UNIVERSITY OF CALIFORNIA, SAN DIEGO

Enabling Techniques for Wide Bandwidth Fractional-*N* Phase Locked Loops

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

Doctor of Philosophy

in Electrical and Computer Engineering (Electronic Circuits and Systems)

by

Sudhakar Pamarti

Committee in charge:

Professor Ian Galton, Chair
Professor Peter Asbeck
Professor Bhaskar Rao
Professor Robert Bitmead
Professor Patrick Fitzsimmons

2003

To my sister

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

LIST OF TABLES

ACKNOWLEDGEMENTS

Foremost, I am thankful to my advisor, Ian Galton, for his support and encouragement through out my graduate student years. I am particularly grateful for any sense of professional discipline that I might have cultivated under his guidance. I am thankful to Professors Peter Asbeck, Bhaskar Rao, Robert Bitmead, and Patrick Fitzsimmons who kindly agreed to serve on my doctoral committee, and Karol Provite, Carolyn Kuttner and Renee Gramlich for their support with the various administrative matters concerning graduate students. I am also grateful to Lars Jansson for his invaluable help and advice with my research.

I am thankful to Jim Thomas, Ian's family, and the past and the present members of Ian's research group – Henrik Jensen, Eric Fogleman, Bill Huff, Jared Welz, Sheng Ye, Eric Siragusa, Asaf Fishov, Alan Lewis, Ashok Swaminathan, Erica J. Poole, and Andrea Panigada – for making me feel welcome, and perfectly at home in San Diego. I am particularly grateful to Jared for our numerous and engrossing discussions about practically everything under the sun. These discussions have had a significant influence on my outlook, technical and otherwise, in more ways than he knows. I am also grateful to Eric Siragusa, Ashok Swaminathan, and Mani Vaidyanathan for curing me of my layout blues by introducing me to the wonderful world of Torrey Pines.

Above all, I am indebted to my dear sister, Lavanya, and my parents, Venkata Subba Rao and Sita Devi Pamarti, for their loving support and encouragement for every endeavor of mine.

VITA

May 14, 1974	Born, Hyderabad, India
1995	B.Tech., Indian Institute of Technology, Kharagpur
1995-1997	Hughes Software Systems, New Delhi
1997-2003	Graduate Student Researcher, Department of Electrical and Computer Engineering, University of California, San Diego
2003	Ph.D., University of California, San Diego

PUBLICATIONS

"A Wideband 2.4 GHz Delta-Sigma Fractional-*N* PLL With 1 Mb/s In-Loop Modulation," *IEEE Journal of Solid State Circuits*, accepted for publication.

"Phase Noise Cancellation Design Tradeoffs in Delta-Sigma Fractional-*N* PLLs," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, under review.

"One-bit Dithering in Single Stage Digital Delta-Sigma Modulators," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, in preparation.

"One-bit Dithering in Multi-stage Delta-Sigma Modulator Based Digital-Analog Conversion," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, in preparation.

ABSTRACT OF THE DISSERTATION

Enabling Techniques for Wide Bandwidth Delta-Sigma Fractional-*N* Phase Locked Loops

by

Sudhakar Pamarti

Doctor of Philosophy in Electrical and Computer Engineering

(Electronic Circuits and Systems)

University of California, San Diego, 2003

Professor Ian Galton, Chair

Delta-sigma fractional-*N* phase locked loops are widely used for frequency synthesis in electronic communication systems. A wide bandwidth makes it possible for the delta-sigma fractional-*N* phase locked loop to perform digitally-controlled frequency modulation at high bit-rates, thereby simplifying transceiver circuitry. Wide bandwidth delta-sigma fractional-*N* phase locked loops offer a multitude of other benefits that contribute to lower costs and a reduced power consumption in the electronic communication products which use these phase locked loops. In spite of the benefits, wide bandwidth delta-sigma fractional-*N* phase locked loops have not gained general acceptance because of their poor phase noise and spurious tone performance, particularly when they are implemented in integrated circuit (IC) form.

This dissertation presents two signal processing techniques – a *phase noise cancellation technique* and a *charge pump linearization technique* – that significantly reduce phase noise and spurious tones in a wide bandwidth delta-sigma fractional-N phase locked loop. Chapter 1 presents a prototype CMOS IC that demonstrates the efficacy of the two techniques – reduction of the phase noise by at least 16 dB, and reduction of spurious tones by at least 8 dB – in a 2.4 GHz delta-sigma fractional-N phase locked loop with 460 kHz wide bandwidth. Chapter 2 presents a theoretical basis for the phase noise cancellation technique and suggests design guidelines to tailor the technique to meet the target requirements of a general wide bandwidth delta-sigma fractional-N phase locked loop. The effectiveness of the phase noise cancellation technique hinges on eliminating limit cycles in the digital delta-sigma modulators, which the technique employs. Chapter 3 presents conditions to theoretically guarantee that one-bit dither eliminates limit cycles in a large class of digital delta-sigma modulators. It also extends the theory to suppress spurious tones in a large class of delta-sigma modulator based digital-to-analog conversion systems.

Chapter 1

A Wideband 2.4 GHz Delta-Sigma Fractional-*N* PLL With 1 Mb/s In-Loop Modulation

Sudhakar Pamarti, Lars Jansson, and Ian Galton

Abstract—A phase noise cancellation technique and a charge pump linearization technique, both of which are insensitive to component errors, are presented and demonstrated as enabling components in a wideband CMOS ΔΣ fractional-*N* PLL. The PLL has a loop bandwidth of 460 kHz and is capable of 1 Mb/s in-loop FSK modulation at center frequencies of 2402 + *k* MHz for *k* = 0, 1, 2, ...78. For each frequency, measured results indicate that the peak spot phase noise reduction achieved by the phase noise cancellation technique is 16 dB or better, and the minimum suppression of factional spurious tones achieved by the charge pump linearization technique is 8 dB or better. With both techniques enabled, the PLL achieves a worst-case phase noise of –121 dBc/Hz at 3 MHz offsets, and a worst-case in-band noise floor of –96 dBc/Hz. The PLL circuitry consumes 34.4 mA from 1.8–2.2V supplies. The IC is realized in a 0.18 μm mixed-signal CMOS process, and has a die size of 2.72mm × 2.47mm.

I. INTRODUCTION

This paper presents a phase noise cancellation technique that relaxes the fundamental tradeoff between phase noise and bandwidth in conventional delta-sigma (ΔΣ) fractional-*N* phase-locked loops (PLLs), and a charge pump linearization technique that improves the spurious performance of wideband fractional-*N* PLLs. Together, the techniques make it practical to significantly increase the bandwidth of ΔΣ fractional-*N* PLLs without degrading phase noise and spurious performance. They are demonstrated in a CMOS ΔΣ fractional-*N* PLL that can be configured as a Bluetooth-

compliant wireless LAN transmitter and a local oscillator for a direct-conversion Bluetooth-compliant receiver. The techniques enable the PLL to achieve the required phase noise and spurious performance specifications with a bandwidth of 460 kHz, which is sufficiently wide to allow in-loop modulation of the required 1 Mb/s transmit signal. Moreover, the wide bandwidth significantly reduces the susceptibility of the voltage-controlled oscillator (VCO) to pulling, and causes the PLL phase noise arising from $1/f$ noise and $1/f^3$ noise in the VCO to be largely attenuated [1, 2]. Unlike other methods of in-loop modulation for wireless transmitters, the phase noise cancellation technique is not sensitive to analog component errors, does not require calibration, and does not require a Type-1 PLL and the associated phase detector complications [3, 4, 5, 6, 7]. The benefit of the charge pump linearization technique is that it does not require dynamic bias adjustment so its bandwidth is not limited by an analog feedback circuit [8]. Although the two techniques complement each other in that they both enhance performance in wideband $\Delta\Sigma$ fractional-N PLLs, they are independent and each can be applied in the absence of the other.

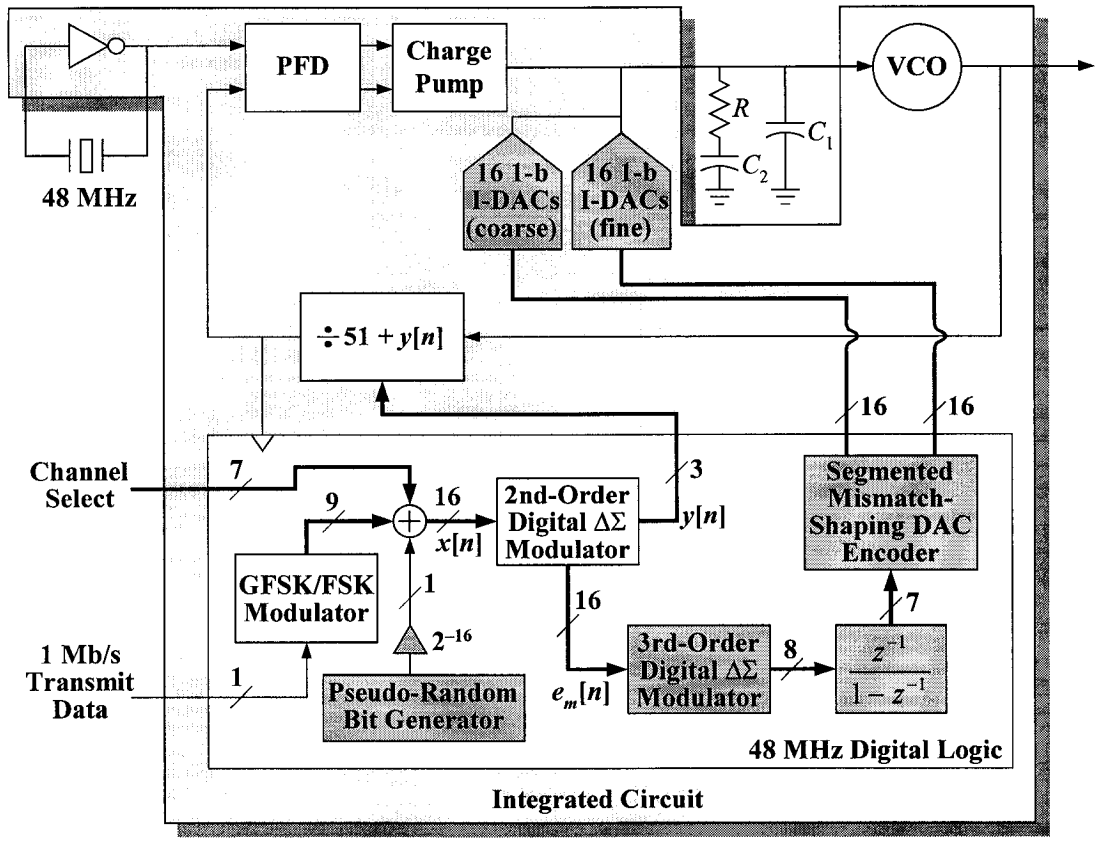

Fig. 1.1: High-level functional diagram of the implemented ΔΣ fractional-*N* PLL.

A high-level block diagram of the implemented PLL is shown in Fig. 1.1. It differs from a conventional ΔΣ fractional-*N* PLL in that the dark gray blocks have been added to implement the phase noise cancellation technique, and the charge pump and phase-frequency detector (PFD) blocks have been modified from their conventional forms to implement the charge pump linearization technique. The details of the PLL are described throughout the remainder of the paper. Sections II and III describe the signal processing details of the phase noise cancellation technique and the charge pump linearization technique, respectively. Section IV presents circuit details, and Section V presents measurement results.

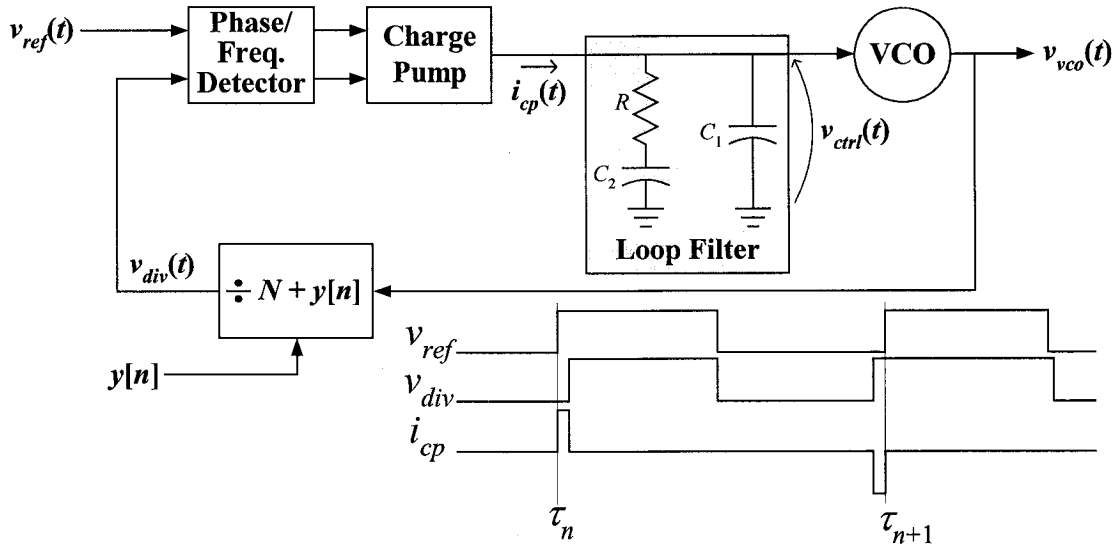

Fig. 1.2: The "core" of a typical fractional-*N* PLL.

II. THE PHASE NOISE CANCELLATION TECHNIQUE

A. Problems with Conventional Fractional-*N* PLLs

The core of a typical fractional-*N* PLL is shown in Fig. 1.2. It consists of a PFD, a charge pump, a loop filter, a VCO, and a frequency divider. The divider output, $v_{div}(t)$, is a two level signal in which the nth and $(n+1)$th rising edges are separated by $N + y[n]$ periods of the VCO output, for $n = 1, 2, 3, \ldots$, where N is a constant integer, and $y[n]$ is a sequence of integers generated by digital logic not shown in the figure. As indicated in the figure for the case where the PLL is locked, if the nth rising edge of the reference signal, $v_{ref}(t)$, occurs before that of $v_{div}(t)$, the charge pump generates a current pulse of nominal amplitude I_{CP} and a duration equal to the time difference between the two edges. Otherwise, the situation is similar except the polarity of the current pulse is reversed.

If $y[n]$ could be set to any desired value between –1 and 1, say α, then the output frequency of the PLL would settle to $(N + \alpha) f_{ref}$, so it would be possible to achieve any output frequency between $(N - 1) f_{ref}$ and $(N + 1) f_{ref}$. Unfortunately, $y[n]$ is restricted to integer values because the divider simply counts rising VCO edges. However, $y[n]$ can be a sequence of integer values that *average* to α. Such a sequence can be written as $y[n] = \alpha + e_m[n]$, where $e_m[n]$ is zero-mean *quantization noise* caused by using integer values in place of the ideal fractional value. In this case, the PLL output frequency settles to $(N + \alpha) f_{ref}$ as desired, although a price is paid in terms of added phase noise.

As shown in [9], in terms of the effect it has on the PLL phase noise, the quantization noise can be modeled as a sequence of additive charge samples, $Q_{e_m}[n]$, that get injected into the loop filter once every reference period. Neglecting a constant offset associated with the initial conditions of the loop filter, it can be shown that $Q_{e_m}[n]$ is well modeled as

$$Q_{e_m}[n] = T_{VCO} I_{CP} \sum_{k=n_0}^{n-1} e_m[k], \tag{1}$$

where T_{VCO} is the period of the VCO output, and $n_0 < n$ is an arbitrary initial time index. The PLL acts on this sequence as a lowpass filter in the process of converting it to output phase noise. Therefore, spectral components of $e_m[n]$ outside the bandwidth of the PLL are suppressed, but those inside the bandwidth of the PLL are amplified through the discrete-time integration in (1) and can add significantly to the overall phase noise of the PLL.

In early fractional-N PLLs the problem of suppressing the PLL phase noise that would otherwise result from $e_m[n]$ has been addressed using a *DAC cancellation path* to suppress $Q_{e_m}[n]$ [10, 11]. Because $y[n]$ is generated digitally, $-Q_{e_m}[n]$ can be calculated by digital circuitry, converted by a DAC to an analog current, and added to the output of the charge pump. If the DAC has sufficient precision and the correct gain, the added signal nearly cancels the component of the charge pump output corresponding to $Q_{e_m}[n]$. In most fractional-N PLLs of this type, $y[n]$ is generated using one or two digital error-accumulator structures designed to ensure that the sum of $e_m[n]$ in (1) is bounded. The resulting $Q_{e_m}[n]$ sequence tends to have a large dynamic range, a high spurious tone content, and significant spectral power within the PLL bandwidth. Therefore, excellent cancellation accuracy is required; if $Q_{e_m}[n]$ is only partially cancelled because of gain errors, distortion, or insufficient dynamic range in the DAC cancellation path, the remaining portion of $Q_{e_m}[n]$ contains in-band noise and spurious tones which can contribute significant phase noise [12, 13]. Consequently, the approach has been used mainly in high-cost applications such as test and measurement equipment wherein component trimming and calibration are practical.

A more recent technique that circumvents the DAC precision and gain matching problems uses a digital $\Delta\Sigma$ modulator with at least second-order quantization noise shaping to generate $y[n]$ such that $Q_{e_m}[n]$ has at least one zero at dc with most of its power concentrated at high frequencies, outside the passband of the PLL [14, 15,

16]. Provided the bandwidth of the PLL is sufficiently narrow, most of the quantization noise is suppressed by the PLL so a DAC cancellation path is not necessary. Such PLLs have come to be known as ΔΣ fractional-*N* PLLs, and have become widely used in consumer-oriented communication devices over the last decade. Nevertheless, the need to suppress out-of-band quantization noise imposes a fundamental bandwidth versus phase noise tradeoff in ΔΣ fractional-*N* PLLs that causes problems in many applications.

One such problem is *VCO pulling*. For example, when a narrowband PLL is used to provide the RF local oscillator for a direct conversion transmitter, even a small amount of parasitic coupling of the transmitted signal to the VCO circuitry tends to corrupt or *pull* the VCO output which, in turn, causes the up-converted transmit signal to be distorted. However, if the bandwidth of the PLL is at least comparable to the modulation bandwidth, the PLL is much less susceptible to this problem because the feedback within the PLL tends to fight the corrupting effects of the modulated transmit signal.

Another problem with narrowband fractional-*N* PLLs is that they often preclude in-loop VCO modulation for direct synthesis of frequency modulated transmit signals. In principle such signals can be generated directly by a ΔΣ fractional-*N* PLL thereby eliminating the need for conventional upconversion stages and much of the attendant analog circuitry. Specifically, if α in the discussion above is replaced by $\alpha + x_m[n]$, where $x_m[n]$ is a zero-mean modulation sequence, the resulting PLL output has a center frequency of $(N + \alpha) f_{ref}$ but is frequency modulated by a lowpass filtered version of $x_m[n]f_{ref}$. The PLL must have a sufficiently narrow bandwidth to suppress

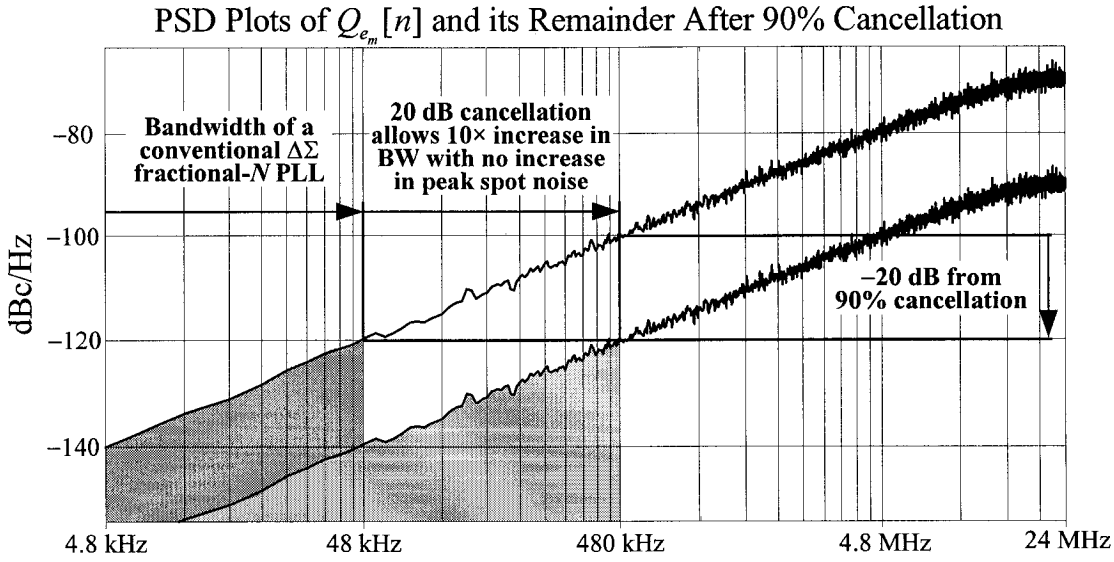

Fig. 1.3: Illustration of increase of the useable PLL bandwidth due to phase noise cancellation.

the phase noise, yet must have a sufficiently wide bandwidth to accommodate the VCO modulation. In many applications, such as the Bluetooth transmitter application used as a demonstration vehicle in this work, it is not possible to simultaneously satisfy both of these requirements using conventional techniques.

B. Phase Noise Cancellation Technique Overview

As shown in Fig. 1.1, the phase noise cancellation technique combines the two fractional-*N* PLL approaches described above. A second-order digital ΔΣ modulator generates $y[n]$ as in a conventional ΔΣ fractional-*N* PLL, *and* a DAC cancellation path attenuates $Q_{e_m}[n]$. As explained below, the combination of the two approaches in conjunction with quantization noise-shaping, mismatch noise-shaping, and 1-bit dither, greatly reduces the respective limitations suffered by each approach in isolation.

Fig. 1.3 illustrates that combining the two approaches makes it possible to

widen the PLL bandwidth relative to that of a conventional ΔΣ fractional-*N* PLL without increasing the peak spot phase noise. The top curve in the figure represents a power spectral density (PSD) plot of $Q_{e_m}[n]$ scaled by the dc value of the PLL phase transfer function between $Q_{e_m}[n]$ and the PLL output, so its units are dBc/Hz referred to the PLL output. The bottom curve represents the PSD, also in units of dBc/Hz referred to the PLL output, of the portion of $Q_{e_m}[n]$ that remains after cancellation where the DAC cancellation path has a 10% gain error but is otherwise ideal. Suppose, as an example, that the peak spot phase noise resulting from quantization noise is to be limited to –120 dBc/Hz. Without the DAC cancellation path, i.e., in the case of a conventional ΔΣ fractional-*N* PLL, it can be seen from the top curve in the figure that the bandwidth of the PLL would have to be limited to 48 kHz. In contrast, it can be seen from the bottom curve in the figure that with the DAC cancellation path the bandwidth of the PLL can be set to 480 kHz. Thus, even with a 10% gain error in the DAC cancellation path, the bandwidth of the PLL can be increased by a factor of 10 without increasing the peak spot phase noise of the PLL.

While combining the two fractional-*N* PLL approaches relaxes both the bandwidth versus phase noise tradeoff and the required gain-accuracy in the DAC cancellation path relative to the two approaches, respectively, in isolation, it does not reduce the dynamic range and linearity requirements of the DAC cancellation path. Furthermore, $Q_{e_m}[n]$ must be nearly free of spurious tones, or else high gain-accuracy would again be required in the DAC cancellation path to properly cancel the spurious tones. These problems are addressed in the implemented PLL by several means. As

described in detail below, delta-sigma re-quantization and a segmented, mismatch-shaping, current pulse DAC are used to obtain high DAC cancellation path dynamic range and linearity, and 1-bit dithering is used to eliminate spurious tones.

C. Phase Noise Cancellation Technique Signal Processing Details

As shown in Fig. 1.1, the architecture consists of a 48 MHz crystal reference source, the PLL core described above, a 48 MHz digital section, a bank of 16 coarse 1-bit current pulse DACs, and a bank of 16 fine 1-bit current pulse DACs. The 48 MHz digital section consists of digital logic in which all registers are clocked on the rising edges of the divider output. It generates $y[n]$ and 32 1-bit sequences that control the two banks of 1-bit current pulse DACs. During each reference period, each 1-bit current pulse DAC generates a positive or negative pulse of current depending upon whether its input bit is high or low. Each pulse has a duration of 4 VCO periods. The nominal magnitudes of the current pulses are $I_{CP}/16$ and $I_{CP}/128$ for the coarse and fine 1-bit current pulse DACs, respectively.

The input to the second-order $\Delta\Sigma$ modulator, $x[n]$, is a 16 bit two's complement number in the range –1 to 1 of the form $x[n] = \alpha + x_m[n] + d[n]$, where $\alpha = (k - 46)/48$ selects the desired Bluetooth channel frequency for $k = 0, 1, ..., 78$, $x_m[n]$ is optional FSK or GFSK modulation, and $d[n]$ is a 1-bit pseudo-random dither sequence. The dither sequence is generated by an on-chip length-22 linear feedback shift register and is scaled such that it represents the least significant bit (LSB) of $x[n]$. The details of the second-order $\Delta\Sigma$ modulator are shown in Fig. 1.4(a). It has unity gain and a quantization step-size of unity, so its output has the form $y[n] = x[n - 2] + e_m[n]$, and

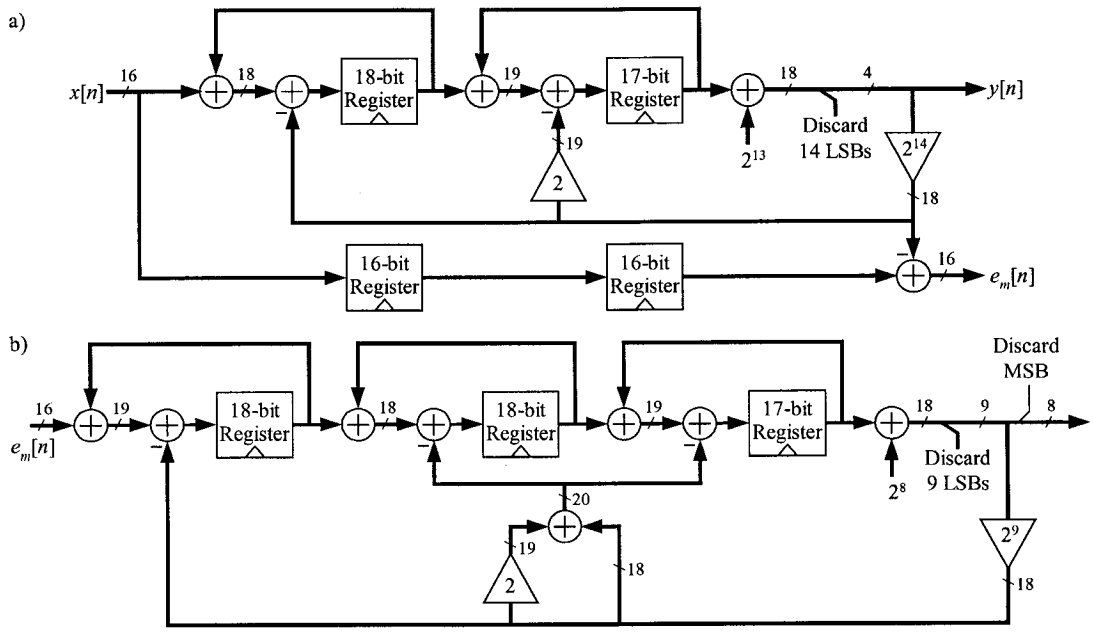

Fig. 1.4: Details of employed digital ΔΣ modulators – (a) The second-order ΔΣ modulator. (b) The third-order ΔΣ modulator.

takes on values in the range: –2, –1, 0, 1, and 2. As proven in [17] and [18], the dither sequence completely eliminates spurious tones in $e_m[n]$, so $e_m[n]$ has the same PSD as white noise passed through a discrete-time filter with two zeros at dc. The discrete-time integration in (1) cancels one of the zeros, so $Q_{e_m}[n]$ has the first-order shaped PSD represented by the top curve in Fig. 1.3. Although the dither behaves as white noise, its magnitude is sufficiently small that its contribution to the PLL phase noise is negligible in the band of interest.

Ideally, the DAC cancellation path would digitally integrate $e_m[n]$ to obtain $Q_{e_m}[n]$ as in (1), and, for each n, inject a current pulse into the loop filter with a width equal to that of the corresponding current pulse from the charge pump and an amplitude chosen such that the total charge carried by the pulse is precisely $Q_{e_m}[n]$.

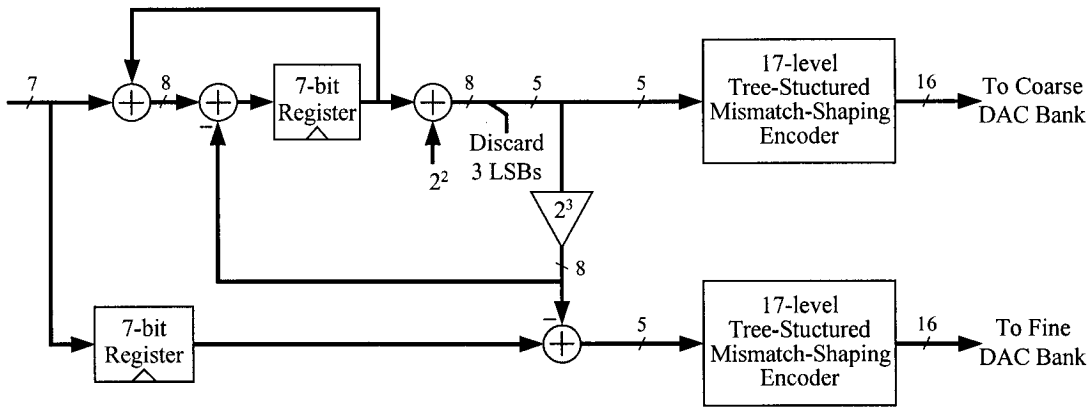

Fig. 1.5: Details of the mismatch-shaping digital encoder.

