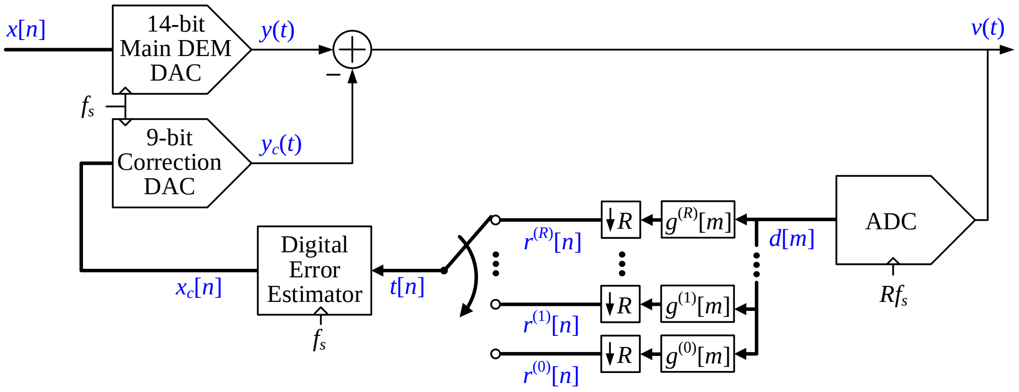


Figure 30: Modified version of Fig. 3a with equivalent behavior.

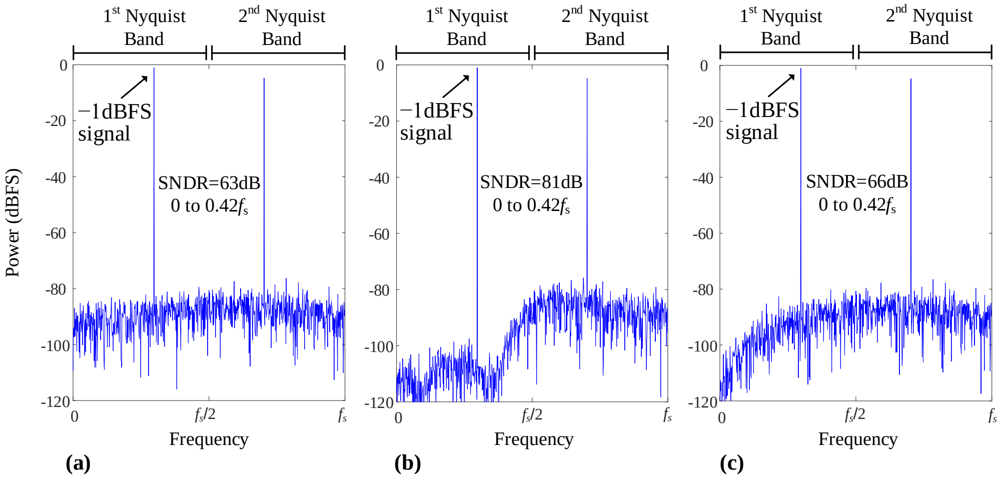


Figure 31: Representative simulated output spectra with a) MNC off, b) MNC on, and c) MNC on but without oversampling.

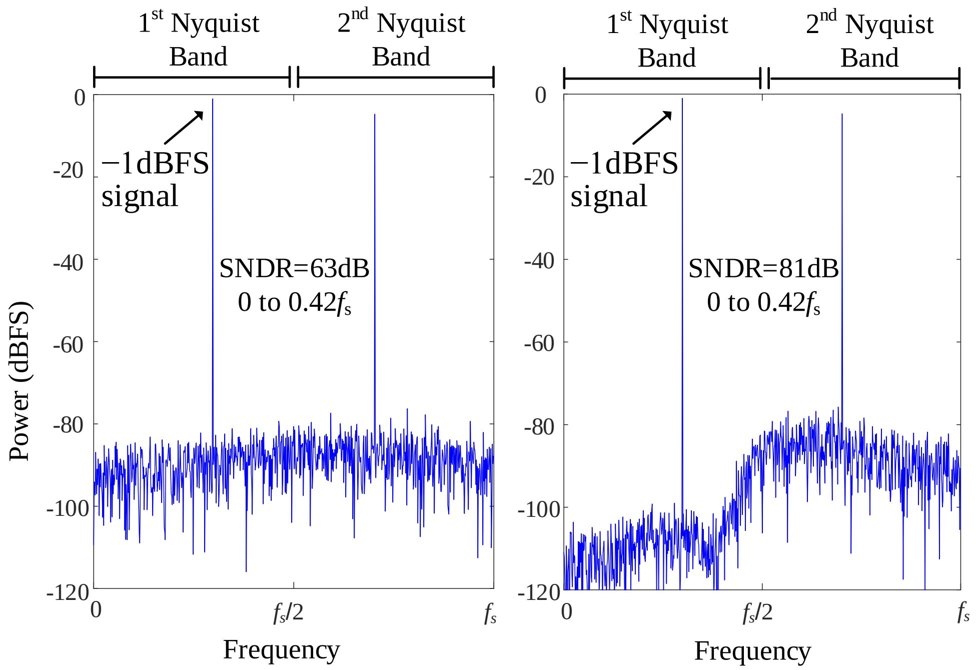


Figure 32: Representative simulated output spectra without/with SMNC.

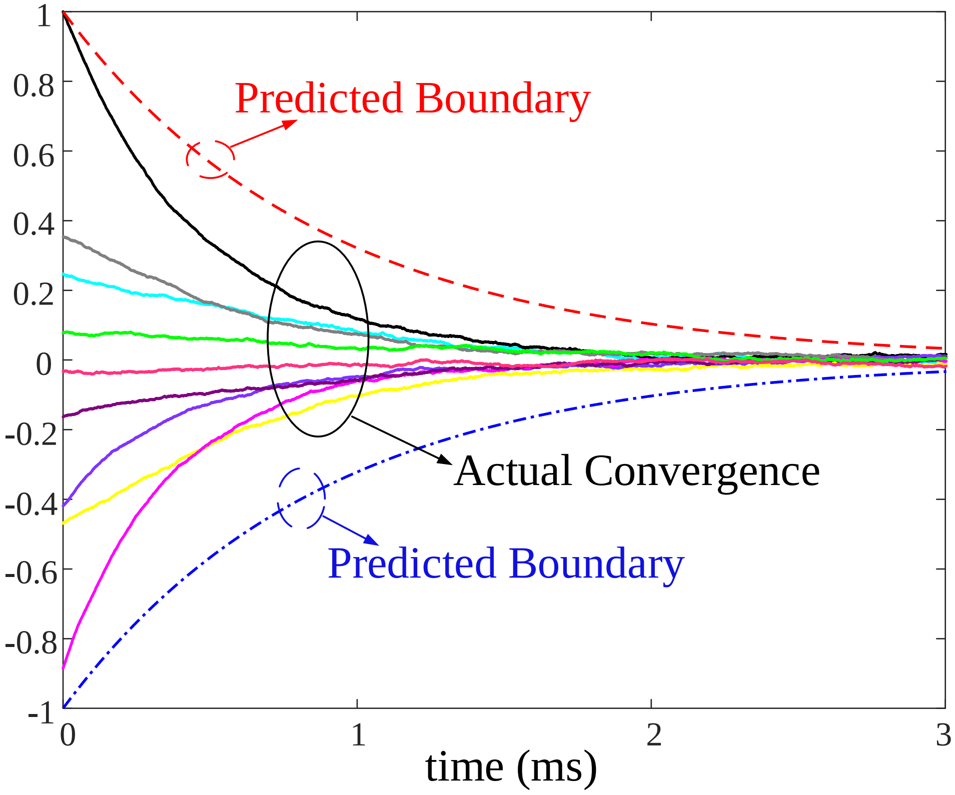


Figure 33: Transient convergence behavior of the SMNC technique.

REFERENCES

1. W. Schofield, D. Mercer, L. St. Onge, “A 16b 400MS/s DAC with <80dBc IMD to 300MHz and 160dBm/Hz noise Power Spectral Density,” *IEEE International Solid State Circuits Conference,* February 2003.

2. Q. Huang, P. A. Francese, C. Martelli, and J. Nielsen,“A 200Ms/s 14b 97 mW DAC in 0.18μm CMOS,” *IEEE International Solid State Circuits Conference,* February 2004.

3. H.-H. Chen, J. Lee, J. Weiner, Y.-K. Chen, and J.-T. Chen, “A 14-bit 150 MS/s CMOS DAC with Digital Background Calibration,” *Symposium on VLSI Circuits*, pp. 51-52, June 2006.

4. M. Clara, W. Klatzer, B. Seger, A. Di Giandomenico, and L. Gori, “A 1.5V 200MS/s 13b 25mW DAC with Randomized Nested Background Calibration in 0.13 μm CMOS,” *IEEE International Solid State Circuits Conference,* February 2007.

5. M. Clara, W. Klatzer, D. Gruber, A. Marak, B. Seger, and W. Pribyl, “A 1.5 V 13 bit 130-300 MS/s Self-calibrated DAC with Active Output Stage and 50 MHz Signal Bandwidth in 0.13μm CMOS,” *European Solid-State Circuits Conference*, pp. 262-265, September, 2008.

6. B. Catteau, P. Rombouts, J. Raman, and L. Weyten, “An on-line calibration technique for mismatch errors in high-speed DACs,” *IEEE Transactions on Circuits and Systems–I, Reg. Papers*, vol. 55, no. 7, pp. 1873–1883, Aug. 2008

7. C.-H. Lin, F. M. L. van der Goes, J. R. Westra, J. Mulder, Y. Lin, E. Arslan, E. Ayranci, X. Liu, K. Bult, “A 12 bit 2.9 GS/s DAC With IM3 < −60 dBc Beyond 1 GHz in 65 nm CMOS,” *IEEE Journal of Solid-State Circuits*, vol. 44, no. 23, pp. 3285-3293, December 2009.

8. Y. Tang, J. Briaire, K. Doris, R. van Veldhoven, P. van Beek, H. Hegt, and A. van Roermund, “A 14 bit 200 MS/s DAC With SFDR >78 dBc, IM3 < −83 dBc and NSD < −163 dBm/Hz Across the Whole Nyquist Band Enabled by Dynamic-Mismatch Mapping,” *IEEE Journal of Solid-State Circuits*, vol. 46, no. 6, pp. 1371-1381, June 2011.

9. S. Spiridon, J. van der Tang, H. Yan, H.-F. Chen, G. Guermandi, X. Liu, E. Arslan, R. van der Goes, K. Bult, “A 375 mW Multimode DAC-Based Transmitter With 2.2 GHz Signal Bandwidth and In-Band IM3<−58 dBc in 40 nm CMOS,” *IEEE Journal of Solid-State Circuits*, vol. 48, no. 7, pp. 1595-1604, July 2013.

10. W.-T. Lin, H.-Y. Huang, and T.-H. Kuo, “A 12-bit 40 nm DAC Achieving SFDR > 70 dB at 1.6 GS/s and IMD < -61dB at 2.8 GS/s With DEMDRZ Technique,” *IEEE Journal of Solid-State Circuits*, Vol. 49, no. 3, pp. 708-717, March 2014.

11. S. M. Lee, D. Seo, S. M. Taleie, D. Kong, M. J. McGowan, T. Song, G. Saripalli, J. Kuo, S. Bazarjani, “A 14b 750MS/s DAC in 20nm CMOS with <−168dBm/Hz noise floor beyond Nyquist and 79dBc SFDR utilizing a low glitch-noise hybrid R-2R architecture,” in *Symp. VLSI Circuits Dig*., June 2015.

12. Engel, M. Clara, H. Zhu, and P. Wilkins, “A 16-bit 10 Gsps currentsteering RF DAC in 65 nm CMOS achieving 65 dBc ACLR multi-carrier performance at 4.5 GHz Fout,” in *Symp. VLSI Circuits Dig*., June 2015.

13. S. Su and M. S.-W. Chen, “A 12-bit 2 GS/s dual-rate hybrid DAC with pulse-error pre-distortion and in-band noise cancellation achieving > 74 dBc SFDR and < −80 dBc IM3 up to 1 GHz in 65 nm CMOS,” *IEEE Journal of Solid-State Circuits,* vol. 51, no. 12, pp. 2963–2978, December 2016.

14. C.-H. Lin, K. L. J. Wong, T.-Y. Kim, G. R. Xie, D. Major, G. Unruh, S. R. Dommaraju, H. Eberhart, A. Venes, “A 16b 6GS/s Nyquist DAC with IMD <−90dBc up to 1.9GHz in 16nm CMOS,” *IEEE International Solid State Circuits Conference*, February 2018.

15. D. Kong and I. Galton, “Adaptive Cancellation of Static and Dynamic Mismatch Error in Continuous-Time DACs,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 2, pp. 421–433, Feb. 2018.

16. D. Kong, K. Rivas-Rivera, I. Galton, “A 600 MS/s DAC with over 87dB SFDR and 77dB peak SNDR Enabled by Adaptive Cancellation of Static and Dynamic Mismatch Error,” *IEEE Journal of Solid-State Circuits*, under review (Submitted manuscript available at http://ispg.ucsd.edu/unpublished-paper/).

17. K. L. Chan, J. Zhu, and I. Galton, “Dynamic Element Matching to Prevent Nonlinear Distortion from Pulse-Shape Mismatches in High-Resolution DACs,” *IEEE Journal of Solid-State Circuits*, vol. 43, no. 9, pp. 2067-2078, September 2008.

18. J. Remple, I. Galton, “The Effects of Inter-Symbol Interference in Dynamic Element Matching DACs,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 1, pp. 14-23, January 2017.

19. P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, 1993.

20. J. E. Mazo, “On the independence theory of equalizer convergence,” *Bell Syst. Tech. J.*, vol. 58, pp. 963-993, May-June 1979.

21. B. Widrow, “Adaptive Filters,” *Aspects of Network and System Theory*, R.E. Kalman and N. DeClaris, eds., Holt, Rinehart and Winston, pp. 563-586, 1971.

22. W. A. Gardner, “Learning characteristics of stochastic-gradient-de- scent algorithms: A general study, analysis, and critique,” *Signal Processing*, vol. 6. pp. 113-133, 1984.

23. G. Taylor and I. Galton, “A Mostly-Digital Variable-Rate Continuous-Time Delta-Sigma Modulator ADC,” *IEEE Journal of Solid-State Circuits*, vol. 45, no. 12, pp. 2634–2646, December 2010.

24. R. A. Horn and C. R Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

CHAPTER 4

A RIGOROUS MEAN-SQUARE CONVERGENCE ANALYSIS OF MISMATCH NOISE CANCELLATION TECHNIQUE

Abstract— Mismatch noise cancellation (MNC) technique is an adaptive digital calibration technique recently proposed and experimentally validated to suppress the static and dynamic mismatch error for continuous-time DACs. This paper presents a rigorous mean-square convergence analysis of MNC technique, which for the first time quantifies the impact of the noise present during calibration on the post-calibration DAC signal-to-noise-ratio (SNR). The results of this paper provide guidance into the design of MNC, and offer insights into the mechanism of other similar adaptive systems.

I. INTRODUCTION

High-speed, high resolution Digital-to-Analog Converters (DACs) with continuous-time outputs are critical in many applications. Each of these DACs operates by interpolating a discrete-time digital sequence into a continuous-time analog output. Ideally, the output of a continuous-time DAC is linearly scaled with its input code during each clock interval. Unfortunately, inadvertent but inevitable clock skew and component mismatches cause non-ideal deviations of both the scale factor and shape of each overall DAC output pulse, which give rise to static mismatch error and dynamic mismatch error, respectively. In practice, both types of mismatch error can significantly limit DAC performance. Various digital calibration techniques have been demonstrated to reduce static mismatch error [1-14], but they do not well-address dynamic mismatch error.

Mismatch noise cancellation (MNC) technique was recently proposed in [15] to address this problem, its effectiveness was experimentally validated in [16] with a prototype DAC IC. MNC incorporates a feedback loop that measures and cancels both static and dynamic mismatch error caused by clock skew and component mismatches over the DAC's first Nyquist band. The mismatch cancellation of MNC is achieved with a correction DAC driven by a digital correction sequence, the sequence is derived from a large number of coefficients, the values of which represent the static and dynamic mismatch profile of the main DAC. MNC measures the main DAC's mismatch error and uses this information to update these coefficients such that the main DAC's static and dynamic mismatch error are suppressed over the first Nyquist band. The convergence behaviors of these coefficients were studied in [15] with a mean convergence analysis, which proved that the statistic mean of each coefficient converges to a steady-state value. However, [15] did not quantify how significant each coefficient can deviate from its statistic mean from time to time due to the large noise inevitably present during calibration, it also did not quantify how these deviations affect the post-calibration DAC performance. To provide the guidance on the design of MNC, it is necessary to answer the following questions: for a given set of design parameters, how to estimate the impact of these noise on the calibration accuracy and the post-calibration DAC performance, and how to choose the design parameters to minimize this impact. This paper presents answers to these questions.

The paper is organized as follows. Section II presents an overview of MNC. Section III studied the noise impact on MNC and proves rigorously that the impact can be made arbitrarily small by design. Section IV presents an analytical bound of the noise contribution to the post-calibration DAC signal-to-noise-ratio (SNR), and presents guidelines to minimize this contribution. Section V presents simulation results, which are in good agreement with the analytical results.

II. OVERVIEW OF MNC

As shown in Fig. 34, MNC consists of a feedback loop around a 14-bit main DAC, the feedback loop adaptively measures and cancels static and dynamic mismatch error within the main DAC's first Nyquist band. The feedback loop consists of an oversampling ADC, a lowpass decimation filter, a digital error estimator and a correction DAC.

The main DAC incorporates a dynamic element matching (DEM) encoder and multiple 1-bit RZ DACs from [17]. The use of RZ 1-bit DACs mitigates the ISI effect. As analyzed in [18], such a DAC converts an input sequence, $x[n]$, into an output waveform given by

$$y(t) = \alpha(t)x\left[n_t\right] + e_{DAC}(t), \tag{137}$$

where $\lfloor n_t \rfloor$denotes the largest integer less than or equal to t/T_s with $T_s = 1/f_s$, and $\alpha(t)$ is a T_s-periodic pulse shaping waveform. Each period of $\alpha(t)$ is scaled by an input code to form a corresponding pulse of $\alpha(t)x[n_t]$, thus $\alpha(t)x[n_t]$ represents the ideal output component of the DAC. The $e_{DAC}(t)$ term represents the DAC's static and dynamic mismatch error, which is ensured by the DEM encoder to be a noise-like waveform free of nonlinear distortions. The $e_{DAC}(t)$ term is referred to as mismatch noise and has the form

$$e_{DAC}(t) = \sum_{k=1}^{L} d_k(t) s_k\left[n_t\right], \tag{138}$$

where each $d_k(t)$ is a T_s-periodic waveform dependent on the main DAC's mismatch profile, and the $s_k[n]$ sequences for k = 1, 2, …, L are pseudo-random sequences uncorrelated with each other and $x[n]$, zero-mean, and restricted to values of −1,0 and 1. Each $s_k[n]$ sequence is generated within the DEM encoder and is known to the system a priori, L is an integer number that is dependent on the structure of the DEM encoder.