Unfortunately, this is difficult to accomplish in practice because the precise width of the charge pump pulse is not known a priori, and the pulse can be very narrow. Instead, a fixed-width current pulse can be used. In this case $Q_{e_m}[n]$ is not cancelled immediately as it is added, so the cancellation process introduces a voltage transient each period at the VCO input. Most of the power associated with the voltage transient is outside of the PLL bandwidth so its contribution to the PLL phase noise tends to be small. In most conventional PLLs with DAC cancellation paths, the pulse width is equal to the reference period [19]. However, in the current work the pulse width is set to four VCO periods to better match the charge pump pulse width thereby reducing the transient at the VCO and decreasing the resulting PLL phase noise contribution.

If $Q_{e_m}[n]$ were calculated directly using $e_m[n]$ in (1), a 15-bit current DAC with a step size of $0.5 \cdot I_{CP} \cdot 2^{-15}$, e.g., 19.5 nA for the implemented PLL, would be required to generate the necessary current pulses. Such a DAC would be very difficult to implement. Instead, as indicated in Fig. 1.1, $e_m[n]$ is re-quantized from 16 bits to 8 bits by a third-order digital $\Delta\Sigma$ modulator, the details of which are shown in Fig.

1.4(b), and the result is digitally integrated and converted to current pulses. The output of the integrator is a 7-bit sequence proportional to $Q_{e_m}[n-1]+e_{rq}[n]$, where $e_{rq}[n]$ is second-order shaped re-quantization noise resulting from digitally integrating the re-quantization noise from the third-order digital ΔΣ modulator. Because of its second-order high-pass shape and small magnitude, $e_{rq}[n]$ does not result in a significant increase in the PLL phase noise. Thus, re-quantization reduces the problem of designing a 15-bit DAC with a minimum step-size of 19.5 nA to that of designing a 7-bit DAC with a minimum step-size of 10 μA. The DAC is implemented by the two banks of 1-bit current pulse DACs. During the nth reference period the input bits to the 1-bit DACs are chosen such that

$$v[n]=\left[8\cdot\sum_{k=1}^{16}\left(v_{c_k}[n]-\tfrac{1}{2}\right)+\sum_{k=1}^{16}\left(v_{f_k}[n]-\tfrac{1}{2}\right)\right]\Delta_v \qquad (2)$$

where $v[n]$ is the output of the digital integrator, $v_{c_k}[n]$ and $v_{f_k}[n]$ are 0 or 1 input values to the kth 1-bit DACs in the coarse and fine DAC banks, respectively, and Δ_v is the LSB weight of $v[n]$.

For most values of $v[n]$, there are several combinations of $v_{c_k}[n]$ and $v_{f_k}[n]$ that satisfy (2). For example, when $v[n] = -63\Delta_v$, any one of the 16 1-bit DAC inputs in each DAC bank can be set to 1 with the rest set to 0. To the extent that the 1-bit DACs in each DAC bank are perfectly matched and the ratio between coarse and fine 1-bit DACs is exactly eight, it does not matter which of the possible input selections is made. In conventional segmented DACs good matching is assumed, so for each value of $v[n]$ only one of the combinations of $v_{c_k}[n]$ and $v_{f_k}[n]$ that satisfy (2) is ever used.

Unfortunately, if the conventional approach had been used in this work, even mismatches of less than 1% among the unit current sources that make up the 1-bit DACs would give rise to harmonic distortion severe enough to prevent the PLL from meeting the target specifications, and reducing the mismatches to much less than 1% in present CMOS technology can be difficult. To circumvent this problem, a segmented mismatch-shaping DAC encoder is used prior to the banks of 1-bit DACs [20, 21, 22].

During the nth reference period, the encoder selects one of the combinations of $v_{c_k}[n]$ and $v_{f_k}[n]$ that satisfy (2) as a function of $v[n]$ such that the error from mismatches introduced by the DAC, referred to as *mismatch-noise*, has first-order highpass spectral shaping with no spurious tones. Consequently, much of the mismatch-noise power is outside the PLL bandwidth. For the implemented PLL, simulations indicate that the target specifications can be met provided the matching of the unit current sources has a standard deviation of no more than 5% which is not difficult to achieve in practice. As shown in Fig. 1.5, the encoder consists of a first-order digital $\Delta\Sigma$ modulator and two 17-level tree-structured mismatch-shaping encoders of the type presented in [23]. The $\Delta\Sigma$ modulator quantizes $v[n]$ to a 17-level sequence which drives the 17-level mismatch-shaping encoder associated with the coarse DAC bank. The quantization noise from the $\Delta\Sigma$ modulator drives the 17-level mismatch-shaping encoder associated with the fine DAC bank.

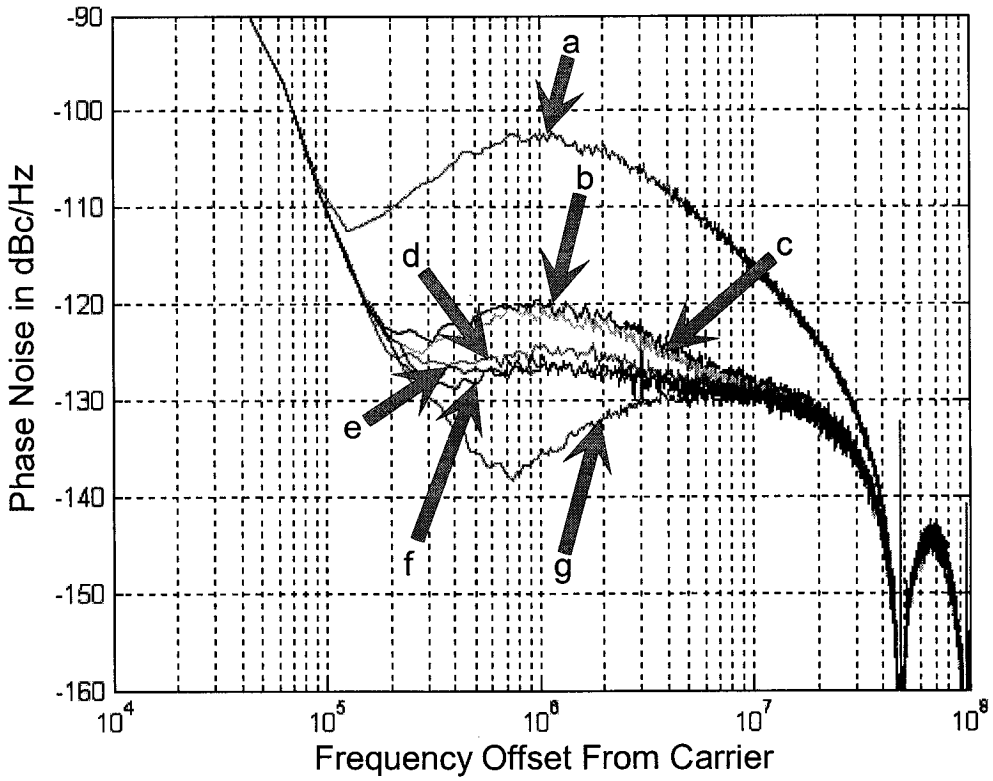

Fig. 1.6: Simulated output phase noise PSD plots of the implemented PLL (a) without the phase noise cancellation technique, (b)-(f) with the phase noise cancellation technique, 5% unit current source errors in the 1-bit DACs, and 12%, 10%, 4%, 2%, and 0% gain mismatches, respectively, and (g) with ideal phase noise cancellation.

Fig. 1.6 shows simulated output phase noise PSD plots corresponding to quantization noise and mismatch-noise for the implemented PLL with various DAC cancellation path gain error levels. The results were generated by an event-driven simulator that accurately models both the discrete-time and continuous-time portions of the system. The unit current source values in the 1-bit current pulse DACs were chosen with random errors such that they have a 5% standard deviation from their nominal value. As indicated in the figure, even with a 12% DAC cancellation path gain error and the relatively poor current source matching (curve "b" in the figure), the

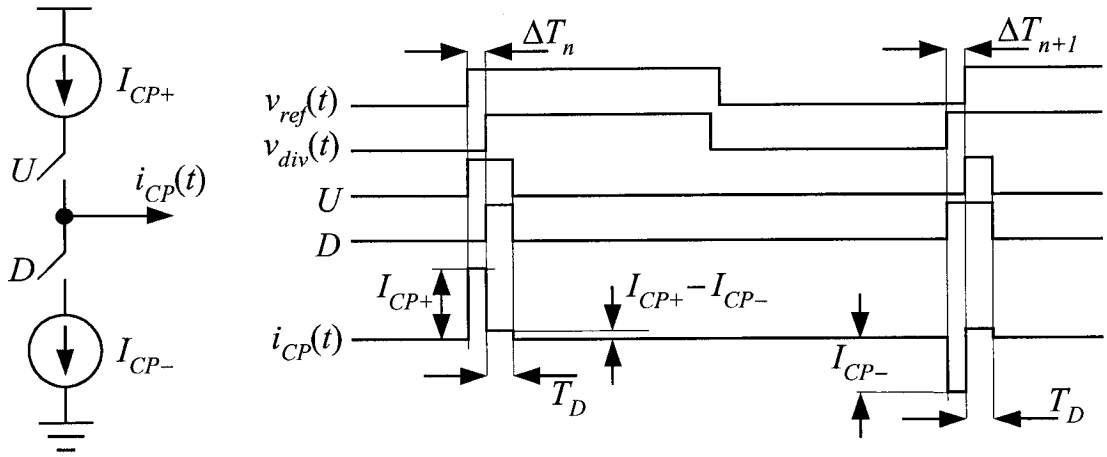

Fig. 1.7: A conventional charge pump and the associated timing diagram.

phase noise cancellation technique reduces the peak spot phase noise by 20 dB, and the spot phase noise at a 3 MHz offset from the carrier is well below the –120 dBc/Hz value required by the Bluetooth specification.

III. THE CHARGE PUMP LINEARIZATION TECHNIQUE

A. The Problem

A conventional charge pump and the associated timing diagram are shown in Fig. 1.7. The rising edges of the PFD outputs, U and D, are triggered by those of $v_{ref}(t)$ and $v_{div}(t)$, respectively. The falling edges of U and D both occur after a delay of T_D following the later of the rising edges of $v_{ref}(t)$ and $v_{div}(t)$. The delay ensures that each current source in the charge pump is turned on for a minimum duration of T_D every reference period to solve the charge pump dead-zone problem [24].

The positive and negative current sources in the charge pump are on when U and D, respectively, are high and are off otherwise. Therefore, neglecting a constant,

the charge carried by $i_{CP}(t)$ during the nth reference period is

$$Q_{CP}[n] = \Delta T_n I_{CP} + \frac{1}{2}|\Delta T_n|\Delta I_{CP}, \quad (3)$$

where ΔT_n is the time difference between the nth rising edges of $v_{div}(t)$ and $v_{ref}(t)$, and I_{CP} and ΔI_{CP} are the average of and difference between the positive and negative current source values, i.e., I_{CP+} and I_{CP-}, respectively. Ideally, I_{CP+} and I_{CP-} are equal, but in practice they differ because of component mismatches and the different voltages across the respective current source transistors. The result is the second term in (3) which is non-linear with respect to ΔT_n.

Unfortunately, the non-linearity induces spurious tones at multiples of αf_{ref} in the PLL phase noise. The problem becomes increasingly severe as the bandwidth of the PLL is increased, because spurious tones that are well out of band and, thus, highly attenuated in a narrowband PLL are less out of band, and, thus, less attenuated in a wideband PLL. A conventional solution is to use analog feedback to equalize I_{CP+} and I_{CP-} [8]. However, in a wideband PLL the charge pump output voltage variations tend to be very abrupt which makes the design of an effective analog compensation circuit difficult.

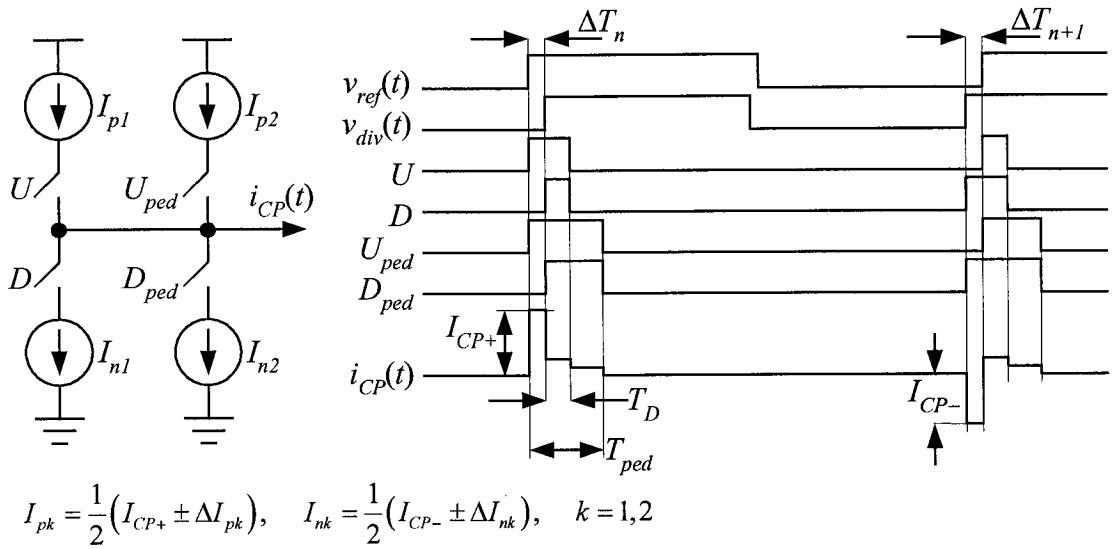

Fig. 1.8: The modified charge pump and the associated timing diagram.

B. The Proposed Technique

The charge pump linearization technique involves modifications to both the PFD and the charge pump. The modified PFD generates U and D signals as in the conventional case, but also generates two new signals, U_{ped} and D_{ped}. As shown in Fig. 1.8, each reference period the rising edges of U_{ped} and D_{ped} are aligned with those of U and D, respectively, but their falling edges both occur after a delay of T_{ped} following the earlier of the rising edges of $v_{ref}(t)$ and $v_{div}(t)$. The charge pump is modified in that the I_{CP+} and I_{CP-} current sources are each split into two nominally identical half-sized current sources. The two halves corresponding to I_{CP+} are switched by U and U_{ped}, and the two halves corresponding to I_{CP-} are switched by D and D_{ped}. The duration, T_{ped}, is referred to as the *pedestal time* and is designed to be longer than the maximum value of $\Delta T_n + T_D$ when the PLL is locked. This maximum value is three VCO periods plus T_D for the implemented PLL, so T_{ped} can be made sufficiently small that its effect on

the noise introduced by the charge pump is negligible.

It can be shown that, neglecting a constant,

$$Q_{CP}[n] = \Delta T_n I_{CP} + \frac{1}{2}|\Delta T_n|\left(\Delta I_p - \Delta I_n\right), \tag{4}$$

where ΔI_p and ΔI_n are differences, arising from component mismatches, between the values of the two positive and the two negative current source halves, respectively. As in the conventional charge pump, the differences give rise to a non-linear term in $Q_{CP}[n]$. However, in contrast to the conventional case, the non-linear term is a result of mismatches between like current courses with identical voltages across their respective transistors. Therefore, the non-linearity introduced by the proposed technique is much less than that introduced by a conventional charge pump and PFD. Although it was not necessary in this project, the non-linearity can be further suppressed by randomly interchanging the signals U and U_{ped}, and D and D_{ped} using a pseudo-random bit sequence.

Table 1.1: Simulated phase noise contributions of the various circuit blocks and the relevant PLL parameters.

Fractional-*N* phase locked loop parameters	
Reference frequency	48 MHz
Loop bandwidth	460 kHz
Zero location	92 kHz
Pole location	2.3 MHz
Charge pump current	1.28 mA
Minimum PFD pulse duration, T_D	0.5 ns – 1 ns
Pedestal duration, T_{ped}	3 ns – 6 ns
Simulated worst case phase contributions at 3 MHz offset:	
Voltage controlled oscillator and buffers	-127 dBc/Hz
Modified phase frequency detector	-132 dBc/Hz
Charge pump and DACs	-134 dBc/Hz
Crystal reference oscillator	-134 dBc/Hz
1.8 Volt – 2.7 Volt converters	-139 dBc/Hz
Loop filter resistor	-147 dBc/Hz
Multi-modulus frequency divider	-153 dBc/Hz

IV. CIRCUIT ISSUES

Overview

The circuit is implemented in the TSMC 0.18 μm 1P6M mixed-signal CMOS process with the thin top-metal option, and installed in a 5mm TQFP 32 pin package. All pads include ESD protection circuitry. The PFD, charge pump, DAC banks, and VCO are designed for a 2.7 V supply. The remaining components are designed for a 1.8 V supply. All the blocks shown in Fig. 1.1 except the crystal and the loop filter capacitors and resistor are implemented on-chip. A VCO output buffer, a VCO divider buffer, a 1.8 V to 2.7 V logic converter block, and a three-wire digital interface are also included on the chip. Separate deep N-wells under the digital logic and critical

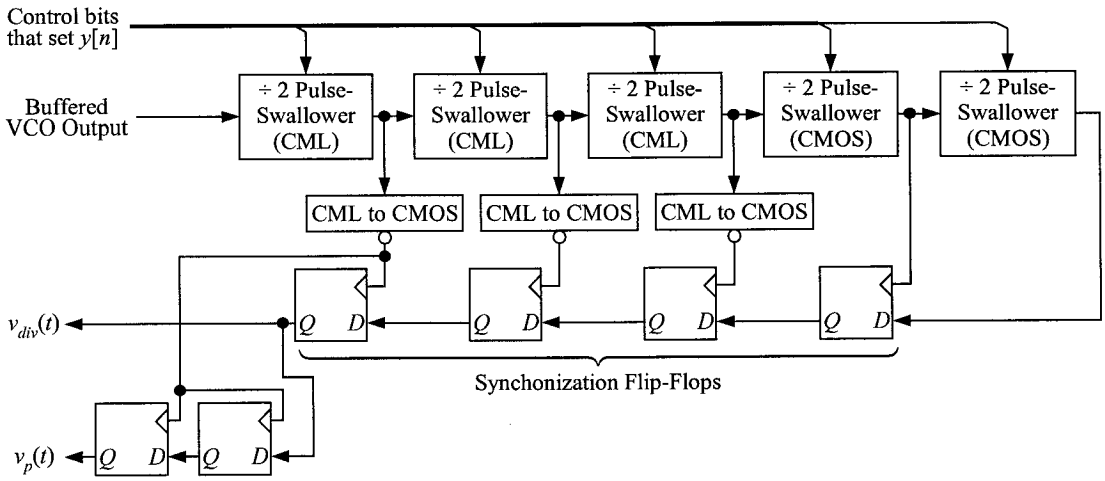

Fig. 1.9: The frequency divider circuit.

analog circuitry, and separate supply domains help prevent digital interference from disturbing analog circuit behavior. A summary of the designed loop parameters and simulated phase noise contributions of the various circuits are shown in Table 1.1.

Frequency Divider

As shown in Fig. 1.9, the core of the divider consists of five divide-by-two, pulse swallowing blocks [25]. The three highest frequency pulse swallowing blocks consist of current-mode-logic (CML), and the other two blocks consist of static CMOS logic. The four synchronization flip-flops ensure that the rising edges of $v_{div}(t)$ are aligned to the appropriate rising edges of the first pulse-swallowing block. Two additional flip-flops are used to derive a DAC pulse termination signal that goes high 4 VCO periods after each rising edge of $v_{div}(t)$.

The reason for synchronizing the rising edges of $v_{div}(t)$ to edges of the first pulse-swallowing block is to reduce *modulus-dependent delay mismatches*, i.e., systematic timing errors in $v_{div}(t)$ that depend upon $y[n]$. Such errors have an effect

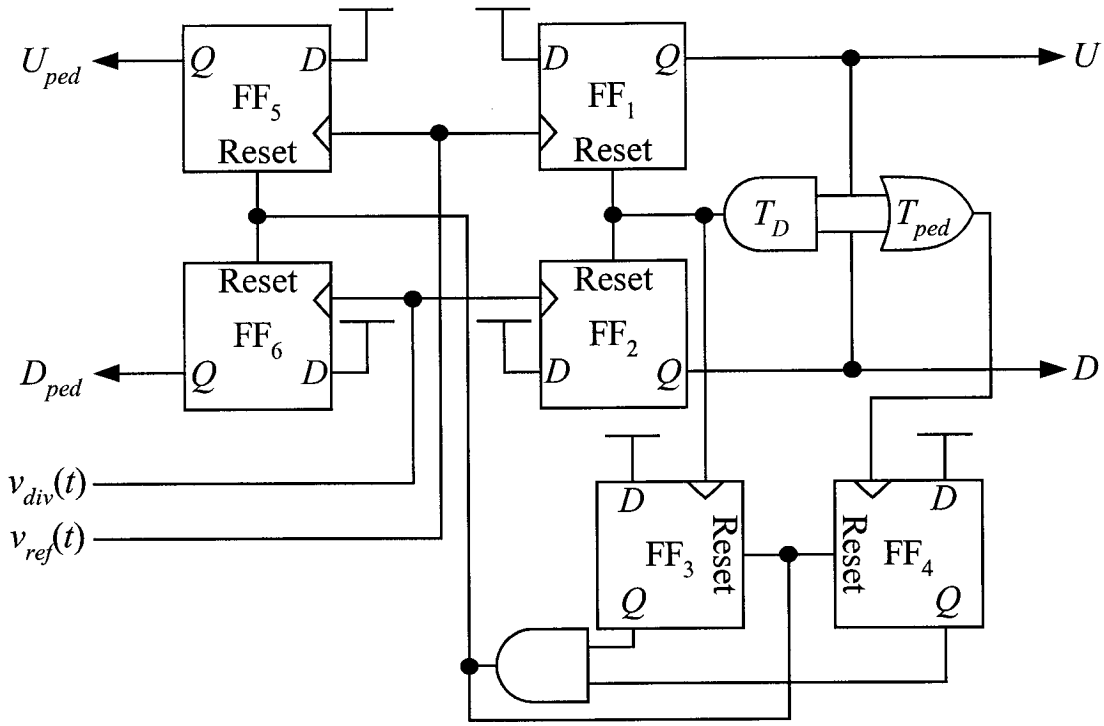

Fig. 1.10: The modified PFD circuit.

similar to charge pump non-linearity in that they induce spurious tones in the PLL phase noise at multiples of αf_{ref}. Simulations of the implemented PLL indicate that modulus-dependent delay mismatches must be restricted to less than 1% of the VCO period, i.e., to less than 4 ps, to suppress the spurious tones to less than –60 dBc. This is achieved by the synchronization flip-flops which successively align the edges of the signal from the last pulse swallowing block to those of the previous pulse swallowing blocks. In principle, only the final flip-flop is necessary, but the other three are included to avoid race conditions.

PFD, Charge Pump, and 1-bit Current Pulse DACs

The PFD is shown in Fig. 1.10. Flip-flops 1 and 2 and the associated *AND* gate generate the U and D signals as in a conventional PFD, and the remaining circuitry

generates the U_{ped} and D_{ped} signals. The circuit is configured such that the rising edges of U_{ped} coincide with those of U, and the rising edges of D_{ped} coincide with those of D. The *AND* gate and *OR* gate driven by U and D have built-in delays of T_D and T_{ped}, respectively. Therefore, during the nth reference period, flip-flops 1 and 2 are reset after a delay of $\Delta T_n + T_D$ following the earlier of the times at which U and D go high, whereas flip-flops 5 and 6 are reset after a delay of the maximum of $\Delta T_n + T_D$ and T_{ped} following the earlier of the times at which U and D go high.

As described in the previous section, T_{ped} is chosen to be longer than the maximum value of $\Delta T_n + T_D$ expected to occur when the PLL is locked, in which case the PFD output signals are as illustrated in Fig. 1.8. When the PLL is in the process of acquiring lock, $\Delta T_n + T_D$ is usually longer than T_{ped}. In this case U coincides with U_{ped} and D coincides with D_{ped}, so the current from the charge pump is the same as in the conventional case. Therefore, the charge pump linearization technique does not affect the behavior of the PLL during acquisition.

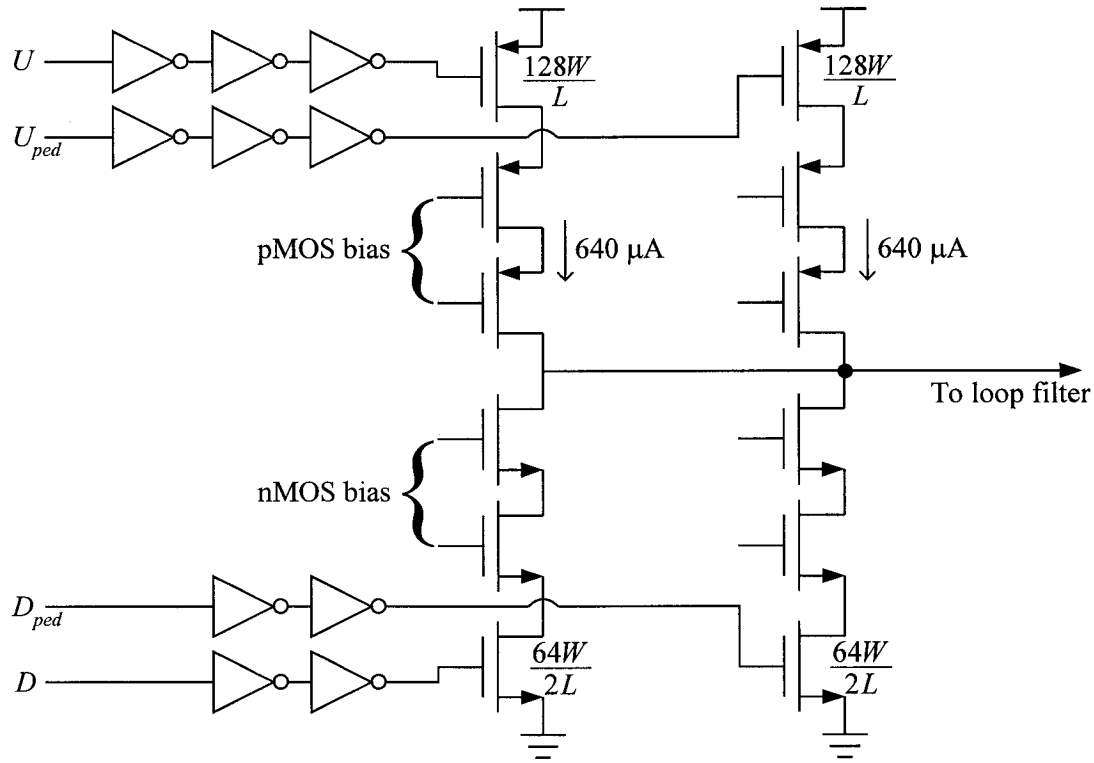

Fig. 1.11: The modified charge pump circuit.

As shown in Fig. 1.11, and explained in the previous section, the charge pump consists of two halves, one controlled by U and D, and the other controlled by U_{ped} and D_{ped}. Each half consists of positive and negative 640 μA cascode current sources with triode MOS switches near the supply rails [26]. The pMOS transistors that make up the switches and cascode current sources have twice the width and half the length of the corresponding nMOS transistors so as to approximately match the loading on the PFD output lines and the switching speeds of the positive and negative current sources. The chains of inverters are scaled to have a common propagation delay so the inverted copies of U and U_{ped} presented to the pMOS switches are properly aligned with the non-inverted copies of D and D_{ped} presented to the nMOS switches [24].

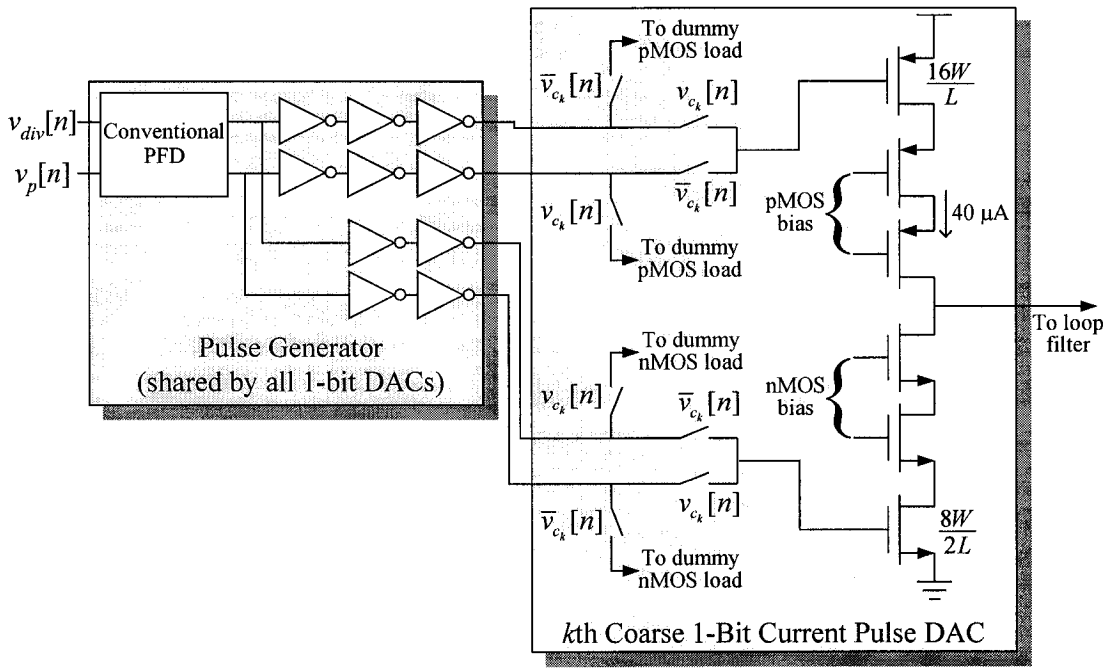

Fig. 1.12: The DAC pulse generator and the *k*th coarse 1-bit current pulse DAC circuits.