The sampling theorem implies that no matter how the mismatch noise, $e_{DAC}(t)$, looks like, there must exist a correction DAC input sequence, $x_c[n]$, which generates a correction DAC output waveform, $y_c(t)$, that cancels $e_{DAC}(t)$ over the first Nyquist band, up to the accuracy of the correction DAC. The non-idealities from the correction DAC can be neglected if it contributes to error below the 14bit quantization noise floor of the main DAC. Given the small full-scale range needed to cancel $e_{DAC}(t)$, this level of accuracy can be easily achieved [15, 16].

The objective of MNC feedback loop is to generate the $x_c[n]$ sequence in real time such that $y_c(t)$ sufficiently approximates and cancels $e_{DAC}(t)$ over the first Nyquist band. To do this, it is necessary to measure $e_{DAC}(t)$ over the first Nyquist band, which requires a digitized version of the main DAC's output waveform over the first Nyquist band. This is achieved with the oversampling ADC and the lowpass decimation filter in Fig. 34. Consequently, the decimation filter's output, $r[n]$, contains a digitized version of the portion of $e_{DAC}(t)$ restricted to the first Nyquist band that is not yet fully canceled by MNC.

It is implied by (138) that $e_{DAC}(t)$ is highly correlated with the $s_k[n]$ sequences. Therefore, if $e_{DAC}(t)$ is sufficiently suppressed by $y_c(t)$ over the first Nyquist band, $r[n]$ would be sufficiently uncorrelated with the $s_k[n]$ sequences. Otherwise, $r[n]$ will be correlated with at least some $s_k[n]$ sequences. The digital estimator in Fig. 34 exploits such properties by correlating $r[n]$ with the $s_k[n]$ sequences and use the correlation results to update $x_c[n]$ in real time. The MNC feedback loop operates continuously such that all the correlation results reduce to 0.

The details of the digital error estimator are shown in Fig. 35. It consists of L channels of $s_k[n]$ residue estimators for $k = 1, 2, \ldots, L$, each of which correlates $r[n]$ with N time-shifted versions of one of the $s_k[n]$ sequences, scales the results by a loop gain constant, K, and accumulates the results into N coefficients, $\alpha_{k,0}[n]$, $\alpha_{k,1}[n]$, …, and $\alpha_{k,N-1}[n]$. The input of each accumulator contains a mean component representing the remaining portion of $d_k(t)$ in (138) not yet cancelled by MNC, and a component representing correlation noise. It follows from the analysis in [15] that MNC causes the accumulator outputs to go up and down to subtract the main DAC's mismatch error, which subsequently causes the mean component at the input of each accumulator to reach 0. If the correlation noise is ignored, this implies that each coefficient at the accumulator outputs would converge to a constant steady-state value. In the presence of correlation noise, each coefficient would still on average converge to this value, except that its instantaneous value would fluctuate above and below it due to noise. Mathematically, this is described by $E(a_{k,m}[n]) \rightarrow a_{k,m}'$ as $n \rightarrow \infty$, where $a_{k,m}'$ is the constant steady state value of $a_{k,m}[n]$, and $E(a_{k,m}[n])$ is the mean value of $a_{k,m}[n]$.

The analysis in [15] did not quantify how much each coefficient fluctuates from its mean as a result of the correlation noise, this fluctuation is often predicted by steady-state mean squared error given by $E\{(a_{k,m}[n]-a_{k,m}')^2\}$ where n is sufficiently large. Therefore, [15] did not quantify the impact of correlation noise on the main DAC's post-calibration performance. Although intuition implies that this impact can be reduced by decreasing the magnitude of the constant loop gain, K, this remains to be proven and quantified. Since the power of correlation noise is dominated by that of the ADC noise and the main DAC's signal component, which is of orders of magnitude larger than the post-calibration noise requirement of the main DAC, it is necessary to quantify this effect and provide useful guidance to minimize this effect.

A rigorous mathematical analysis of MNC convergence in presence of correlation noise is presented in Section III. The results of the analysis are used in Section IV to develop means to predict and minimize the impact of correlation noise on the post-calibration DAC performance.

III. MEAN-SQUARE CONVERGENCE ANALYSIS

The impact of correlation noise on the accuracy of the MNC coefficients are quantified by the steady-state mean squared error given by $E\{(a_{k,m}[n]- a_{k,m}')^2\}$ where n is sufficiently large, $k = 1, 2, \ldots, L$ and $m = 0, 1, \ldots, N-1$. These $L\times N$ mean squared error can be evaluated with a vector-based analysis. Let $\mathbf{a}_k[n]$ denote an N-dimensional vector given by $\mathbf{a}_k[n] = [a_{k,0}[n], a_{k,1}[n+1], \ldots, a_{k,N-1}[n+N-1]]^T$, and let $\mathbf{a}_k'$ denote the constant steady-state value given by $[a_{k,0}', a_{k,1}', \ldots, a_{k,N-1}']^T$, it follows that $a_{k,m}[n]-a_{k,m}'$ corresponds to the mth entry of $\mathbf{a}_k[n-m]-\mathbf{a}_k'$. Denote $\mathbf{z}_k[n] = \mathbf{a}_k[n]-\mathbf{a}_k'$, the analysis reduces to evaluating L vectors given by $\mathbf{z}_k[n]$ for $k = 1, 2, \ldots, L$, and the mean squared error associated with them. Sub-section A derives the difference equation of $\mathbf{z}_k[n]$ for each k, and sub-section B evaluates the mean squared error associated with them.

A. System-Related Difference Equations

It follows from the analysis in [15] that each MNC coefficient, $a_{k,m}[n]$, follows the difference equations given by

$$a_{k,m}[n] = a_{k,m}[n-1] + Ku_{k,m}[n] \tag{139}$$

and

$$\begin{aligned} u_{k,m}[n] = s_k[n-m+P-Q]\Bigg(& r_{ideal}[n] + \sum_{l=1}^{L}\sum_{i=-\infty}^{\infty} b_l[i]s_l[n-i] \\ & - \sum_{l=1}^{L}\sum_{i=-\infty}^{\infty}\sum_{j=0}^{N-1} h_c[i]a_{l,j}[n-i]s_l[n+P-i-j]\Bigg), \end{aligned} \tag{140}$$

where $r_{ideal}[n]$ is the decimation filter output sequence that would have occurred in the absence of both $e_{DAC}(t)$ and the correction DAC feedback loop, $b_k[n]$ represents the impulse response of an equivalent causal discrete-time linear time-invariant (LTI) system between the $s_k[n]$ sequence and the decimation filter output with the correction DAC feedback path disabled, and $-h_c[n]$ represents the impulse response of another causal discrete-time LTI system between the correction DAC input and the decimation filter output (the –1 factor is used to simplify the analysis). Furthermore, $b_k[n] = 0$ and $h_c[n] = 0$ for all $n < 0$ for causality and also for $n = 0$ to prevent the feedback loop from being delay-free. The definition of $r_{ideal}[n]$ implies that it represents the portion of ADC noise and main DAC's signal component at the output of the decimation filter.

Without loss of generality, $a_{k,m}[n]$ for each k and m is evaluated at $n \geq 0$ from an initial condition at $n \leq -1$. To simplify the notation, let us define a modified $s_k[n]$ sequence for each k, given by $S_k[n]$ as

$$S_k[n] = s_k[n+P], \tag{141}$$

where P is the delay term in each $s_k[n]$ residue estimator of Fig. 35. Replacing n with n+m, replacing i with i+m−P in the double sum of (140) and replacing i with i+m−j in the triple sum of (140), finally applying (141) yields

$$a_{k,m}[n+m] = a_{k,m}[n+m-1] + Ku_{k,m}[n+m]. \tag{142}$$

and

$$\begin{aligned} u_{k,m}\left[n+m\right] &= S_k\left[n-Q\right]\sum_{l=1}^{L}\left(\sum_{i=-\infty}^{\infty} b_l[i+m-P]S_l\left[n-i\right]\right. \\ &\left.-\sum_{j=0}^{N-1}\sum_{i=-\infty}^{\infty} h_c[i+m-j]a_{l,j}[n-i+j]S_l[n-i]\right) + S_k\left[n-Q\right]r_{ideal}\left[n+m\right] \end{aligned} \tag{143}$$

Combining (142) and (143) for m = 0, 1, …, N−1 and applying the definition of $\mathbf{a}_k[n]$ = = $[a_{k,0}[n], a_{k,1}[n+1], \ldots, a_{k,N-1}[n+N-1]]^T$ yields

$$\mathbf{a}_k[n] = \mathbf{a}_k[n-1] + K\sum_{i\geq-(N-2)}\sum_{l=1}^{L} S_k\left[n-Q\right]S_l\left[n-i\right]\left(\mathbf{b}_{l\_i} - \mathbf{H}_{\mathbf{c}\_i}\mathbf{a}_l[n-i]\right) + KS_k\left[n-Q\right]\mathbf{r}[n], \tag{144}$$

where

$$\mathbf{b}_{l\_i} = \begin{bmatrix} b_l[i-P] \\ b_l[i-P+1] \\ \vdots \\ b_l[i-P+N-1] \end{bmatrix},\ \mathbf{r}[n] = \begin{bmatrix} r_{ideal}[n] \\ r_{ideal}[n+1] \\ \vdots \\ r_{ideal}[n+N-1] \end{bmatrix},$$

$$\mathbf{H}_{\mathbf{c}\_i} = \begin{bmatrix} h_c[i] & h_c[i-1] & \cdots & h_c[i-N+1] \\ h_c[i+1] & h_c[i] & \cdots & h_c[i-N+2] \\ \vdots & \vdots & \ddots & \vdots \\ h_c[i+N-1] & h_c[i+N-2] & \cdots & h_c[i] \end{bmatrix}. \tag{145}$$

The lower limit of i in (144) is set to –(N−2), this is because it follows from $b_k[n]$ and $h_c[n]$ = 0 for all $n \leq 0$ that $\mathbf{b}_{k\_i}$ and $\mathbf{H}_{\mathbf{c}\_i}$ are both **0** for i < –(N−2). Equation (144) holds for all $n \geq 0$ with the initial condition at $n \leq -1$.

It follows from the definition of $\mathbf{z}_k[n] = \mathbf{a}_k[n] - \mathbf{a}_k'$ that

$$\mathbf{b}_{k\_i} - \mathbf{H}_{\mathbf{c}\_i}\mathbf{a}_k[n-i] = (\mathbf{b}_{k\_i} - \mathbf{H}_{\mathbf{c}\_i}\mathbf{a}_k') - \mathbf{H}_{\mathbf{c}\_i}\mathbf{z}_k[n-i]. \tag{146}$$

Replacing $\mathbf{a}_k[n]$ and $\mathbf{a}_k[n-1]$ in (144) with $\mathbf{z}_k[n]+\mathbf{a}_k'$ and $\mathbf{z}_k[n-1]+\mathbf{a}_k'$, respectively and then substituting (146) with k replaced by l into (144) yields

$$\mathbf{z}_k[n] = \mathbf{z}_k[n-1] - K \sum_{i\geq-(N-2)} \sum_{l=1}^{L} S_k\left[n-Q\right] S_l\left[n-i\right] \mathbf{H}_{\mathbf{c}\_i} \mathbf{z}_l[n-i] + K\mathbf{e}_{k,n}, \tag{147}$$

where each $\mathbf{e}_{k,n}$ is a zero-mean additive noise vector given by

$$\mathbf{e}_{k,n} = S_k\left[n-Q\right]\mathbf{r}[n] + \sum_{i\geq-(N-2)} \sum_{l=1}^{L} S_k\left[n-Q\right] S_l\left[n-i\right] (\mathbf{b}_{l\_i} - \mathbf{H}_{\mathbf{c}\_i}\mathbf{a}_l'). \tag{148}$$

The difference equation (147) holds for all $n \geq 0$ with the initial condition of $\mathbf{z}_k[n] = \mathbf{a}_k[n]-\mathbf{a}_k'$ for $n \leq -1$. It was proven in [15] that the $\mathbf{b}_{l\_i}-\mathbf{H}_{\mathbf{c}\_i}\mathbf{a}_l'$ terms in (148) for $i = Q$ and all l are $\mathbf{0}$.

It follows from the definition of $\mathbf{r}[n]$ that $S_k[n-Q]\mathbf{r}[n]$ in (148) represents the portion of $\mathbf{e}_{k,n}$ contributed by the ADC noise and main DAC's signal component, while the remaining portion represents a small portion of the main DAC's mismatch error that MNC cannot fully correct.

B. Evaluating Mean-Square Error

Let us define the following *RMS norm* for any N-dimensional real vector $\mathbf{v} = [v_j]$,

$$\left\|\mathbf{v}\right\| = \sqrt{\sum_{j=1}^{N} E\left[{v_j}^2\right]}. \tag{149}$$

The RMS norm is a useful metric to evaluate the mean squared error of each coefficient, because it follows from the definition of $\mathbf{z}_k[n]$ that $\|\mathbf{z}_k[n]\|^2$ represents the sum of the mean squared error of $a_{k,0}[n]$, $a_{k,1}[n]$, …, and $a_{k,N-1}[n]$.

Let us constrain K to be positive, and define the following system related parameter,

$$h_K = \min_{0<\alpha\leq K}\left(1-\left\|\mathbf{I}-\alpha\mathbf{H}_{\mathbf{c}\_Q}\right\|_2\right)/\alpha, \tag{150}$$

where $\|\mathbf{I}-\alpha\mathbf{H}_{\mathbf{c}\_Q}\|_2$ is the *spectral norm* of $\mathbf{I}-\alpha\mathbf{H}_{\mathbf{c}\_Q}$. For any $N\times N$ deterministic real matrix $\mathbf{D}$, the spectral norm of $\mathbf{D}$ is given by

$$\| \mathbf{D} \|_2 = \max_{\mathbf{v} \in \mathbb{R}^N, \|\mathbf{v}\|_2 \neq 0} \left(\|\mathbf{Dv}\|_2 / \|\mathbf{v}\|_2 \right), \tag{151}$$

where $\|\mathbf{Dv}\|_2$ and $\|\mathbf{v}\|_2$ are the Euclidean norm of the vectors **Dv** and **v**, respectively, which is equal to the square root of the largest eigenvalue of $\mathbf{D}^{\mathrm{H}}\mathbf{D}$, where $\mathbf{D}^{\mathrm{H}}$ denotes the conjugate transpose of **D**.

The results of the analysis in this section is summarized as follows.

Theorem 1: If the magnitude of $h_{\mathrm{c}}[n]$ is bounded by an exponentially decaying curve as n increases, and if there exists a positive value of K satisfying $h_K > 0$, then for any positive number ε, there must exist a positive number δ such that

$$\limsup_{n\to\infty} \left\| \mathbf{z}_k[n] \right\|^2 < \varepsilon \text{ for all } 0 < K < \delta \tag{152}$$

and $k = 1, 2, \ldots, L$, provided that there exists a bounded positive integer M such that $c_{\min}[n] \neq 0$ occurs at least once in any consecutive M samples, where $c_{\min}[n]$ is the minimum value of $E\{S_k^2[n]\}$ over $k = 1, 2, \ldots, L$ at time index n.

The limit superior of $\|\mathbf{z}_k[n]\|^2$ in (152) represents the steady-state mean squared error. It follows from (152) that the steady-state mean squared error can be made arbitrarily small by reducing the magnitude of K.

Theorem 2: If the magnitude of $h_{\mathrm{c}}[n]$ is bounded by an exponentially decaying curve as n increases, and if there exists a positive value of K such that $h_K > 0$ is satisfied, and if the statistics of $\mathbf{r}[n]$ and $S_k[n]$ for each k do not change over time, then for any positive number ε, there must exist a positive number δ such that for all $0 < K < \delta$, the limit superior of $\|\mathbf{z}_k[n]\|^2$ satisfies

$$\limsup_{n\to\infty} \left\| \mathbf{z}_k[n] \right\|^2 \le K\alpha \left(\left\| \mathbf{r}[n] \right\| + \sum_{i \ge -(N-2)} \sum_{l=1}^{L} \left\| \mathbf{b}_{l\_i} - \mathbf{H}_{\mathbf{c}\_i} \mathbf{a}_l \,' \right\| \right)^2, \tag{153}$$

where

$$\alpha \leq \frac{1}{2h_K} + \varepsilon \,, \tag{154}$$

provided that $c_{\min}[n] > 0$ for all $k = 1, 2, \ldots, L$.

The hypothesis that the statistics of $\mathbf{r}[n]$ and the $S_k[n]$ sequences are constant over time is reasonable in the case of foreground MNC calibration.

In practice, the contribution of the $\|\mathbf{r}[n]\|$-associated term in (153) is much larger than the other terms, thus if ε is sufficiently small, the bound in (153)-(154) reduces to $KN\sigma_r^2/(2h_K)$, where $\sigma_r^2 = \|\mathbf{r}[n]\|^2/N$ represents the power of the ADC noise and main DAC's signal component after being filtered by the decimation filter. However, Theorem 2 does not rigorously quantify how small δ needs to be such that (153)-(154) is satisfied for a given ε, the answer to this question requires sophisticated and tedious computation. Instead, the analysis yields a guideline of K, which states that for (153)-(154) to be satisfied for a reasonably small value of ε, it requires

$$0 < K << 2h_K \,/\, A, \tag{155}$$

where $$A = L \sum_{i \geq -(N-2)} \left\|\mathbf{H}_{\mathbf{c}\_i}\right\|_2^2 + \sum_{q=Q+1}^{2Q+N-1} \left\|\mathbf{H}_{\mathbf{c}\_q}\right\|_2 \left\|\mathbf{H}_{\mathbf{c}\_(2Q-q)}\right\|_2 + (1+Q)\left\|\mathbf{H}_{\mathbf{c}\_Q}\right\|_2^2 + \sum_{i=-(N-2)}^{0} \left\|\mathbf{H}_{\mathbf{c}\_i}\right\|_2^2 . \tag{156}$$

If the impulse response of the MNC feedback path is a delayed delta function satisfying $h_c[n] = 1$ if $n = Q$ and $h_c[n] = 0$ otherwise, (155)-(156) reduces to a simple equation of

$$0 < K << \frac{2}{2NL + 2N + Q} \tag{157}$$

The results of (155)-(156) together with $h_K > 0$ provide guidelines in the choice of MNC design parameters. It is found that (155)-(156) is almost always satisfied in practice (e.g., K used in the prototype DAC IC in [16] is close to $2h_K/3000A$). Although the derivation of (155)-(156) is not rigorous and the definition of << 1 is relatively vague, the simulation results in Section V confirmed that in almost all practical cases, the upper bound of (153)-

(154) can be used to estimate the impact of correlation noise on the post-calibration DAC SNR within an accuracy of 1dB.

The proof of Theorem 1 and 2 is presented in the remainder of this section.

Proof of Theorem 1:

The $S_k[n-Q]S_l[n-i]\mathbf{H}_{\mathbf{c}\_i}$ matrix and the additive noise vector $\mathbf{e}_{k,n}$ in (147) are both associated with the $S_k[n-Q]S_l[n-i]$ variables, each of these variables are restricted within the range of -1 and 1 and are associated with samples of the $S_k[n]$ sequences. In this paper, these type of variables are referred to as *modulation variables*. More accurately, the modulation variables, denoted as s, is defined as a single or a product of variables, each variable must be within the range of -1 and 1 and is either a deterministic scaling factor or associated with samples of $S_k[n]$ sequences for $1 \leq k \leq L$ at a single time index n, and different variables cannot be associated with samples of $S_k[n]$ sequences of the same time index. For example, $S_1[n-3]$, $(S_3^2[n-1]S_1^2[n-1])(S_1^2[n-2])$ and $0.5(S_1^2[n-3]–1/2)(S_2^2[n-6])$ are all modulated variables, because each of them is either a single or a product of variables satisfying the above definition.

It follows from (148) that $\mathbf{e}_{k,n}$ is a sum of vectors, each vector in the form of $s_j\mathbf{q}_j$, where s_j is in the form of a modulation variable, and $\mathbf{q}_j$ is a vector independent of any sample of $S_k[n]$ sequences for $1 \leq k \leq L$. This type of vector is referred to as *modulated vector* in the paper. It is possible for a modulated vector to be a function of K, but it is required that for any bounded value of K, the sum of the RMS norm of each $\mathbf{q}_j$ vector in any modulated vector is bounded.

Equation (147) is in a “non-causal” form since $\mathbf{z}_k[n]$ at time index n is affected by $\mathbf{z}_k[n-i]$ for $i \leq 0$, this is an artifact of the analysis. It follows from Lemma 1 in Appendix A that (147) can be converted into an alternative causal form as

$$\mathbf{z}_k\left[n\right]=\mathbf{z}_k\left[n-1\right]-\sum_{q=1}^{3}K^q\mathbf{H}_{q,k}\left(\mathbf{z}\left[n-1\right]\right)+K\mathbf{e}_{k,n}+K^2\mathbf{v}_{k,n}\ \text{ for } k=1,2,...,L, \tag{158}$$

for all $n \geq 0$, with the initial condition for $n \leq -1$ identical to that of (147), where each $\mathbf{H}_{q,k}(\mathbf{z}[n-1])$ for q = 1, 2, 3 has the form of $\mathbf{H}(\mathbf{z}[n-1])$ defined in Lemma 1, and $\mathbf{v}_{k,n}$ is a modulated vector. The $\mathbf{H}_{1,k}(\mathbf{z}[n-1])$ term represents the double sum term in (147) except that each state variable with time index larger than $n-1$ is replaced with the same state variable with time index $n-1$, i.e.,

$$\mathbf{H}_{1,k}\left(\mathbf{z}\left[n-1\right]\right)=\sum_{i\geq 1}\sum_{l=1}^{L}\mathbf{H}_{k,l,i,n}\mathbf{z}_l\left[n-i\right], \tag{159}$$

where

$$\mathbf{H}_{k,l,i,n}=\begin{cases}\sum\limits_{j=-(N-2)}^{1}S_k[n-Q]S_l[n-j]\mathbf{H}_{\mathbf{c}\_j}, & \text{if } i=1,\\ S_k[n-Q]S_l[n-i]\mathbf{H}_{\mathbf{c}\_i}, & \text{otherwise.}\end{cases} \tag{160}$$

The $\mathbf{H}_{2,k}(\mathbf{z}[n-1])$ term is derived by summing up $S_k[n-Q]S_l[n-j]\mathbf{H}_{\mathbf{c}\_j}\cdot\mathbf{H}_{1,l}(\mathbf{z}[m-1])$ over $1 \leq l \leq L$, $-(N-2) \leq j \leq 0$, $n \leq m \leq n-j$, and then replacing each state variable with time index larger than $n-1$ by the same state variable with time index $n-1$. Each $\mathbf{H}_{q,k}(\mathbf{z}[n-1])$ for q = 1, 2, 3 was defined in Lemma 1 for $n \geq 0$. For the completeness of analysis, let us define each $\mathbf{H}_{q,k}(\mathbf{z}[n-1])$ for $n \leq -1$ as $\mathbf{0}$.

Equation (158) has multiple state variables $\mathbf{z}_k[n]$ for k = 1, 2, …, L. The value of $\mathbf{z}_k[n]$ at any given time $n \geq 0$ is a linear combination of the initial condition at $n \leq -1$ and the additive noise vectors, $K\mathbf{e}_{k,p}+K^2\mathbf{v}_{k,p}$ for p = 0, 1, …, n. If each $K\mathbf{e}_{k,n}+K^2\mathbf{v}_{k,n}$ in (158) is replaced with $\mathbf{0}$, (158) reduces to

$$\mathbf{x}_k\left[n\right]=\mathbf{x}_k\left[n-1\right]-\sum_{q=1}^{3}K^q\mathbf{H}_{q,k}\left(\mathbf{x}\left[n-1\right]\right)\ \text{ for } k=1,2,...,L \tag{161}$$

and $n \geq 0$, with an initial condition of $\mathbf{x}_k[n] = \mathbf{z}_k[n]$ for $n \leq -1$, where $\mathbf{H}_{q,k}(\mathbf{x}[n-1])$ is given by $\mathbf{H}_{q,k}(\mathbf{z}[n-1])$ with each state variable, $\mathbf{z}_l[n-i]$, replaced by $\mathbf{x}_l[n-i]$ of the same index l. Therefore, $\mathbf{x}_k[n]$ represents the portion of $\mathbf{z}_k[n]$ contributed by its initial condition. The definition of $\mathbf{z}_k[n] = \mathbf{a}_k[n]-\mathbf{a}_k'$ for $n \le -1$ and the initial condition of $a_{k,m}[n]$ in (139)-(140) imply that $\mathbf{z}_k[n]$ for all $n \le -1$ are modulated vectors and $\mathbf{z}_k[n] = -\mathbf{a}_k'$ for all $n \le -N-1$. Therefore, by defining $\mathbf{H}_{q,k}(\mathbf{z}[n-1]) = \mathbf{0}$ for all $n \le -1$, it follows that $\mathbf{x}_k[n]$ in (161) resulting from the above initial condition can be viewed as the sum of N+1 portions, each portion is given by (161) evaluated from a different initial condition at $n \le j$ in the form of

$$\mathbf{x}_k[n] = \begin{cases} \mathbf{v}_k[n], & \text{if } n = j, \\ \mathbf{0}, & \text{if } n < j, \end{cases} \quad \text{for } k = 1, 2, ..., L, \tag{162}$$

where $\mathbf{v}_k[n] = \mathbf{z}_k[n]-\mathbf{z}_k[n-1]$ for $j = -1, -2, \ldots, -N$ and $\mathbf{v}_k = -\mathbf{a}_k'$ for $j = -\infty$. The definition of $\mathbf{H}_{q,k}(\mathbf{z}[n-1]) = \mathbf{0}$ for all $n \le -1$ ensures that $\mathbf{x}_k[n]$ with the initial condition given by (162) evaluates to $\mathbf{v}_k[j]$ at each time of $j < n \le -1$.

Similarly, the contribution of $K\mathbf{e}_{k,p}+K^2\mathbf{v}_{k,p}$ at time index p to $\mathbf{z}_k[n]$ at time index n can be estimated by evaluating (161) from a different initial condition at $n \le p$ given by

$$\mathbf{x}_k[n] = \begin{cases} K\mathbf{e}_{k,p} + K^2\mathbf{v}_{k,p}, & \text{if } n = p, \\ \mathbf{0}, & \text{if } n < p, \end{cases} \quad \text{for } k = 1, 2, ..., L. \tag{163}$$

Therefore, evaluating $\mathbf{x}_k[n]$ in (161) with each initial condition in (162)-(163) and applying superposition yields $\mathbf{z}_k[n]$.

It follows that each non-zero vector of $\mathbf{x}_k[n]$ in (162)-(163) is a modulated vector, and its modulation variables are either deterministic 1, or is a single or a product of samples of $S_k[n]$ sequences for $1 \le k \le L$, thus each of the above initial condition can be expressed as a sum of finite number of *basis initial condition*, each of which has the form

$$\mathbf{x}_k[n] = s_{\text{base}} \begin{cases} \mathbf{q}_{\text{base}(k)}, & \text{if } n = j_{\text{ini}}, \\ \mathbf{0}, & \text{if } n < j_{\text{ini}}, \end{cases} \quad \text{for } k = 1, 2, ..., L, \tag{164}$$

where j_{ini} represents j in (162) or p in (163), it can be any integer including negative infinity, s_{base} is either deterministic 1, or is a single or a product of samples of $S_k[n]$ sequences for $1 \le k \le L$ with the time indexes restricted within the range of $j_{\text{ini}}-Q < n < j_{\text{ini}}+N$ (samples with time indexes below this range are grouped into $\mathbf{q}_{\text{base}(k)}$, while samples with time indexes above this range have not yet appeared in the system thus do not influence the initial condition), while $\mathbf{q}_{\text{base}(k)}$ is a vector independent of any samples of $S_k[n]$ sequences for $1 \le k \le L$ except for those at time index $n \le j_{\text{ini}}-Q$. In addition, s_{base} in (164) associated with $\mathbf{x}_k[n]$ are identical for all k, while $\mathbf{q}_{\text{base}(k)}$ can be dependent on k. It is required that $E(s^2_{\text{base}}) \neq 0$, and $\|\mathbf{q}_{\text{base}(k)}\|$ can not 0 for all k, otherwise (164) reduces to a trivial case of all 0.

Equation (161), (164), and $\mathbf{H}_{q,k}(\mathbf{z}[n-1]) = \mathbf{0}$ for all $n \le -1$ imply that $\mathbf{x}_k[n]$ for all $j_{\text{ini}} \le n \le \max(j_{\text{ini}}, -1)$ are identical. Therefore, it is sufficient to analyze $\mathbf{x}_k[n]$ for all $n \ge \max(j_{\text{ini}}+1, 0)$. Let $x_{\max}[n]$ denote

$$x_{\max}[n] = \max_{1 \le k \le L} \|\mathbf{x}_k[n]\|, \tag{165}$$

and let $c_k[n]$ denote

$$c_k[n] = E\left\{S_k{}^2[n] s_{\text{base}}{}^2\right\} / E\left\{s_{\text{base}}{}^2\right\}, \tag{166}$$

where s_{base} is the scaling variable in the basis initial condition given by (164). It follows from (166) that

$$c_k[n] = c\left\{S_k{}^2[n]\right\} \le 1 \text{ for all } k, n. \tag{167}$$

Furthermore, since s_{base} is not associated with any sample of $S_k[n]$ sequences at time index $n \ge j_{\text{ini}}+N$, it follows that

$$c_k[n] = E\left\{S_k{}^2[n]\right\} \text{ for } n \ge j_{\text{ini}} + N. \tag{168}$$

Under the hypothesis of Theorem 1 that there exists a bounded positive integer M such that $c_{\min}[n] \neq 0$ occurs at least once in any consecutive M samples, it follows from Lemma 3 and the properties of $c_k[n-Q]$ in (167)-(168) that $\|\mathbf{x}_k[n]\|^2$ for any $n \ge \max(j_{\text{ini}}+1, 0)$ evaluated from any basis initial condition in (164) must be bounded by an exponential decaying curve given by

$$\|\mathbf{x}_k[n]\|^2 \le x_{\max}{}^2[j_{\text{ini}}](1 - ch_K K)^{\lfloor (n-\max(j_{\text{ini}}+1,0)-Q-N)/M \rfloor} \tag{169}$$

for $1 \le k \le L$, provided that the magnitude of K is less than a certain value, where c is a positive constant independent of K, but is associated with the value of $c_{\min}[n]$ averaged over the entire period of convergence. Let the value of K be chosen smaller than this upper limit. It follows from (158) that $\mathbf{z}_k[n]$ at time index $n \geq 0$ is a linear function of its initial condition and the additive noise terms $K\mathbf{e}_{k,p}+K^2\mathbf{v}_{k,p}$ for $p = 0, 1, \ldots, n$. The contribution of the initial condition to $\mathbf{z}_k[n]$ can be obtained by evaluating $\mathbf{x}_k[n]$ in (161) with the initial condition in (162). The initial condition in (162) can be expressed as a sum of finite number of basis initial conditions in the form of (164), and it follows from (169) that $\|\mathbf{x}_k[n]\|^2 \rightarrow 0$ as $n \rightarrow \infty$ from each of these basis initial conditions, thus the rule of superposition from Lemma 3 implies that $\|\mathbf{x}_k[n]\|^2 \rightarrow 0$ as $n \rightarrow \infty$ from the overall initial condition. This implies that the contribution of the system's initial condition to $\|\mathbf{z}_k[n]\|^2$ dies out exponentially as $n \rightarrow \infty$, which by itself satisfies (152).

Similarly, the contribution of each $K\mathbf{e}_{k,p}+K^2\mathbf{v}_{k,p}$ to $\mathbf{z}_k[n]$ can be obtained by evaluating $\mathbf{x}_k[n]$ in (161) with the initial condition in (163). The initial condition in (163) for a given p also can be written as a sum of finite number of basis initial conditions in the form of (164), where $j_{\text{ini}} = p$ and $s_{\text{base}}\mathbf{q}_{\text{base}(k)}$ represents a portion of $K\mathbf{e}_{k,p}+K^2\mathbf{v}_{k,p}$. Therefore, the contribution of each $s_{\text{base}}\mathbf{q}_{\text{base}(k)}$ component of $K\mathbf{e}_{k,p}+K^2\mathbf{v}_{k,p}$ to $\mathbf{z}_k[n]$ follows from (161) and (164) with $j_{\text{ini}} = p$. Each $\mathbf{H}_{q,k}(\mathbf{x}[n-1])$ for $q = 1, 2, 3$ in (161) has the form of $\mathbf{H}_s(\mathbf{x}[n-1])$, the definition of $\mathbf{H}_s(\mathbf{x}[n-1])$ implies that each state variable, $\mathbf{x}_l[n-i]$, in $\mathbf{H}_{q,k}(\mathbf{x}[n-1])$ is scaled by samples of $S_k[n]$ sequences with time indexes strictly larger than $n-i-Q$. It follows from (164) with $j_{\text{ini}} = p$ that all the state variables with time indexes smaller than p are $\mathbf{0}$, thus this portion of state variables are removed from $\mathbf{H}_{q,k}(\mathbf{x}[n-1])$ to simplify the analysis. Consequently, all state variables left in $\mathbf{H}_{q,k}(\mathbf{x}[n-1])$ are scaled by samples of $S_k[n]$ sequences with time indexes larger than $p-Q$. Furthermore, (148) implies that each $s_{\text{base}}\mathbf{q}_{\text{base}(k)}$ component of $K\mathbf{e}_{k,p}$ is given by either $KS_k[p-Q]\mathbf{r}[p]$ or $KS_k[p-Q]S_l[p-i](\mathbf{b}_{l\_i}-\mathbf{H}_{\mathbf{c}\_i}\mathbf{a}_l')$ with $i \neq Q$. These observations imply that the contributions of $KS_k[p-Q]\mathbf{r}[p]$ for different p to $\mathbf{z}_k[n]$ are uncorrelated, and the contributions of $KS_k[p-Q]S_l[p-i](\mathbf{b}_{l\_i}-\mathbf{H}_{\mathbf{c}\_i}\mathbf{a}_l')$ for different p to $\mathbf{z}_k[n]$ are also uncorrelated, thus the contribution from the same component of $K\mathbf{e}_{k,p}$ for $p = 0, 1, \ldots, n$ to $\|\mathbf{z}_k[n]\|^2$ is given by the sum of the power of individual contribution. The contribution of the $KS_k[p-Q]\mathbf{r}[p]$ component of $K\mathbf{e}_{k,p}$ to $\|\mathbf{z}_k[n]\|^2$ is bounded by (169) with $x_{\max}{}^2[p] = K^2c_{\max}[p-Q]\|\mathbf{r}[p]\|$, while the contribution of the $KS_k[p-Q]S_l[p-i]$ $(\mathbf{b}_{l\_i}-\mathbf{H}_{\mathbf{c}\_i}\mathbf{a}_l')$ component of $K\mathbf{e}_{k,p}$ is bounded by (169) with $x_{\max}{}^2[p] = K^2c_{\max}[p-Q]$ $\|S_l[p-i](\mathbf{b}_{l\_i}-\mathbf{H}_{\mathbf{c}\_i}\mathbf{a}_l')\|$, where $c_{\max}[p-Q]$ is the maximum value of $E\{S_k{}^2[p-Q]\}$ over k. Summing up all these components, it follows that the contribution of $K\mathbf{e}_{k,p}$ for $p = 0, 1, \ldots, n$ to $\|\mathbf{z}_k[n]\|^2$ is given by $o(K)$, where $o(K)$ is bounded and in the order of K.

Unfortunately, this zero-correlation property does not apply for each component of $K^2\mathbf{v}_{k,p}$. Instead, Lemma 2 is used to bound the overall contribution of $K^2\mathbf{v}_{k,p}$ for $p = 0, 1, \ldots, n$ to $\|\mathbf{z}_k[n]\|$ by adding up the individual contribution in magnitude. The contribution of each $s_{\text{base}}\mathbf{q}_{\text{base}(k)}$ component to $\|\mathbf{z}_k[n]\|$ is given by $\|\mathbf{x}_k[n]\|$, which is bounded by the square root of the right side of (169), and since each $s_{\text{base}}\mathbf{q}_{\text{base}(k)}$ component of $K^2\mathbf{v}_{k,p}$ is scaled by K^2, this implies that the overall contribution of $K^2\mathbf{v}_{k,p}$ to $\|\mathbf{z}_k[n]\|$ and $\|\mathbf{z}_k[n]\|^2$ are in the order of $o(K)$ and $o(K^2)$, respectively.

Combining all the contributions to $\|\mathbf{z}_k[n]\|^2$ yields (152).

□

Proof of Theorem 2

The proof of Theorem 1 shows that there exists a positive upper limit of K below which the contribution of the system's initial condition to $\|\mathbf{z}_k[n]\|^2$ dies out to 0 as $n \to \infty$, thus it remains to analyze the contribution from $K\mathbf{e}_{k,p}+K^2\mathbf{v}_{k,p}$. The proof of Theorem 1 also shows that the overall contribution of $K^2\mathbf{v}_{k,p}$ for $p = 0, 1, \ldots, n$ to $\|\mathbf{z}_k[n]\|^2$ at any time n is in the order of $o(K^2)$, which decreases with K at a rate faster than that of (153)-(154).

Therefore, it is sufficient to analyze the overall contribution of $K\mathbf{e}_{k,p}$ for $p = 0, 1, \ldots, n$ to $\|\mathbf{z}_k[n]\|^2$.

It follows from (148) that each component of $K\mathbf{e}_{k,p}$ is given by $KS_k[p-Q]\mathbf{r}[p]$ or $KS_k[p-Q]\ S_l[p-i](\mathbf{b}_{l\_i}-\mathbf{a}_l')$ for $i \neq Q$ and $l = 1, 2, \ldots, L$, each of which has the form of $s_{\text{base}}\mathbf{q}_{\text{base}(k)}$ in (164). The contribution of each $s_{\text{base}}\mathbf{q}_{\text{base}(k)}$ component to $\|\mathbf{z}_k[n]\|^2$ is given by $\|\mathbf{x}_k[n]\|^2$, where $\mathbf{x}_k[n]$ follows (161) and (164) with $j_{\text{ini}} = p$. The subsequent analysis applies (182) of Lemma 3 to yield a tight bound of $\|\mathbf{x}_k[n]\|^2$. Lemma 3 states that for any given $0 < \beta < 1$, (182) holds if K is chosen small enough.