Figure 1.12 shows a simplified circuit diagram of the *k*th coarse 1-bit current pulse DAC and the pulse generator shared by all the 1-bit current pulse DACs. The switched current sources in the coarse and fine 1-bit current pulse DACs, respectively, are 40 μA and 5 μA scaled down versions of the those in the charge pump. The pulse generator contains a copy of the conventional portion of the PFD described above and four chains of scaled inverters similar to those that drive the charge pump switches. The PFD is driven by the two divider output signals, so each reference period its top output goes high for a duration of 4 VCO periods plus T_D, and its bottom output goes high for a duration of only T_D. Inverted and non-inverted copies of these signals are presented to each 1-bit current pulse DAC to drive the pMOS and nMOS switches, respectively. In each case, the 1-bit DAC input causes one of these signals to be presented to the MOS switch and the other to be presented to a dummy MOS switch.

The purpose of the dummy MOS switches is to maintain data-invariant loads on the pulse generator output lines.

Voltage Controlled Oscillator

The on-chip VCO is a negative-g_m CMOS *LC* oscillator designed to have a center frequency of 2.448 GHz. It incorporates a differential inductor implemented as a square spiral of metal layers 5 and 6 sandwiched together. A MOS varactor provides 200 MHz/V tuning over a 1 V range. The differential VCO outputs are ac-coupled to two resistively loaded differential source-coupled buffers: one to drive the divider and one to drive 50 Ω loads off the chip. A configuration option allows for the use of an off-chip VCO in place of the on-chip VCO; the on-chip VCO can be disabled and a direct connection is provided from a pin to the input of the divider buffer.

Loop Filter

The loop filter components are: R = 641 Ω, C_1 = 100 pF, and C_2 = 2.4 nF, of which R, C_2, and 60 pF of C_1 are off-chip. The remaining 40 pF of C_1 is on-chip to help reduce the voltage variations caused by fast charge pump current switching through the inductive bond wires. Given that the divider modulus is approximately 51, the VCO gain is 200 MHz/V, and the nominal charge pump current magnitude is 2×640 μA, these component values give rise to a PLL bandwidth of approximately 460 kHz.

48 MHz Digital Logic

The 48 MHz digital logic was implemented using a standard cell library

available to the authors in which the transistors have a minimum gate length of 0.25 μm. While a more compact and lower power design would have been possible with a standard cell library optimized for the 0.18 μm process, the project schedule did not permit such optimization.

V. MEASUREMENT RESULTS

Three copies of the IC were tested on separate circuit boards. The performance of each part was verified for all 79 Bluetooth channels with the phase noise cancellation and charge pump linearization techniques individually and simultaneously enabled and disabled with and without FSK modulation. On each Bluetooth channel and each part, the phase noise cancellation technique was found to reduce the spot phase noise by 16 dB or better, and the charge pump linearization technique was found to reduce the spurious tone floor by 8 dB or better. With both techniques enabled, each part was found to achieve a worst-case phase noise of –121 dBc/Hz at 3 MHz offsets, a worst-case spurious tone level of –54 dBc, and a worst-case in-band noise floor of –96 dBc/Hz. The measured results are summarized in Table 1.2, and a die photograph is shown in Fig. 1.13.

Table 1.2: Performance summary.

Design Details:			
Technology	TSMC 0.18μm 1P6M CMOS		
Package	5 × 5 mm^2, 32 pin TQFP		
Die Area	2.72 × 2.47 mm^2 (includes pads and ESD protection)		
Frequency Range	2402 + k MHz, $k = 0, 1, 2, \ldots, 78$		
Crystal Reference	48 MHz		
Loop Bandwidth	460 kHz		
Measured Current Consumption, V_{DD}, and Area by Block:			
VCO	3.0 mA @ 1.9 V	0.4 mm^2	67 mW
PFD, CP & DAC	10.3 mA @ 1.9 V	0.04 mm^2	
Divider	6.7 mA @ 1.8 V	0.4 mm^2	
Digital Logic	8.8 mA @ 2.2 V	0.68 mm^2	
Internal VCO Buffer	5.6 mA @ 1.8 V	0.013 mm^2	
Crystal Oscillator Buffer	4.1 mA @ 2.2 V	0.04 mm^2	21 mW
External VCO Buffer	6.9 mA @ 1.8 V	0.015 mm^2	
Measured Worst Case Performance With Both Techniques *Enabled*:	On-chip VCO:	Off-chip VCO:	
Phase Noise @ 50 kHz	–96 dBc/Hz	–96 dBc/Hz	
Phase Noise @ 3 MHz	–121 dBc/Hz	–127 dBc/Hz	
Largest Fractional Spur @ < 2 MHz	–54 dBc @ 1 MHz	–45 dBc @ 1 MHz	
Largest Fractional Spur @ ≥ 2 MHz	–56 dBc @ 2 MHz	–58 dBc @ 2 MHz	
Largest Fractional Spur @ ≥ 3 MHz	–57 dBc @ 3 MHz	–62 dBc @ 3 MHz	
Reference Spur	–65 dBc	–66 dBc	
Measured Worst Case Performance With Both Techniques *Disabled*:	On-chip VCO:	Off-chip VCO:	
Phase Noise @ 50 kHz	–96 dBc/Hz	–96 dBc/Hz	
Phase Noise @ 3 MHz	–107 dBc/Hz	-107 dBc/Hz	
Largest Fractional Spur @ < 2 MHz	–40 dBc @ 1 MHz	–35 dBc @ 1 MHz	
Largest Fractional Spur @ ≥ 2 MHz	–44 dBc @ 2 MHz	–46 dBc @ 2 MHz	
Largest Fractional Spur @ ≥ 3 MHz	–49 dBc @ 3 MHz	–52 dBc @ 3 MHz	
Measured Performance With Both Techniques *Disabled*:	On-chip VCO:	Off-chip VCO:	
Effect of charge pump linearization technique on phase noise	None observed	None observed	
Effect of ΔΣ noise cancellation technique on fractional spur level	None observed	2-9 dB spur reduction	

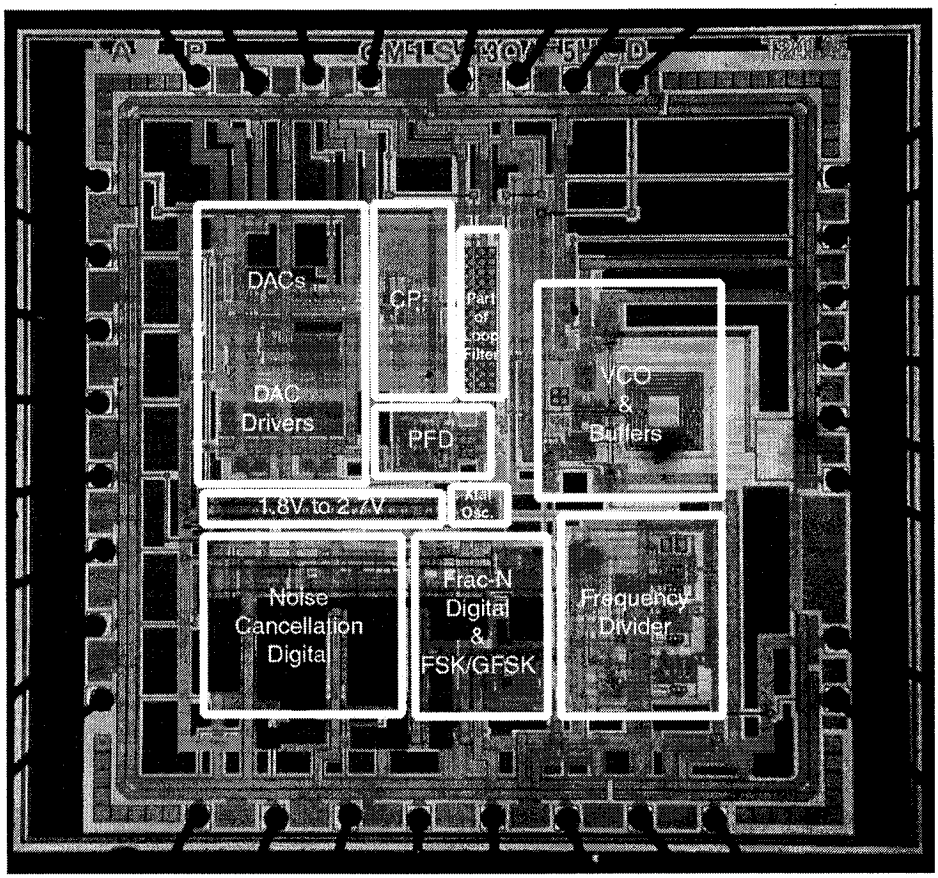

Fig. 1.13: Die photograph.

Figs. 1.14 and 1.15 show representative PSD plots measured with the PLL set to the 2.431 GHz Bluetooth channel and the phase noise cancellation and charge pump linearization techniques both enabled and both disabled. Fig. 1.14 shows PSD plots of the PLL output signal and phase noise with the PLL operating without modulation. Fig. 1.15 shows PSD plots of the PLL output signal for the PLL operating with 1 Mb/s FSK modulation. In both cases the phase noise improvement resulting from the techniques is evident. Similar results are seen for each part on every Bluetooth channel.

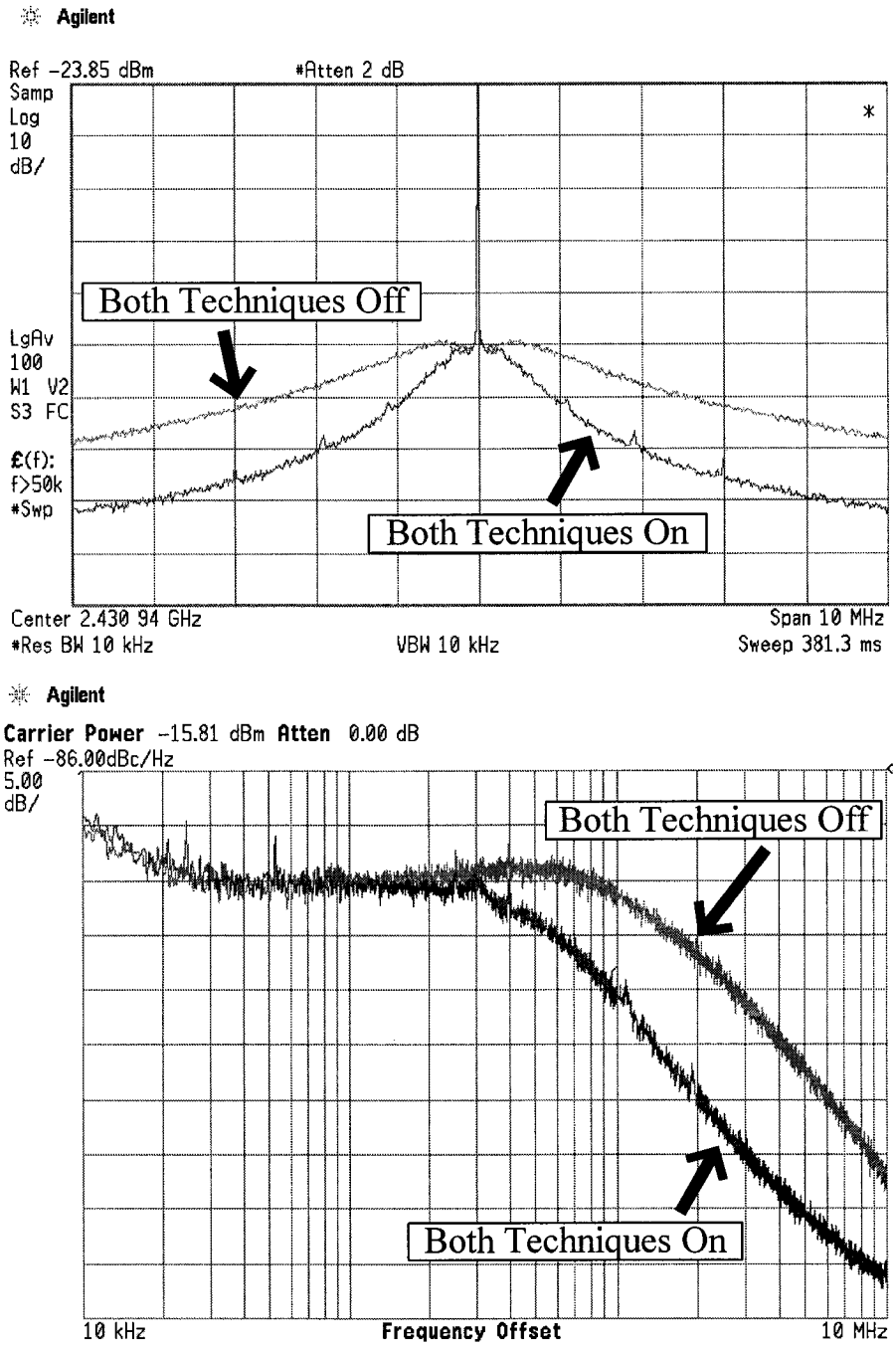

Fig. 1.14: Measured PSD plots of the output signal and phase noise of the PLL tuned to 2.431 GHz without modulation.

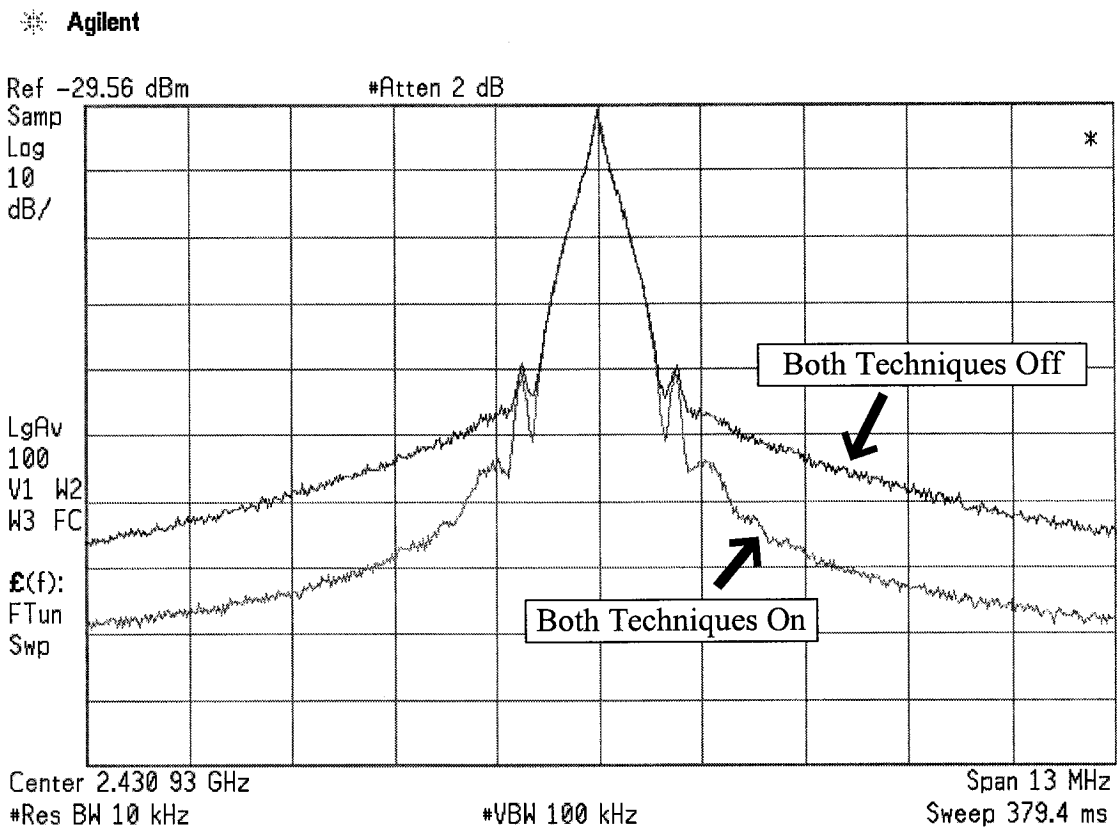

Fig. 1.15: Measured PSD plot of the output signal of the PLL tuned to 2.431 GHz with 1 Mb/s FSK modulation.

Fig. 1.16 shows an eye pattern from the PLL with 1 Mb/s FSK transmit modulation and both techniques enabled measured by down-converting the PLL output signal to an intermediate frequency through a spectrum analyzer and frequency demodulating the result using a vector analyzer. The minimum frequency deviation is approximately 120 kHz and the zero-crossing error is less than one-eighth of the symbol period as required by the application. Again, almost identical results were observed for each part on every Bluetooth channel.

The spurious tone reduction achieved by the charge pump linearization technique is most easily observed when the PLL is tuned to Bluetooth channels that are

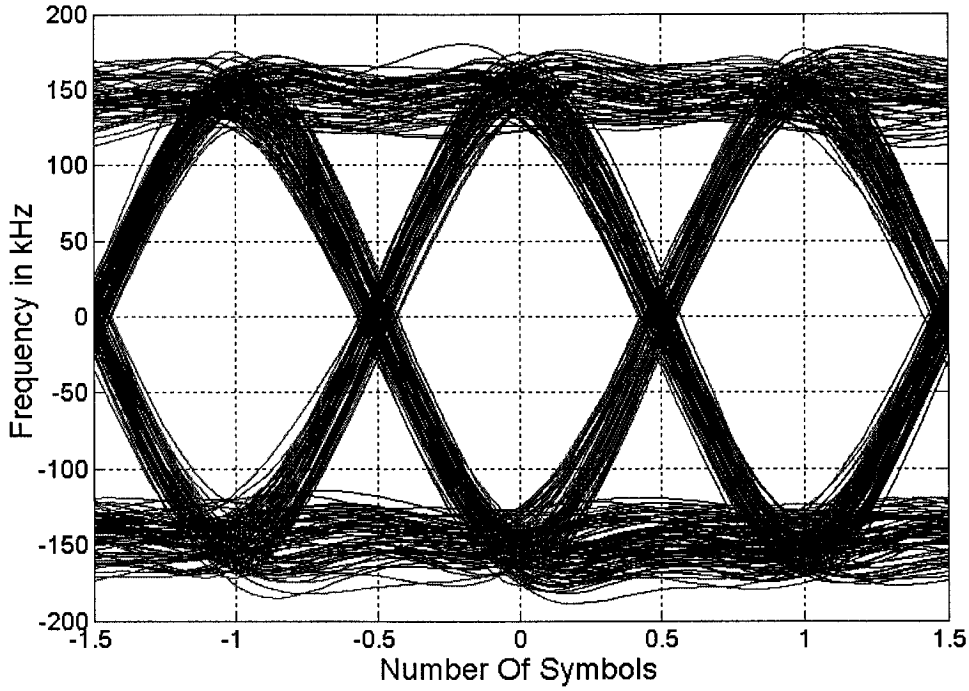

Fig. 1.16: Measured eye pattern corresponding to the output signal shown in Fig. 1.17.

close to integer multiples of the 48 MHz reference frequency. In such cases α is small, so the spurious tones in the PLL phase noise resulting from non-linearity, which occur at multiples of αf_{ref}, are not highly attenuated by the lowpass transfer function of the PLL. Fig. 1.17 shows PSD plots of the PLL output signal with and without the charge pump linearization technique enabled for such a case, i.e., for the VCO tuned to 2.453 GHz so that $\alpha f_{ref} = 5$ MHz. The overlaid plots are intentionally displaced in frequency to make the spurious tone reduction visible.

As mentioned in the previous section, the VCO and charge pump were designed to operate from a 2.7 V supply with a VCO center frequency of 2.448 GHz, but the measured VCO center frequency turned out to be 2.25 GHz. To force the VCO

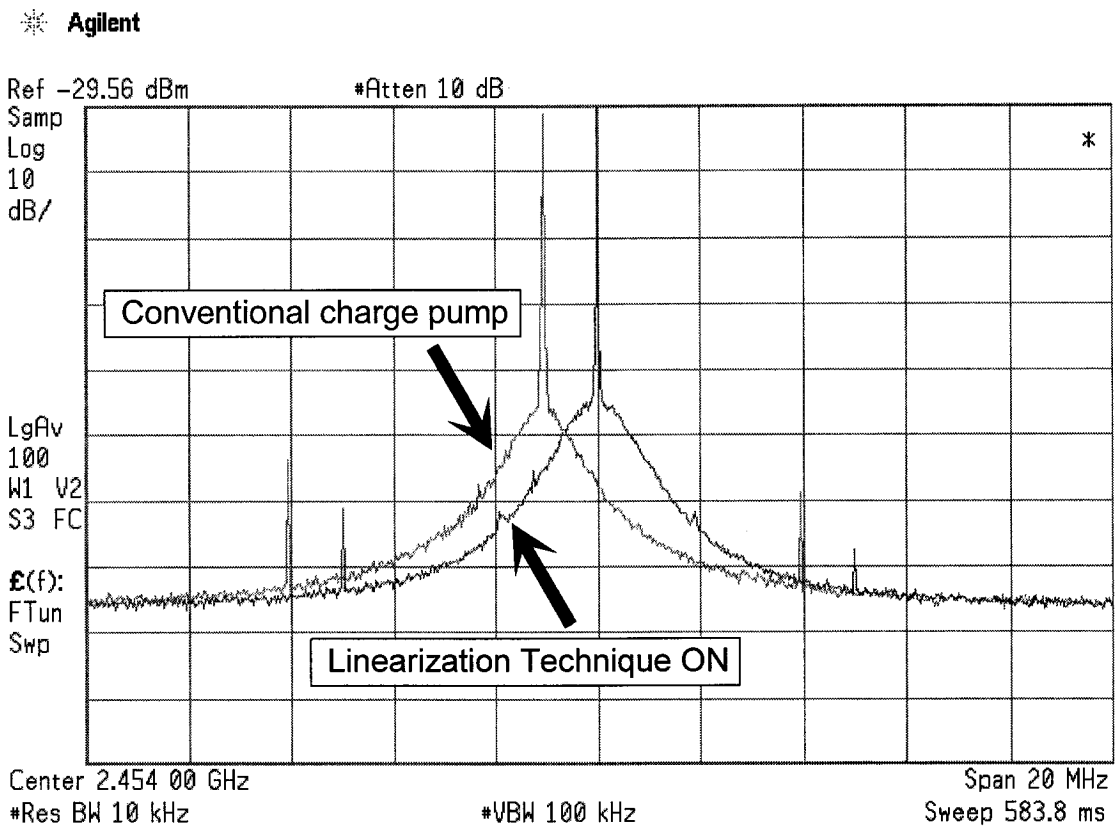

Fig. 1.17: Measured PSD plots of the PLL tuned to 2.453 GHz with the charge pump linearization technique enabled and disabled.

into the Bluetooth frequency range the VCO and charge pump, which share the same power supply lines, had to be run from a 1.9 V supply during testing. It is likely that this increased the phase noise by at least 3 dB and increased distortion because several critical transistors were forced into their triode regions. Nevertheless, as described above and summarized in Table 1.2 the IC performed well.

Each of the tested parts met the Bluetooth phase noise and eye pattern specifications on all channels. They also met the Bluetooth spurious tone specifications except for a small number of channels on which the spurious tones were at most 3 dB above the specification. The slightly elevated spurious tone level is a

result of having to run the VCO and charge pump from a 1.9 V supply instead of the 2.7 V supply for which it was designed. In support of this assertion, the PLL configured with an off-chip VCO and the charge pump operating from a 2.7 V supply was found to meet all required specifications on all channels (see Table 1.2).

The circuitry was designed conservatively to help ensure first-silicon success and clearly demonstrate the phase noise cancellation and charge pump linearization techniques. In particular, as tabulated in Table 1.1, large noise margins were used in the designing the circuits to ensure that the phase noise below 5 MHz would be dominated by residual $\Delta\Sigma$ quantization noise and spurious tones resulting from non-linearities. Consequently, the measured in-band phase noise is much lower than required to meet the Bluetooth specifications. While this design strategy has served the purpose of demonstrating the phase noise cancellation and charge pump linearization techniques, the current consumption of the PLL could be reduced significantly by optimizing the analog circuitry so that its in-band noise contribution is closer to the Bluetooth specification.

VI. CONCLUSION

A phase noise cancellation technique and a charge pump linearization technique have been proposed and demonstrated as enabling components in a wideband CMOS $\Delta\Sigma$ fractional-N PLL configured as a Bluetooth wireless LAN transmitter. The phase noise cancellation technique relaxes the fundamental tradeoff between phase noise and bandwidth in conventional $\Delta\Sigma$ fractional-N PLLs and does not require tight component matching or calibration. Theoretical and experimental

results have been presented that indicate the technique enables a ten-fold increase in PLL bandwidth without an increase in spot phase noise. The charge pump linearization technique provides a simple means of improving the spurious performance of wideband fractional-*N* PLLs that avoids the bandwidth limitations of previously presented techniques involving analog feedback circuits.

CHAPTER ACKNOWLEDGEMENTS

The text of Chapter 1 is set to appear as a regular paper in the *IEEE Journal of Solid State Circuits*. The dissertation author was the primary researcher. Ian Galton supervised the research which forms the basis of the chapter. Lars Jansson assisted in the partial design of the integrated circuit. The author is grateful to Eric Fogleman, Kishore Seendripu, Eric Siragusa, Andrea Spandonis, Ashok Swaminathan, Kevin Wang, Jared Welz, and Sheng Ye for their assistance with and advice regarding this project.

REFERENCES

1. F. L. Martin, et. al., "A wideband 1.3 GHz PLL for transmit remodulation suppression," *IEEE International Solid-State Circuits Conference*, Digest of Technical Papers, vol. 44, pp. 164-165, Feb. 2001.
2. G. Chang, et. al. "A direct-conversion single-chip radio-modem for Bluetooth," *IEEE International Solid-State Circuits Conference*, Digest of Technical Papers, vol. 45, Feb. 2002.
3. M. H. Perrott, T. L. Tewksbury III, C. G. Sodini, "A 27-mW CMOS fractional-N synthesizer using digital compensation for 2.5-Mb/s GFSK modulation,"

IEEE Journal of Solid-State Circuits, vol. 32, no. 12, pp. 2048-2059, Dec. 1997.

4. N. Filiol, et. al., "A 22 mW Bluetooth RF transceiver with direct RF modulation and on-chip IF filtering," *IEEE International Solid-State Circuits Conference*, Digest of Technical Papers, vol. 44, pp. 202-203, Feb. 2001.

5. D. R. McMahill, C. G. Sodini, "A 2.5-Mb/s GFSK 5.0-Mb/s 4-FSK Automatically Calibrated Σ–Δ Frequency Synthesizer," *IEEE Journal of Solid State Circuits*, vol. 37, no. 1, pp. 18-26, January 2002.

6. D. R. McMahill, C. G. Sodini, "Automatic calibration of modulated frequency synthesizers," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 49, no. 5, pp. 301-311, May 2002.

7. S. Willingham, et. al., "An integrated 2.5GHz ΣΔ frequency synthesizer with 5μs settling and 2Mb/s closed loop modulation," *IEEE International Solid-State Circuits Conference*, Digest of Technical Papers, vol. 43, pp. 200-201, Feb. 2000.

8. J. S. Lee, et. al., "Charge pump with perfect current matching characteristics in phase-locked loops," *Electronic Letters*, vol. 36, no. 23, pp. 1907-1908, November 2000.

9. M. H. Perrott, M. D. Trott, C. G. Sodini, "A Modeling Approach for D-S Fractional-N Frequency Synthesizers Allowing Straightforward Noise Analysis," *IEEE Journal of Solid State Circuits*, vol. 37, no. 8, pp. 1028-38, August 2002.

10. G. C. Gillette, "Digiphase Synthesizer," *Proc of the 23rd Annual Frequency Control Symposium*, pp. 201-210, 1969.

11. N. B. Braymer, "Frequency synthesizer," United States Patent no. 3,555,446, January 12, 1971.

12. W. F. Egan, *Frequency Synthesis by Phase Lock*, second edition, Wiley Interscience, 2000.

13. J. A. Crawford, *Frequecy Synthesizer Design Handbook*, Artech House Inc., 1994.

14. B. Miller, B. Conley, "A multiple modulator fractional divider," *Annual IEEE Symposium on Frequency Control*, vol. 44, pp. 559-568, March 1990.

15. B. Miller, B. Conley, "A multiple modulator fractional divider," *IEEE Transactions on Instrumentation and Measurement*, vol. 40, no. 3, pp. 578-583, June 1991.

16. T. A. Riley, M. A. Copeland, T. A. Kwasniewski, "Delta-sigma modulation in fractional-N frequency synthesis," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 5, pp. 553-559, May, 1993.

17. I. Galton, "One-bit dithering in delta-sigma modulator-based D/A conversion," *Proc. of the IEEE International Symposium on Circuits and Systems*, vol. 2, pp. 1310-1313, May 1993.

18. I. Galton, "Granular quantization noise in a class of delta-sigma modulators," *IEEE Transactions on Information Theory*, vol. 40, no. 3, pp. 848-859, 1994.

19. N. King, "Phase locked loop variable frequency generator" United States Patent no. 4,204,174, May 20, 1980.

20. I. Galton, "Spectral shaping of circuit errors in digital-to-analog converters," *IEEE Trans. Circuits Syst. II*, vol. 44, no. 10, pp. 808-17, Oct. 1997.

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

22. A. Fishov, E. Siragusa, J. Welz, E. Fogleman, I. Galton, "Segmented mismatch-shaping D/A conversion," *Proc. of the IEEE International Symposium on Circuits and Systems*, May 2002.

23. E. Fogleman, I. Galton, W. Huff, H. T. Jensen, "A 3.3V single-poly CMOS audio ADC delta-sigma modulator with 98dB peak SINAD and 105dB peak SFDR," *IEEE Journal of Solid State Circuits*, vol. 35, no. 3, pp. 297-307, March, 2000.

24. B. Razavi, *Design of Analog CMOS Integrated Circuits*, first ed., McGraw Hill, 2001, pp. 562-566.

25. C. S. Vaucher, D. Kasperkovitz, "A Wide-Band Tuning System For Fully Integrated Satellite Receivers," *IEEE Journal of Solid State Circuits*, vol. 33, no. 7, pp. 987-997, July 1998.

26. W. Rhee, "Design of High-Performance CMOS Charge Pumps In Phase Locked Loops," *Proc. of the IEEE International Symposium on Circuits and Systems*, vol. 2, pp. 545-548, 1999.

Chapter 2

Phase Noise Cancellation Design Tradeoffs in Delta-Sigma Fractional-*N* PLLs

Sudhakar Pamarti and Ian Galton

Abstract— A theoretical analysis of a recently proposed phase noise cancellation technique that relaxes the fundamental tradeoff between phase noise and bandwidth in ΔΣ fractional-*N* PLLs is presented. The limits imposed by circuit errors and PLL dynamics on the phase noise and loop bandwidth that can be achieved by PLLs incorporating the technique are quantified. Design guidelines are derived that enable customization of the technique in terms of PLL target specifications.

I. INTRODUCTION

A phase noise cancellation technique is presented in [1] that employs a digital-to-analog converter (DAC) cancellation path to suppress the phase noise arising from quantization error in a delta-sigma (ΔΣ) fractional-*N* phase locked loop (PLL). The technique has been shown to allow a ten-fold increase in the PLL bandwidth without increasing the spot phase noise arising from ΔΣ modulator quantization noise for a specific PLL architecture and application: a 2.4 GHz second-order ΔΣ fractional-*N* PLL with a 460 kHz minimum bandwidth and 1 Mb/s in-loop FSK modulation for a Bluetooth wireless LAN compliant direct conversion transceiver. This paper presents a theoretical analysis of the phase noise cancellation technique with the goal of facilitating its application to realize other wide bandwidth, low noise ΔΣ fractional-*N* PLLs.

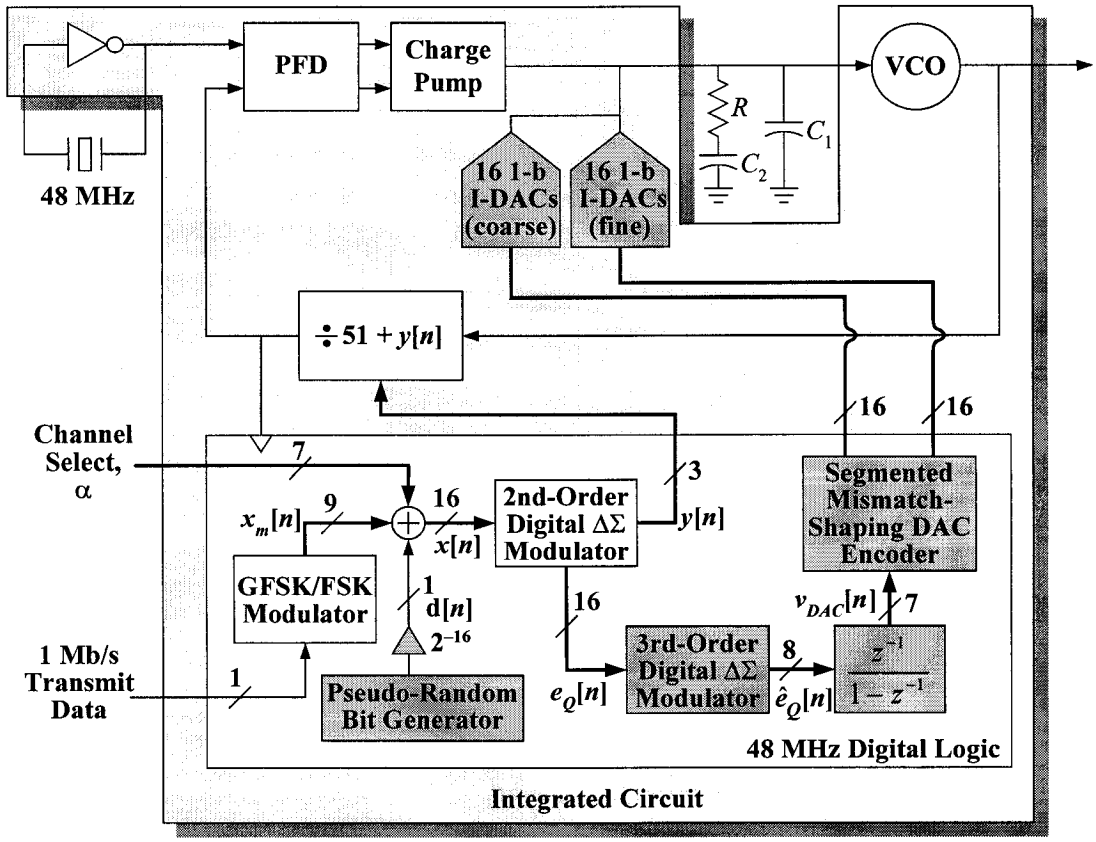

Fig. 2.1: A high level functional diagram of the ΔΣ FNPLL presented in [1].

Circuit errors and the dynamics of the PLL impose limits on the phase noise and bandwidth achievable using the phase noise cancellation technique. The technique employs quantization noise shaping, mismatch noise shaping, and one-bit dithering to significantly reduce these limits compared to prior art [2, 3, 4, 5, 6]. The paper quantifies the effects of noise shaping and dithering on the PLL phase noise, presents guidelines to customize the phase noise cancellation technique as a function of the PLL target specifications, and presents guidelines to reduce the hardware complexity of the technique without adversely effecting the PLL phase noise.

Section II presents an overview of the phase noise cancellation technique and

describes the various ways in which it can be customized. Sections III and IV analyze the limits imposed on the effectiveness of the phase noise cancellation technique by circuit gain errors and PLL dynamics, respectively. Sections V and VI present methods for reducing the hardware complexity of the technique.

II. Overview Of Phase Noise Cancellation Technique

A high level functional diagram of the integrated circuit (IC) presented in [1] is reproduced in Fig. 2.1. It includes all the components of a conventional second-order $\Delta\Sigma$ fractional-N PLL and some additional components which constitute the phase noise cancellation technique. These additional components are indicated by the shaded blocks in the figure. The segmented mismatch shaping DAC encoder and the two banks of 1-bit current DACs together constitute a DAC, which is henceforth referred to as the *cancellation DAC*. In the absence of the phase noise cancellation technique, the quantization noise, $e_Q[n]$, from the second-order digital $\Delta\Sigma$ modulator effectively injects a charge sample, $Q_Q[n]$, into the loop filter each reference period, thereby perturbing the VCO and causing phase noise. The cancellation technique suppresses this phase noise by nominally injecting $-Q_Q[n]$ into the loop filter. Aside from a constant offset, the sequence of charge samples is well modeled as

$$Q_Q[n] = I_{CP} T_{VCO} \sum_{k=n_0}^{n-1} e_Q[k], \tag{1}$$

where I_{CP} is the nominal charge pump current, T_{VCO} is the nominal period of the PLL output, and $n_0 < n$ is an arbitrary starting time index. The phase noise cancellation technique generates an estimate of $-Q_Q[n]$ by digitally computing $e_Q[n]$, reducing its

bit-width using the third-order digital $\Delta\Sigma$ modulator, accumulating the result, and using the cancellation DAC to generate proportional analog charge samples. The cancellation DAC generates charge samples by injecting appropriately scaled current pulses which are four VCO periods wide. The combination of the third-order digital $\Delta\Sigma$ modulator, the integrator, and the cancellation DAC is referred to as the *DAC cancellation path*. Note that while the pseudo-random bit generator is a part of some conventional $\Delta\Sigma$ fractional-N PLLs, it is shaded in the figure to emphasize its essential role in the phase noise cancellation technique. Later sections of the paper describe the role in detail.

The goal of the phase noise cancellation technique is to remove all of $Q_Q[n]$ without introducing other sources of error. However, gain mismatches between the charge pump and cancellation DAC cause a portion of $Q_Q[n]$ to be left behind in the loop filter every reference period. Similarly, requantization of $e_Q[n]$ in the cancellation path, mismatches among the 1-bit current DACs, and 1-bit dithering contribute additional error charge along with that left behind by imperfect cancellation of $Q_Q[n]$. In spite of these imperfections, the system in Fig. 2.1 achieves a low phase noise[1] while maintaining a minimum bandwidth of 460 kHz. In order to achieve the same peak spot phase noise without the phase noise cancellation technique, a PLL bandwidth of no more than 50 kHz would be required. This bandwidth extension is the principal benefit of the phase noise cancellation technique. The success of the

[1] For example, at 3 MHz from the PLL center frequency the phase noise is –127 dBc/Hz.

technique results from several architectural choices:

- use of a second-order digital ΔΣ modulator to choose the frequency division ratios,
- use of cancellation DAC current pulses with durations of 4 VCO periods,
- use of a third-order digital ΔΣ modulator with which to requeantize $e_Q[n]$ to 8 bits,
- use of a segmented mismatch shaping DAC encoder, and
- use of one-bit dither.

As is shown in the remainder of the paper, the first two choices determine the bandwidth and phase noise performance limits of the cancellation technique. The other choices reduce the hardware complexity of the DAC cancellation path while ensuring that phase noise due to the requantization error, dither, and mismatches among the 1-bit DACs is free of spurious tones and otherwise negligible.

The analysis offers design guidelines for how to customize the phase noise cancellation technique to ΔΣ fractional-N PLLs of other specifications. For instance, one might use a second-order digital ΔΣ modulator to requantize $e_Q[n]$ instead of a third-order ΔΣ modulator, or requantize $e_Q[n]$ to 4-bits instead of 8-bits. The analysis is performed in the context of a system that uses the same general architecture as shown in Fig. 2.1 but possibly differs in the parameters of the PLL and the above mentioned choices. Expressions are derived that predict the power spectral density (PSD) of the PLL phase noise caused by errors in the DAC cancellation path. These expressions are explicit functions of most of the above mentioned choices. For

example, one of the expressions is a function of the duration of DAC current pulses. A designer can use the expressions to pick values for the above choices that ensure that the PLL phase noise is small enough to meet specific requirements. To avoid burdening the designer with too many equations, qualitative recommendations are presented to serve as design guidelines in customizing the phase noise cancellation technique.

For ease of reference, the digital ΔΣ modulator used to choose the sequence of division ratios is henceforth called a *fractional modulator*, and the digital ΔΣ modulator which requantizes $e_Q[n]$ is called the *requantization modulator*.

III. FRACTIONAL MODULATOR ORDER

Any mismatch between the charge pump current and the cancellation DAC current causes phase noise in the PLL output. This phase noise tends to dominate the contributions of other errors in the cancellation path such as requantization and mismatches among the 1-bit DACs. This section studies the impact of the order of the fractional modulator, L, on the phase noise caused by the mismatch. The requantization of $e_Q[n]$ is ignored to simplify analysis.

Suppose that I_{DAC} in Amperes is the nominal gain of the cancellation DAC, and $x_{DAC}[n]$ is its (unitless) input sequence. Then, the cancellation DAC generates current pulses which have nominal current values $i_{DAC}[n] = -I_{DAC}x_{DAC}[n]$. Since the requantization of $e_Q[n]$ is ignored, it follows from Fig. 2.1 that $x_{DAC}[n]$ is just the sum of all the past values of $e_Q[n]$:

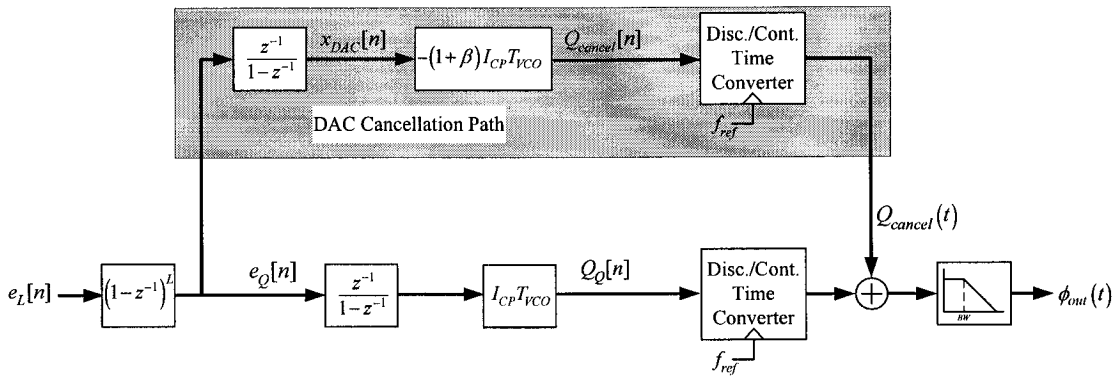

Fig. 2.2: A model for the cancellation technique including a gain error in the cancellation path.

$$x_{DAC}[n] = \sum_{k=n_0}^{n-1} e_Q[k].$$

Suppose that T_{DAC} is the nominal duration of the cancellation DAC current pulses. Therefore, the charge added to the loop filter by the cancellation path is

$$Q_{cancel}[n] = i_{DAC}[n] \cdot T_{DAC} = -I_{DAC}T_{DAC} \sum_{k=n_0}^{n-1} e_Q[k].$$

It follows from (1) that to cancel $Q_Q[n]$, T_{DAC} and I_{DAC} must satisfy $I_{DAC}T_{DAC} = I_{CP}T_{VCO}$. Suppose that there is a normalized mismatch, β, between the charge pump current and I_{DAC} *i.e.*, the cancellation DAC has a gain of $(1 + \beta)I_{DAC}$ instead of I_{DAC}. Then,

$$Q_{cancel}[n] = -(1+\beta) I_{CP}T_{VCO} \sum_{k=n_0}^{n-1} e_Q[k]. \qquad (2)$$

Therefore, $Q_{cancel}[n] \neq -Q_Q[n]$, and a portion of $Q_Q[n]$ remains in the loop filter and causes phase noise. The order of the fractional modulator, L, determines the severity of this effect.

Phase noise contribution

Fig. 2.2 presents a signal processing model that predicts the phase noise as a function of β. Note that except for the shaded portion, the model is well known [7, 8]. The shaded portion represents the DAC cancellation path when requantization of $e_Q[n]$ is ignored, as given by equation (2). The model output is the PLL phase noise. The lowpass filter in the model represents the response of the PLL to charge samples added to its loop filter. It is expressed as $(2\pi/I_{CP}T_{VCO})A_\phi(s)$ where $A_\phi(s)$ is the well known, closed-loop transfer function from the reference to the output in a PLL, normalized to unity gain in the pass band [7, 9]. It is determined by the parameters of the PLL, and its –3dB cut off frequency is the bandwidth of the PLL. For instance, the PLL core in Fig. 2.1 results in

$$A_\phi(s) \approx \frac{1}{1+s/K+s^2/\sqrt{b}K^2}, \quad \text{where } K=\frac{I_{CP}RK_{VCO}}{2\pi(N+a)} \text{ and } b=1+\frac{C_2}{C_1}.$$

The effect of adding a sequence of charge samples to the loop filter is modeled by converting the sequence into a continuous-time signal and applying the result to the input of the low pass filter. Since $Q_Q[n]$ and the cancellation charge samples, $Q_{cancel}[n]$, are both added to the loop filter, they are converted into continuous-time charge signals, summed and applied to the input of the low pass filter. The relation between $Q_Q[n]$ and the quantization noise from the fractional modulator, $e_Q[n]$, which is given by (1), is explicitly shown in the model. The quantization noise, $e_Q[n]$, is modeled as an additive error source, $e_L[n]$, passing through L discrete differentiators.

As suggested in [10], in non-overloading $\Delta\Sigma$ modulators of order $L \geqslant 2$, a one-

bit dither signal added to the least significant bit (LSB) of the input ensures that $e_L[n]$ is white, and uniformly distributed over the range –0.5 to 0.5 with a variance of 1/12. The pseudo-random bit generator shown in Fig. 2.1 provides this one-bit dither[2]. For $e_L[n]$ to have these properties, it is essential that the fractional modulator has enough output levels such that its internal quantizer never overloads. For instance, in [1] a fractional modulator of order $L = 2$ with a five-level quantizer is used to achieve an input no-overload range of –0.5 to 0.5.

The PSD of the PLL phase noise due to the gain mismatch follows from the model:

$$S_{\phi}^{\beta}(j2\pi f) = \beta^2 \frac{\pi^2}{3f_{ref}} \left| 2\sin\left(\frac{\pi f}{f_{ref}}\right) \right|^{2(L-1)} \left| A_{\phi}(j2\pi f) \right|^2 \quad \text{rad}^2/\text{Hz}, \tag{3}$$

where f_{ref} is the reference frequency and f is the frequency offset relative to the PLL center frequency. Note that with $\beta = 1$, equation (3) reduces to the well known expression for the PSD of the phase noise due to quantization noise in a conventional Lth order $\Delta\Sigma$ fractional-N PLL [7, 8]. Equation (3) can be used to determine the value of L that satisfies the phase noise and bandwidth specifications for an expected β. For example, in the system in Fig. 2.1, f_{ref} = 48 MHz, the normalized mismatch is expected to be 10% *i.e.*, $\beta = 0.1$, and the required bandwidth is 460 kHz [1]. The target specifications require that the PLL phase noise is less than –127 dBc/Hz at a 3 MHz offset from the PLL center frequency. The poles and zeros of $A_{\phi}(s)$ were chosen to

[2] The one-bit dither also causes phase noise, which is usually negligible and is considered in a later section.

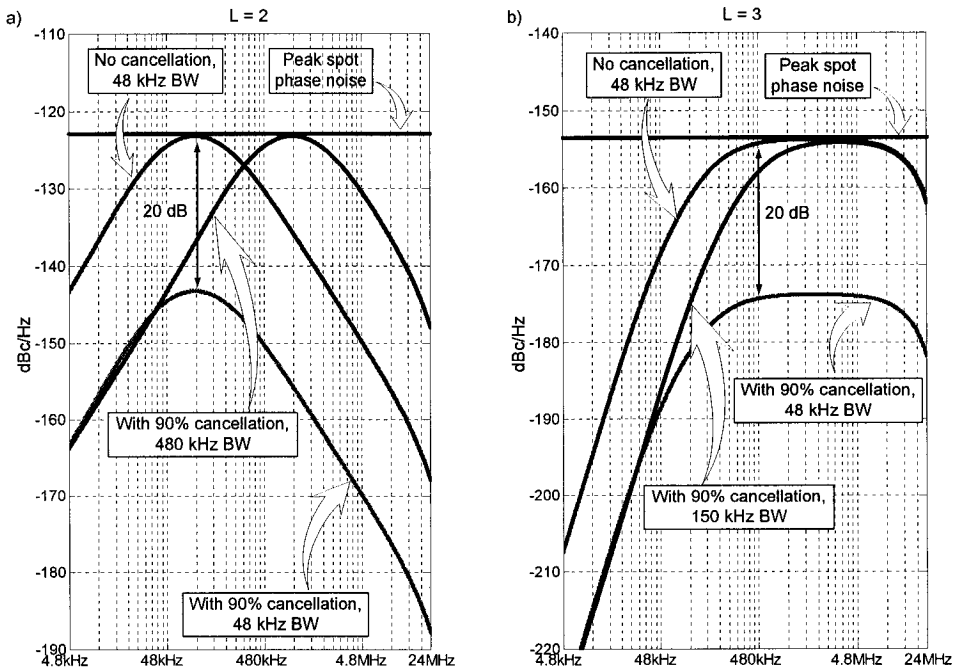

Fig. 2.3: Illustration of bandwidth extension made possible by the phase noise cancellation technique.

ensure that the PLL has a 67 degree phase margin. Consequently, $|A_\phi(j2\pi f)|$ is about – 19 dB at f = 3 MHz. Substituting these values[3] into (3) indicates that while L = 2 meets the phase noise specification, L = 1 does not. Appendix C presents a convenient set of equations for the design of a L^{th} order ΔΣ fractional-N PLL with a required loop bandwidth, phase margin, and acceptable quantization induced phase noise.

Recommended fractional modulator order

Often, the above calculation has to be repeated for a number of offset frequencies, f, to choose an acceptable value for L. Moreover, many of the PLL

parameters which effect the above calculation (*e.g.*, the poles and zeros of $A_\phi(s)$, and the reference frequency) are also choices available to the designer, necessitating many iterations of the above calculation to complete the design. Therefore, choosing a value for L is not always as straightforward as in the above example. This section simplifies the problem by showing that $L = 2$ or 3 are the best choices for many $\Delta\Sigma$ fractional-N PLLs.

It follows from (3) that the phase noise cancellation technique reduces the spot phase noise caused by $e_Q[n]$ in a $\Delta\Sigma$ fractional-N PLL by $-20\log_{10}|\beta|$ dB. The reduced phase noise can be traded off to increase the PLL bandwidth. Fig. 2.3 illustrates the tradeoff for the system depicted in Fig. 2.1. The PLL parameters are chosen such that $A_\phi(s)$ effectively has two poles–one at its passband edge and the other at roughly five times the bandwidth. The top and bottom curves in Fig. 2.3(a) are plots of $S_\phi^\beta(j2\pi f)$ where $A_\phi(s)$ has a 48 kHz bandwidth, $L = 2$, and $\beta = 1$ and 0.1, respectively. In other words, they respectively represent phase noise PSDs[4] in a second-order $\Delta\Sigma$ fractional-N PLL without the DAC cancellation path and with a 90% accurate DAC cancellation path. The 20 dB reduction implies that the $A_\phi(s)$ can now have a 10-fold wider bandwidth, namely 480 kHz, and still maintain the same peak spot phase noise, as indicated by the middle curve in the figure. A similar bandwidth extension is possible for a third-order $\Delta\Sigma$ fractional-N PLL, as illustrated by Fig. 2.3(b). However, note that the bandwidth extension in this case is only 3-fold.

---

[3] Note that the PSD expression must be divided by 2π to convert to dBc/Hz values.

Similarly going to a higher order than $L = 3$ further reduces the bandwidth extension offered by the technique. Since a wide bandwidth is desirable for a variety of reasons, the achievable bandwidth extension is considered to be the principal benefit of the phase noise cancellation technique [1]. Choices $L = 2$ or 3 offer the greatest bandwidth extension without complicating the requirements of other components of the PLL.

Suppose that without the cancellation technique, $A_\phi(s)$ has a bandwidth BW_{old} and achieves a certain peak spot phase noise. Suppose that the phase noise reduction allows $A_\phi(s)$ to have a wider bandwidth BW_{new} while maintaining the same peak spot phase noise. The achievable bandwidth extension is then defined as $\lambda = BW_{new}/BW_{old}$. The achievable bandwidth extension is expected to depend on L, and the locations of the poles and zeros of $A_\phi(s)$. However, as shown in Appendix A, an approximate but reasonable estimate for the achievable bandwidth extension is

$$\lambda \approx \left|1/\beta\right|^{\frac{1}{(L-1)}}, \tag{4}$$

which is independent of $A_\phi(s)$. For instance, for $\beta = 0.1$, 10-fold, 3-fold and 2-fold bandwidth extension is possible for $\Delta\Sigma$ fractional-N PLLs with $L = 2$, 3 and 4 respectively. This is illustrated by the plots shown in Fig. 2.3. It follows from (4) that only a small bandwidth extension is achieved for orders $L > 3$.

Fractional modulators of order $L > 3$ are undesirable for other reasons as well. It can be shown [11] that they need more output levels than lower order modulators to

---

[4] The PSD plots in all the figures in this paper are scaled to dBc/Hz values.

ensure that $e_L[n]$, and, hence, the PLL phase noise have no spurious tones. This complicates the design of the frequency divider because more output levels imply a wider range of frequency division ratios. The resulting charge pump current pulses are also wider and contribute more charge pump noise. Another problem arises because charge pump current pulses do not occur uniformly in time–the start of a charge pump pulse sometimes coincides with a rising edge transition of the reference signal, while at other times it coincides with the rising edge transition of the divider output signal. This time-variant behavior has the effect of applying a non-linearity to the quantization noise, $e_Q[n]$. Consequently, high frequency components of $e_Q[n]$ fold to lower frequencies and increase close-in phase noise. The effect is aggravated for $L > 3$ because the spectrum of $e_Q[n]$ is such that it has more power in the higher frequencies.

Fractional modulators of order $L = 1$ are not recommended either since they cause a lot of phase noise close to the PLL center frequency. This is evident from the absence of any zeros at dc in the expression for $S_\phi^\beta(j2\pi f)$ in equation (3) when $L = 1$. The reason is that when $L = 1$, the phase noise caused by $e_Q[n]$, even after cancellation, does not have the familiar high-pass spectral shape. The example calculation at the end of previous subsection, which picked $L = 2$ over $L = 1$, illustrates this claim.

IV. DAC CURRENT PULSE DURATION

The duration of the DAC current pulse, T_{DAC}, affects the PLL phase noise in two ways. First, any static error in the DAC current pulse duration causes incomplete

removal of $Q_Q[n]$, just like the gain mismatch, β, considered in the previous section. Second, the non-zero width of the DAC current pulses allows $Q_Q[n]$ to disturb the VCO before being removed by the DAC current pulses. This phenomenon is described in detail later. First, the error in T_{DAC} is considered. Requantization error is ignored in the following discussion.

Gain errors due to imperfect pulse timing

Suppose that the ΔΣ fractional-*N* PLL changes from one center frequency to a new center frequency such that the nominal period of the VCO changes from T_{VCO} to T_{VCO}^*. In this case, the charge effectively added to the loop filter by $e_Q[n]$ is:

$$Q_Q[n] = I_{CP}T_{VCO}^* \sum_{k=n_0}^{n-1} e_Q[k].$$

As described previously, to remove $Q_Q[n]$ it is necessary to ensure that $I_{CP}T_{VCO}^* = I_{DAC}T_{DAC}$. If T_{DAC} does not change with T_{VCO}, this does not happen. Consequently, a portion of $Q_Q[n]$ is left in the loop filter, similar to the effect of a normalized gain mismatch, β. The recommended solution is to make T_{DAC} equal an integer number of VCO periods, i.e., $T_{DAC} = M_{DAC}T_{VCO}$ where M_{DAC} is an integer. Then, as T_{VCO} varies so does T_{DAC} and $I_{CP}T_{VCO}^* = I_{DAC}T_{DAC}$ is satisfied. The frequency divider, which operates by counting an integer number of VCO periods, can be easily modified to generate a pulse whose duration is equal to a specified integer number of VCO periods.

Even so, inevitable timing errors in the circuitry that generates the cancellation

DAC current pulse cause its duration to be $(T_{DAC} + \Delta T_{DAC})$ resulting in a normalized gain error of $\Delta T_{DAC}/T_{DAC}$ in the cancellation path. The PLL phase noise contributed by this gain error can be predicted by adding $\Delta T_{DAC}/T_{DAC}$ to β in (3):

$$\beta_{eff} = \beta + \frac{\Delta T_{DAC}}{T_{DAC}}.$$

Usually, these timing errors do not scale with T_{DAC}. For example, suppose the circuitry that generates the cancellation pulse has a timing error of at least 20 ps[5], and that the DAC current pulse is four VCO periods wide. At 2.4 GHz, this results in a normalized gain error of 1.2%. A simple way to ensure that these timing errors do not limit the PLL phase noise is to choose a wide cancellation DAC current pulse to ensure that $\Delta T_{DAC}/T_{DAC} \ll \beta$, but as described below this causes other problems.

[5] This is not a pessimistic estimate considering that a typical inverter delay in 1.8 Volt, 0.18μm CMOS technology is about 60ps.

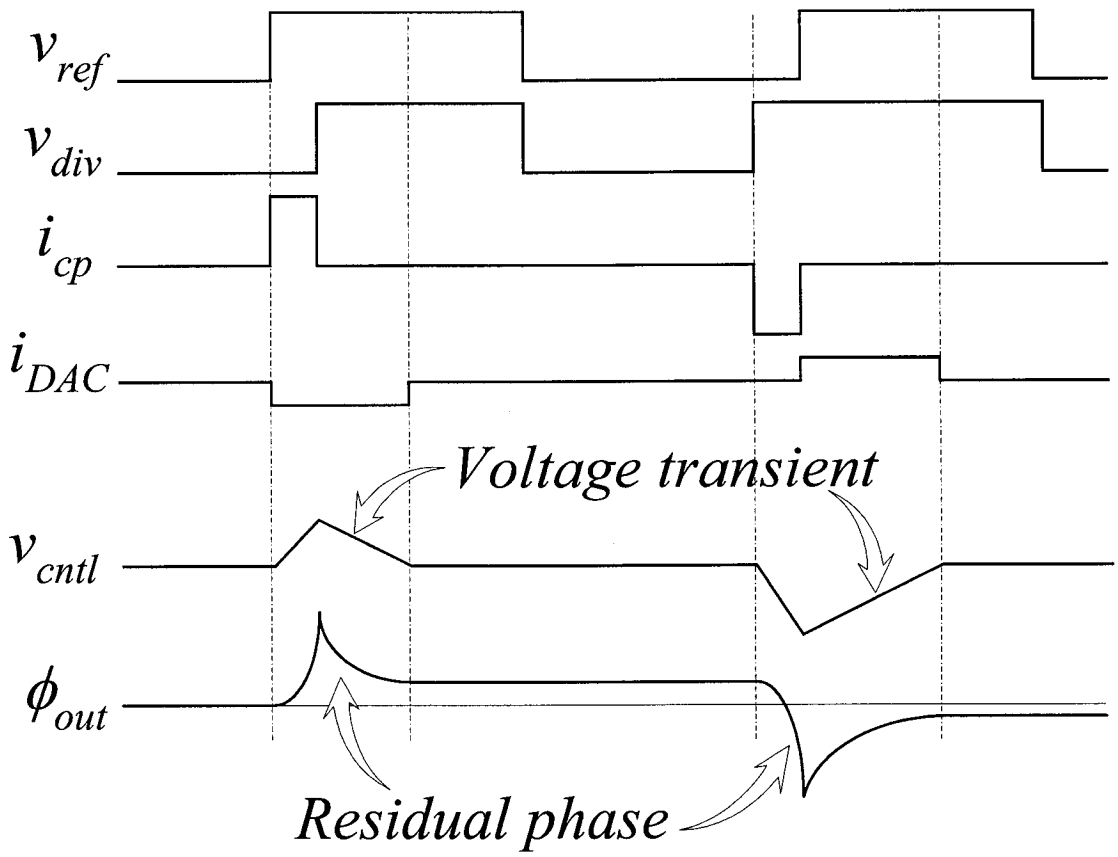

Fig. 2.4: Mechanism of imperfect phase noise cancellation.