Let us first discuss a hypothetical case where β in (182) is replaced by 0, it follows that (182) reduces to

$$\frac{\|\mathbf{x}_k[n]\|^2}{\|\mathbf{x}_k[n-1]\|^2} \leq \begin{cases} 1-2c_k[n-Q]Kh_K, & \text{if } n-Q \geq p, \\ 1, & \text{otherwise.} \end{cases} \tag{170}$$

It follows from (170) and the properties of $c_k[n]$ in (167)-(168) that the contribution of each $s_{\text{base}}\mathbf{q}_{\text{base}(k)}$ component of $K\mathbf{e}_{k,p}$ to $\|\mathbf{z}_k[n]\|^2$ is bounded by $\|\mathbf{x}_k[n]\|^2 \leq \|\mathbf{x}_k[p]\|^2(1-2E\{S_k^2[n]\}Kh_K)^{(n-p-Q-N)}$ for $0 \leq p \leq n$, where $E\{S_k^2[n]\}$ is time-invariant and satisfies $E\{S_k^2[n]\} > 0$ for $k = 1, 2, \ldots, L$ as implied by the hypothesis of Theorem 2, $\|\mathbf{x}_k[p]\|^2 = K^2\|S_k[p-Q]\mathbf{r}[p]\|^2$ for $s_{\text{base}}\mathbf{q}_{\text{base}(k)} = KS_k[p-Q]\mathbf{r}[p]$ and $\|\mathbf{x}_k[p]\|^2 = K^2\|S_k[p-Q]S_l[p-i](\mathbf{b}_{l\_i}-\mathbf{a}_l')\|^2$ for $s_{\text{base}}\mathbf{q}_{\text{base}(k)} = KS_k[p-Q]S_l[p-i](\mathbf{b}_{l\_i}-\mathbf{a}_l')$. It follows from the proof of Theorem 1 that the contributions of the same $s_{\text{base}}\mathbf{q}_{\text{base}(k)}$ component of $K\mathbf{e}_{k,p}$ for $p = 0, 1, \ldots, n$ to $\mathbf{z}_k[n]$ are guaranteed to be uncorrelated. Summing up each component's contribution for $p = 0, 1, \ldots, n$ in power and over i , l in magnitude yields (153), where $\alpha = K(1+o(K))/(2h_K)$, thus for any given positive ε, (154) holds if K is chosen small enough.

However, as stated in Lemma 3, the validity of (182) requires the value of β to be larger than 0. If (182) with $\beta > 0$ is used to estimate the bound, it follows from intuition that for any positive Ɛ, (153)-(154) would still hold if both β and K are small enough, this is proven in Appendix B, which yields (153) with $\alpha = K(1+o(K))(1+(c_{\max}[n]/c_{\min}[n])\beta/(1-\beta))/(2h_K)$, where $c_{\min}[n]$ and $c_{\max}[n]$ are the minimum and maximum values of $E\{S_k^2[n]\}$ over k, respectively. Since (182) is guaranteed by choosing K to be small enough, it follows that (153)-(154) hold provided that K is chosen small enough and $c_{\min}[n] > 0$.

□

The proof of Theorem 2 implies that in order for (153)-(154) to be satisfied for a small positive number, Ɛ, (182) needs to be satisfied for a small positive number, β. It follows from the proof of Lemma 3 that this can be achieved by choosing K to be small enough such that λ given by (204) satisfies $0 < \lambda < \beta$. It follows from the expression of λ that if the second-order terms are ignored, K is required to satisfy (155).

IV. DISCUSSIONS

This section derives useful information from the analysis of section III to quantify the contribution of correlation noise to the post-calibration DAC SNR. It also presents design guidelines to minimize this noise contribution.

The value of $\|\mathbf{z}_k[n]\|^2$ for $n \to \infty$, where $\|\mathbf{z}_k[n]\|$ is the RMS norm of $\mathbf{z}_k[n]$ defined in (149), represents the overall steady-state mean squared error of $a_{k,m}[n]$ for $m = 1, 2, \ldots, N-1$. The results of Section III shows that if $h_K > 0$ is satisfied, then the steady-state mean squared error can be arbitrarily reduced by decreasing the magnitude of K as given by (152), and if the magnitude of K is chosen “sufficiently small”, the steady-state mean squared error in MNC foreground calibration is pessimistically bounded by (154). It further provides a guideline in the choice of K given by (155).

Let us illustrate these results with a design example. Suppose the MNC feedback path's impulse response, $h_c[n]$, can be approximated as a delayed delta function and Q is chosen as the delay, i.e., $h_c[n] = 1$ if $n = Q$ and $h_c[n] = 0$ otherwise, it follows from the guideline in (157) that for a prototype design of $N = 9$, $L = 35$ and $Q = 21$ from [16], K needs to satisfy $0 < K << 0.003$ as indicated by (155). This can be easily achieved without compromising the convergence rate derived from [15]. Indeed, the actual value of K used in the prototype design is of the order of 10^{-6} and thus is much smaller than this upper bound.

In practice, any deviation in the shape of $h_c[n]$ from a delayed delta function will cause the value of h_K to decrease from 1, and it follows from Theorem 1's hypothesis that it is important to keep h_K positive. This can be achieved by keeping the bandwidth of the feedback path high enough (preferred to be above the Nyquist frequency of the DAC) and choose Q as the closest integer to the average delay of the feedback path.

The results of (153)-(154) can be directly applied to estimate the contribution of correlation noise to post-calibration DAC SNR. It is worth noting that since the MNC coefficients are frozen at the end of the foreground calibration and subsequently used in the normal operation, the actual noise component on each coefficient will vary from calibration to calibration. This also means that with a given set of design parameters, the post-calibration DAC SNR contributed by correlation noise may slightly vary from calibration to calibration. The analysis presented in this paper provides an estimation of the average contribution of correlation noise to post-calibration DAC SNR across different runs of calibrations, but not the SNR variation across different runs of calibrations. Fortunately, it is found through simulations that this variation is very small (<1dB), which is largely due to the averaging effect from the large number of coefficients used for generating mismatch cancellation waveform.

The DAC input sequence during normal operation is usually different from that used during MNC foreground calibration, thus each $c_k[n] = E\{S_k^2[n]\}$ during normal operation is different from the corresponding value during the foreground calibration. Let $c_{k\_\mathrm{post}}[n]$ denote the value of $E\{S_k^2[n]\}$ during normal operation. Furthermore, since the impact of the $\|\mathbf{r}[n]\|$-associated term in (153) is much larger than the other terms, only this dominant term is used for estimation. It follows from (153)-(154) with ε = 0 that the average contribution of correlation noise to post-calibration DAC SNR can be pessimistically bounded by

$$\mathrm{SNR}_\mathrm{c} \geq 10\log\left(P_\mathrm{sig} \Bigg/ \left(\frac{a_\mathrm{shape} KN \sum_{k=1}^{L} c_{k\_\mathrm{post}} \sigma_r{}^2}{2h_K} \right)\right), \tag{171}$$

where the subscript c in the notation of SNR_c is used to imply that it represents the SNR contributed by the correlation noise only, not the overall DAC SNR that would otherwise include quantization noise, etc. The P_sig term is the DAC signal power over the first Nyquist band, c_{k_post} is the time average of $c_{k\_\mathrm{post}}[n]$, a_shape represents the roll-off introduced by the correction DAC's pulse shape given by

$$a_\mathrm{shape} = f_s \int_{-f_s/2}^{f_s/2} |A_c(j2\pi f)|^2 \, df, \tag{172}$$

where $A_c(j2\pi f)$ is the Fourier Transform of the correction DAC's pulse shape limited within a $1/f_s$-period. It follows that if the correction DAC's pulse shape is an ideal RZ rectangular pulse with 80% duty cycle and magnitude of 1, then $a_\mathrm{shape} = 0.54$ and is independent of f_s.

The value of N in (172) represents the number of coefficients in each $s_k[n]$ residue estimator shown in Fig. 35. It follows from [15] that a relatively large N (e.g., N = 9) is necessary to achieve significant cancellation static and dynamic mismatch error. However, (172) implies that a larger N results in a higher sensitivity of the post-calibration DAC SNR to the correlation noise. Fortunately, this problem can be addressed by reducing the magnitude of K, as it follows from (172) that each reduction of K by half results in a 3 dB improvement of SNR_c.

V. SIMULATION RESULTS

Three sets of simulations are performed to validate the contribution of correlation noise on the post-calibration DAC SNR predicted by (172). Both main and correction DACs incorporate RZ 1-bit DACs with an 80% duty cycle, which results in $a_{\text{shape}} = 0.54$. The clock rates of both main and correction DACs are 600 MHz, while the clock rate of the oversampling ADC is 3 GHz.

The simulation is performed with MNC first operating in foreground mode, and wait sufficiently long until all the coefficients fluctuate around their steady-state mean values. The values of these coefficients are subsequently frozen and used to generate mismatch cancellation waveform for the main DAC during normal operation. In the foreground mode, the main DAC's input is toggled back and forth between -2389.5Δ and -2388.5Δ, where Δ is the step size of the main DAC. This choice of the input sequence results in $s_k[n]$ sequences with a low percentage of zero values, which ensures rapid MNC loop convergence [15]. During the normal operation, a full-scale input signal of 179.4 MHz with a peak-to-peak swing of $2^{14}\Delta$ is applied to the input of the main DAC, and the DAC SNR contributed by correlation noise is evaluated.

Each simulation sweeps one of the three parameters of (172), i.e., σ_r, N, and K. The value of L is fixed as 35, which is determined by the structure of DEM encoder used in the design. The value of Q is chosen as 21, which is the closest integer to the average delay of

MNC feedback path. The MNC feedback path is properly designed such that h_K is positive, the value of which, as extracted from separate simulation, is 0.6 and found to be nearly independent of K in practical cases where K is small. Each c_{k_post} of (172) is also estimated from separate simulations.

Fig. 36 compares (172) with the simulated post-calibration DAC SNR contributed by correlation noise for $K = 2\times10^{-5}$, $N = 9$ and different values of σ_r, where σ_r is varied by varying the quantization step size of the ADC. Fig. 37 presents the same comparison for $N = 9$, $\sigma_r = 15\Delta$, and different values of K. Fig. 38 presents the same comparison for $K = 2\times10^{-5}$, $\sigma_r = 15\Delta$, and different values of N. In all three cases, the analytical results closely agree with the simulation results. The simulations also quantitatively demonstrate that the noise impact on post-calibration DAC SNR can be reduced by decreasing the magnitude of K.

APPENDIX A

Lemma 1: The difference equation (147), derived from (139)-(140), can be converted into (158)-(160) without change of the initial condition. Each $\mathbf{H}_{q,k}(\mathbf{z}[n-1])$ term in (158) for $q = 1, 2, 3$ has the form of $\mathbf{H}(\mathbf{z}[n-1])$, where $\mathbf{H}(\mathbf{z}[n-1])$ is defined as a sum of *a finite number* of components, each component again is a sum term in the form of $\mathbf{D}(\mathbf{z}[n-1])$ given by

$$\mathbf{D}\left(\mathbf{z}\left[n-1\right]\right)=\sum_{i\geq1}s_{n,i}\mathbf{D}_i\mathbf{z}_l\left[n-i\right] \tag{173}$$

where l represents any integer between 1 and L, $s_{n,i}$ is a modulation variable associated with time indexes larger than $n-i-Q$, $\mathbf{D}_i$ is an $N\times N$ deterministic real matrix. For any bounded value of K, the spectral norm of each $\mathbf{D}_i$ is bounded by an exponentially decaying curve as i increases.

The $\mathbf{v}_{k,n}$ term, similar to $\mathbf{e}_{k,n}$, is an additive noise and has the form of a modulated vector. All the modulation variables of $\mathbf{H}_{q,k}(\mathbf{z}[n-1])$ and $\mathbf{v}_{k,n}$ are a single or a product of samples of $S_k[n]$ sequences with the time index smaller than $n+N$ and must contain $S_k[n-Q]$ with index k.

Proof of Lemma 1:

It follows from the definition of $\mathbf{H}(\mathbf{z}[n-1])$ that the portion of the double sum term in (147) (excluding the scaling factor K) associated with state variables with time indexes smaller than n has the form of $\mathbf{H}(\mathbf{z}[n-1])$.

Let $\mathbf{H}_i$ for $-(N-2) \leq i \leq N-2$ represent any $N\times N$ deterministic real matrix that satisfies the following property: each entry of $\mathbf{H}_i$ with the row index u and column index v with $u-v \leq -i$ must be 0, it follows that $\mathbf{H}_{\mathbf{c}\_i}$ in (147) for $-(N-2) \leq i \leq N-2$ has the form of $\mathbf{H}_i$.

Replacing n with $n+p-i$ in (147) and multiplying both sides by $\mathbf{H}_i$, and finally grouping all the terms associated with $\mathbf{z}_l[n-j]$ for $j \geq 1$ and $1 \leq l \leq L$ into $\mathbf{H}(\mathbf{z}[n-1])$, it follows that $\mathbf{H}_i\mathbf{z}_k[n+p-i]$ for any $-(N-2) \leq p \leq 0$ and $-(N-2) \leq i \leq p$ can be written as

$$\begin{aligned} \mathbf{H}_i\mathbf{z}_k[n+p-i] &= \mathbf{H}_i\mathbf{z}_k[n+p-i-1] \\ &+ \sum_{j=-(N-2)}^{p-i}\sum_{l=1}^{L} KS_k[n+p-i-Q]S_l[n+p-i-j]\mathbf{H}_i\mathbf{H}_{\mathbf{c}\_j}\mathbf{z}_l[n+p-i-j] \\ &+K\mathbf{H}\left(\mathbf{z}[n-1]\right)+K\mathbf{H}_i\mathbf{e}_{k,n+p-i}. \end{aligned} \tag{174}$$

The ranges of p and i imply that the upper limit of the index j in (174) is at most $N-2$, thus $\mathbf{H}_{\mathbf{c}\_j}$ has the form of $\mathbf{H}_j$. The properties of $\mathbf{H}_i$ matrix imply that $\mathbf{H}_i\mathbf{H}_{\mathbf{c}\_j}$ in (174) has the form of $\mathbf{H}_{i+j-1}$. Replacing $S_l[n+p-i-j]\mathbf{H}_i\mathbf{H}_{\mathbf{c}\_j}\mathbf{z}_l[n+p-i-j]$ in (174) by $S_l[n+p-q-1]\mathbf{H}_q\,\mathbf{z}_l[n+p-q-1]$ with $q = i+j-1$, and letting $\mathbf{z}_{k(n,i,p)}$ and $\mathbf{z}_{l(n,q,p)}$ denote $\mathbf{H}_i\mathbf{z}_k[n+p-i]$ and $\mathbf{H}_q\mathbf{z}_l[n+p-q]$, respectively, yields

$$\mathbf{z}_{k(n,i,p)} = \mathbf{z}_{k(n,i,p-1)} + K\sum_{q=-(N-2)}^{p-1}\sum_{l=1}^{L} S_k[n+p-i-Q]S_l[n+p-q-1]\mathbf{z}_{l(n,q,p-1)} + K\mathbf{H}(\mathbf{z}[n-1]) + K\mathbf{H}_i\mathbf{e}_{k,n+p-i}. \tag{175}$$

The lower index of q in (175) is set to $-(N-2)$, this is because the properties of $\mathbf{H}_q$ implies that $\mathbf{H}_q = \mathbf{0}$ for all $q \le -(N-1)$. The following analysis proves that for each $-(N-2) \le p \le 0$, $\mathbf{z}_{k(n,i,p)}$ has the form

$$\mathbf{z}_{k(n,i,p)} = \mathbf{H}_i\mathbf{z}_k[n-1] + K\mathbf{H}(\mathbf{z}[n-1]) + K\mathbf{w}_{k(n,i,p)} \text{ for} -(N-2) \le i \le p \text{ and } 1 \le k \le L, \tag{176}$$

where $\mathbf{H}_i$ is the same $\mathbf{H}_i$ matrix in the expression of $\mathbf{z}_{k(n,i,p)}$, and $\mathbf{w}_{k(n,i,p)}$ is a modulated vector. The proof is done through mathematical induction. The *inductive step* is, if (176) holds for $p = J-1$ where $-(N-2) \le J-1 \le -1$, then (176) must hold for $p = J$. The induction *base step* is (176) holds for $p = -(N-2)$. The induction base step follows from (175), this is explained as follows. With $-(N-2) \le i \le p$ and $p = -(N-2)$, it follows that the double sum term in (175) vanishes, and the $\mathbf{z}_{k(n,i,p-1)}$ term in (175), given by $\mathbf{H}_i\mathbf{z}_k[n+p-1-i]$, can be written as $\mathbf{H}_i\mathbf{z}_k[n-1]$. Furthermore, the $\mathbf{H}_i\mathbf{e}_{k,n+p-i}$ term in (175) is a modulated vector because $\mathbf{e}_{k,n}$ in (147) is a modulated vector. The proof of the inductive step also follows from (175), this is explained as follows. The double sum term in (175) with $p = J$ contains a finite number of terms associated with $\mathbf{z}_{k(n,q,J-1)}$, each $\mathbf{z}_{k(n,q,J-1)}$ has the index q restricted within the range of $-(N-2) \le q \le J-1$, this range of q in $\mathbf{z}_{k(n,q,J-1)}$ is identical to that of i given by (176) with $p = J-1$. Since the inductive step is built on the premise that (176) holds for $p = J-1$, it follows that $\mathbf{z}_{k(n,q,J-1)}$ can be expressed in the form of (176) with $p = J-1$. Furthermore, the ranges of i, p, q imply that the modulation variable, $S_k[n+p-i-Q]S_l[n+p-q-1]$, in (175) is associated with time index larger than $n-1-Q$. These properties imply that the double sum term in (175) itself has the form of $K\mathbf{H}(\mathbf{z}[n-1])+K\mathbf{w}$, where $\mathbf{w}$ is a modulated vector. Similarly, the remaining portion of (175) also has this form. Combining these results yields (176).