Non-zero DAC current pulse width

Wide cancellation pulses are not very effective in canceling the phase noise contributions of narrow charge pump pulses. Even if they remove $Q_Q[n]$ completely, they disturb the VCO in doing so and cause phase noise. This phenomenon is illustrated in Fig. 2.4, where, for the sake of simplicity, dither and modulation signals are ignored, it is assumed that the loop filter comprises just one capacitor *i.e.*, $R = C_I = 0$ in Fig. 2.1, and the PLL is assumed to be in frequency and phase lock. The waveforms labeled i_{CP} and i_{DAC} in Fig. 2.4 represent the current pulses that are added

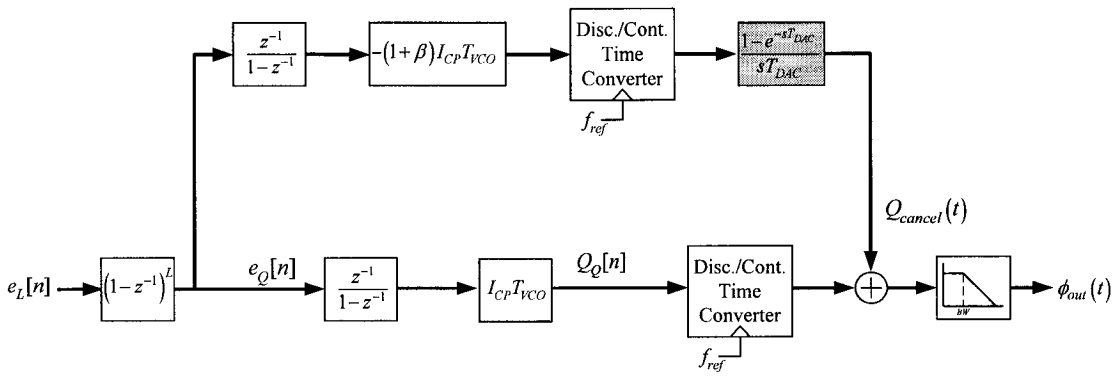

Fig. 2.5: A model for the cancellation technique including the effect of finite DAC pulse width.

to the loop filter by the charge pump and the cancellation path respectively. The waveforms labeled v_{cntl} and ϕ_{out}, respectively, represent the input voltage of the VCO and the PLL phase noise, $\phi_{out}(t)$. Assuming that the cancellation path has no gain error, the charge added by i_{DAC}, *i.e.*, $Q_{cancel}[n]$, exactly cancels out that added by i_{CP}, *i.e.*, $Q_Q[n]$, as illustrated by v_{cntl} returning to its original value at the end of each cancellation DAC pulse. However, the ramp-like voltage transients in v_{cntl} disturb the VCO. These disturbances are accumulated into a residual phase, as shown in the figure.

If the charge pump and the cancellation DAC pulses were of the same width, or better, if they were both impulses, the phase noise cancellation would have been complete. While the very narrow charge pump pulses can be modeled as impulses, the same is not true for the wide cancellation DAC pulses. As the figure suggests, the wider the cancellation DAC pulses, the larger the voltage transients and the residual phase. The effect of the non-zero width of the cancellation DAC pulses can be incorporated into the model by adding a zero-order hold block in the cancellation path,

as indicated by the shaded block in Fig. 2.5. Appendix B justifies this modification and clarifies the inherent assumptions. Using the approximation[6] $e^{-x} \approx 1 - x + x^2/2$, the zero-order hold block can be reduced to a left plane zero, $(1 + sT_{DAC}/2)$. Consequently, an expression for the PSD of the residual PLL phase noise can be obtained from the model:

$$S_{\phi}^{\beta,T_{DAC}}(j2\pi f) \approx \left\{\beta^2 + (\pi f T_{DAC})^2\right\} \frac{\pi^2}{3 f_{ref}} \left| 2\sin\left(\frac{\pi f}{f_{ref}}\right) \right|^{2(L-1)} \left| A_{\phi}(j2\pi f) \right|^2 \quad \text{rad}^2/\text{Hz}, \quad (5)$$

where it is assumed that $|\beta| \ll 1$. The effect of non-zero T_{DAC} is represented by the $\pi f T_{DAC}$ term in the expression.

---

[6] The approximation is good for frequencies, $f \ll 2/T_{DAC}$. For instance, in the system in Fig. 2.1, in which $T_{DAC} = 4T_{VCO}$, the approximation is good for frequencies, $f \ll f_{VCO}/2$.

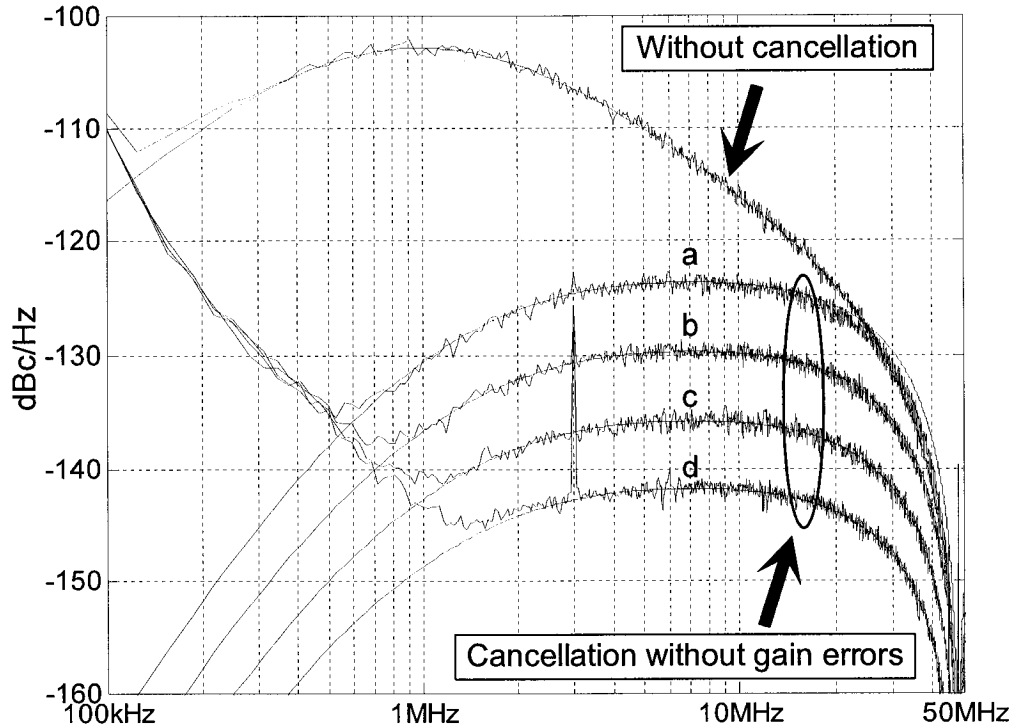

Fig. 2.6: Predicted and simulated phase noise PSD for the cancellation technique for DAC pulses of duration (a) 32 (b) 16 (c) 8 and (d) 4 VCO periods.

The validity of equation (5) is demonstrated in Fig. 2.6 in which plots of $S_{\phi}^{\beta,T_{DAC}}(j2\pi f)$ are compared to simulated phase noise PSDs. The simulations correspond to a second-order $\Delta\Sigma$ fractional-N PLL with a 480 kHz bandwidth and an ideal cancellation path (*i.e.*, a DAC cancellation path without requantization or component errors). The smooth curves are plots of $S_{\phi}^{\beta,T_{DAC}}(j2\pi f)$ for T_{DAC} = 4, 8, 16 and 32 VCO periods, and the ragged curves are simulated phase noise PSDs for the same set of values of T_{DAC}. For comparison, simulated and theoretical plots of phase noise PSDs for the same PLL, but without the cancellation path, are included. The

$1/f^2$ phase noise deviations exhibited by the simulated PSDs at low frequency offsets are due to the one-bit dither at the input of the fractional modulator in Fig. 2.1.

The small fractional spur visible in the simulated curves is not predicted by (5). It is caused by the non-uniform occurrence of the charge pump pulses. The start of a charge pump pulse varies from one reference period to another; while sometimes it coincides with the rising edge transition of the reference waveform, at other times it coincides with the rising edge transition of the divider output waveform. This time-variant behavior is responsible for the spurious tone. The spurious tone occurs in the conventional ΔΣ fractional-*N* PLL as well, but is masked by the phase noise caused by $e_Q[n]$. When the phase noise cancellation technique removes most of $e_Q[n]$, this spurious tone is uncovered. The spurious tone is, however, so small that it is often dominated by spurious tones caused by other non-linearities in the PLL.

Equation (5) can be used to choose a T_{DAC} which satisfies the phase noise and bandwidth specifications for an expected normalized gain mismatch, β. Alternatively, T_{DAC} may be chosen such that $\pi f_{crit} T_{DAC} \ll \beta$, where f_{crit} is the critical frequency offset at which it is most difficult to meet the phase noise requirements of a particular ΔΣ fractional-*N* PLL. For instance, for the system in [1], f_{crit} = 3 MHz, and the expected gain mismatch is 10% *i.e.*, β = 0.1. Therefore, the constraint implies that $T_{DAC} \ll$ 10 ns or about 26 VCO periods. The choice used in the system is T_{DAC} = 4 VCO periods. The corresponding phase noise is indicated by the bottom most curve in Fig. 2.6 which is about 30 dB below the phase noise requirement of –127 dBc/Hz of the system.

Recommended cancellation DAC current pulse duration

Before recommending a choice for T_{DAC}, it is useful to enumerate the inferences of the two preceding subsections:

- T_{DAC} must be an integer number of VCO periods, $T_{DAC} = M_{DAC}*T_{VCO}$,
- T_{DAC} must be large enough so that $\Delta T_{DAC}/T_{DAC} < \beta$
- T_{DAC} must be small enough that $\pi f_{crit} T_{DAC} \ll \beta$

The recommended duration is $T_{DAC} = M_{DAC}*T_{VCO}$, where M_{DAC} is an integer chosen as a compromise between the last two constraints. For instance, suppose that the expected normalized mismatch is 10% *i.e.*, $\beta = 0.1$, the timing error is $\Delta T_{DAC} = 40$ ps, the nominal VCO period, T_{VCO}, is approximately 400 ps, and the critical frequency offset is f_{crit} = 3 MHz. Then the last two constraints, respectively, require that $M_{DAC} \gg 1$ and $M_{DAC} \ll 26$. In [1] a good compromise was found to be $M_{DAC} = 4$.

V. REQUANTIZATION

The purpose of requantizing $e_Q[n]$ is to reduce the required performance of the cancellation DAC. For instance, in Fig. 2.1, if $e_Q[n]$ were not requantized, the cancellation DAC would have to be a 15-bit DAC. Moreover, its LSB would correspond to a current on the order of a few nA. Requantization allows the use of only a 7-bit DAC with an LSB corresponding to 10 μA. The penalty is an increase in the PLL phase noise.

Suppose that a unity gain, Mth order digital ΔΣ modulator requantizes $e_Q[n]$

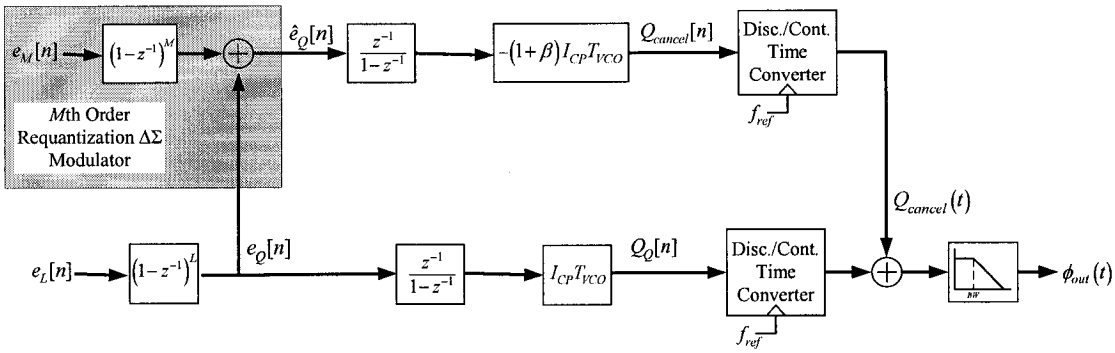

Fig. 2.7: A model for the cancellation technique including the requantization ΔΣ modulator.

and that the requantized sequence, $\hat{e}_Q[n]$, has a least significant bit (LSB) of Δ_{RQ}. For instance, in Fig. 2.1, $\hat{e}_Q[n]$ is an 8-bit number taking on values in the range –2 to 2 corresponding to an LSB of 1/64. The requantization error, $\hat{e}_Q[n]-e_Q[n]$, causes an error charge to be added to the loop filter every reference period. The amount of the phase noise contributed by requantization is determined by M and Δ_{RQ}. The relationship is derived below ignoring the effects of non-zero cancellation pulse widths to simplify the analysis.

Phase noise contribution

The effect of requantization on the PLL phase noise is incorporated into the model in Fig. 2.7 by adding a requantization error term as indicated by the shaded portion in Fig. 2.7. The requantization error is modeled as an additive source, $e_M[n]$, passing through M discrete differentiators. Using analyses similar to those presented in [10], it can be shown that one-bit dither added to the input of the fractional modulator ensures that $e_M[n]$ is white, uncorrelated with $e_L[n]$ and its delayed versions,

is uniformly distributed from $-0.5\Delta_{RQ}$ to $0.5\Delta_{RQ}$, and has a variance of $\Delta_{RQ}^2/12$. For this to be true, it is essential for the *M*th order ΔΣ modulator to have enough output levels such that its internal quantizer never overloads. An expression for the PSD of the phase noise contributed by the requantization error follows from the model:

$$S_\phi^{RQ}(j2\pi f) \approx \frac{\Delta_{RQ}^2 \pi^2}{3 f_{ref}} \left| 2\sin\left(\frac{\pi f}{f_{ref}}\right) \right|^{2(M-1)} \left| A_\phi(j2\pi f) \right|^2, \qquad (6)$$

where it has been assumed that β is much less than unity. Equation (6) can be used to determine values of M and Δ_{RQ} which satisfy the phase noise and bandwidth specifications.

Recommended M and Δ_{RQ}

The recommended choices are $M = L$ or $L + 1$, where L is the order of the fractional modulator, and the requantization LSB satisfy $\Delta_{RQ} < \beta$. As shown below, these choices ensure that the phase noise caused by requantization error is negligible compared to that caused by DAC cancellation path gain mismatch. This in turn ensures that requantization does not limit the phase noise performance of the ΔΣ fractional-*N* PLL.

In the absence of requantization, non-zero cancellation pulse width effects, and other DAC errors, the lowest phase noise which the cancellation technique can guarantee is $S_\phi^\beta(j2\pi f)$ given in (3). Therefore, choosing M and Δ_{RQ} such that $S_\phi^{RQ}(j2\pi f) < S_\phi^\beta(j2\pi f)$, ensures that requantization error does not limit phase noise performance. In this respect, it is useful to compare the two quantities:

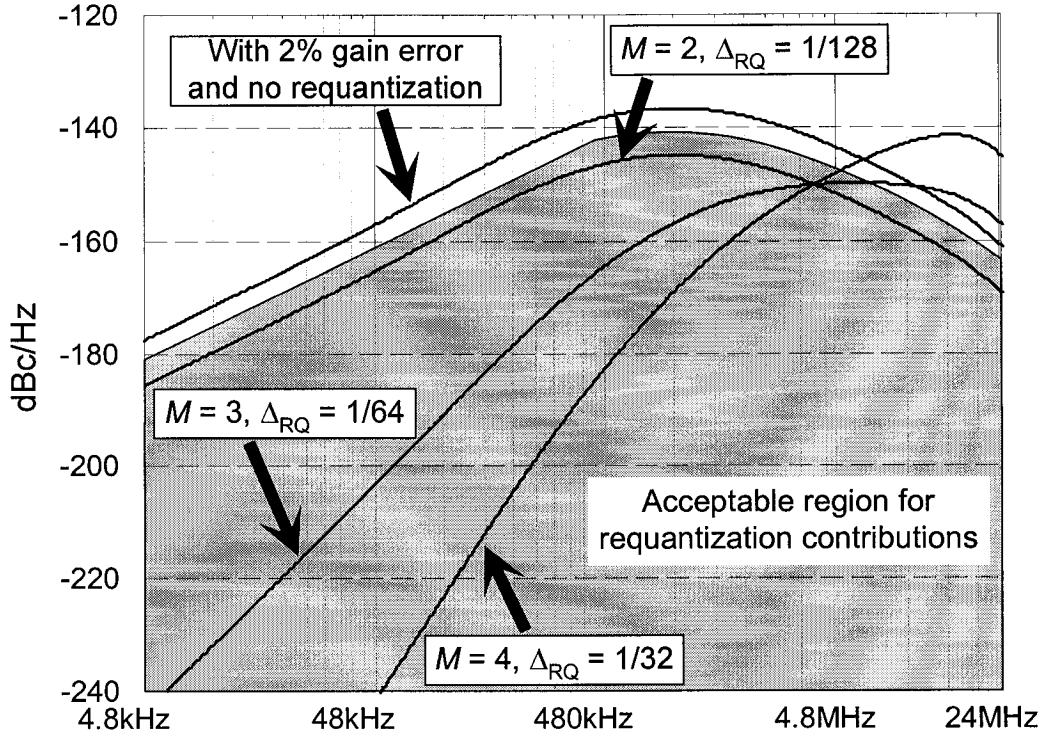

Fig. 2.8: Illustration of the effects of requantization on the phase noise of the PLL output.

$$\frac{S_\phi^{RQ}(j2\pi f)}{S_\phi^{\beta}(j2\pi f)} \approx \frac{\Delta_{RQ}^2}{\beta^2}\left|2\sin\left(\frac{\pi f}{f_{ref}}\right)\right|^{2(M-L)} \tag{7}$$

One choice that ensures that the $S_\phi^{RQ}(j2\pi f) < S_\phi^{\beta}(j2\pi f)$ is $M = L$ and $\Delta_{RQ} < \beta$. However, by using $M > L$, it might be possible to requantize more coarsely so as to further reduce the required performance of the cancellation DAC.

The possibility is illustrated in Fig. 2.8 which corresponds to a 480 kHz bandwidth, second-order ΔΣ fractional-N PLL with the phase noise cancellation technique. The top curve is the expected PLL phase noise due to a 2% gain error and no requantization in the cancellation path. Requantization will not noticeably increase

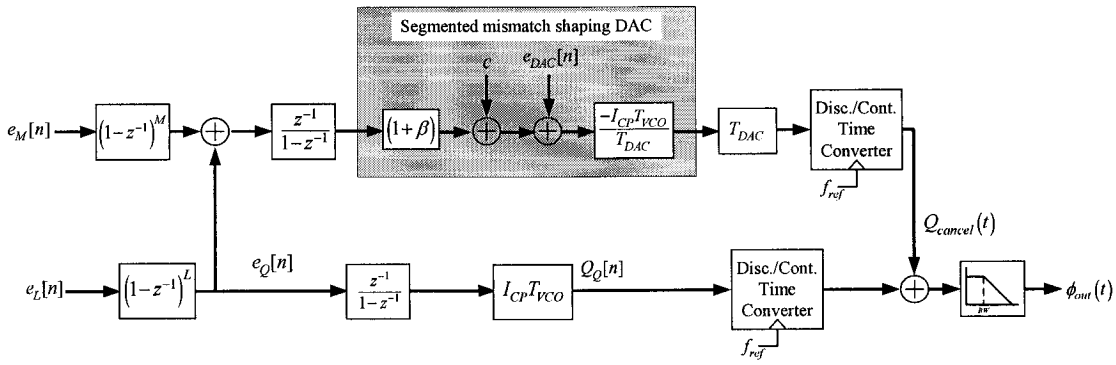

Fig. 2.9: A model for the cancellation technique including the segmentation of the DAC.

the PLL phase noise if $S_\phi^{RQ}(j2\pi f)$ is restricted to the shaded region, which starts about 3 dB below the top curve. Plots of $S_\phi^{RQ}(j2\pi f)$ for orders $M = L$, $L + 1$ and $L + 2$ (*i.e.*, $M = 2$, 3 and 4), and for specific values of Δ_{RQ} are included. In each case, the largest Δ_{RQ} was chosen that ensures that $S_\phi^{RQ}(j2\pi f)$ lies mostly within the shaded region. The choices $M > L$ allow coarser quantization, but for high frequencies the requantization contributions are larger than those due to the gain error alone. At least for $M = L + 1$, this is not particularly worrisome since $S_\phi^{RQ}(j2\pi f)$ is still much less than the peak spot phase noise.

VI. MISCELLANEOUS FACTORS

Segmented mismatch shaping DAC encoder

The combined output of the two DAC banks can be modeled using an offset, a gain error, and a normalized additive error source, $e_{DAC}[n]$, as shown by the shaded blocks in Fig. 2.9. The DAC error, $e_{DAC}[n]$, is caused by mismatches among the 1-bit

DAC elements. It causes phase noise. The constant offset has no noticeable effect on the PLL phase noise, the gain error has already been considered in Section III.

The segmented mismatch shaping encoder controls the operation of the DAC banks such that $e_{DAC}[n]$ is uncorrelated with the input to the DAC, spurious-free and has a zero at dc. It follows from the model that the contribution of $e_{DAC}[n]$ to the PLL phase noise PSD is

$$S_{\phi}^{DAC}(j2\pi f) = \frac{1}{f_{ref}} S_{DAC}\left(e^{j2\pi f T_{ref}}\right) \cdot \left|A_{\phi}(j2\pi f)\right|^2, \tag{8}$$

where $S_{DAC}\left(e^{jw}\right)$ is the PSD of $e_{DAC}[n]$. The zero at dc ensures that $S_{DAC}\left(e^{jw}\right)$ and hence $S_{\phi}^{DAC}(j2\pi f)$ has very little power in frequencies close to the PLL center frequency.

The segmented mismatch shaping encoder exploits redundancy in the DAC banks to guarantee that $e_{DAC}[n]$ has the aforementioned properties. While multiple methods of realizing the encoder have been reported [12, 13, 14, 15], none of them offer closed form expressions for $S_{DAC}\left(e^{jw}\right)$. Therefore, simulations are relied upon to determine the degree of mismatch among the DAC elements that can be tolerated. As reported in [16], it may be possible to derive closed form expressions for $S_{DAC}\left(e^{jw}\right)$ if detailed statistics of the quantization noise are available. Another alternative is to use reported bounds on the power in low frequency bands [17] to make some approximate quantitative predictions about tolerable mismatches.

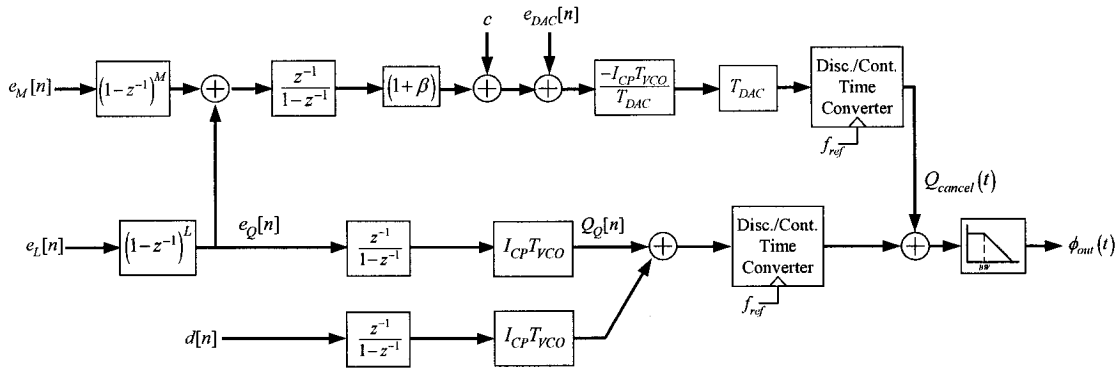

Fig. 2.10: A model of the cancellation technique including dither.

Number of input bits in the fractional modulator

The digital hardware complexity of the cancellation path can be reduced by allocating only a few bits to the input of the fractional modulator. The reason is that both the fractional and requantization modulator have data paths which are at least as wide as the input of the fractional modulator. However, a lower limit is imposed on the number of input bits by one-bit dither employed by the cancellation technique.

Suppose that K bits are allocated to the input of the fractional modulator. Therefore, one-bit dither added to the LSB of the input of the fractional modulator contributes undesirable FM modulation of $\pm f_{ref}/2^K$. The models presented so far have neglected the effect of dither in comparison with other sources of error. However, if K is small, the undesired FM modulation could degrade the signal-to-noise ratio of the transmitted frequency modulation signal. Even in the absence of frequency modulation, it causes f^{-2} noise in the PLL output phase. While these effects are well understood by prior art, they are usually dominated by the other sources of noise in the

ΔΣ fractional-N PLL.

Note that the one-bit dither can not be altogether eliminated since its absence would cause strong fractional spurs in the PLL phase. It is however tempting to modify $Q_{cancel}[n]$ to cancel contributions from dither as well. This promises to allow the use of as few bits, K, as possible. The possibility becomes evident by considering Fig. 2.10, which includes one-bit dither in the signal processing model. The dither can be added to $e_Q[n]$ before requantization by the Mth order ΔΣ modulator. Rather surprisingly, including the dither in the cancellation path causes spurious tones to reappear in the PLL phase noise. It negates the claims made in previous sections about the spurious-free nature of $e_M[n]$ and $e_{DAC}[n]$. Simulations corroborate this counter-intuitive phenomenon and it can be proved following analyses similar to those in [10]. However, the proof is not included in this paper.

VII. CONCLUSION

A theoretical analysis of the phase noise cancellation technique applied to a ΔΣ fractional-N PLL has been presented. The influence of circuit errors on the effectiveness of the phase noise cancellation technique has been analyzed and quantified. A fundamental lower limit on the phase noise imposed by the use of a current DAC for the phase noise cancellation has been derived. Recommendations have been made that enable customization of the phase noise cancellation technique in response to specific PLL target specifications.

APPENDIX A

The achievable bandwidth extension depends on L and the location of the poles and zeros of $A_\phi(s)$. Suppose that in a conventional Lth order $\Delta\Sigma$ fractional-N PLL, $A_\phi(s) = A_{old}(s)$ where $A_{old}(s)$ is a low pass filter of bandwidth BW_{old}. It can be approximately[7] represented as

$$A_{old}(s) = \prod_{k=1}^{R} \frac{1}{1 + s/2\pi f_k},$$

where f_k is the k-th pole frequency, R is the number of poles, and $f_1 = BW_{old}$. It is assumed for now that $A_{old}(s)$ has no complex poles. Suppose that when the phase noise cancellation technique is applied, then $A_\phi(s) = A_{new}(s)$ where $A_{new}(s)$ is a low pass filter of bandwidth BW_{new}. Define the achievable bandwidth extension as $\lambda \triangleq BW_{new}/BW_{old}$. It is also assumed that the poles of $A_{new}(s)$ are all scaled by λ. This is a reasonable assumption since it would impart the same phase margin to the core PLL. Therefore, it can be represented as

$$A_{new}(s) = \prod_{k=1}^{R} \frac{1}{1 + s/2\pi\lambda f_k}.$$

Now, the phase noise contributed by $e_Q[n]$ without cancellation technique is

$$S_\phi^{old}(j2\pi f) = \frac{\pi^2}{3 f_{ref}} \left| 2\sin\left(\frac{\pi f}{f_{ref}}\right) \right|^{2(L-1)} \left| A_{old}(j2\pi f) \right|^2.$$

It has two parts – $A_{old}(j2\pi f)$ and un-filtered phase noise, which increases at the rate of $20*(L-1)$ dB/decade till $0.5*f_{ref}$. To prevent the spot phase noise for $f < 0.5*f_{ref}$ from

becoming too large, the Lth order $\Delta\Sigma$ fractional-N PLL has at least $(L-1)$ poles in $A_{old}(s)$. Then, $S_{\phi}^{old}(j2\pi f)$ reaches its maximum value when $f = f_{L-1}$, or in other words, it peaks when $(L-1)$ poles of $A_{old}(s)$ "kick in". Assuming that $f_{L-2} \neq f_{L-1}$, $|A_{old}(j2\pi f)|$ can be approximated as:

$$\left|A_{old}(j2\pi f)\right| \approx \prod_{k=1}^{L-1} \frac{f_k}{f} \qquad \forall f \gg f_{L-1}.$$

Using the above approximation and using $\sin(x) \approx x$ for small x *i.e.*, for $f \ll f_{ref}$, it follows that the peak spot phase noise is approximately,

$$\max\left\{S_{\phi}^{old}(j2\pi f)\right\} \approx \frac{\pi^2}{3f_{ref}}\left(\frac{2\pi}{f_{ref}}\right)^{2(L-1)} \prod_{k=1}^{L-1} f_k^2.$$

Proceeding similarly it can be shown that the peak spot phase noise for the system with the phase noise cancellation technique is approximately,

$$\max\left\{S_{\phi}^{new}(j2\pi f)\right\} \approx \beta^2 \frac{\pi^2}{3f_{ref}}\left(\frac{2\pi}{f_{ref}}\right)^{2(L-1)} \prod_{k=1}^{L-1} (\lambda f_k)^2.$$

The achievable bandwidth extension is obtained by equating the above two peak spot phase noise values. Equating them results in $1 \approx \beta^2 \lambda^{2(L-1)}$ from which it follows that the achievable bandwidth extension is $\lambda \approx |1/\beta|^{1/(L-1)}$. Note that the argument can be extended to include complex poles in $A_{\phi}(s)$ provided that the $(L-1)$th pole is itself not a complex pole.