It follows from the definition of $\mathbf{z}_{n,i,p} = \mathbf{H}_i\mathbf{z}_k[n+p-i]$ that each $\mathbf{H}_{\mathbf{c}\_i}\mathbf{z}_k[n-i]$ component of (147) for $-(N-2) \le i \le 0$ has the form of $\mathbf{z}_{n,i,p}$ with $p = 0$ and $\mathbf{H}_i=\mathbf{H}_{\mathbf{c}\_i}$, thus has the form of (176). Substituting (176) with $p = 0$ and $\mathbf{H}_i=\mathbf{H}_{\mathbf{c}\_i}$ for $-(N-2) \le i \le 0$ into (147) and grouping all the K^2 terms into $K^2\mathbf{H}(\mathbf{z}[n-1])$ yields

$$\mathbf{z}_k\left[n\right] = \mathbf{z}_k\left[n-1\right] - K\mathbf{H}_{1,k}\left(\mathbf{z}\left[n-1\right]\right) + K^2\mathbf{H}(\mathbf{z}[n-1]) + K\mathbf{v}_{k,n,0} \quad \text{for } k = 1, 2, ..., L, \tag{177}$$

where $\mathbf{H}_{1,k}(\mathbf{z}[n-1])$ is given by (159) and has the form of $\mathbf{H}(\mathbf{z}[n-1])$, and $\mathbf{v}_{k,n,0}$ is a modulated vector. For any finite integer $j \le 0$, recursively expanding (177) and grouping all the terms scaled by K^2 into $K^2\mathbf{H}(\mathbf{z}[n-1])$, yields

$$\mathbf{z}_k\left[n-j\right] = \mathbf{z}_k\left[n-1\right] - K\mathbf{H}_{k(n-j)}\left(\mathbf{z}\left[n-1\right]\right) + K^2\mathbf{H}(\mathbf{z}[n-1]) + K\mathbf{v}_{k,n,j} \quad \text{for } k = 1, 2, ..., L, \tag{178}$$

where $\mathbf{H}_{k(n-j)}(\mathbf{z}[n-1])$ is derived by summing up $\mathbf{H}_{1,k}(\mathbf{z}[m-1])$ for $n \le m \le n-j$ and then replacing each state variable with time index larger than $n-1$ by the same state variable with time index $n-1$. It follows that $\mathbf{H}_{k(n-j)}(\mathbf{z}[n-1])$ also has the form of $\mathbf{H}(\mathbf{z}[n-1])$, and each additive noise term $\mathbf{v}_{k,n,j}$ is a modulated vector. The reason why the recursive expansion of (177) yields (178) follows from mathematical induction. The induction base step, i.e., (178) is satisfied for $j = 0$, directly follows from (177). The inductive step, i.e., (178) is satisfied for $j = -q$ provided that (178) is satisfied for $j = -(q-1)$ for $q \ge 1$, follows by replacing n with $n+1$ in (178) and then applying (177) to expand a finite number of terms that is associated with $\mathbf{z}_k[n]$ at time index n with respect to state variables at time index $n-1$ or smaller.

Substituting (178) for $-(N-2) \le j \le 0$ and $1 \le k \le L$ into (147) yields (158). Since all modulation variables of $\mathbf{H}(\mathbf{z}[n-1])$ and $\mathbf{e}_{k,n}$ in (147) is either $S_k[n-Q]$ or a product of $S_k[n-Q]$ and some samples of $S_k[n]$ sequences, it follows that each $\mathbf{H}_{q,k}(\mathbf{z}[n-1])$ term and each modulated vector in (158)-(160) must also follow these properties. Furthermore, it follows from (139)-(141) that $a_{k,m}[n+m]$ in the expression of $\mathbf{z}_k[n]$ must not be dependent on samples of $S_k[n]$ sequences with time indexes equal or larger than $n+N$, thus all the modulation variables in (158) are only associated with time indexes smaller than $n+N$.

□

Lemma 2: Any N-dimensional real vectors **v** and **w** satisfy

$$\|\mathbf{v}\| - \|\mathbf{w}\| \leq \|\mathbf{v}+\mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|. \tag{179}$$

Proof of Lemma 2:

It follows from the definition of ||**v**+**w**||^2 that

$$\|\mathbf{v}+\mathbf{w}\|^2 = \sum_{j=0}^{N-1} E\left[(v_j + w_j)^2\right] \leq \sum_{j=0}^{N-1}\left(\sqrt{E\left[v_j^{\ 2}\right]} + \sqrt{E\left[w_j^{\ 2}\right]}\right)^2. \tag{180}$$

It follows from the Cauchy–Schwarz inequality that (180) is further upper bounded by (||**v**||+||**w**||)^2, which is same upper bound of (179). Replacing **w** with −**w** and then replacing **v** with **v**+**w** yields the lower bound of (179).

□

Lemma 3: If the magnitude of $h_c[n]$ is bounded by an exponentially decaying curve as n increases, and if there exists a positive K such that $h_K > 0$ is satisfied, then for any $0 < \beta < 1$, there must exist a positive value of $K_{\max}$ such that for any $0 < K \leq K_{\max}$, $x_{\max}^2[n]$ given by (165) with $\mathbf{x}_k[n]$ given by (161) and (164) must satisfy the following exponential trajectory,

$$\frac{x_{\max}{}^2[n]}{x_{\max}{}^2[n-1]} \leq \begin{cases} 1-(1-\beta)h_K K_{(\min,n-Q)}, \text{ if } n-Q \geq j_{\text{ini}}, \\ 1+\beta h_K K_{(\max,n-Q)}, \quad \text{otherwise,} \end{cases} \tag{181}$$

for all $n \geq \max(j_{\text{ini}}+1,\ 0)$, where $K_{(\min,n-Q)}$ and $K_{(\max,n-Q)}$ are the minimum and maximum value of $2c_k[n-Q]K$ over k, where $c_k[n-Q]$ is given by (166) with n replaced by $n-Q$, and each $\|\mathbf{x}_k[n]\|^2$ for $k = 1, 2, \ldots, L$ has a tighter upper bounds of

$$\|\mathbf{x}_k[n]\|^2 \leq \begin{cases} \left(1-h_K K_{(k,n-Q)}\right)\|\mathbf{x}_k[n-1]\|^2 + \beta h_K K_{(k,n-Q)} x_{\max}{}^2[n-1], \text{ if } n-Q \geq j_{\text{ini}}, \\ \|\mathbf{x}_k[n-1]\|^2 + \beta h_K K_{(k,n-Q)} x_{\max}{}^2[n-1], \quad \text{otherwise,} \end{cases} \tag{182}$$

where $K_{(k,n-Q)} = 2c_k[n-Q]K$ for each k.

Proof of Lemma 3:

The proof of Lemma 3 uses mathematical induction. The inductive step, which is proven shortly, is: for any $0 < \beta < 1$, there must exist a range of K given by $0 < K \le K_{\max}$ such that for each K within this range and each $n \ge \max(j_{\text{ini}}+1, 0)$, if

$$\mathbf{x}_{\max}{}^{2}[J] / \mathbf{x}_{\max}{}^{2}[J-1] \ge 1 - 4K \left\| \mathbf{H}_{\mathbf{c}\_Q} \right\|_{2} \qquad (183)$$

holds for all $J < n$, then the conditions of Lemma 3's hypothesis are sufficient to ensure that (183) must hold for $J = n$, and (181)-(182) hold at time n.

The induction base step, i.e., (183) holds for all $J \le \max(j_{\text{ini}}, -1)$, follows directly from the basis initial condition in (164) and $\mathbf{H}_{q,k}(\mathbf{x}[n-1]) = \mathbf{0}$ for $n \le -1$. Therefore, provided that the inductive step is true and K is chosen within $0 < K \le K_{\max}$, it follows from the induction that (183) and (181)-(182) hold for all integers $n \ge \max(j_{\text{ini}}+1, 0)$. Hence, it remains to show that the inductive step is true. This is shown in the remainder of the proof.

Lemma 1 implies that each $\mathbf{H}_{q,k}(\mathbf{x}[n-1])$ term in (161) for $q = 1, 2, 3$ has the form of $\mathbf{H}(\mathbf{x}[n-1])$, where $\mathbf{H}(\mathbf{x}[n-1])$ is given by $\mathbf{H}(\mathbf{z}[n-1])$ defined in Lemma 1 with each state variable, $\mathbf{z}_l[n-i]$, replaced by $\mathbf{x}_l[n-i]$, and $\mathbf{H}(\mathbf{x}[n-1])$ satisfies all the associated properties of $\mathbf{H}(\mathbf{z}[n-1])$ described in Lemma 1.

For any modulation variable, s, let $s[n]$ denote all portions of s associated with time index n multiplied by any deterministic scaling factor in its expression (if s is not associated with time index n, then $s[n]$ is equal to the deterministic scaling factor), thus $s[n]$ is restricted to values between -1 and 1 by definition. For example, if $s = 0.5(S_1[6]^2-1/2)S_2[6]S_5[7]$, then $s[6] = 0.5(S_1[6]^2-1/2)S_2[6]$. Let $c\{s[n]\}$ denote

$$c\left\{s[n]\right\} = E\left\{s[n] s_{\text{base}}{}^{2}\right\} / E\left\{s_{\text{base}}{}^{2}\right\}, \qquad (184)$$

where s_{base} is the scaling variable in the basis initial condition given by (164) and is a single or a product of samples of $S_k[n]$ sequences. A special case of $c\{s[n]\}$ with $s[n] = S_k{}^2[n]$ was previously defined by $c_k[n]$ in (166), i.e., $c\{S_k{}^2[n]\} = c_k[n]$. It follows from these definitions that if $s[n] = S_k^2[n]-c_k[n]$, or if $s[n]$ is both zero mean and in the form of a single or a product of samples of $S_k[n]$ sequences at time n, then $c\{s[n]\} = 0$.

It follows from these properties that each modulation variable of $\mathbf{H}_{1,k}(\mathbf{x}[n-1])$ in (161), given by $s = S_k[n-Q]S_l[n-i]$ as according to (159)-(160), satisfies $c\{s[n-Q]\} = c_k[n-Q]$ for $l = k$ and $i = Q$, and $c\{s[n-Q]\} = 0$ otherwise. The component of $\mathbf{H}_{1,k}(\mathbf{x}[n-1])$ associated with $S_k[n-Q]S_l[n-i]$ for $l = k$ and $i = Q$ is given by $S_k^2[n-Q]\mathbf{H}_{\mathbf{c}\_Q}\mathbf{x}_k[n-Q]$. Let $\mathbf{A}_k(\mathbf{x}[n-1])$ represent $\mathbf{H}_{1,k}(\mathbf{x}[n-1])-c_k[n-Q]\mathbf{H}_{\mathbf{c}\_Q}\mathbf{x}_k[n-Q]$, it follows that $\mathbf{A}_k(\mathbf{x}[n-1])$ is equal to $\mathbf{H}_{1,k}(\mathbf{x}[n-1])$ except that the $S_k^2[n-Q]\mathbf{H}_{\mathbf{c}\_Q}\mathbf{x}_k[n-Q]$ component is replaced by $(S_k^2[n-Q]-c_k[n-Q])\mathbf{H}_{\mathbf{c}\_Q}\mathbf{x}_k[n-Q]$, where the new modulation variable, $0.5(S_k^2[n-Q]-c_k[n-Q])$, satisfies $c\{s[n-Q]\} = 0$ (a scaling factor of 0.5 is grouped into this modulation variable to simplify the subsequent analysis).

It follows from $\mathbf{H}_{1,k}(\mathbf{x}[n-1]) = \mathbf{A}_k(\mathbf{x}[n-1])+c_k[n-Q]\mathbf{H}_{\mathbf{c}\_Q}\mathbf{x}_k[n-Q]$ that (161) contains a term given by $\mathbf{x}_k[n-1]-Kc_k[n-Q]\mathbf{H}_{\mathbf{c}\_Q}\mathbf{x}_k[n-Q]$. To explicitly capture the effect of $\mathbf{x}_k[n-Q] = \mathbf{0}$ for $n-Q < j_{\text{ini}}$ as implied by (164), $\mathbf{x}_k[n-Q]$ is replaced by $u_{\text{ini}}[n-Q]\mathbf{x}_k[n-Q]$, where $u_{\text{ini}}[n-Q] = 1$ if $n-Q \geq j_{\text{ini}}$ and 0 otherwise, thus

$$\begin{aligned}\mathbf{x}_k\left[n-1\right] - Kc_k[n-Q]\mathbf{H}_{\mathbf{c}\_Q}\mathbf{x}_k\left[n-Q\right] &= \left(\mathbf{I} - Kc_k[n-Q]u_{\text{ini}}[n-Q]\mathbf{H}_{\mathbf{c}\_Q}\right)\mathbf{x}_k[n-1] \\ &+ Kc_k[n-Q]u_{\text{ini}}[n-Q]\mathbf{H}_{\mathbf{c}\_Q}\sum_{i=1}^{Q-1}\left(\mathbf{x}_k[n-i]-\mathbf{x}_k[n-i-1]\right),\end{aligned} \tag{185}$$

where $\mathbf{I}$ is the $N\times N$ identity matrix. If $n-i < \max(j_{\text{ini}}+1, 0)$, it follows from $1 \leq i \leq Q-1$ that either $u_{\text{ini}}[n-Q] = 0$ or $\mathbf{x}_k[n-i]-\mathbf{x}_k[n-i-1] = \mathbf{0}$. Otherwise, the $\mathbf{x}_k[n-i]-\mathbf{x}_k[n-i-1]$ term in (185) can be expanded by (161) with n replaced by $n-i$. Therefore, the upper index, $Q-1$, of the sum term in (185) can be replaced by $Q' = \min(Q-1, n-\max(j_{\text{ini}}+1, 0))$. Applying (161) to expand $\mathbf{x}_k[n-i]-\mathbf{x}_k[n-i-1]$ and then substituting the result back into (161) yields

$$\mathbf{x}_k[n] = (\mathbf{I} - K\mathbf{I}_{(k,n)})\mathbf{x}_k[n-1] - K\mathbf{u}_k[n-1] - K^2\left(\mathbf{v}_k[n-1]+\mathbf{w}_k[n-1]\right) + K^3\mathbf{H}\left(\mathbf{x}\left[n-1\right]\right),$$

$$\tag{186}$$

where

$$\mathbf{I}_{(k,n)} = c_k[n-Q]u_{\text{ini}}[n-Q]\mathbf{H}_{\mathbf{c}\_Q}, \tag{187}$$

$$\mathbf{u}_k[n-1] = \mathbf{A}_k\left(\mathbf{x}[n-1]\right), \tag{188}$$

$$\mathbf{v}_k[n-1] = \sum_{i=1}^{Q'} \mathbf{I}_{(k,n)}\mathbf{I}_{(k,n-i)}\mathbf{x}_k[n-i-Q], \tag{189}$$

$$\mathbf{w}_k[n-1] = \sum_{i=1}^{Q'} \mathbf{I}_{(k,n)}\mathbf{u}_k[n-i-1] + \mathbf{H}_{2,k}\left(\mathbf{x}\left[n-1\right]\right), \tag{190}$$

where all the term scaled by K^3 is grouped into $K^3\mathbf{H}(\mathbf{x}[n-1])$, and $\mathbf{H}_{2,k}(\mathbf{x}[n-1])$ in (190) corresponds to the same term in (161). It follows that (188)-(190) each has the form of $\mathbf{H}(\mathbf{x}[n-1])$. Taking RMS norm on both sides of (186) and applying Lemma 4 yields

$$a_0 - \sum_{j=1}^{6} a_j K^j \le \left\|\mathbf{x}_k[n]\right\|^2 \le a_0 + \sum_{j=1}^{6} a_j K^j, \tag{191}$$

where a_0, a_1, and a_2 are given by

$$a_0 = \left\|(\mathbf{I} - K\mathbf{I}_{(k,n)})\mathbf{x}_k[n-1]\right\|^2, \tag{192}$$

$$a_1 = \left|(\mathbf{I} - K\mathbf{I}_{(k,n)})\mathbf{x}_k[n-1], \mathbf{u}_k\left[n-1\right]\right|, \tag{193}$$

$$\begin{aligned} a_2 = {} & \left|\mathbf{u}_k\left[n-1\right], \mathbf{u}_k\left[n-1\right]\right| + \left|(\mathbf{I} - K\mathbf{I}_{(k,n)})\mathbf{x}_k[n-1], \mathbf{w}_k\left[n-1\right]\right| \\ & + \left|(\mathbf{I} - K\mathbf{I}_{(k,n)})\mathbf{x}_k[n-1], \mathbf{v}_k\left[n-1\right]\right|. \end{aligned} \tag{194}$$

As proven shortly, each of these terms can be bounded by

$$a_0 \le \left(1 - Kh_K c_k[n-Q]u_{\text{ini}}[n-Q]\right)^2 \left\|\mathbf{x}_k[n-1]\right\|^2, \tag{195}$$

$$a_0 \ge \left(1 - K\left\|\mathbf{H}_{\mathbf{c}\_Q}\right\|_2 c_k[n-Q]u_{\text{ini}}[n-Q]\right)^2 \left\|\mathbf{x}_k[n-1]\right\|^2, \tag{196}$$

$$a_1K + a_2K^2 \le K^2 c_k[n-Q]\left(A - \left\|\mathbf{H}_{\mathbf{c}\_Q}\right\|_2^{\;2} + o(K)\right) x_{\max}{}^2[n-1], \tag{197}$$

where A is given by (156), $o(K)$ is a bounded and in the order of K. The proof of (195) and (196) is presented as follows.

If $c_k[n-Q]u_{\text{ini}}[n-Q] \neq 0$, the definitions of h_K and $\mathbf{I}_{(k,n)}$ in (150) and (187), and $0 < Kc_k[n-Q]u_{\text{ini}}[n-Q] \le K$ imply that

$$\left(1-\left\|\mathbf{I}-K\mathbf{I}_{(k,n)}\right\|_2\right)/\left(Kc_k[n-Q]u_{\text{ini}}[n-Q]\right)\geq h_K, \tag{198}$$

which yields

$$\left\|\mathbf{I}-K\mathbf{I}_{(k,n)}\right\|_2 \leq 1-Kh_K c_k[n-Q]u_{\text{ini}}[n-Q]. \tag{199}$$

If $c_k[n-Q]u_{\text{ini}}[n-Q] = 0$, then $\|\mathbf{I}-K\mathbf{I}_{(k,n)}\|_2 = 1$ and thus (199) still holds. Applying Lemma 5 and (199) to (192) yields (195).

It follows from Lemma 2 and Lemma 5 that

$$\sqrt{a_0} = \left\|(\mathbf{I}-K\mathbf{I}_{(k,n)})\mathbf{x}_k[n-1]\right\| \geq \left\|\mathbf{x}_k[n-1]\right\| - \left\|K\mathbf{I}_{(k,n)}\right\|_2 \left\|\mathbf{x}_k[n-1]\right\|.$$

This and the definition of $\mathbf{I}_{(k,n)}$ in (187) yield (196).

The proof of (197) is presented as follows.

For any finite integer $q \leq 0$, recursively expanding (161) and grouping all the terms scaled by K^2 into $K^2\mathbf{H}(\mathbf{x}[n-1])$ yields

$$\mathbf{x}_k\left[n-q\right] = \mathbf{x}_k\left[n-1\right] - K\mathbf{H}_{k(n-q)}\left(\mathbf{x}\left[n-1\right]\right) + K^2\mathbf{H}(\mathbf{x}[n-1]), \tag{200}$$

where $\mathbf{H}_{k(n-q)}(\mathbf{x}[n-1])$ is derived by summing up $\mathbf{H}_{1,k}(\mathbf{x}[m-1])$ for $n \leq m \leq n-q$ and then replacing each state variable with time index larger than $n-1$ by the same state variable with time index $n-1$. It follows that $\mathbf{H}_{k(n-q)}(\mathbf{x}[n-1])$ also has the form of $\mathbf{H}(\mathbf{x}[n-1])$. The recursive expansion applies the same mathematical induction used in Lemma 1 to expand (177) into (178).

For any integers i, I satisfying $1 \leq i \leq I$ and $n-I \geq j_{\text{ini}}$, replacing n with $n-(I-1)$ and q with $i-(I-1)$ in (200) yields

$$\mathbf{x}_k\left[n-i\right] = \mathbf{x}_k\left[n-I\right] - K\mathbf{H}_{k(n-i)}(\mathbf{x}[n-I]) + K^2\mathbf{H}(\mathbf{x}[n-I]), \tag{201}$$

where $\mathbf{H}_{k(n-i)}(\mathbf{x}[n-I])$ is derived by summing up $\mathbf{H}_{1,k}(\mathbf{x}[m-1])$ for $n-(I-1) \leq m \leq n-i$ and then replacing each state variable with time index larger than $n-I$ by the same state variable with time index $n-I$, and $\mathbf{H}_{k(n-i)}(\mathbf{x}[n-I])$ has the form of $\mathbf{H}(\mathbf{x}[n-I])$.

Many terms described before have the form of $\mathbf{H}(\mathbf{x}[n-1])$. The basis initial condition indicates that $\mathbf{x}_l[n-i] = \mathbf{0}$ for $n-i < j_{\text{ini}}$, thus all the components of $\mathbf{H}(\mathbf{x}[n-1])$ associated with

$\mathbf{x}_l[n-i]$ for $n-i < j_{\text{ini}}$ can be removed, this is implicitly done in the subsequent analysis. The definition of $\mathbf{H}(\mathbf{x}[n-1])$ indicates that its modulation variables in the scaling factor of $\mathbf{x}_l[n-i]$ are associated with time indexes larger than $n-i-Q$, this implies that all the modulation variables of $\mathbf{H}(\mathbf{x}[n-1])$ are only associated with time indexes larger than $j_{\text{ini}}-Q$. Furthermore, many modulation variables introduced, denoted as s, can be factored into the product of two modulation variables, s_1 and s_2, where $s_1 \in \{S_k[n-Q], c_k[n-Q], 0.5(S_k^2[n-Q]-c_k[n-Q])\}$, s_2 can either be a deterministic scaling factor or a variable associated with samples of $S_k[n]$ sequences. It follows from (184), (166) and $|s_2| \leq 1$ that s must satisfy $c\{s[n-Q]\} \leq c_k[n-Q]$. This is because if $s_1 = S_k[n-Q]$, then $c\{s[n-Q]\} \leq c\{|S_k[n-Q]|\} = c\{S_k^2[n-Q]\} = c_k[n-Q]$. If $s_1 = c_k[n-Q]$, then $c\{s[n-Q]\} \leq c\{c_k[n-Q]\} = c_k[n-Q]$. If $s_1 = 0.5(S_k^2[n-Q]-c_k[n-Q])$, then $c\{s[n-Q]\} \leq c\{0.5S_k^2[n-Q]\}+ c\{0.5c_k[n-Q]\} = c_k[n-Q]$.

The inner product between two terms, each has the form of $\mathbf{H}(\mathbf{x}[n-1])$, is given by $<\mathbf{H}(\mathbf{x}[n-1]), \mathbf{H}(\mathbf{x}[n-1])>$, and can be expanded into a sum of components, each has the form $<s\mathbf{D}_{q1,i1}\mathbf{x}_{l1}[n-i_1], \mathbf{D}_{q1,i2}\mathbf{x}_{l2}[n-i_2]>$ for i_1, $i_2 \geq 1$ and a finite number of q_1 and q_2. Since both $||\mathbf{D}_{q1,i1}||_2$ and $||\mathbf{D}_{q2,i2}||_2$ by definition are bounded by exponentially decaying curves as i_1 and i_2 increase, the sum of $||\mathbf{D}_{q1,i1}||_2||\mathbf{D}_{q2,i2}||_2$ for all i_1, i_2, q_1, q_2 must be bounded. Let $<\mathbf{H}_n>$ denote any portion of these components where each of its modulation variable can be factored into $s_1 \in \{S_k[n-Q], c_k[n-Q], 0.5(S_k^2[n-Q]-c_k[n-Q])\}$ and another modulation variable, it follows that their modulation variables satisfy $c\{s[n-Q]\} \leq c_k[n-Q]$. Furthermore, let $<\mathbf{H}'_n>$ denote any portion of these terms with the above constraint satisfied and an additional constraint: for any $<s\mathbf{D}_{q1,i1}\mathbf{x}_{l1}[n-i_1], \mathbf{D}_{q2,i2}\mathbf{x}_{l2}[n-i_2]>$ component of $<\mathbf{H}'_n>$ with indexes i_1 and i_2, there must exist an integer p such that $c\{s[n-p]\} = 0$ is satisfied, where $n-p > j_{\text{ini}}-Q$ and both $p-i_1$ and $p-i_2$ cannot be unbounded positive integer.

Let us first evaluate the value of a_1. It follows from the definition of a_1 in (193) that it is the magnitude of

$$\left\langle (\mathbf{I}-K\mathbf{I}_{(k,n)})\mathbf{x}_k[n-1],\ \mathbf{A}_k\left(\mathbf{x}[n-1]\right)\right\rangle. \tag{202}$$

To simplify the notations, let I denote $I_{n,n-Q}$, where $I_{n,n-Q}$ is given by $I_{n,p}$ defined in Lemma 6 with p replaced by $n–Q$. Given that $n \geq \max(j_{\mathrm{ini}}+1, 0)$, it follows from the definition of $I_{n,n-Q}$ that $0 < I \leq Q+N$ and $n-I \geq j_{\mathrm{ini}}$. Let $\mathbf{v}_0$ denote $(\mathbf{I}-K\mathbf{I}_{(k,n)})\mathbf{x}_k[n-I]$, let $\mathbf{w}_0$ denote $\mathbf{A}_k(\mathbf{x}[n-1])$ with each of its state variable $\mathbf{x}_l[n-j]$ for $j = 1, 2, \ldots, I-1$ replaced by $\mathbf{x}_l[n-I]$, this and (201) imply that the left-side term of (202) can be written as $\mathbf{v}_0+K\mathbf{v}_1+K^2\mathbf{v}_2$, and the right-side term of (202) can be written as $\mathbf{w}_0+K\mathbf{w}_1+K^2\mathbf{w}_2$, where each of $\mathbf{v}_i$ for $0 \leq i \leq 2$ and $\mathbf{w}_j$ for for $0 \leq j \leq 2$ has the form of $\mathbf{H}(\mathbf{x}[n-1])$. Furthermore, $\mathbf{v}_1$ is given by $-(\mathbf{I}-K\mathbf{I}_{(k,n)})\mathbf{H}_{k(n-1)}(\mathbf{x}[n-I])$, $\mathbf{w}_1$ is given by $\mathbf{A}_k(\mathbf{x}[n-1])$ with each of its state variable $\mathbf{x}_l[n-i]$ replaced by $-\mathbf{H}_{l(n-i)}(\mathbf{x}[n-I])$ for $1 \leq i < I$ and replaced by $\mathbf{0}$ for $i \geq I$, where the definitions of $\mathbf{H}_{k(n-1)}(\mathbf{x}[n-I])$ and $\mathbf{H}_{l(n-i)}(\mathbf{x}[n-I])$ follow from that of $\mathbf{H}_{k(n-i)}(\mathbf{x}[n-I])$ in (201). Expanding (202) results in $<\mathbf{v}_0, \mathbf{w}_0>$ plus a finite number of terms in the form of $K^q<\mathbf{v}_i, \mathbf{w}_j>$ for $q = 1, 2, 3, 4$. Since the modulation variables in $<\mathbf{v}_0, \mathbf{w}_0>$ are given by those of $\mathbf{A}_k(\mathbf{x}[n-1])$ that satisfy $c\{s[n-Q]\} = 0$, it follows from Lemma 6 and $I = I_{n,n-Q}$ that $|\mathbf{v}_0, \mathbf{w}_0| = 0$. All the other terms associated with K^q for $q = 2, 3, 4$ can be lumped into $K^2<\mathbf{H}_n>$, it follows from (218) in Lemma 7 that they are bounded by $Ko(K)c_k[n-Q]x_{\max}^2[n-1]$, where $o(K)$ is bounded and in the order of K. The remaining term is given by $K<\mathbf{v}_0, \mathbf{w}_1>+K<\mathbf{v}_1, \mathbf{w}_0>$. Expanding $K<\mathbf{v}_0, \mathbf{w}_1>+K<\mathbf{v}_1, \mathbf{w}_0>$ and applying the definition of $\mathbf{A}_k(\mathbf{x}[n-1])$ and $\mathbf{H}_{l(n-i)}(\mathbf{x}[n-I])$, where both $\mathbf{A}_k(\mathbf{x}[n-1]) = \mathbf{H}_{0,k}(\mathbf{x}[n-1])-c_k[n-Q]\mathbf{H}_{\mathbf{c}\_Q}\mathbf{x}_k[n-Q]$ and $\mathbf{H}_{l(n-i)}(\mathbf{x}[n-I])$ are associated with $\mathbf{H}_{0,k}(\mathbf{x}[n-1])$ defined in (159)-(160), a careful analysis of the resulting terms shows that a portion of them are given by $K<s\mathbf{D}_1\mathbf{x}_l[n-i_1], \mathbf{D}_2\mathbf{x}_l[n-i_2]>$, where $s = S_k^2[n-Q]S_k^2[n-q]$, $\mathbf{D}_1 = (\mathbf{I}-K\mathbf{I}_{(k,n)})\mathbf{H}_{\mathbf{c}\_(2Q-q)}$, $\mathbf{D}_2 = \mathbf{H}_{\mathbf{c}\_q}$ for $q = Q+1, Q+2, \ldots, Q+I-1$, where i_1, $i_2 \geq I$ and are bounded. It follows that these modulation variables all satisfy $c\{S[n-Q]\} \le c_k[n-Q]$, thus Lemma 6, $I = I_{n,n-Q}$, and the matrix property of $\|\mathbf{D}_1\mathbf{D}_2\|_2 \le \|\mathbf{D}_1\|_2\|\mathbf{D}_2\|_2$ imply that each term's magnitude is bounded by $Kc_k[n-Q]\|\mathbf{H}_{\mathbf{c}\_(2Q-q)}\|_2\|\mathbf{H}_{\mathbf{c}\_q}\|_2 x_{\max}^2[n-1](1+o(K))$, where $o(K) = (1-4K\|\mathbf{H}_{\mathbf{c}\_Q}\|_2)^{-(i_1+i_2-2)/2}-1$ is bounded since both i_1 and i_2 are bounded. The remaining portion of $K\langle\mathbf{v}_0, \mathbf{w}_1\rangle$ and $K\langle\mathbf{v}_1, \mathbf{w}_0\rangle$ both have the form of $K\langle\mathbf{H}'_n\rangle$, it follows from (219) in Lemma 7 that their magnitudes are bounded by $Ko(K)c_k[n-Q]x_{\max}^2[n-1]$.

Let us evaluate the value of a_2. It follows from Lemma 5 that a_2 is bounded by the sum of the magnitudes of $\langle\mathbf{u}_k[n-1], \mathbf{u}_k[n-1]\rangle$, $\langle(\mathbf{I}-K\mathbf{I}_{(k,n)})\mathbf{x}_k[n-1], \mathbf{w}_k[n-1]\rangle$ and $\langle(\mathbf{I}-K\mathbf{I}_{(k,n)})\mathbf{x}_k[n-1], \mathbf{v}_k[n-1]\rangle$. Expanding each of them and applying the definition of $\mathbf{A}_k(\mathbf{x}[n-1])$ and $\mathbf{H}_{1,k}(\mathbf{x}[n-1])$, a careful analysis of the resulting terms shows that a portion of $\langle\mathbf{u}_k[n-1], \mathbf{u}_k[n-1]\rangle$ consists of a component given by $\langle s(2\mathbf{H}_{\mathbf{c}\_Q})\mathbf{x}_k[n-Q], \mathbf{H}_{\mathbf{c}\_Q}\mathbf{x}_k[n-Q]\rangle$, where $s = (S_k^2[n-Q]-c_k[n-Q])^2/2$, and a few components given by $\langle S_k^2[n-Q]S_l^2[n-i]\mathbf{H}_{\mathbf{c}\_i}\mathbf{x}_l[n-j], \mathbf{H}_{\mathbf{c}\_i}\mathbf{x}_l[n-j]\rangle$ for $l = 1, 2, \ldots, L$, where $i = j$ for $j \ge 2$ (other than $l, j = k, Q$) and $i = -(N-2), -(N-3), \ldots, 1$ for $j = 1$. By definition, this portion of $\langle\mathbf{u}_k[n-1], \mathbf{u}_k[n-1]\rangle$ has the form of $\langle\mathbf{H}_n\rangle$, thus (218) of Lemma 7 implies that its magnitude is bounded by $(\rho+o(K))c_k[n-Q]x_{\max}^2[n-1]$, where ρ is the sum of $L\|\mathbf{H}_{\mathbf{c}\_i}\|_2^2$ for $i \ge -(N-2)$ and $\|\mathbf{H}_{\mathbf{c}\_Q}\|_2^2$. A portion of $\langle(\mathbf{I}-K\mathbf{I}_{(k,n)})\mathbf{x}_k[n-1], \mathbf{w}_k[n-1]\rangle$ consists of a total of $N-1$ components associated with $\mathbf{H}_{2,k}(\mathbf{x}[n-1])$ in (190) given by $\langle S_k^2[n-Q]S_k^2[n-i]\mathbf{H}_{\mathbf{c}\_i}\mathbf{x}_k[n-1], \mathbf{H}_{\mathbf{c}\_i}\mathbf{x}_k[n-1]\rangle$ for $-(N-2) \le i \le 0$, by definition, this portion has the form of $\langle\mathbf{H}_n\rangle$, thus (218) of Lemma 7 implies that its magnitude is bounded by $(\rho+o(K))c_k[n-Q]x_{\max}^2[n-1]$, where ρ is the sum of $\|\mathbf{H}_{\mathbf{c}\_i}\|_2^2$ for $-(N-2) \le i \le 0$. The remaining portion of $\langle\mathbf{u}_k[n-1], \mathbf{u}_k[n-1]\rangle$ and $\langle(\mathbf{I}-K\mathbf{I}_{(k,n)})\mathbf{x}_k[n-1], \mathbf{w}_k[n-1]\rangle$ both have the form of $\langle\mathbf{H}'_n\rangle$, thus their magnitudes are both bounded by $o(K)c_k[n-Q]x_{\max}^2[n-1]$.

It follows from (189) and Lemma 4 that $|(\mathbf{I}-K\mathbf{I}_{(k,n)})\mathbf{x}_k[n-1], \mathbf{v}_k[n-1]|$ is bounded by the sum of $|(\mathbf{I}-\mathbf{I}_{(k,n)})\mathbf{x}_k[n-1], \mathbf{I}_{(k,n)}\mathbf{I}_{(k,n-i)}\mathbf{x}_k[n-i-Q]|$ for $1 \le i \le Q'$. It follows from Lemma 8, Lemma 5, $\|\mathbf{I}-K\mathbf{I}_{(k,n)}\|_2 \le 1$ from (199) and $\|\mathbf{I}_{(k,n)}\mathbf{I}_{(k,n-i)}\|_2 \le \|\mathbf{I}_{(k,n)}\|_2\|\mathbf{I}_{(k,n-i)}\|_2 \le c_k[n-Q]\|\mathbf{H}_{\mathrm{c}\_Q}\|_2{}^2$ from (187) that each of them is bounded by $c_k[n-Q]\|\mathbf{H}_{\mathbf{c}\_Q}\|_2{}^2\|\mathbf{x}_k[n-1]\|\cdot\|\mathbf{x}_k[n-i-Q]\|$, combining them for $1 \le i \le Q'$ and further applying (183) and $Q' \le Q-1$ yields the upper bound, $(Q-1)c_k[n-Q]\|\mathbf{H}_{\mathbf{c}\_Q}\|_2{}^2 x_{\max}{}^2[n-1](1+o(K))$, where $o(K)$ is bounded because i is bounded.

Combining all these results yields (197).

Similar reasoning implies that a_jK^j for j = 3, 4, 5, 6 in (191) are each bounded by the magnitude of $K^3|\langle\mathbf{H}_n\rangle|$, which is again bounded by $c_k[n-Q]K^2o(K)$ according to (218) of Lemma 7, where $o(K)$ is bounded and in the order of K. Substituting this and (195)-(197) into (191), and grouping all the terms in the order of at least K^2 into $\lambda h_K K_{(k,n-Q)}$, where $K_{(k,n-Q)} = 2c_k[n-Q]K$, yields the lower bound

$$\|\mathbf{x}_k[n]\|^2 \ge \begin{cases} \left(1-\|\mathbf{H}_{\mathbf{c}\_Q}\|_2 K_{(k,n-Q)}\right)\|\mathbf{x}_k[n-1]\|^2 - \lambda h_K K_{(k,n-Q)} x_{\max}{}^2[n-1], \text{if } n-Q \ge j_{\mathrm{ini}}, \\ \|\mathbf{x}_k[n-1]\|^2 - \lambda h_K K_{(k,n-Q)} x_{\max}{}^2[n-1], \qquad \text{otherwise,} \end{cases} \tag{203}$$

and the upper bounds in (182) with β replaced by λ, where

$$\lambda = K\left(A+o(K)\right)/\left(2h_K\right) \tag{204}$$

and A is given by (156), $o(K)$ is a bounded term in the order of K.

Let p be chosen as the index such that $\|\mathbf{x}_p[n]\| = x_{\max}[n]$, substituting this into (182) with β replaced by λ yields

$$\frac{x_{\max}{}^2[n]}{x_{\max}{}^2[n-1]} \le \begin{cases} 1-(1-\lambda)h_K K_{(p,n-Q)}, & \text{if } n-Q \ge j_{\mathrm{ini}}, \\ 1+\lambda h_K K_{(p,n-Q)}, & \text{otherwise.} \end{cases} \tag{205}$$

Combining (205) with $K_{(\min,n-Q)} \le K_{(p,n-Q)} \le K_{(\max,n-Q)}$ yields (181) with β replaced by λ.

Let l be chosen as the index that satisfies $\|\mathbf{x}_l[n-1]\| = x_{\max}[n-1]$, substituting this into (203) yields

$$\|\mathbf{x}_l[n]\|^2 \ge \begin{cases} x_{\max}{}^2[n-1]\left(1-\left(\left\|\mathbf{H}_{\mathbf{c}\_Q}\right\|_2+\lambda h_K\right)K_{(l,n-Q)}\right), & \text{if } n-Q \ge j_{\text{ini}}, \\ x_{\max}{}^2[n-1]\left(1-\lambda h_K K_{(l,n-Q)}\right), & \text{otherwise.} \end{cases} \tag{206}$$

By definition, $x_{\max}[n] \ge \|\mathbf{x}_l[n]\|$ and $K_{(l,n-Q)} \le K_{(\max,n-Q)}$, substituting these into (206) yields

$$\frac{x_{\max}{}^2[n]}{x_{\max}{}^2[n-1]} \ge \begin{cases} 1-\left(\left\|\mathbf{H}_{\mathbf{c}\_Q}\right\|_2+\lambda h_K\right)K_{(\max,n-Q)}, & \text{if } n-Q \ge j_{\text{ini}}, \\ 1-\lambda h_K K_{(\max,n-Q)}, & \text{otherwise.} \end{cases} \tag{207}$$

The definition of h_K in (150) implies that it does not decrease as K decreases. Under the hypothesis of Lemma 3, let us choose K small enough such that $h_K > 0$ is satisfied. It follows from (150) and the matrix property of $\|\mathbf{I}-K\mathbf{H}_{\mathbf{c}\_Q}\|_2 \ge \|\mathbf{I}\|_2 - K\|\mathbf{H}_{\mathbf{c}\_Q}\|$ that

$$0 < h_K \le \left(1-\left\|\mathbf{I}-K\mathbf{H}_{\mathbf{c}\_Q}\right\|_2\right)/K \le \left\|\mathbf{H}_{\mathbf{c}\_Q}\right\|_2. \tag{208}$$

It follows from (208) that if λ satisfies $0 < \lambda < 1$, then (207) is tighter than (183) with $J = n$, thus (183) is satisfied for $J = n$. Furthermore, for any $0 < \beta < 1$, if λ satisfies $0 < \lambda < \beta$, (181)-(182) with β replaced by λ are tighter than (181)-(182), thus (181)-(182) are also satisfied. It remains to show that there exists a positive number, $K_{\max}$, such that $0 < \lambda < \beta$ is satisfied for all $0 < K \le K_{\max}$. The proof of this directly follows from the expression of λ in (204).