---

[7] Type-II PLLs have an in-band pole-zero doublet, which is ignored in this argument.

APPENDIX B

In the system in Fig. 2.1, once every reference period, the phase noise cancellation technique generates a pulse of current, $i_{DAC}[n]$, which has a duration of T_{DAC} seconds, starting from the rising edge transition of the divider output waveform. The resulting waveform can be denoted as

$$Q_{cancel}(t) = \sum_{n=0}^{\infty} i_{DAC}[n]\{u(t - t_{div}[n]) - u(t - t_{div}[n] - T_{DAC})\},$$

where $u(t)$ is the unit step function, the PLL is assumed to start at $n = 0$, and $t_{div}[n]$ is the time when the frequency divider finishes its division cycle and produces a rising edge transition. The rising edge transitions of the divider output waveform are not uniformly spaced in time. However, for the purposes of this model, it can be approximated as $t_{div}[n] \simeq nT_{ref}$ and the above equation can be modified to:

$$Q_{cancel}(t) = p(t) * \hat{Q}_{cancel}(t)$$

$$\text{where } p(t) = \frac{1}{T_{DAC}}\{u(t) - u(t - T_{DAC})\},$$

$$\text{and } \hat{Q}_{cancel}(t) = \sum_{n=0}^{\infty} i_{DAC}[n] \cdot T_{DAC} \cdot \delta(t - nT_{ref}).$$

The impulse train $\hat{Q}_{cancel}(t)$ is the same as the output of the discrete-time to continuous-time converter acting on $Q_{cancel}[n]$ in Fig. 2.5. Its convolution with $p(t)$ is represented by a multiplication in the Laplace domain by the Laplace transform of $p(t)$:

$$\frac{1 - e^{-sT_{DAC}}}{sT_{DAC}}.$$

This is the transfer function of the well known zeroth-order hold block.

APPENDIX C

The most important requirements of a L^{th} order $\Delta\Sigma$ fractional-N PLL are a certain loop bandwidth, f_{BW} in Hz, a minimum phase margin, PM degrees, and an upper limit on the phase noise within the loop bandwidth, $S_{near}(f)$, and at critical frequencies in the stop band, $S_{far}(f)$, both in dBc/Hz. Standard techniques and design equations are well known in prior art to guarantee that phase noise caused by circuit noise in the PLL is less than $S_{near}(f)$ and $S_{far}(f)$ respectively inside and outside the loop bandwidth [8, 18]. Owing to the high-pass shaping of the phase noise caused by the digital $\Delta\Sigma$ quantization, it is necessary to ensure that outside the loop bandwidth, quantization induced phase noise is less than $S_{far}(f)$. This appendix presents a convenient set of equations which can be used to choose the various nominal parameters of the PLL – the gain of the VCO, K_{VCO} in Hz/Volt, the charge pump current, I_{CP} in Amperes, and the loop filter component values, R, C_1, and C_2 in Ohms and Farads – such that the PLL has the requisite bandwidth, phase margin and that the quantization induced phase noise is less than $S_{far}(f)$.

As shown in [7] and [9], the closed-loop transfer function from the reference to the output in the PLL, normalized to unity gain in the pass band is:

$$A_{\phi}(s) = \frac{T(s)}{1+T(s)},$$

where $T(s)$ is the loop transmission of the phase locked loop, and is given by:

$$T(s)=\frac{K(1+s\tau_2)}{s^2\tau_2(1+s\tau_p)},$$

$$where\ \ K=\left(\frac{b-1}{b}\right)\frac{I_{CP}K_{VCO}R}{N+\alpha},\quad \tau_p=\tau_2/b,\quad \tau_2=RC_2.$$

It follows from the above two equations that,

$$A_\phi(s)=\frac{1+s\tau_2}{1+s\tau_2+s\tau_2/K+s\tau_2\tau_p/K}.$$

A design equation for the loop bandwidth can be derived by approximating the denominator of $A_\phi(s)$ as a product of three real poles. As shown below, the zero and the pole of the loop filter are chosen to be sufficiently apart in frequency ($b > 10$) and so that $1/\tau_2 < K < 1/\tau_p$, to ensure that the PLL has good phase margin. Since

$$(1+s\tau_2)(1+s/K)(1+s\tau_p)=1+s\tau_2\left(1+\frac{1}{\sqrt{b}}+\frac{1}{b}\right)+s^2\frac{\tau_2}{K}\left(1+\frac{1}{\sqrt{b}}+\frac{1}{b}\right)+s^3\frac{\tau_2\tau_p}{K},$$

and for $b > 10$, the RHS of the above equation approximates the denominator of $A_\phi(s)$, $A_\phi(s)$ reduces to

$$A_\phi(s)\simeq\frac{1}{(1+s/K)(1+s\tau_p)}. \qquad (9)$$

Therefore, the bandwidth of the PLL is approximately,

$$f_{BW}=\frac{K}{2\pi}=\frac{b-1}{b}\cdot\frac{I_{CP}K_{VCO}R}{2\pi(N+\alpha)}. \qquad (10)$$

The phase margin of the PLL follows from the expression for $T(s)$:

$$PM=\tan^{-1}(\Omega_u\tau_2)-\tan^{-1}(\Omega_u\tau_2/b),$$

where $\Omega_u \approx K$ rad/s is the unity gain frequency of $T(s)$. This equation can be used to choose the zero of the loop filter, τ_2, and b to guarantee a required minimum phase

margin. However, multiple choices of τ_2 and b are capable of ensuring the same phase margin value, complicating the design process. An optimum choice is obtained since for a given b, the most phase margin is obtained when

$$\frac{\partial(PM)}{\partial K}=0\Rightarrow\frac{\sqrt{b}}{\tau_2}=K=\frac{1}{\sqrt{b}\tau_2},$$

i.e., when the zero of the loop filter, the PLL loop bandwidth, and the non-dc pole of the loop filter are in geometric mean with a progression factor $\sqrt{b}$. Under this condition, the phase margin is,

$$PM=\tan^{-1}\left(\frac{\sqrt{b}-1/\sqrt{b}}{2}\right). \tag{11}$$

Therefore, equation (11) can be used to determine the appropriate value of b which ensures the required phase margin. Note that as claimed earlier, $b > 10$ results in good phase margin. Equations (10), and (11), along with the definitions of τ_2 and b provide a convenient set of design equations to choose the components of the PLL:

$$\begin{aligned} RC_2&=\frac{\sqrt{b}}{K}=\frac{\sqrt{b}}{2\pi f_{BW}},\\ C_1&=\frac{C_2}{b-1}. \end{aligned} \tag{12}$$

The explicit expression for $A_\phi(s)$ given in (9) can be used in conjunction with (3) to determine if the quantization induces phase noise is less than the allowable limit, $S_{far}(f)$. The allowable phase noise limits, $S_{far}(f)$, are often the toughest for offset frequencies, $f > 1/2\pi\tau_p$. Therefore, the design process is simplified by approximating the magnitude response of $A_\phi(s)$ for frequencies $f > 1/2\pi\tau_p$:

$$\left|A_\phi(j2\pi f)\right| \simeq \frac{f_{BW}}{f^2\tau_p} = \frac{\sqrt{b}f_{BW}^2}{f^2}, \qquad \forall \frac{1}{2\pi\tau_p} < f.$$

It follows from (3) and the above approximation that the phase noise caused by ΔΣ quantization for offset frequencies, $f > 1/2\pi\tau_p$, is

$$S_\phi^\beta(j2\pi f) = \beta^2 \frac{\pi^2 b}{3f_{ref}} \left|2\sin\left(\frac{\pi f}{f_{ref}}\right)\right|^{2(L-1)} \left|\frac{f_{BW}}{f}\right|^4 \qquad rad^2/Hz. \tag{13}$$

Suppose that the PLL output is represented as

$$x(t) = A\sin\left(2\pi f_{PLL}t + \phi_0 + \phi_{PLL}(t)\right),$$

where A and ϕ_0 are arbitrary constants determining the amplitude of the PLL output and an initial phase, and $\phi(t)$ is the PLL phase noise whose one-sided power spectral density is given by (13). It follows that the one-sided power spectrum of the PLL output is,

$$S_{xx}(j2\pi f) = A^2\pi\cdot\delta(f - f_{PLL}) + \frac{1}{4}A^2\cdot S_\phi^\beta(j2\pi f),$$

where $\delta(f)$ is the Dirac-delta function. Therefore, the PLL phase noise expressed relative to the carrier power in units of dBc/Hz is:

$$\begin{aligned} S(j2\pi f) &= 10\cdot\log_{10}\left[\frac{S_\phi^\beta(j2\pi f)}{4\pi}\right] \\ &= 10\cdot\log_{10}\left|\frac{\pi b\beta^2}{f_{ref}}\cdot\left(\frac{f_{BW}}{f}\right)^4\cdot\sin^{2(L-1)}\left(\frac{\pi f}{f_{ref}}\right)\right| \qquad dBc/Hz. \end{aligned} \tag{14}$$

This expression can be compared with $S_{far}(f)$ to determine the bandwidth, the order, L, and the reference frequency to meet the phase noise requirements of the PLL.

CHAPTER ACKNOWLEDGEMENTS

The text of this chapter, in partial or in full, is under review for publication in the *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*. The dissertation author is the primary researcher. Ian Galton supervised the research which is the subject of this chapter. The author is grateful to Sheng Ye, Ashok Swaminathan, and Eric Siragusa for their help in reviewing this chapter.

REFERENCES

1. S. Pamarti, L. Jansson, I. Galton, "A Wideband 2.4 GHz Delta-Sigma Fractional-*N* PLL with 1 Mb/s In-Loop Modulation," *IEEE Journal of Solid State Circuits*, to appear.

2. G. C. Gillette, "Digiphase Synthesizer," *Proceedings of 23rd Annual Frequency Control Symposium*, pp. 201-210, 1969.

3. N. B. Braymer, "Frequency synthesizer," United States Patent no. 3,555,446, January 12, 1971.

4. N. King, "Phase locked loop variable frequency generator" United States Patent no. 4,204,174, May 20, 1980.

5. J. A. Crawford, *Frequency Synthesizer Design Handbook*, Artech House Inc., 1994.

6. W. F. Egan, *Frequency Synthesis by Phase Lock*, second edition, Wiley Interscience, 2000.

7. I. Galton, "Delta-sigma fractional-*N* phase-locked loops," *Phase-Locking in High-Performance Systems : From Devices to Architectures*, Edited by B. Razavi, John Wiley & Sons, February 2003.

8. M. H. Perrott, M. D. Trott, C. G. Sodini, "A Modeling Approach for D-S Fractional-N Frequency Synthesizers Allowing Straightforward Noise Analysis," *IEEE Journal of Solid State Circuits*, vol. 37, no. 8, pp. 1028-38, August 2002.

9. F. M. Gardner, *Phaselock Techniques*, second ed., John Wiley & Sons, 1979.

10. I. Galton, "One-bit dithering in delta-sigma modulator-based D/A conversion," *Proc. of the IEEE International Symposium on Circuits and Systems*, 1993.

11. S. Norsworthy, R. Schreier, G. C. Temes, *Delta-Sigma Data Converters: Theory, Design, and Simulation*, IEEE Press, 1996.

12. I. Galton, "Spectral shaping of circuit errors in digital-to-analog converters," *IEEE Trans. Circuits Systems-II*, vol. 44, no. 10, pp. 808-17, Oct. 1997.

13. R. Adams, K. Q. Nguyen, "A 113-dB SNR Oversampling DAC with Segmented Noise-Shaped Scrambling", *IEEE JSSCC*, vol. 33, no. 12, pp. 1871-1878, Dec. 1998.

14. A. Fishov, E. Siragusa, J. Welz, E. Fogleman, I. Galton, "Segmented mismatch-shaping D/A conversion," *Proc. of the IEEE International Symposium on Circuits and Systems*, May 2002.

15. J. Welz, I. Galton, "Necessary and sufficient conditions for mismatch shaping in multi-bit digital-to-analog converters," *IEEE Transactions on Circuits and Systems – II: Analog and Digital Signal Processing*, vol. 49, no. 12, pp. 748-759, December 2002.

16. J. Welz, I. Galton, "The mismatch-shaping noise PSD from a tree-structured DAC in a second-order delta-sigma modulator with a midscale input," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, May 7-11, 2001, pp. 2625-2628.

17. J. Welz, I. Galton, "A tight signal-band power bound on mismatch noise in a mismatch-shaping DAC," under review in *IEEE Transactions on Information Theory*.

18. J. Craninckx, M. S. J. Steyaert, "A fully integrated CMOS DCS-1800 frequency synthesizer," *IEEE Journal of Solid State Circuits*, vol. 33, pp. 2054-2065, December 1998.

Chapter 3

One-bit Dithering in Digital Delta-Sigma Modulators

Sudhakar Pamarti, Jared Welz, and Ian Galton

Abstract— Theoretical sufficient conditions which ensure that one-bit least significant bit dither eliminates limit cycles and resultant spurious tones in general single stage and multi-stage digital delta-sigma (ΔΣ) modulators are presented. A large class of popular ΔΣ modulators in which one-bit dither eliminates limit cycles are identified by applying the sufficient conditions. Means of imparting spectral shape to the dither while eliminating limit cycles are presented.

I. INTRODUCTION

Digital delta-sigma (ΔΣ) modulators are widely used in high precision over-sampled digital-to-analog (D/A) converters and fractional-*N* phase locked loops. They are however susceptible to periodic limit cycles, causing significant spurious tones in the power spectra of the outputs of such systems. As observed in [1], spurious tones in some digital ΔΣ modulators may be suppressed at the expense of a simple linear feedback shift register and little extra digital logic by adding one-bit dither to the least significant bit (LSB) of the ΔΣ modulator input.

This paper presents theoretical conditions which help determine if one-bit dither eliminates limit cycles in a given digital ΔΣ modulator. These conditions promise to be of immense value to the designer who had so far only two options to suppress spurious tones in a digital ΔΣ modulator:

- Adding large amounts of dither to span the quantization step size [2] or,
- Relying on simulations to choose ΔΣ modulators that might result in low spurious tones.

As illustrated in this paper, armed with the presented conditions, the designer would be able to pick either a single stage or a multi-stage digital ΔΣ modulator suitable to the target application, add one-bit dither to it and be assured of the suppression of spurious tones in the digital ΔΣ modulator output. This paper applies the sufficient conditions to determine if one-bit dither suppresses spurious tones in many of the popular digital ΔΣ modulators. It also suggests means to spectrally shape the one-bit dither so that it does not interfere with the signal that is being converted into the analog domain.

Structure of the paper

Section I derives the aforementioned sufficient conditions in the form of a single theorem in the context of a generic ΔΣ modulator with a single requantizer. Section II illustrates the application of the sufficient conditions to some popular ΔΣ modulators which are special cases of the generic delta-sigma modulator. Section III suggests how to impart frequency domain shaping to the dither signal while eliminating spurious tones. Section IV extends the results to present sufficient conditions to use dither to remove spurious tones in Multi-stAge noise SHaping (MASH) architectures. Appendices A and B prove the theorem in Section I in a general context such that the results can be reused in Sections III and IV. Appendix C

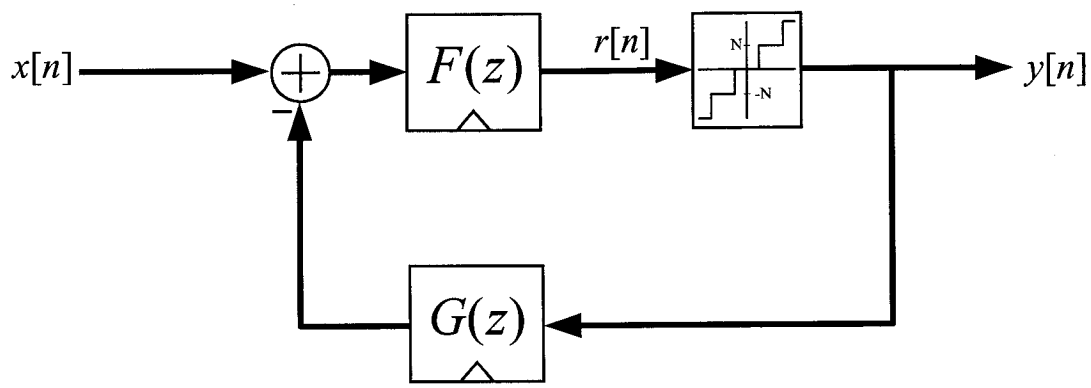

Fig. 3.1: A generic single stage digital ΔΣ modulator.

provides proofs for assorted theorems in Sections I and IV.

I. Dither in Single Stage Delta-Sigma Modulators

A. Need for dither

Fig. 3.1 shows a generic digital ΔΣ modulator, which emphasizes the negative feedback nature of digital ΔΣ modulation. The system consists of a causal forward transmission filter, $F(z)$, followed by a non-overloading digital requantizer, and a feedback filter, $G(z)$, which filters the output of the requantizer and feeds it back. The impulse responses of $F(z)$, and $G(z)$, denoted $f[n]$ and $g[n]$ respectively, are integer valued. The requantizer is a mid-tread requantizer[1] of step size N. The non-overloading mid-tread requantization operation is defined as:

$$y[n] = N\left\lfloor \frac{r[n]}{N} + \frac{1}{2} \right\rfloor,$$

[1] The results presented in this paper are applicable to mid-rise requantization with minor modifications; hence, they are not discussed here.

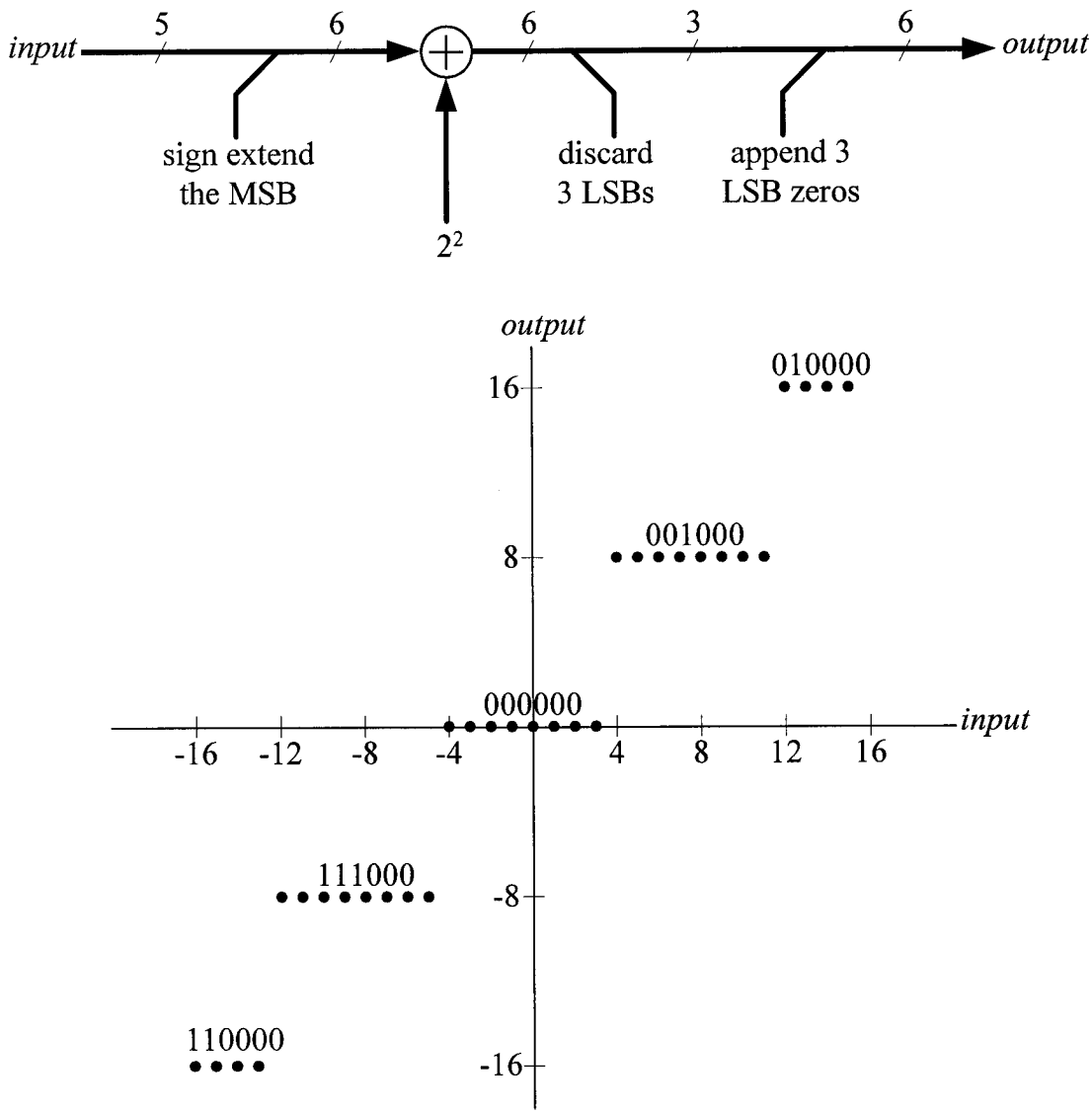

Fig. 3.2: Sample mid-tread requantization of a binary, 2's complement sequence.

where $r[n]$ and $y[n]$ are respectively the input and output of the requantizer shown in Fig. 3.1. Fig. 3.2 illustrates mid-tread requantization of binary 2's-complement representations of integers for $N=8$. Traditionally, the requantization operation is modeled without any approximation using an additive source of error, $e[n]$:

$$e[n] \triangleq y[n]-r[n]=\frac{N}{2}-\left\langle \frac{r[n]}{N}+\frac{1}{2}\right\rangle, \tag{1}$$

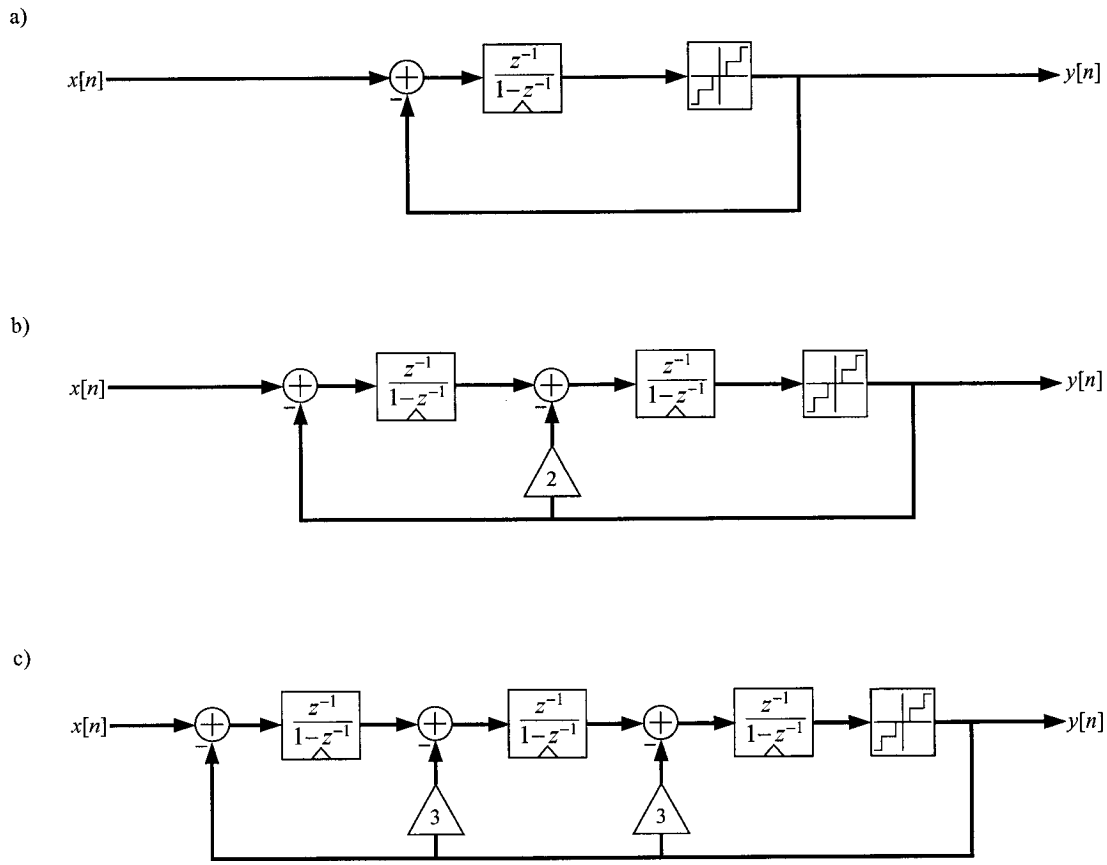

Fig. 3.3: Example digital ΔΣ modulators – (a) The first-order ΔΣ modulator (b) The second-order dual-loop ΔΣ modulator (c) The third-order ΔΣ modulator.

where $\langle x \rangle = x - \lfloor x \rfloor$. Consequently, the Z-transform of the output of the delta sigma modulator can be related to the Z-transforms of the signal, and the additive error source, $e[n]$, as:

$$Y(z) = X(z)\underbrace{\frac{F(z)}{1+F(z)G(z)}}_{STF(z)} + E(z)\underbrace{\frac{1}{1+F(z)G(z)}}_{NTF(z)}. \qquad (2)$$

The second term in (2) is referred to in published literature as the *quantization noise* of the digital ΔΣ modulator. The filters, $F(z)$ and $G(z)$, are usually chosen such that the noise transfer function, $NTF(z)$, in (2) de-emphasizes the quantization noise power in a

Table 3.1: Details of the example ΔΣ modulators in Fig. 3.3.

Example Multi-bit Single Quantizer Digital Delta-Sigma Modulators					
Name of ΔΣ modulator	Figure reference	Forward transmission filter		Feedback filter	
		Transfer function, F(z)	Impulse response, f[n]	Transfer function, G(z)	Impulse response, g[n]
1st Order	2(a)	$\frac{z^{-1}}{1-z^{-1}}$	$u[n-1]$	1	$\delta[n]$
2nd Order	2(b)	$\left(\frac{z^{-1}}{1-z^{-1}}\right)^2$	$(n-1)u[n-2]$	$2z-1$	$2\delta[n+1]$ $-\delta[n]$
3rd Order	2(c)	$\left(\frac{z^{-1}}{1-z^{-1}}\right)^3$	$\frac{(n-2)(n-1)}{2}u[n-3]$	$3z^2-3z+1$	$3\delta[n+2]$ $-3\delta[n+1]$ $-\delta[n]$

certain band of frequencies occupied by the ΔΣ modulator input. Many of the digital ΔΣ modulators reported in literature are special cases of the generic form of Fig. 3.1. Fig. 3.3 shows some popular ΔΣ modulators (1st, 2nd and 3rd order multi-bit low pass delta-sigma modulators) and Table 3.1 shows that the corresponding forward transmission and feedback filters are integer valued as described above. The filters, $F(z)$ and $G(z)$, in each of these example digital ΔΣ modulators are such that quantization noise is high-pass shaped.

The nature of the quantization noise – whether it has spurious tones or not *etc.*, – depends on the statistics of $e[n]$. It is often assumed that $e[n]$ is white, independent of the ΔΣ modulator input, $x[n]$, and uniformly distributed over $\{-N/2+1,...,0,...,N/2\}$. Such assumptions enable the designer to use equation (2)

and make quantitative predictions crucial to D/A system design *e.g.*, the shape and magnitude of the power spectral density of the output of the delta-sigma modulator, signal-to-noise ratio in a given frequency band *etc.* However, as shown below, $e[n]$ depends strongly on the ΔΣ modulator input, $x[n]$, and often has significant spurious tones rendering the above assumptions baseless and flawed.