□

Lemma 4 For any N-dimensional real vectors $\mathbf{v} = [v_j]$**,** $\mathbf{w} = [w_j]$ and $\mathbf{u} = [u_j]$**,** let <**v**, **w**> denote an inner product

$$\left\langle \mathbf{v},\mathbf{w}\right\rangle = \sum_{j=1}^{N} E\left[v_j w_j\right] \tag{209}$$

And let |**v**, **w**| denote the magnitude of <**v**, **w**>, the following properties

$$\|\mathbf{u}\|^2 = |\mathbf{u},\mathbf{u}| \tag{210}$$

and

$$|\mathbf{u},\mathbf{w}| - |\mathbf{v},\mathbf{w}| \le |\mathbf{u}+\mathbf{v},\mathbf{w}| \le |\mathbf{u},\mathbf{w}| + |\mathbf{v},\mathbf{w}|, \tag{211}$$

must hold. The proof of Lemma 4 follows from the definition of inner product and RMS norm, and the property of <**u**+**v**, **w**> = <**u**, **v**>+<**u**, **w**>.

□

Lemma 5: Any N-dimensional real vector **v** and any $N\times N$ deterministic real matrix **D** must satisfy

$$\left\|\mathbf{Dv}\right\| \leq \left\|\mathbf{D}\right\|_2 \left\|\mathbf{v}\right\|. \tag{212}$$

Proof of Lemma 5:

It follows from (151) that $\|\mathbf{Dv}\|_2 \leq \|\mathbf{D}\|_2\|\mathbf{v}\|_2$, where $\|\mathbf{v}\|_2$ is the Euclidean norm of **v**. Since **D** is deterministic, it follows that

$$E\left\{\left\|\mathbf{Dv}\right\|_2^{\ 2}\right\} \leq \left\|\mathbf{D}\right\|_2^{\ 2} E\left\{\left\|\mathbf{v}\right\|_2^{\ 2}\right\}. \tag{213}$$

By definition, $E\{\|\mathbf{v}\|_2\}$ and $E\{\|\mathbf{Dv}\|_2\}$ is the RMS norm of **v** and **Dv**, respectively, this yields (212).

□

Lemma 6: Let $I_{n,p}$ = min(n−p+N, n−j_{ini}), where j_{ini} is the starting time index of the basis initial condition in (164). Let u denote $<s\mathbf{D}_1\mathbf{x}_{l1}[n-i_1], \mathbf{D}_2\mathbf{x}_{l2}[n-i_2]>$, where $\mathbf{x}_{l1}[n]$ and $\mathbf{x}_{l2}[n]$ are any two state variables in the difference equations (161) and (164), $\mathbf{D}_1$ and $\mathbf{D}_2$ are any $N\times N$ deterministic real matrixes, s represents any modulation variable. For any $n \geq \max(j_{\text{ini}}+1, 0)$, $p > j_{\text{ini}}-Q$, and i_1, $i_2 \geq I_{n,p}$,

$$\left|u\right| \leq c\left\{s[p]\right\}\left\|\mathbf{D}_1\right\|_2\left\|\mathbf{D}_2\right\|_2 x_{\max}{}^2[n-1](1-4K\left\|\mathbf{H}_{\mathbf{c}\_Q}\right\|_2)^{-(i_1+i_2-2)/2} \tag{214}$$

holds provided that (183) holds for all $q < n$, where $c\{s[p]\}$ is given by (184) with $s[p]$ representing all portions of s associated with time index p.

Proof of Lemma 6:

Given that $p > j_{\text{ini}}-Q$, it follows from the definition of $\mathbf{q}_{\text{base}(k)}$ in (164) that $\mathbf{q}_{\text{base}(l)}$ for any $1 \leq l \leq L$ is not associated with any form of $s[p]$ at time index p. Therefore, if the s_{base}

term in the basis initial condition of (164) is replaced by deterministic 1, neither $\mathbf{x}_l[n-i]$ for $n-i \leq p-N$ nor $n-i \leq j_{\text{ini}}$ are associated with any form of $s[p]$ at time index p. This also means that $\mathbf{x}_l[n-i]$ for any $i \geq I_{n,p}$ must not be associated with any form of $s[p]$ at time index p, let $\mathbf{v}_l'[p]$ denote this vector. If the s_{base} term in (164) is included, $\mathbf{x}_l[n-i]$ is simply $\mathbf{v}_1'[p]$ multiplied by s_{base} and thus

$$\mathbf{D}\mathbf{x}_l[n-i] = \mathbf{D}\mathbf{v}_l\,'[p]s_{\text{base}} = s_{\text{base}}[p]\mathbf{w}_l\,'[p], \tag{215}$$

where $\mathbf{w}_l'[p] = \mathbf{D}\mathbf{v}_l'[p]s_{\text{base}}/s_{\text{base}}[p]$, $s_{\text{base}}[p]$ represents all portions of s_{base} associated with time index p (let $s_{\text{base}}[p]$ = deterministic 1 if s_{base} is not associated with time index p), thus neither $\mathbf{w}_l'[p]$ nor $(s/s[p])\mathbf{w}_l'[p]$ are associated with $s[p]$ with time index p. This, (215), and Lemma 8 imply

$$\begin{aligned} &\left|s\mathbf{D}_1\mathbf{x}_{l1}[n-i_1], \mathbf{D}_2\mathbf{x}_{l2}[n-i_2]\right| = \left|s[p]s_{\text{base}}{}^2[p](s/s[p])\mathbf{w}_{l1}\,'[p], \mathbf{w}_{l2}\,'[p]\right| \\ &= \left|E\left\{s[p]s_{\text{base}}{}^2[p]\right\}\right| \cdot \left|(s/s[p])\mathbf{w}_{l1}\,'[p], \mathbf{w}_{l2}\,'[p]\right| \leq \left|E\left\{s[p]s_{\text{base}}{}^2[p]\right\}\right| \cdot \left\|\mathbf{w}_{l1}\,'[p]\right\| \cdot \left\|\mathbf{w}_{l2}\,'[p]\right\|. \end{aligned} \tag{216}$$

The last step of (216) applies the property of $\|(s/s[p])\mathbf{w}_l'[p]\| \leq \|\mathbf{w}_l'[p]\|$, which holds because $-1 \leq s/s[p] \leq 1$. By definition, $E\{s_{\text{base}}^2[p]\} \neq 0$ and $s_{\text{base}}[p]$ is independent of $\mathbf{w}_l'[p]$, this and (215) imply that $\|\mathbf{w}_l'[p]\| = \|\mathbf{D}\mathbf{x}_l[n-i]\|/(E\{s_{\text{base}}^2[p]\})^{1/2}$. Furthermore, $s[p]s_{\text{base}}^2[p]$ is independent of $(s_{\text{base}}'[p])^2$, thus $E\{s[p]s_{\text{base}}^2[p]\}/E\{s_{\text{base}}^2[p]\} = E\{s[p]s_{\text{base}}^2\}/\ E\{s_{\text{base}}^2\} = c\{s[p]\}$ This and Lemma 5 imply that (216) is further upper bounded by

$$c\left\{s[p]\right\}\left\|\mathbf{D}_1\right\|_2\left\|\mathbf{D}_2\right\|_2\left\|\mathbf{x}_{l2}[n-i_1]\right\| \cdot \left\|\mathbf{x}_{l2}[n-i_2]\right\|. \tag{217}$$

The definition of $I_{n,p}$ with $n \geq \max(j_{\text{ini}}+1, 0)$ implies that $I_{n,p} > 0$, which also implies that $i_1 > 0$ and $i_2 > 0$. Substituting (165) and (183) into (217) yields (214).

□

Lemma 7: For <$\mathbf{H}_n$> and <$\mathbf{H}'_n$> defined in the proof of Lemma 3, let their state variables satisfy the difference equations (161) and (164). For any $n \geq \max(j_{\text{ini}}+1, 0)$, there must exist a range of K given by $0 < K \leq \mathcal{E}$, for each K within this range, if (183) is satisfied for all $q < n$, then

$$\left|\left\langle \mathbf{H}_n \right\rangle\right| \leq \left(\rho + o(K)\right) c_k[n-Q] x_{\max}{}^2[n-1] \tag{218}$$

and

$$\left|\left\langle \mathbf{H}_n{}' \right\rangle\right| \leq o(K) c_k[n-Q] x_{\max}{}^2[n-1] \tag{219}$$

are both satisfied, where ρ is the sum of $\|\mathbf{D}_{q1,i1}\|_2\|\mathbf{D}_{q2,i2}\|_2$ for all i_1, i_2, q_1 and q_2 in $<\mathbf{H}_n>$, which is bounded by definition, $o(K)$ is a bounded term in the order of K.

Proof of Lemma 7:

The definition of $<\mathbf{H}_n>$ or $<\mathbf{H}'_n>$ imply that each component in its sum term has the form of $<s\mathbf{D}_{q1,i1}\mathbf{x}_{l1}[n-i_1], \mathbf{D}_{q2,i2}\mathbf{x}_{l2}[n-i_2]>$ for i_1, $i_2 \geq 1$ and a finite number of q_1 and q_2. Lemma 4 implies that $|<\mathbf{H}_n>|$ or $|<\mathbf{H}'_n>|$ is upper bounded by the sum of the magnitude of each of these components. The definitions of $<\mathbf{H}_n>$ and $<\mathbf{H}'_n>$ imply that $c\{s[n-Q]\} \leq c_k[n-Q]$.

The proof of (218) is presented as follows. For any given $<s\mathbf{D}_{q1,i1}\mathbf{x}_{l1}[n-i_1], \mathbf{D}_{q2,i2}\mathbf{x}_{l2}[n-i_2]>$ component of $<\mathbf{H}_n>$, if i_1, $i_2 \geq I_{n,n-Q}$, where $I_{n,n-Q}$ is defined in Lemma 6 with p replaced by $n-Q$ and is bounded by definition, then it follows from Lemma 6 that its magnitude is upper bounded by $c_k[n-Q]\|\mathbf{D}_{q1,i1}\|_2\|\mathbf{D}_{q2,i2}\|_2(1-4K\|\mathbf{H}_{\mathbf{c}\_Q}\|_2)^{-(i1+i2-2)/2}x_{\max}{}^2[n-1]$. If at least one of i_1 and i_2 is smaller than $I_{n,n-Q}$, let I_1 be the larger value of i_1 and $I_{n,n-Q}$, and let I_2 be the larger value of i_2 and $I_{n,n-Q}$, it follows from this definition that both I_1-i_1 and I_2-i_2 are bounded, thus it follows from (201) that $<s\mathbf{D}_{q1,i1}\mathbf{x}_{l1}[n-i_1], \mathbf{D}_{q2,i2}\mathbf{x}_{l2}[n-i_2]>$ can be written as $<s\mathbf{D}_{q1,i1}\mathbf{x}_{l1}[n-I_1], \mathbf{D}_{q2,i2}\mathbf{x}_{l2}[n-I_2]>$ plus at most additional 3 terms, each of these additional terms has the form of $K^j<s\mathbf{D}_{q1,i1}\mathbf{H}(\mathbf{x}[n-I_1]), \mathbf{D}_{q2,i2}\mathbf{H}(\mathbf{x}[n-I_2])>$ for $j \geq 1$, where $\mathbf{H}(\mathbf{x}[n-I_1])$ and $\mathbf{H}(\mathbf{x}[n-I_2])$ each has the form of $\mathbf{H}(\mathbf{x}[n-1])$ with n replaced by $n-(I_1-1)$ and $n-(I_2-1)$, respectively.

Let us analyze $<s\mathbf{D}_{q1,i1}\mathbf{H}(\mathbf{x}[n-I_1]), \mathbf{D}_{q2,i2}\mathbf{H}(\mathbf{x}[n-I_2])>$ first, this term can be further expanded into $<S\mathbf{D}_{q1,i1}\mathbf{D}_{q3,i3}\mathbf{x}[n-(I_1-1)-i_3], \mathbf{D}_{q2,i2}\mathbf{D}_{q4,i4}\mathbf{x}[n-(I_2-1)-i_4]>$ for $i_3 \geq 1$ and $i_4 \geq 1$ and a finite number of q_3 and q_4, where $\mathbf{D}_{q3,i3}$ and $\mathbf{D}_{q4,i4}$ are deterministic matrixes from $\mathbf{H}(\mathbf{x}[n-I_1])$ and $\mathbf{H}(\mathbf{x}[n-I_2])$, respectively. Since each new modulation variable, S, is scaled by the original modulation variable, s, thus S is also the product of $s_1 \epsilon \{S_k[n-Q], c_k[n-Q], (S_k^2[n-Q]-c_k[n-Q])/2\}$ and another modulation variable and satisfies $c\{S[n-Q]\} \leq c_k[n-Q]$, and since $(I_1-1)+i_3 \geq I_{n,n-Q}$ and $(I_2-1)+i_4 \geq I_{n,n-Q}$, it follows from Lemma 6 and the matrix property of $\|\mathbf{D}_1\mathbf{D}_2\|_2 \leq \|\mathbf{D}_1\|_2\|\mathbf{D}_2\|_2$ that $|S\mathbf{D}_{q1,i1}\mathbf{D}_{q3,i3}\mathbf{x}[n-(I_1-1)-i_3], \mathbf{D}_{q2,i2}\mathbf{D}_{q4,i4}\mathbf{x}[n-(I_2-1)-i_4]| \leq c_k[n-Q]\|\mathbf{D}_{q1,i1}\|_2\|\mathbf{D}_{q2,i2}\|_2\|\mathbf{D}_{q3,i3}\|_2\|\mathbf{D}_{q4,i4}\|_2(1-4K\|\mathbf{H}_{\mathbf{c}\_Q}\|_2)^{-(i3+I1+i4+I2-4)/2}x_{\max}^2[n-1]$. Since $\|\mathbf{D}_{q3,i3}\|_2$ and $\|\mathbf{D}_{q4,i4}\|_2$ are both bounded by exponentially decaying curves as i_3 and i_4 increase, it follows that if K is positive and smaller than a certain value, both $\|\mathbf{D}_{q3,i3}\|_2(1-4K\|\mathbf{H}_{\mathbf{c}\_Q}\|_2)^{-i3/2}$ and $\|\mathbf{D}_{q4,i4}\|_2(1-4K\|\mathbf{H}_{\mathbf{c}\_Q}\|_2)^{-i4/2}$ are also bounded by exponentially decaying curves as i_3 and i_4 increase, and the sum of them over all $i_3, i_4 \geq 1$ and a finite number of q_3 and q_4 are bounded by $bx_{\max}^2[n-1]$, where b is bounded. Since both I_1-i_1 and I_2-i_2 are bounded, it follows that $|s\mathbf{D}_{q1,i1}\mathbf{H}(\mathbf{x}[n-I_1]), \mathbf{D}_{q2,i2}\mathbf{H}(\mathbf{x}[n-I_2])|$ is bounded by $b(1+o(K))c_k[n-Q]\|\mathbf{D}_{q1,i1}\|_2\|\mathbf{D}_{q2,i2}\|_2(1-4K\|\mathbf{H}_{\mathbf{c}\_Q}\|_2)^{-(i1+i2)/2}x_{\max}^2[n-1]$.

Combining these results imply that $|s\mathbf{D}_{q1,i1}\mathbf{x}[n-i_1], \mathbf{D}_{q2,i2}\mathbf{x}[n-i_2]|$ is bounded by $(1+o(K))c_k[n-Q]\|\mathbf{D}_{q1,i1}\|_2\|\mathbf{D}_{q2,i2}\|_2(1-4K\|\mathbf{H}_{\mathbf{c}\_Q}\|_2)^{-(i1+i2)/2}x_{\max}^2[n-1]$, where $o(K)$ is bounded and in the order of K. Since $\|\mathbf{D}_{1,i1}\|_2$ and $\|\mathbf{D}_{2,i2}\|_2$ are also bounded by exponentially decaying curves as i_1 and i_2 increase, thus if K is positive and smaller than a certain value, summing up these results for all $i_1, i_2 \geq 1$ and a finite number of q_1 and q_2 yields the bound in (218).

The proof of (219) is presented as follows. The definition of $<\mathbf{H}'_n>$ implies that for each $<s\mathbf{D}_{q1,i1}\mathbf{x}[n-i_1], \mathbf{D}_{q2,i2}\mathbf{x}[n-i_2]>$ component in it, there must exist an integer p satisfying $c\{s[n-p]\} = 0$, where $n-p > j_L-Q$, and both $p-i_1$ and $p-i_2$ cannot be unbounded positive integer. Let I denote the larger value of $I_{n,n-p}$ and $I_{n,n-Q}$. If both $i_1 \geq I$ and $i_2 \geq I$ are satisfied, then it follows from Lemma 6 and $c\{s[n-p]\} = 0$ that the magnitude of this term is 0. Otherwise, let I_1 be the larger value of i_1 and I, and let I_2 be the larger value of i_2 and I. Since both $p-i_1$ and $p-i_2$ cannot be unbounded positive integers, it follows that both I_1-i_1 and I_2-i_2 must be bounded, thus it follows from (201) that $<s\mathbf{D}_{1,i1}\mathbf{x}[n-i_1], \mathbf{D}_{2,i2}\mathbf{x}[n-i_2]>$ can be expanded into the sum of at most 4 terms, the first term is $<s\mathbf{D}_{1,i1}\mathbf{x}[n-I_1], \mathbf{D}_{2,i2}\mathbf{x}[n-I_2]>$, the magnitude of which is 0 as implied by Lemma 6, and the other terms each have the form of $K^j<s\mathbf{D}_{q1,i1}\mathbf{H}(\mathbf{x}[n-I_1]), \mathbf{D}_{q2,i2}\,\mathbf{H}(\mathbf{x}[n-I_2])>$ for $j = 1$ or 2. It follows from the reasoning used in the proof of (218) that the overall magnitude of these other terms is bounded by $o(K)c_k[n-Q]||\mathbf{D}_{q1,i1}||_2||\mathbf{D}_{q2,i2}||_2(1-4K||\mathbf{H}_{\mathbf{c}\_Q}||_2)^{-(i1+i2)/2}x_{\max}{}^2[n-1]$, and summing up all these results for all i_1, $i_2 \geq 1$ and a finite number of q_1 and q_2 yields (219).