Since the input to the requantizer in Fig. 3.1 can be expressed as

$$r[n] = x[n] * f[n] - y[n] * g[n] * f[n],$$

where "*" is the convolution operator, it follows from equation (1) that

$$e[n] = \frac{N}{2} - N\left\langle \frac{x[n] * f[n] - y[n] * f[n] * g[n]}{N} + \frac{1}{2} \right\rangle. \tag{3}$$

Owing to the digital nature of the system, the input and output of the ΔΣ modulator, $x[n]$ and $y[n]$, can be assumed to take on integer values only. Moreover, note that the mid-tread requantization implies that $y[n]$ takes on only integer multiples[2] of N. Since $f[n]$, $g[n]$ are also integer valued by assumption, (3) can be simplified as:

$$e[n] = \frac{N}{2} - N\left\langle \frac{x[n] * f[n]}{N} + \frac{1}{2} \right\rangle. \tag{4}$$

The above equation supports the earlier claim that $e[n]$ is a non-linear function of the ΔΣ modulator input. It has significant spurious content as well. For example, if the input of the 2nd order ΔΣ modulator were a small constant, the two accumulators, represented by $f[n]$ in (4), cause $e[n]$ to repeat at a certain period determined by the

---

[2] **Note:** Consider the example where $r[n] = 25$ is truncated by a mid-tread quantizer of step size $N = 8$. In some published work, the output of the quantizer is implied to be

...footnote continued on next page

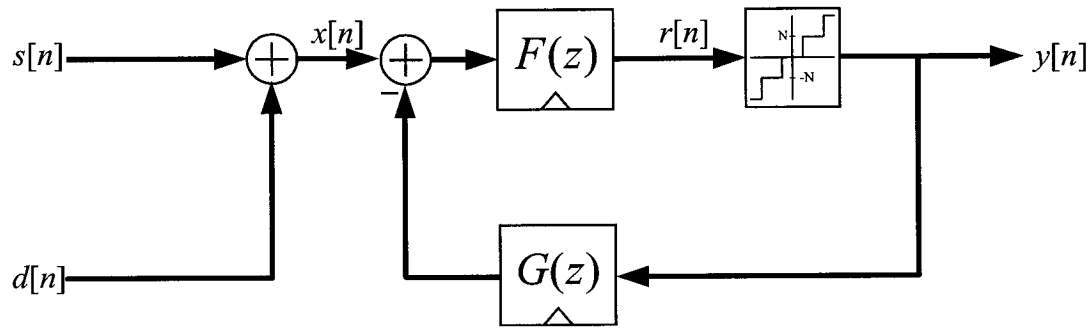

Fig. 3.4: A generic dithered digital ΔΣ modulator.

constant input. These repetitions are called limit cycles and cause spurious tones in the power spectral densities of the quantization noise and the ΔΣ modulator output. If the input were a periodic signal, similar periodicities are exhibited.

If a one-bit random sequence, $d[n]$, were added to the ΔΣ modulator input as shown in Fig. 3.4, $e[n]$ may no longer repeat periodically, thereby suppressing spurious tones. The sequence $s[n]$ represents the desired signal that needs to be requantized using the ΔΣ modulator. The sequence $d[n]$ comprises one-bit samples which are independent of themselves and the desired signal. The sequence $d[n]$ is called *dither* and its addition to the input of the ΔΣ modulator is called *dithering*. Fig. 3.5 illustrates the effects of one-bit dither on the quantization noise of digital ΔΣ modulators. Fig. 3.5(a) shows simulated power spectral densities of the quantization noise of the 2nd order digital ΔΣ modulator shown in Fig. 3.3(b) without and with one-bit dither, for radian frequencies from 0 to π. The elimination of spurious tones when one-bit dither is used is evident from the figure. Fig. 3.5(b) shows similar plots for the 3rd order ΔΣ

---

the value $y[n] = 3$ instead of $y[n] = 24 = 3*8$. For purposes of simplicity, this paper
...footnote continued on next page

modulator shown in Fig. 3.3(c).

Additive dither, $d[n]$, can annul the statistical dependence of $e[n]$ on the $\Delta\Sigma$ modulator input, and validate the assumptions of uniformity and whiteness of $e[n]$ for a large class of $\Delta\Sigma$ modulators. The rest of the section analyzes requantization within the digital $\Delta\Sigma$ modulator in the presence of the one-bit dither. The analysis is in the context of the generic dithered digital $\Delta\Sigma$ modulator shown in Fig. 3.4.

Before proceeding further, it would be useful to clarify some nomenclature. This paper refers to the additive error source $e[n]$ as *requantization error*. This is not to be confused with quantization noise.

---

uses the latter convention.

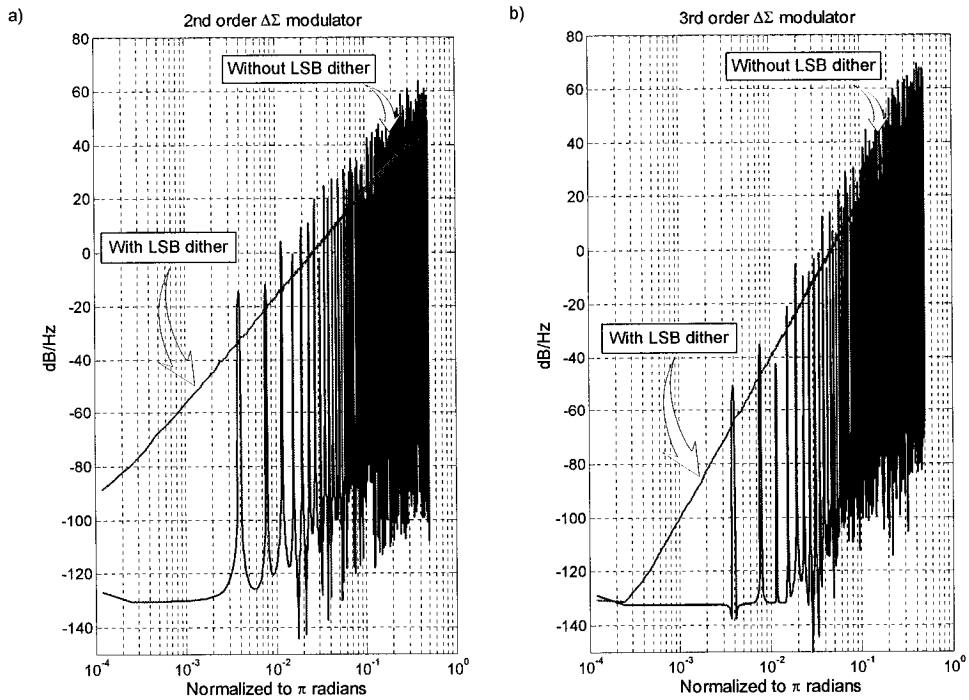

Fig. 3.5: Elimination of spurious tones using one-bit dither in digital ΔΣ modulators.

B. Dithered Requantization

Suppose that the dither, $d[n]$, is a sequence of independent identically distributed (*iid*) random variables. Suppose that $d[n]$ is independent of the desired signal as well. Assume that each sample of this *iid* dither takes on one of two consecutive integer values with equal probability:

$$P(d[n]=m)=\begin{cases}0.5, & m=0,\\ 0.5, & m=1.\end{cases} \tag{5}$$

Such dither (henceforth referred to as *LSB dither*) can be readily implemented using a long binary, maximal length pseudo-random sequence using a simple linear feedback shift register.

Let n_0 be the sample time when the system is "turned on" *i.e.* all the signals in the system are assumed to be zero for $n \leq n_0, n \in \mathbb{Z}$ It follows from Fig. 3.4 and by substituting $x[n] = s[n] + d[n]$ in (4) that for each $n \geq n_0, n \in \mathbb{Z}$, the requantization error can now be written as:

$$e[n] = \frac{N}{2} - N\left\langle \frac{z[n]}{N} + \frac{1}{2} \right\rangle$$
$$\text{where } z[n] = \sum_{m=n_0}^{n} s[m]f[n-m] + \sum_{m=n_0}^{n} d[m]f[n-m] \tag{6}$$

The second term in the expression for $z[n]$ in (6) is a linear combination of a number of *iid* random variables. As time progresses, this term includes an increasingly large number of *iid* random variables. Intuitively, this very random term gives $z[n]$ the desired properties of uniformity, whiteness and independence from all $s[n]$. This section derives sufficient conditions under which, as more time elapses since the "start-up" of the system, the requantization error $e[n]$ converges in distribution to a sequence, $\tilde{e}[n]$, with the following properties:

- *Uniformity*: For any finite integer n, $\tilde{e}[n]$ is uniformly distributed over the range of values $\{-N/2 + 1,\ldots,0,\ldots,N/2\}$
- *Signal-independence*: $\tilde{e}[n]$ is independent of $s[l]$ for any $n,l \in \mathbb{Z}$ and $(n - l)$ finite.
- *Dither-independence*: $\tilde{e}[n]$ is independent of $d[l]$ for any $n,l \in \mathbb{Z}$ and $(n - l)$ finite.
- *Pair-wise independence*: $\tilde{e}[n]$ is independent of $\tilde{e}[n-p]$ for all integers $p \neq 0$.

- *Whiteness*: $\tilde{e}[n]$ is wide sense stationary and uncorrelated with $\tilde{e}[n-p]$ for all integers $p \neq 0$.

Note: Henceforth, "the requantization error $e[n]$ has a property x" will mean "as $n_0 \to -\infty$, $e[n]$ converges in distribution to $\tilde{e}[n]$ which satisfies property x" where x is one of the above five properties. Moreover, all these above properties are together referred to as *desired properties*.

Once it is proved that $e[n]$ has the *desired properties* for a particular dithered $\Delta\Sigma$ modulator, expressions for the power spectral density (PSD) of $y[n]$, SNR *etc.*, can be derived from (2). To illustrate, the output of the 2nd order digital $\Delta\Sigma$ modulator shown in Fig. 3.3(b) is

$$y[n] = s[n-2] + d[n-2] + e[n] - 2e[n-1] + e[n-2].$$

If the requantization error, $e[n]$, has the desired properties then, the terms in the RHS of the above expression are independent of each other. Therefore, not only does the PSD of $y[n]$ have no spurious tones, it can also be analytically shown to be

$$S_{yy}\left(e^{jw}\right) = S_{ss}\left(e^{jw}\right) + \sigma_{dd}^2 + \left|1 - e^{-jw}\right|^4 \sigma_{ee}^2,$$

where $S_{ss}(e^{jw})$ is the PSD of the desired signal[3], and σ_{dd}^2, σ_{ee}^2 are respectively the mean square values of the LSB dither and the requantization error. Note that the 2nd order high pass shaping of the requantization error is evident from the above equation. Similarly, the PSD of the output of a general digital $\Delta\Sigma$ modulator, $y[n]$, can be shown using (2) to be:

---

[3] It is assumed that the desired signal is wide sense stationary.

$$S_{yy}\left(e^{jw}\right)=\left|STF\left(e^{jw}\right)\right|^2 S_{ss}\left(e^{jw}\right)+\left|STF\left(e^{jw}\right)\right|^2 \sigma_{dd}^2+\left|NTF\left(e^{jw}\right)\right|^2 \sigma_{ee}^2 \qquad (7)$$

where $\sigma_{dd}^2=1/4$ and $\sigma_{ee}^2=\left(N^2-1\right)/12$. Theorem 1 presents conditions on $f[n]$ which are sufficient for the properties of *uniformity*, *signal independence*, *dither independence*, *pair-wise independence* and *whiteness*. It also derives expressions for the time-averaged mean, auto-covariance of the requantization error and shows that $\sigma_{ee}^2=\left(N^2-1\right)/12$.

Theorem 1: Suppose that for every integer $p > 0$, and any integers k_1, k_2, such that $k_1 + k_2 \neq 0$, and $0 \leqslant k_1, k_2 \leqslant N-1$, at least one of the following is true:

1. The sequence $(k_1 f[r]+k_2 f[r+p]) \bmod N$ does not converge to zero as $r \to \infty$
2. A non-negative integer $r_{1,2} \neq p$ exists such that

   $$\left(k_1 f[r_{1,2}]+k_2 f[r_{1,2}+p]\right) \bmod N = N/2$$

3. A non-negative integer $r_2 < p$ exists such that $\left(k_2 f[r_2]\right) \bmod N = N/2$

Suppose also that for at least one $p > 0$, the first condition is true for all integers k_1, k_2, such that $k_1+k_2 \neq 0$, and $0 \leq k_1, k_2 \leq N-1$. Then, requantization error, $e[n]$, has the properties of *uniformity*, *signal-independence*, *dither-independence*, *pair-wise independence* and *whiteness*. Moreover, $e[n]$ has time-averaged mean and auto-covariance of

$$M_e \triangleq \lim_{L\to\infty}\frac{1}{L}\sum_{n=n_0}^{L+n_0-1} e[n] = \frac{1}{2},$$
$$C_{ee}(p) \triangleq \lim_{L\to\infty}\frac{1}{L}\sum_{n=n_0}^{L+n_0-1}\big(e[n]-M_e\big)\big(e[n-p]-M_e\big) = \underbrace{\frac{N^2-1}{12}}_{\sigma_{ee}^2}\delta[p], \tag{8}$$

where $\delta[k]$ is the Kronecker[4] delta function.

Proof: Setting $A(z) = H(z) = F(z)$, and applying Theorem A1 and corollary 1 of Theorem A1 for every $p > 0$, and then applying Theorem A2 proves that in terms of ensemble statistics, $e[n]$ has the properties of *uniformity*, *signal-independence*, *dither-independence*, *pair-wise independence* and *whiteness*. The application of Theorem A3 proves that $e[n]$ has these properties in a time-averaged sense as well, with mean and auto-covariance as given by (8).

■

While seemingly cumbersome, the conditions imposed on $f[n]$ by Theorem 1 can often be easily verified. The next section illustrates how this can be achieved for a few popular delta-sigma modulators.

II. Application of Results to Popular Delta-Sigma Modulators

All the ΔΣ modulators considered in this section are assumed to have enough output levels to ensure that their requantizers do not overload. Moreover, the requantization step size is assumed to be a power of 2, $N = 2^M$, where M is a positive

integer. This assumption is not in the least restrictive as it is a particularly popular choice for binary, two's-complement implementations of the digital ΔΣ modulators.

To apply the results from the previous section to determine if LSB dither can ensure that the requantization error in a particular delta-sigma modulator has the *desired properties*, the following needs to verified:

- the delta-sigma modulator meets the constraints of the presented generic form, and
- the forward transmission filter's impulse response, $f[n]$, satisfies the conditions of Theorem 1.

This procedure shall be illustrated in detail for a 3rd order, multi-bit, digital ΔΣ modulator. Results for other popular ΔΣ modulators will be quoted with the respective proofs pushed to Appendix C.

A. Multi-bit Third Order Delta-Sigma Modulator

The forward transmission filter, $F(z)$, the feedback filter, $G(z)$, and their impulse responses are given in Table 3.1 and both $f[r]$ and $g[r]$ are clearly integer valued. If the requantizer has enough output levels to avoid overload, then this delta-sigma modulator satisfies the constraints of the generic form.

Claim: Sequence $f[r]$ satisfies condition 1 of Theorem 1 for all $p > 0$.

Proof: Suppose to the contrary that condition 1 is not satisfied for some $p > 0$. Then

---

[4] $\delta[k] = \begin{cases} 1, & k = 0, k \in \mathbb{Z}; \\ 0, & k \neq 0, k \in \mathbb{Z}; \end{cases}$

for that particular p and some $r_0 \in \mathbb{Z}$,

$$k_1 \frac{[r-2][r-1]}{2} + k_2 \frac{[r+p-2][r+p-1]}{2} = mN, \quad m \in \mathbb{Z}, \quad \forall r \geq r_0 \geq 3. \tag{9}$$

The set of equations, (9), in the three "unknowns" k_1, k_2, and p can be reduced by considering equations, (9), for any three consecutive values of r greater than or equal to r_0. This results in the following set of solutions for k_1, k_2, and p:

$$\begin{aligned} k_1 + k_2 &= m_1 N, \\ k_2 p &= m_2 N, \\ p &= 1 + 2m_3, \\ \text{where} \quad & m_1, m_2, m_3 \in \mathbb{Z}. \end{aligned} \tag{10}$$

However since $k_2 < N = 2^M$, and $p = 1 + 2m_3$ is odd, the equations, (10), can not all be simultaneously true. So the claim is proved by contradiction.

■

Since the conditions of Theorem 1 are satisfied, the one-bit LSB dither guarantees that $e[n]$ has properties of *uniformity*, *signal independence*, *dither independence*, *pair-wise independence* and *whiteness*. Moreover, its time-averaged mean and auto-covariance are given by equation (8). It should be noted that this result is not restricted to the 3rd order ΔΣ modulator shown in Fig. 3.3(c). It is applicable to any ΔΣ modulator with the same $F(z)$, $G(z)$ and a non-overloading mid-tread requantizer whose step-size is a power of 2.

B. Multi-bit L^{th} Order Delta-Sigma Modulator

The forward transmission filter, $F(z)$ and the feedback filter, $G(z)$ of the L^{th} order delta-sigma modulator can be shown to be

$$F_L(z)=\left(\frac{z^{-1}}{1-z^{-1}}\right)^L;G_L(z)=(1-z)^L-z^L. \tag{11}$$

The corresponding impulse responses, $f_L[r]$ and $g_L[r]$ are integer valued. If the delta-sigma modulator has enough output levels to avoid overload, it satisfies the constraints of the generic delta-sigma modulator.

Theorem C1 proves that $f_1[r]$ does not satisfy the conditions of Theorem 1. Therefore, one-bit LSB dither **does not** guarantee that the requantization error in a non-overloading, 1st order delta-sigma modulator has the *desired properties*. On the other hand, Theorem C2 proves that $f_2[r]$ satisfies all the conditions of Theorem 1. Similarly, Theorem C3 proves that $f_L[r]$ satisfies the conditions of Theorem 1 for $L>3$. Therefore, one-bit LSB dither **does** guarantee that the requantization error, $e[n]$, in a non-overloading, L^{th} order digital $\Delta\Sigma$ modulator ($L\geq 2$) has the *desired properties*. Moreover, its time-averaged mean and auto-covariance are given by equation (8).

The above results suggest that if the one-bit *iid* dither undergoes two or more integrations on the way to the requantizer, then the error $e[n]$ has the *desired properties*. By considering successive samples of $z[n]$ in equation (6), an interesting insight into this result can be obtained. It follows from (6) that,

$$Z(z)=F(z)\cdot\left[S(z)+D(z)\right],$$

where $Z(z)$, $S(z)$, and $D(z)$ are the Z-transforms of $z[n]$, $s[n]$, and $d[n]$ respectively. Multiplying the above expression with $(1-z^{-1})$ results in:

$$(1-z^{-1})Z(z)=\underbrace{(1-z^{-1})F(z)}_{Q(z)}\left[S(z)+D(z)\right],$$

The relation between the samples of $z[n]$ at two consecutive time indices, $n_* - 1$, and n_*, can be obtained by computing the inverse Z-transform of the above equation and substituting $n = n_*$:

$$z[n_*] = z[n_* - 1] + \sum_{m=n_0}^{n} d[m]q[n-m]\Big|_{n=n_*} + s[n] * q[n]\big|_{n=n_*},$$

where $q[n]$ is the impulse response of $Q(z)$. If the filtering undergone in the forward path of the $\Delta\Sigma$ modulator by the one-bit dither *i.e.*, $F(z)$, is only one (delayed) integration, then $q[n] = \delta[n]$, and therefore, $z[n_*]$ and $z[n_* - 1]$ are "separated" by only one new random variable, $d[n_*]$, which has the range of only an LSB –

$$z[n_*] = z[n_* - 1] + d[n_*] + (\text{signal dependent terms}).$$

Since the requantization error at any time index n is a non-linear function of $z[n]$,

$$e[n] = \frac{N}{2} - \left\langle \frac{z[n]}{N} + \frac{1}{2} \right\rangle,$$

$e[n_*]$ and $e[n_* - 1]$ may not be independent even an infinite time after the system's "start-up". However, if the dither undergoes two or more integrations on its way to the requantizer, as more elapses since "start-up", $e[n_*]$ and $e[n_* - 1]$ are separated by an increasing number of random variables. For instance, if dither undergoes two integrations,

$$z[n_*] = z[n_* - 1] + \sum_{m=n_0}^{n_*} d[m] + (\text{terms dependent on } s[n]).$$

The increasing number of random variables influencing the requantization operation imparts the *desired properties* to $\Delta\Sigma$ modulators of order $L \geq 2$. This insight inspires

the exploration of dithered ΔΣ modulator architectures, which impart frequency domain shaping to the dither signal.

III. SHAPED DITHER ARCHITECTURES

While the previous sections suggest that *iid* dither can be used to remove spurious tones in many ΔΣ modulators, the dither is present in the output of the delta-sigma modulator along with the desired signal; it undergoes the same overall filtering as the desired signal, as shown in (7). This limits the in-band SNR, particularly when the number of bits in the input signal is small. For instance, consider the 2nd order, non-overloading ΔΣ modulator shown in Fig. 3.3(b) with an over-sampling ratio of 8 and a 10-bit wide input signal $s[n]$. As mentioned before, the output of the 2nd order ΔΣ modulator can be shown to be

$$y[n] = s[n-2] + \underbrace{d[n-2]}_{\text{dither}} + \underbrace{e[n] - 2e[n-1] + e[n-2]}_{\text{filtered requantization error}}.$$

With the LSB being unity, the amplitude of the largest sine-wave input that can be handled by this ΔΣ modulator is approximately $2^{10}/2 = 512$, corresponding to a power of $0.5*512^2 = 131072$. On the other hand, the total dither power is $\sigma_{dd}^2 = 1/4$ spread uniformly over the whole spectrum. The over-sampling ratio of 8 implies that dither limits the in-band SNR to approximately

$$10\log(131072) - 10\log\left(\frac{(1/4)}{8}\right) \approx 66 \text{ dB}.$$

Increasing the number of input bits to 14 implies that the in-band SNR is limited by dither to 78 dB instead. The amount of undesirable dither power can be made

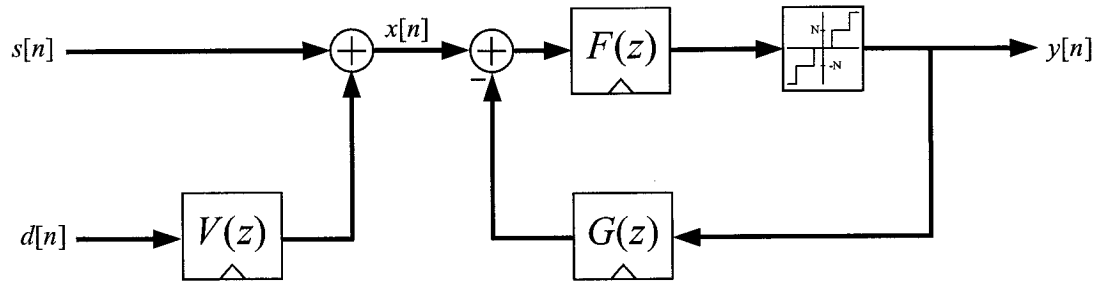

Fig. 3.6: A scheme to introduce shaped dither into the generic ΔΣ modulator.

arbitrarily small by increasing the number of bits representing the input of the delta-sigma modulator. However this could be wasteful in terms of the additional digital circuitry needed. Moreover, digital ΔΣ modulators are employed in fractional-N phase locked loops where the *modulator error* undergoes integration [3]. In such situations, the integrated *iid* dither could severely degrade the phase noise of the phase locked loop output.

An attractive alternative to increasing the number of input bits is to force most of the dither power out of the frequency band of interest. For instance, in a low-pass delta-sigma modulator, if a high-pass spectral shape could be imparted to the dither in the output, the degradation of the in-band SNR due to the dither would be less severe. In general, desired shaping could be imparted to the dither in the output of the delta-sigma modulator by filtering LSB dither before adding the result to the desired signal as shown in Fig. 3.6. The resultant delta-sigma modulator output is

$$Y(z) = \underbrace{STF(z)S(z)}_{\text{desired signal output}} + \underbrace{V(z)STF(z)D(z)}_{\text{shaped dither in output}} + \underbrace{NTF(z)E(z)}_{\text{shaped quantization error}} \tag{12}$$

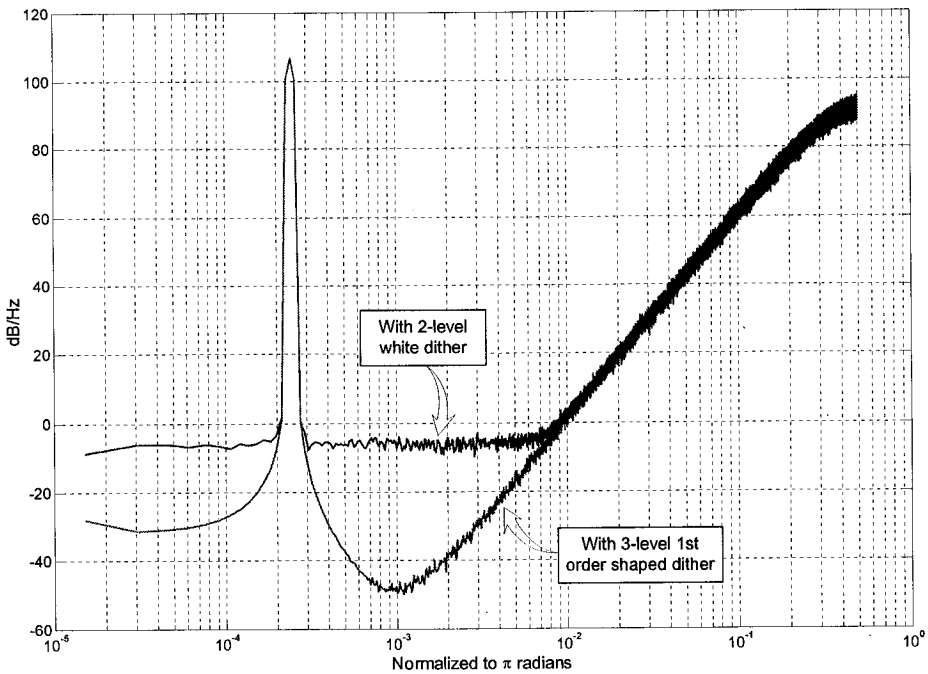

Fig. 3.7: Illustration of increase of the in-band SNR using shaped dither.

where $V(z)STF(z)$ is the net shaping imparted to the dither. Such systems shall be referred to as *shaped dither systems*. For instance, a simple discrete differentiator $V(z)=(1-z^{-1})$ can impart high-pass shaping to the dither. The filtered dither spans only 3 LSBs and barely increases the required digital circuitry making it a very attractive choice. In the case of the aforementioned examples, 1st order shaping implies that dither limits the in-band SNR to 79 dB instead of 66 dB with a 10-bit wide input, and limits the in-band SNR to 91 dB instead of 78 dB with a 14-bit wide input. Fig. 3.7 illustrates the effect of shaping the dither on a 3rd order digital ΔΣ modulator shown in Fig. 3.3(c). The figure shows simulated power spectral densities of the output of the 3rd order digital ΔΣ modulator for the cases of $V(z) = 1$ and

$V(z)=(1-z^{-1})$ respectively, for radian frequencies from 0 to π. The noise floor due to the dither is imparted a high pass shape with $V(z)=(1-z^{-1})$, while no spurious tones are introduced.

The relevant question is whether the requantization error still has the properties of *uniformity*, *signal independence*, *dither independence*, *pair-wise independence* and *whiteness* and if so, for which delta-sigma modulators. If $e[n]$ were to indeed have the *desired properties*, the PSD of $y[n]$ can be derived from (12) to be

$$S_{yy}(e^{jw})=\left|STF(e^{jw})\right|^2 S_{ss}(e^{jw})+\left|STF(e^{jw})\right|^2\left|V(e^{jw})\right|^2\sigma_{dd}^2+\left|NTF(e^{jw})\right|^2\sigma_{ee}^2 \qquad (13)$$

where $\sigma_{dd}^2=1/4$ and $\sigma_{ee}^2=(N^2-1)/12$ represent the variance of the *iid* dither sequence and the requantization error respectively.

To determine if $e[n]$ has the *desired properties*, we need to proceed just as in Section II. Equation (6) can be rewritten as

$$\begin{aligned} e[n]&=\frac{N}{2}-N\left\langle\frac{z[n]}{N}+\frac{1}{2}\right\rangle \\ \text{where}\quad z[n]&=\sum_{m=n_0}^{n}s[m]f[n-m]+\sum_{m=n_0}^{n}d[m]h[n-m], \end{aligned} \qquad (14)$$

where $h[n]$ is the impulse response of the fictitious filter $H(z)\triangleq V(z)F(z)$. The following theorem presents sufficient conditions on $h[n]$ that ensure that the requantization error has the *desired properties* and has a time-averaged variance $\sigma_{ee}^2=(N^2-1)/12$.

Theorem 2: Suppose the impulse response, $h[n]$, satisfies the conditions imposed on $f[n]$ by Theorem 1. Then the requantization error, $e[n]$, of the *shaped dither system* has

the properties of *uniformity*, *signal independence*, *dither independence*, *pair-wise independence* and *whiteness*. Moreover, $e[n]$ has time-averaged mean and auto-covariance given by equation (8).

Proof: Setting $A(z) = F(z)$, and $H(z) = V(z)F(z)$, and applying Theorems A1 and corollary 1 of Theorem A1 for every $p > 0$, then applying Theorem A2, and finally Theorem A3 proves the result.

■

The results derived in Section II can be readily extended to determine if a given shaped dither system *i.e.* $H(z)$, satisfies the conditions of Theorem 2. The following corollary presents a class of shaped dither systems that have requantization error with the *desired properties*.

Corollary: If $H(z) = z^{-L}\left(1 - z^{-1}\right)^{-L}$ for some integer $L \geq 2$, and $N = 2^M$, where M is a positive integer then, the requantization error, $e[n]$, has the properties of *uniformity*, *signal-independence*, *dither independence*, *pair-wise independence* and *whiteness*. Moreover, its time-averaged mean and auto-covariance are given by equation (8).

Proof: It has been proved in Section II that the impulse response of $H(z)$ satisfies the conditions of Theorem 1 for $L \geq 2$. Consequently, Theorem 2 becomes applicable and the result follows.