□

Lemma 8: Any N-dimensional real vectors $\mathbf{v} = [v_j]$ and $\mathbf{w} = [w_j]$ must satisfy

$$\left|\mathbf{v},\mathbf{w}\right| \leq \left\|\mathbf{v}\right\| \cdot \left\|\mathbf{w}\right\|. \tag{220}$$

Proof of Lemma 8:

$$\left|\mathbf{v},\mathbf{w}\right| \leq \sum_{j=1}^{N}\left|E\left[v_j w_j\right]\right| \leq \sum_{j=1}^{N}\sqrt{E\left[v_j{}^2\right]}\sqrt{E\left[w_j{}^2\right]}. \tag{221}$$

The Cauchy–Schwarz inequality further yields (220).

□

APPENDIX B

For any given $0 < \beta < 1$, suppose K is chosen small enough such that the results of Lemma 3 hold. The contribution of each component of $K\mathbf{e}_{k,p}$ to $||\mathbf{z}_k[n]||^2$ is given by $||\mathbf{x}_k[n]||^2$ in (182) with $j_{\text{ini}} = p$. Since $||\mathbf{x}_k[p]||^2$ is either $K^2||S_k[p-Q]\mathbf{r}[p]||^2$ or $K^2||S_l[p-i]S_k[p-Q]$

($\mathbf{b}_{l\_i}$−$\mathbf{a}_l$')‖^2 with $i \neq Q$, and the statistics of $S_k[n]$ for each k is time invariant, these and (165) imply that $x_{\max}{}^2[p] = (c_{\max}/c_k)\|\mathbf{x}_k[p]\|^2$, where $c_k = E\{S_k{}^2[n]\}$, $c_{\max}$ is the maximum value of c_k over k.

The properties of (167)-(168) imply that $K_{(k,n-Q)} = 2c_kK$ for $n-Q \geq p+N$, this and (182) imply

$$\|\mathbf{x}_k[n]\|^2 \leq \|\mathbf{x}_k[p]\|^2 \left(1-2c_kKh_K\right)^{n-p-Q-N} + f_p(\beta), \tag{222}$$

where $f_p(\beta)$ represents the terms contributed by $\beta \neq 0$ given by

$$f_p(\beta) = 2\beta c_kKh_K \sum_{i=p}^{n-1} (1-2c_kKh_K)^{n-i-Q-N} x_{\max}{}^2[i]. \tag{223}$$

To simplify the notation, let $1-2c_kKh_K = a$, $1-2(1-\beta)c_{\min}Kh_K = b$, it follows from (181) that $x_{\max}{}^2[i]$ is upper bounded by $(1+o(K))b^{i-p}x_{\max}{}^2[p]$, where $o(K)$ is bounded and in the order of K, and it further follows from $x_{\max}{}^2[p] = (c_{\max}/c_k)\|\mathbf{x}_k[p]\|^2$ that $x_{\max}{}^2[i]$ is bounded by $(c_{\max}/c_k)(1+o(K))b^{i-p}\|\mathbf{x}_k[p]\|^2$. Substituting this into (223) yields

$$f_p(\beta) \leq 2\beta c_{\max}Kh_K \left(1+o(K)\right)\|\mathbf{x}_k[p]\|^2 \sum_{i=p}^{n-1} a^{n-i}b^{i-p}. \tag{224}$$

The sum of the first term on the right side of (222) for p = 0, 1, 2, …, n is $(1+o(K))\|\mathbf{x}_k[p]\|^2/(2c_kKh_K)$, which corresponds to the case of $\beta = 0$ analyzed before. The same sum of the second term, $f_p(\beta)$, is bounded by the sum of the right side of (224) for p = 0, 1, 2, …, n, which is again bounded by

$$\sum_{p=0}^{n} f_p(\beta) \leq 2\beta c_{\max}Kh_K \left(1+o(K)\right)\|\mathbf{x}_k[p]\|^2 \sum_{p=0}^{n} a^p \sum_{p=0}^{n} b^p. \tag{225}$$

It follows from the definition of a and b that (225) is bounded by $(1+o(K))(c_{\max}/c_{\min})\|\mathbf{x}_k[p]\|^2\beta/(1-\beta)(2c_kKh_K)$. Combining these results implies that the sum of (222) for p = 0, 1, 2, …, n is $(1+o(K))(1+(c_{\max}/c_{\min})\beta/(1-\beta))\|\mathbf{x}_k[p]\|^2/(2c_kKh_K)$, and summing up them with each $\|\mathbf{x}_k[p]\|^2$ replaced by its actual expression yields (153) with $\alpha = K(1+o(K))(1+(c_{\max}/c_{\min})\beta/(1-\beta))/(2h_K)$.

ACKNOWLEDGEMENTS

Chapter 4, in part is currently being prepared for submission for publication of the material. D. Kong, I. Galton. The dissertation author is the primary investigator and author of this paper. Professor Ian Galton supervised the research which forms the basis of this paper.

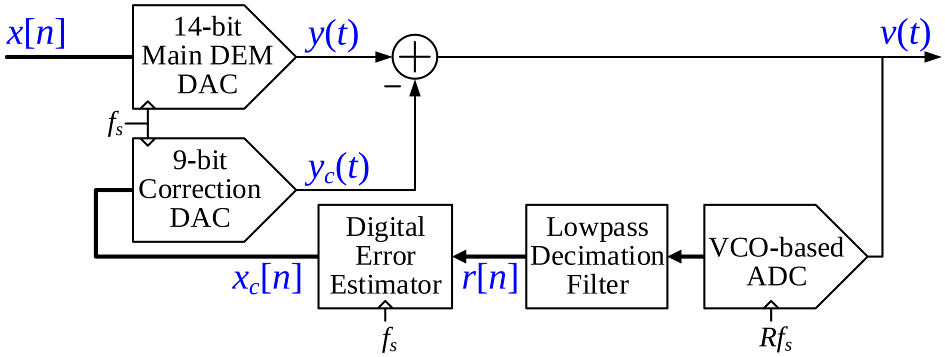


Figure 34: High-level signal processing diagram of MNC.

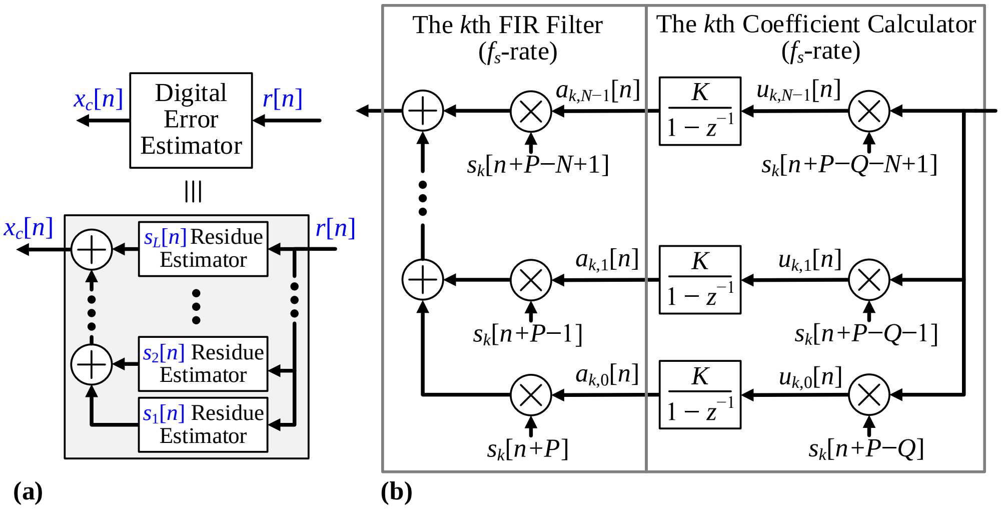


Figure 35: a) High-level structure of the digital error estimator, and b) signal processing details of each $s_k[n]$ residue estimator.

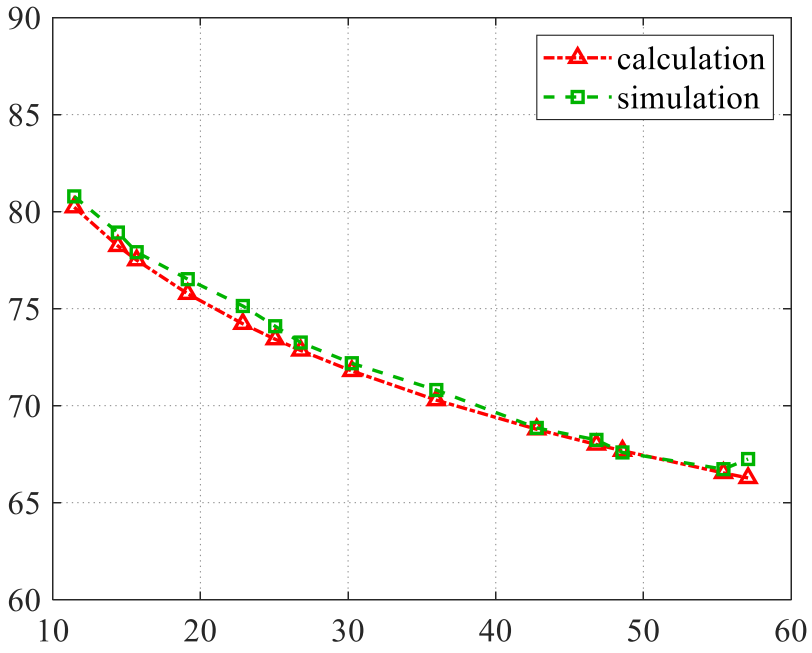


Figure 36: DAC SNR$_c$ for different values of σ_e.

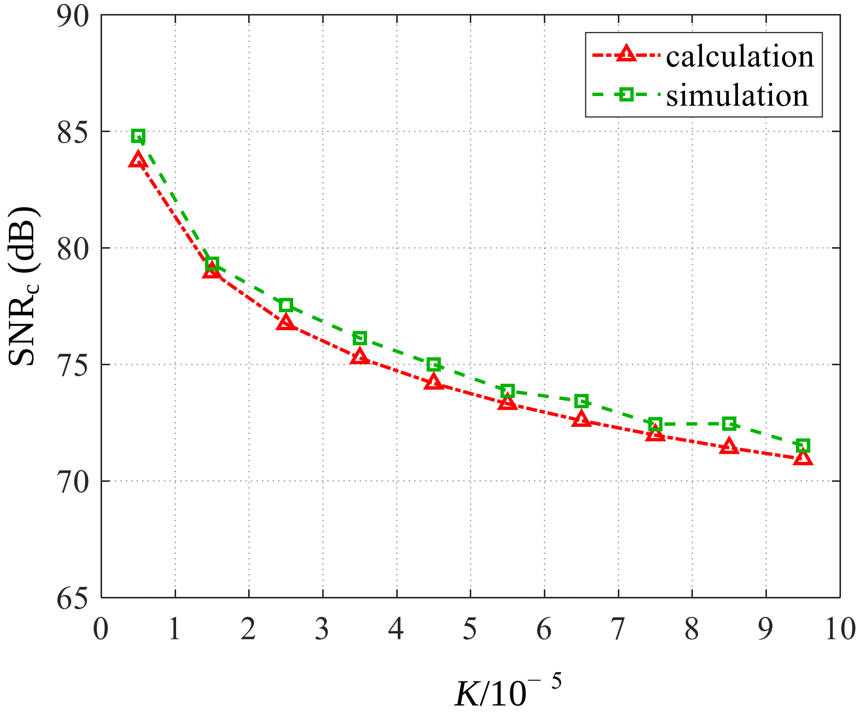


Figure 37: DAC SNR$_c$ for different values of K.

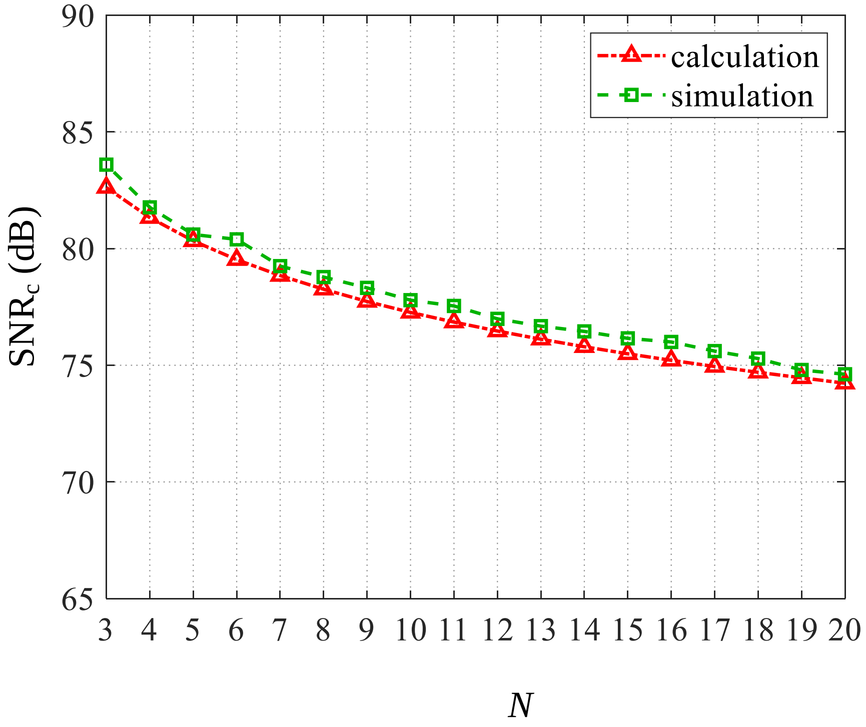


Figure 38: DAC SNR$_c$ for different values of N.

REFERENCES

1. W. Schofield, D. Mercer, L. St. Onge, “A 16b 400MS/s DAC with <80dBc IMD to 300MHz and 160dBm/Hz noise Power Spectral Density,” *IEEE International Solid State Circuits Conference,* February 2003.

2. Q. Huang, P. A. Francese, C. Martelli, and J. Nielsen,“A 200Ms/s 14b 97 mW DAC in 0.18µm CMOS,” *IEEE International Solid State Circuits Conference,* February 2004.

3. H.-H. Chen, J. Lee, J. Weiner, Y.-K. Chen, and J.-T. Chen, “A 14-bit 150 MS/s CMOS DAC with Digital Background Calibration,” *Symposium on VLSI Circuits*, pp. 51-52, June 2006.

4. M. Clara, W. Klatzer, B. Seger, A. Di Giandomenico, and L. Gori, “A 1.5V 200MS/s 13b 25mW DAC with Randomized Nested Background Calibration in 0.13 µm CMOS,” *IEEE International Solid State Circuits Conference,* February 2007.

5. M. Clara, W. Klatzer, D. Gruber, A. Marak, B. Seger, and W. Pribyl, “A 1.5 V 13 bit 130-300 MS/s Self-calibrated DAC with Active Output Stage and 50 MHz Signal Bandwidth in 0.13µm CMOS,” *European Solid-State Circuits Conference*, pp. 262-265, September, 2008.

6. B. Catteau, P. Rombouts, J. Raman, and L. Weyten, “An on-line calibration technique for mismatch errors in high-speed DACs,” *IEEE Transactions on Circuits and Systems–I, Reg. Papers*, vol. 55, no. 7, pp. 1873–1883, Aug. 2008

7. C.-H. Lin, F. M. L. van der Goes, J. R. Westra, J. Mulder, Y. Lin, E. Arslan, E. Ayranci, X. Liu, K. Bult, “A 12 bit 2.9 GS/s DAC With IM3 < −60 dBc Beyond 1 GHz in 65 nm CMOS,” *IEEE Journal of Solid-State Circuits*, vol. 44, no. 23, pp. 3285-3293, December 2009.

8. Y. Tang, J. Briaire, K. Doris, R. van Veldhoven, P. van Beek, H. Hegt, and A. van Roermund, “A 14 bit 200 MS/s DAC With SFDR >78 dBc, IM3 < −83 dBc and NSD < −163 dBm/Hz Across the Whole Nyquist Band Enabled by Dynamic-Mismatch Mapping,” *IEEE Journal of Solid-State Circuits*, vol. 46, no. 6, pp. 1371-1381, June 2011.

9. S. Spiridon, J. van der Tang, H. Yan, H.-F. Chen, G. Guermandi, X. Liu, E. Arslan, R. van der Goes, K. Bult, “A 375 mW Multimode DAC-Based Transmitter With 2.2 GHz Signal Bandwidth and In-Band IM3<−58 dBc in 40 nm CMOS,” *IEEE Journal of Solid-State Circuits*, vol. 48, no. 7, pp. 1595-1604, July 2013.

10. W.-T. Lin, H.-Y. Huang, and T.-H. Kuo, “A 12-bit 40 nm DAC Achieving SFDR > 70 dB at 1.6 GS/s and IMD < -61dB at 2.8 GS/s With DEMDRZ Technique,” *IEEE Journal of Solid-State Circuits*, Vol. 49, no. 3, pp. 708-717, March 2014.

11. S. M. Lee, D. Seo, S. M. Taleie, D. Kong, M. J. McGowan, T. Song, G. Saripalli, J. Kuo, S. Bazarjani, "A 14b 750MS/s DAC in 20nm CMOS with <−168dBm/Hz noise floor beyond Nyquist and 79dBc SFDR utilizing a low glitch-noise hybrid R-2R architecture," in *Symp. VLSI Circuits Dig*., June 2015.

12. Engel, M. Clara, H. Zhu, and P. Wilkins, "A 16-bit 10 Gsps currentsteering RF DAC in 65 nm CMOS achieving 65 dBc ACLR multi-carrier performance at 4.5 GHz Fout," in *Symp. VLSI Circuits Dig*., June 2015.

13. S. Su and M. S.-W. Chen, "A 12-bit 2 GS/s dual-rate hybrid DAC with pulse-error pre-distortion and in-band noise cancellation achieving > 74 dBc SFDR and < −80 dBc IM3 up to 1 GHz in 65 nm CMOS," *IEEE Journal of Solid-State Circuits,* vol. 51, no. 12, pp. 2963–2978, December 2016.

14. C.-H. Lin, K. L. J. Wong, T.-Y. Kim, G. R. Xie, D. Major, G. Unruh, S. R. Dommaraju, H. Eberhart, A. Venes, "A 16b 6GS/s Nyquist DAC with IMD <−90dBc up to 1.9GHz in 16nm CMOS," *IEEE International Solid State Circuits Conference*, February 2018.

15. D. Kong and I. Galton, "Adaptive Cancellation of Static and Dynamic Mismatch Error in Continuous-Time DACs," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 2, pp. 421–433, Feb. 2018.

16. D. Kong, K. Rivas-Rivera, I. Galton, "A 600 MS/s DAC with over 87dB SFDR and 77dB peak SNDR Enabled by Adaptive Cancellation of Static and Dynamic Mismatch Error," *IEEE Journal of Solid-State Circuits*, under review (Submitted manuscript available at http://ispg.ucsd.edu/unpublished-paper/).

17. K. L. Chan, J. Zhu, and I. Galton, "Dynamic Element Matching to Prevent Nonlinear Distortion From Pulse-Shape Mismatches in High-Resolution DACs," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 9, pp. 2067-2078, September 2008.

18. J. Remple, I. Galton, "The Effects of Inter-Symbol Interference in Dynamic Element Matching DACs," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 1, pp. 14-23, January 2017.