■

For instance, since $V(z) = (1 - z^{-1})$ and $F(z) = z^{-3}(1 - z^{-1})^{-3}$ satisfy the conditions of Theorem 2, it follows that the requantization error of a 3rd order $\Delta\Sigma$ modulator ($STF(z) = z^{-3}$) has the *desired properties* and that the dither has a 1st order

high-pass shape in the modulator output. Since 3rd order ΔΣ modulators are popular choices with ΔΣ fractional-N PLLs, this particular shaped dither system would prove very useful as it simultaneously removes spurious tones and ensures that the in-band phase noise is small.

The results presented here are tabulated in Table 3.2 along with other results from the succeeding sections. The numbers in the "maximum dither shaping" column indicate the most shaping that can be imparted to *iid* dither before adding the result to the signal and still ensure that the requantization error has the *desired properties.* For instance, Table 3.2 suggests that a 4th order ΔΣ modulator can have a maximum of 2nd order shaped dither *i.e.*, the *iid* dither can be filtered by at most two differentiators and it would still ensure that $e[n]$ is white.

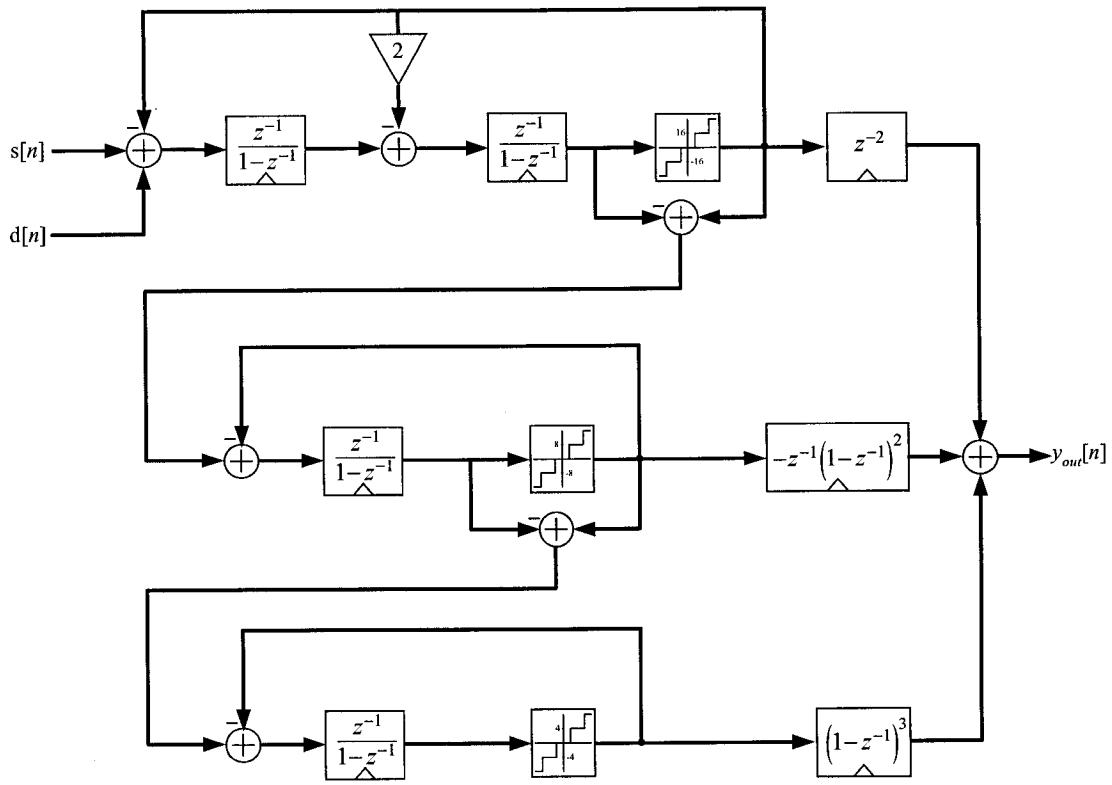

Fig. 3.8: An example 2-1-1 MASH ΔΣ modulator.

IV. DITHERED MASH ARCHITECTURES

A large class of digital ΔΣ modulators not covered by the generic form of Section I is the MASH (Multi-stAge noise SHaping) architecture [4, 5]. High order digital ΔΣ modulators are often realized by cascading multiple lower order digital ΔΣ modulators *e.g.*, the 2-1-1 MASH architecture shown in Fig. 3.8. Such modulators requantize the input signal in steps - coarse quantization of the input, then finer quantization of the error of the first quantization *etc.*, – the outputs of the individual lower order ΔΣ modulators are then combined to form the final output signal. This

section will show that one-bit LSB dither (even shaped dither as illustrated in Fig. 3.6) can ensure that the requantization errors from such a cascaded structure also have the desired properties of *uniformity*, *signal-independence*, *pair-wise independence* and *whiteness*.

Fig. 3.9 shows a generic MASH architecture, which is a cascade of K individual $\Delta\Sigma$ modulator stages. One-bit *iid* dither is filtered by a shaping filter, $V(z)$, and added to the desired signal $s[n]$ to form the input of the MASH system. The i^{th} $\Delta\Sigma$ modulator comprises of forward transmission and feedback filters, $F_i(z)$ and $G_i(z)$, and a mid-tread requantizer of step size N_i, where N_i is a positive integer. The mid-tread requantizer internal to the i^{th} $\Delta\Sigma$ modulator is henceforth referred to as the i^{th} *requantizer*. Consequently, the output of the i^{th} $\Delta\Sigma$ modulator, $y_i[n]$, only takes on values that are integer multiples of N_i. The requantization error from the i^{th} requantizer, $e_i[n]$, is computed by subtracting its input $r_i[n]$, from its output $y_i[n]$, and fed as an input to the succeeding *i.e.*, the $(i + 1)^{st}$ $\Delta\Sigma$ modulator as shown in Fig. 3.9. The outputs, $y_i[n]$, are combined using a bank of post-processing filters $D_i(z)$, i = 1,2,…,K to produce a single output $y_{out}[n]$. The Z-transforms of the requantizer outputs and of $y_{out}[n]$ are related as follows:

$$Y_{out}(z) = \sum_{i=1}^{K} Y_i(z) D_i(z) \tag{15}$$

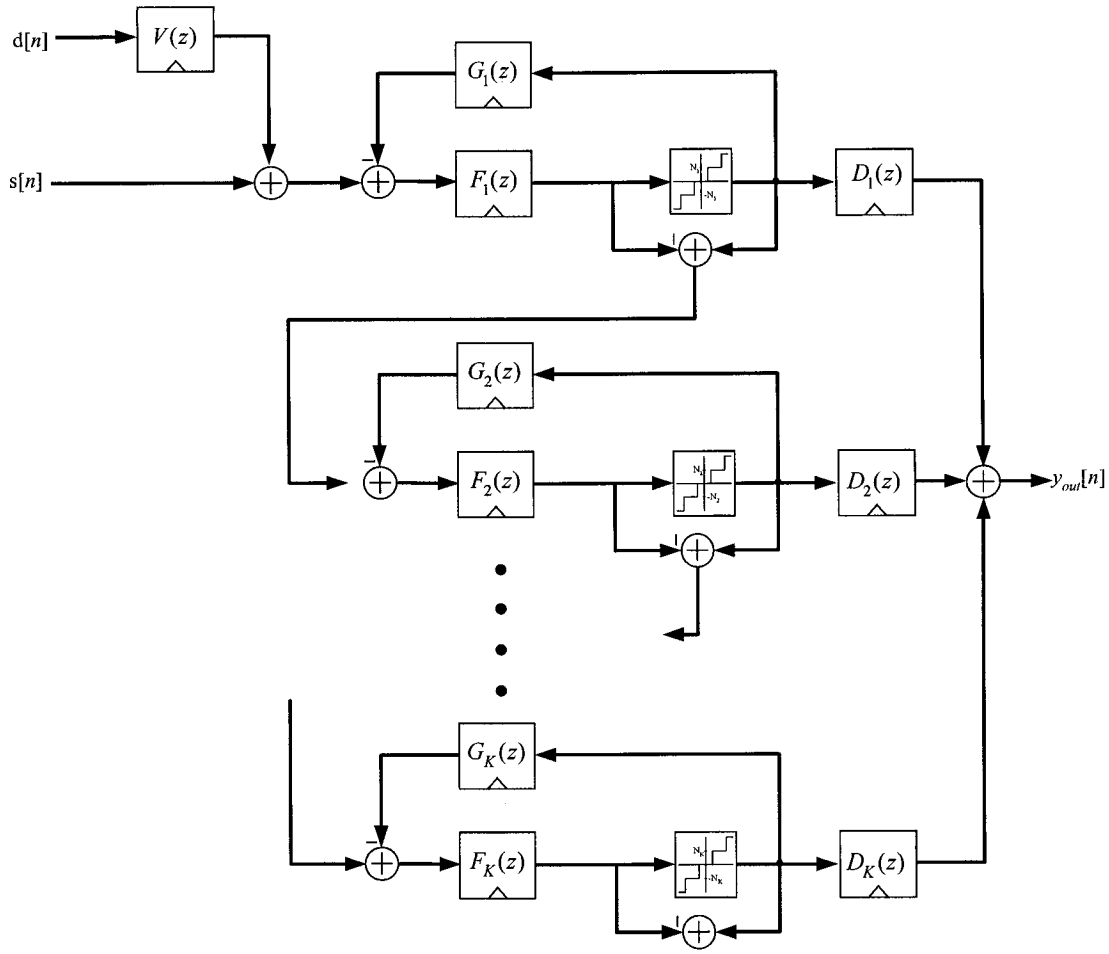

Fig. 3.9: A generic MASH ΔΣ modulator with *K* stages.

The post-processing filters are chosen such that equation (15) reduces to

$$Y_{out}(z) = \underbrace{\prod_{i=1}^{K} STF_i(z)}_{STF_{eq}(z)} \cdot S(z) + \prod_{i=1}^{K} STF_i(z) \cdot V(z) \cdot D(z) + \underbrace{\prod_{i=1}^{K} NTF_i(z)}_{NTF_{eq}(z)} \cdot E_K(z) \qquad (16)$$

where $STF_i(z)$ and $NTF_i(z)$ are defined as in equation (2) for the i^{th} delta-sigma modulator respectively. Such systems are henceforth referred to as *dithered MASH systems*. The generic form is constrained to satisfy the following conditions:

- None of the requantizers overload.

- The requantizer step sizes are such that $N_j = m_{ji}*N_i$, where m_{ji} is a positive integer for every $1 \le j < i \le K$
- The impulse response of every forward transmission filter and every feedback filter, $f_i[n]$ and $g_i[n]$ respectively, is integer valued.
- The impulse response of the dither shaping filter, $v[n]$, is integer valued.
- The impulse response of every post-processing filter $D_i(z)$, *i.e.*, $d_i[n]$, is integer valued.

For instance, the 2-1-1 MASH shown in Fig. 3.8 fits into this description with $K = 3$, $F_1(z) = z^{-2}, (1-z^{-1})^{-2}, G_1(z) = 2z-1$, $F_2(z) = F_3(z) = z^{-1}(1-z^{-1})$, $G_2(z) = G_3(z) = 1$, $D_1(z) = z^{-2}$, $D_2(z) = -z^{-1}(1-z^{-1})^2$, and $D_3(z) = (1-z^{-1})^3$, where the respective impulse responses are all clearly integer valued. It follows that the Z-transform of the output is

$$Y_{out}(z) = z^{-4}S(z) + (1-z^{-1})^4 E_3(z).$$

Digital Reconstruction

Often, the post-processing of the requantizer outputs depicted in equation (15) is performed in the digital domain itself and the result is then sent to a D/A converter. If the K^{th} requantization error, $e_K[n]$, has the *desired properties* then, linear system theory can be applied to (16) and the PSD of $y_{out}[n]$ can be shown to be

$$S_{yy}(e^{jw}) = |STF_{eq}(e^{jw})|^2 S_{ss}(e^{jw}) + |STF_{eq}(e^{jw})|^2 |V(e^{jw})|^2 \sigma_{dd}^2 + |NTF_{eq}(e^{jw})|^2 \sigma_{e_K e_K}^2 \quad (17)$$

where $\sigma_{e_K e_K}^2 = (N_K^2 - 1)/12$ is the variance of the K^{th} requantization error $e_K[n]$. The

following theorem presents sufficient conditions which, when satisfied by the filters in the *dithered MASH system*, ensure that $e_K[n]$ has the *desired properties*.

Theorem 3: Define $T_K(z) \triangleq (-1)^{K-1} F_1(z) \cdot F_2(z) ... F_{K-1}(z) \cdot F_K(z) \cdot V(z)$. If the impulse response of $T_K(z)$, *i.e.*, $t_K[n]$, satisfies the conditions imposed on $f[n]$ by Theorem 1, then the K^{th} requantization error, $e_K[n]$, has the properties of *uniformity*, *signal-independence*, *dither independence*, *pair-wise independence* and *whiteness*.

Proof: See Appendix C.

■

The results derived in Section II can be readily extended to determine which $T_K(z)$ satisfy the conditions of Theorem 3. The following corollary presents a class of *dithered MASH systems* that has requantization $e_K[n]$ error with the *desired properties*.

Corollary: Suppose $T_K(z) = z^{-L}\left(1 - z^{-1}\right)^{-L}$ for some integer $L \geq 2$ then the requantization error, $e_K[n]$, has the properties of *uniformity*, *signal-independence*, *pair-wise independence* and *whiteness*.

Proof: Same as the proof of corollary to Theorem 2.

■

Table 3.2: Guidelines for dithering common digital ΔΣ modulators – "Maximum dither shaping" means the highest order of high-pass shaped dither in the modulator input which still ensures tone-less requantization error.

Dithering in Common Non-overloading, Low-pass ΔΣ Modulators:				
Single Stage Architectures:				
Name of ΔΣ modulator	Figure Reference	Forward Transmission Filter, F(z)	Feedback Filter, G(z)	Allowed high-pass shaping on dither
1st Order	2(a)	$\frac{z^{-1}}{1-z^{-1}}$	1	None
2nd Order	2(b)	$\left(\frac{z^{-1}}{1-z^{-1}}\right)^2$	$2z-1$	0
3rd Order	2(c)	$\left(\frac{z^{-1}}{1-z^{-1}}\right)^3$	$3z^2-3z+1$	1
L^th Order, L>3	-	$\left(\frac{z^{-1}}{1-z^{-1}}\right)^L$	$\left(1-z^{-1}\right)^L - z^L$	$L-2$
MASH Architectures with Digital Recombination:				
Type of MASH	$F_1(z)\cdot F_2(z)\cdot\ldots\cdot F_K(z)$		Allowed high-pass shaping on dither	
1-1	$z^{-2}\left(1-z^{-1}\right)^{-2}$		0	
1-1-1, 1-2, 2-1	$z^{-3}\left(1-z^{-1}\right)^{-3}$		1	
1-1-1-1, 1-2-1, 2-1-1, 1-1-2, 2-2, 1-3, 3-1	$z^{-4}\left(1-z^{-1}\right)^{-4}$		2	
$m_1-m_2-\ldots-m_K$, $m_1+\ldots+m_K=L>4$	$z^{-L}\left(1-z^{-1}\right)^{-L}$		$L-2$	

For instance LSB dither with $V(z) = 1$ in a multi-bit, 1-1 MASH system with mid-tread requantizer step sizes of $N_1 = 2^{M_1}$ and $N_2 = 2^{M_2}$ where $M_1 > M_2$ are all positive integers ensures that the 2nd requantization error has the aforementioned

desired properties. Table 3.2 tabulates some of these results for ready reference.

Analog Reconstruction

In some special cases, the filtered requantizer outputs in (15) are sent to individual D/A converters whose outputs are then added to produce the overall analog output $y_{out}[n]$. In such circumstances, gain mismatches between the otherwise ideal individual D/A converters would imply that the overall analog output contains a linear combination of the K requantizer outputs

$$Y_{out}(z) = STF_{eq}(z)\cdot S(z) + V(z)\cdot STF_{eq}(z)\cdot D(z) + \sum_{i=1}^{K} B_i(z) E_i(z) \qquad (18)$$

where $STF_{eq}(z)$ and $B_i(z)$, i = 1,2,...,K are some arbitrary filters determined by the particular MASH system and gain mismatches. Suppose the following are true:

- Each of the requantization errors, $e_i[n]$, has the properties of uniformity, signal-independence, dither independence, pair-wise independence and whiteness and,
- As $n_0 \to -\infty$ sequences $e_i[n]$ converge to sequences $\tilde{e}_i[n]$ such that $\tilde{e}_i[n]$ is independent of $\tilde{e}_j[n+p]$ for all $i \neq j$ and $p \in \mathbb{Z}$.

Then, the overall analog output, $y_{out}[n]$, has no spurious tones. Furthermore, linear system theory can be applied and the PSD of $y_{out}[n]$ can be shown to be

$$S_{yy}(e^{jw}) = \left|STF_{eq}(e^{jw})\right|^2 S_{ss}(e^{jw}) + \left|STF_{eq}(e^{jw})\right|^2 \left|V(e^{jw})\right|^2 \sigma_{dd}^2 + \sum_{i=1}^{K} \left|B_i(e^{jw})\right|^2 \sigma_{e_i e_i}^2 \qquad (19)$$

where $\sigma_{e_i e_i}^2 = (N_i^2 - 1)/12, i = 1,...,K$ is the variance of the i^{th} requantization error. Note that even if one of the requantization errors, $e_i[n]$, were to not have the *desired* properties, then the last term in (18), and hence $y_{out}[n]$, could exhibit spurious tones.

The following theorem presents sufficient conditions (on the filters in the MASH system), which ensure that all $e_i[n]$ have the aforementioned properties.

Theorem 4: Define $T_i(z) = (-1)^{i-1} F_1(z) \cdot F_2(z) \ldots \cdot F_{i-1}(z) \cdot F_i(z) \cdot V(z)$.

Part (i): If the impulse response of $T_i(z)$ *i.e.*, $t_i[n]$ satisfies the conditions imposed on $f[n]$ by Theorem 1 then, the i^{th} requantization error, $e_i[n]$, has the properties of *uniformity*, *signal independence*, *dither independence*, *pair-wise independence* and *whiteness*.

Part (ii): For all integers p, and k_1, k_2 such that $k_1 + k_2 \neq 0$, $0 \leq k_1 \leq N_i - 1$, and $0 \leq k_2 \leq N_j - 1$, suppose that the sequence $(k_1 t_i[r] + k_2 t_j[r+p]) \bmod N$ does not converge to zero as $r \to \infty$. Then as $n_0 \to -\infty$ sequences $e_i[n]$, $e_j[n]$ converge to sequences $\tilde{e}_i[n]$, $\tilde{e}_j[n]$ such that $\tilde{e}_i[n]$ is independent of $\tilde{e}_j[n+p]$.

Proof: See Appendix C.

■

For instance, it can be shown using the corollary to Theorem 3 that LSB dither with $V(z) = (1 - z^{-1})$ in a 3-1-1 MASH system with mid-tread requantizer step sizes of $N_1 = 2^{M_1}$, $N_2 = 2^{M_2}$ and $N_3 = 2^{M_3}$ where $M_1 \geq M_2 \geq M_3$ are all positive integers ensures that all the requantization errors have the *desired properties*. On the contrary, LSB dither with $V(z) = 1$ in a 1-1 MASH system with mid-tread requantizer step sizes of $N_1 = 2^{M_1}$ and $N_2 = 2^{M_2}$ where $M_1 \geq M_2$, can only ensure that $e_2[n]$ (and **not** $e_1[n]$) has the *desired properties*.

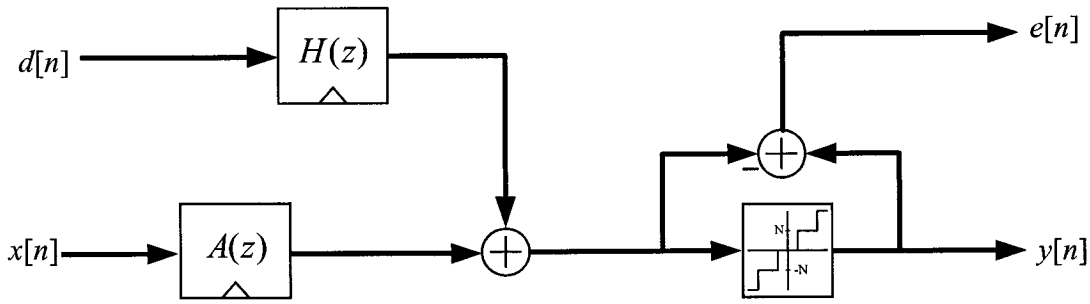

Fig. 3.10: Framework for the theoretical analysis of dithered quantization.

V. CONCLUSION

Theoretical sufficient conditions which ensure that one-bit least significant bit dither eliminates limit cycles and resultant spurious tones in general single stage and MASH digital delta-sigma ($\Delta\Sigma$) modulators were presented. A large class of popular $\Delta\Sigma$ modulators in which one-bit dither eliminates limit cycles are identified by applying the sufficient conditions. Means of imparting spectral shape to the dither while eliminating limit cycles were presented.

APPENDIX A

The theorems and results in this section apply to the system shown in Fig. 3.10. The impulse responses of the filters $A(z)$, and $H(z)$ *i.e.*, $a[n]$, and $h[n]$, are integer valued. The samples of the dither sequence, $d[n]$, are independent of all samples of the desired signal, $s[n]$, of each other, and are identically distributed, with a probability distribution given by equation (5). The desired signal, $s[n]$ takes on values in the range $\{-S/2 + 1,\ldots,0,\ldots,S/2\}$, where S is an even, positive integer. The requantization errors are given by:

$$e[n] = \frac{N}{2} - N\left\langle \frac{z[n]}{N} + \frac{1}{2} \right\rangle$$
$$\text{where} \quad z[n] = \underbrace{\sum_{m=n_0}^{n} s[m]a[n-m]}_{s_a[n]} + \sum_{m=n_0}^{n} d[m]h[n-m]. \tag{20}$$

Theorem A1: For any given integer $p > 0$, suppose that for any integers k_1, k_2 such that $k_1 + k_2 \neq 0$, and $0 \le k_1, k_2 \le N-1$, at least one of the following is true:

1. The sequence $(k_1 h[r] + k_2 h[r+p]) \bmod N$ does not converge to zero as $r \to \infty$.
2. A non-negative integer $r_{1,2} \neq p$ exists such that

   $\left(k_1 h[r_{1,2}] + k_2 h[r_{1,2}+p]\right) \bmod N = N/2$.
3. A non-negative integer $r_2 < p$ exists such that $\left(k_2 h[r_2]\right) \bmod N = N/2$.

Then, as $n_0 \to -\infty$, the requantization error samples at finite time indices n, $n-p > n_0$, namely $e[n]$ and $e[n–p]$, converge in distribution respectively to uniformly distributed random variables, $\tilde{e}[n]$ and $\tilde{e}[n-p]$, such that $\tilde{e}[n]$ is independent of $s[n-p]$, $d[n-p]$, and $\tilde{e}[n-p]$.

Explanation: The properties of $\tilde{e}[n]$ and $\tilde{e}[n-p]$ mentioned in the theorem are mathematically formulated as follows:

Uniformity

$$\begin{aligned} &\tilde{e}[n], \tilde{e}[n-p] \in \{-N/2+1, \ldots, 0, \ldots, N/2\}, \\ &P\left(\tilde{e}[n] = m_e\right) = P\left(\tilde{e}[n-p] = m_e\right) = 1/N, \end{aligned} \tag{21}$$

Signal independence

$$P\left(s[n-p] = m_s, \tilde{e}[n] = m_e\right) = P\left(s[n-p] = m_s\right) \cdot P\left(\tilde{e}[n] = m_e\right), \tag{22}$$

Dither independence

$$P\left(d[n-p]=m_d,\tilde{e}[n]=m_e\right)=P\left(d[n-p]=m_d\right)\cdot P\left(\tilde{e}[n]=m_e\right), \tag{23}$$

Pair-wise independence

$$P\left(\tilde{e}[n-p]=m_{ep},\tilde{e}[n]=m_e\right)=P\left(\tilde{e}[n-p]=m_{ep}\right)\cdot P\left(\tilde{e}[n]=m_e\right), \tag{24}$$

where, m_e, $m_{ep} \in \{-N/2+1,\ldots,0,\ldots,N/2\}$, $m_s \in \{-S/2+1,\ldots,0,\ldots,S/2\}$, and $m_d \in \{0, 1\}$.

The theorem states that if at least one of the conditions 1-3 are satisfied, then $e[n]$ and $e[n-p]$ respectively converge in distribution to $\tilde{e}[n]$ and $\tilde{e}[n-p]$ which have the above properties *i.e.* the following equations are satisfied:

Convergence to uniformity

$$\begin{aligned} P\left(e[n]=m_e\right) &\xrightarrow{n_0\to-\infty} P\left(\tilde{e}[n]=m_e\right), \\ P\left(e[n-p]=m_e\right) &\xrightarrow{n_0\to-\infty} P\left(\tilde{e}[n-p]=m_e\right), \end{aligned} \tag{25}$$

Convergence to signal independence

$$P\left(s[n-p]=m_s,e[n]=m_e\right)\xrightarrow{n_0\to-\infty}P\left(s[n-p]=m_s,\tilde{e}[n]=m_e\right), \tag{26}$$

Convergence to dither independence

$$P\left(d[n-p]=m_d,e[n]=m_e\right)\xrightarrow{n_0\to-\infty}P\left(d[n-p]=m_d,\tilde{e}[n]=m_e\right), \tag{27}$$

Convergence to pair-wise independence

$$P\left(e[n-p]=m_{ep},e[n]=m_e\right)\xrightarrow{n_0\to-\infty}P\left(\tilde{e}[n-p]=m_{ep},\tilde{e}[n]=m_e\right), \tag{28}$$

where, m_e, $m_{ep} \in \{-N/2+1,\ldots,0,\ldots,N/2\}$, $m_d \in \{0, 1\}$, and $m_s \in \{-S/2+1,\ldots,0,\ldots,S/2\}$.

Proof: The goal is to prove that given (20)–(24), if at least one of the conditions 1–3 is satisfied then, equations (25)–(28) are true. Note that (26), and (22) together imply that

$$P\left(s[n-p]=m_s, e[n]=m_e\right) \xrightarrow{n_0 \to -\infty} P\left(s[n-p]=m_s\right) \cdot P\left(\tilde{e}[n]=m_e\right).$$

Summing the above equation over all values of m_s, proves the first equation in (25). Since the conditions of the theorem are independent of n, if $e[n]$ converges in distribution to a uniform random variable, so does $e[n\text{-}p]$, thereby proving the second equation of (25) as well. Therefore, it is sufficient to prove that equations (26)–(28) are true. As shown below, it is accomplished by considering the convergence of the two-dimensional joint characteristic functions of $e[n]$ and $s[n–p]$, $d[n–p]$, $e[n–p]$ as $n_0 \to -\infty$. The proofs of the three equations are similar except for a few details. The common aspects of the three proofs are first presented, followed by the specific details which differentiate the proofs.

Common aspects

Equations (26)–(28) are particular cases of the following generalized equation:

$$P\left(A=i, B=j\right) \xrightarrow{n_0 \to -\infty} P\left(C=i, D=j\right), \tag{29}$$

where A = $s[n–p]$, $d[n–p]$, or $e[n–p]$, B = $e[n]$, C = $s[n–p]$, $d[n–p]$, or $\tilde{e}[n-p]$, D = $\tilde{e}[n]$; while the index $j \in \{-N/2 + 1,\ldots,0,\ldots,N/2\}$, the index $i \in \{-L/2 + 1,\ldots,0,\ldots,L/2\}$,

where L = S, 2, or N corresponding to whether A = $s[n–p]$, $d[n–p]$, or $e[n–p]$. For instance, A = $s[n–p]$, L = S reduces (29) to the equation (26). It follows from Lemma B1 that to prove (29), it is sufficient to prove that:

$$\Phi_{A,B}\left(\frac{2\pi k_L}{L}, \frac{2\pi k_N}{N}\right) \xrightarrow{n_0 \to -\infty} \Phi_{C,D}\left(\frac{2\pi k_L}{L}, \frac{2\pi k_N}{N}\right), \ \forall 0 \le k_L < L, 0 \le k_N < N, \tag{30}$$

where $\Phi_{A,B}(w_1, w_2)$ and $\Phi_{C,D}(w_1, w_2)$ are the joint characteristic functions of the

random variables A, B and C, D respectively. The RHS of (30) is readily derived from the properties of $\hat{e}[n]$ listed in equations (21)–(24):

$$\begin{aligned}&\Phi_{\tilde{e}[n]}\left(\frac{2\pi k_N}{N}\right)=\Phi_{\tilde{e}[n-p]}\left(\frac{2\pi k_N}{N}\right)=\delta[k_N],\\&\Phi_{C,\tilde{e}[n]}\left(\frac{2\pi k_L}{L},\frac{2\pi k_N}{N}\right)=\Phi_C\left(\frac{2\pi k_L}{L}\right)\cdot\delta[k_N]\quad\forall 0\le k_L<L,0\le k_N<N.\end{aligned}\tag{31}$$

Further, since $B = e[n]$ is related to $z[n]$ through the fractional operator, $\langle x\rangle = x-\lfloor x\rfloor$, as shown in (20), it follows from Lemma B2 that,

$$\Phi_{A,e[n]}\left(w_1,\frac{2\pi k_N}{N}\right)=\Phi_{A,z[n]}\left(w_1,\frac{-2\pi k_N}{N}\right)\qquad\forall 0\le k_N<N,w_1\in[0,2\pi).\tag{32}$$

Substituting equations (31), (32) into (30) implies that, to prove (29) true, it is sufficient to prove that:

$$\Phi_{A,z[n]}\left(\frac{2\pi k_L}{L},\frac{-2\pi k_N}{N}\right)\xrightarrow{n_0\to-\infty}\Phi_C\left(\frac{2\pi k_L}{L}\right)\cdot\delta[k_N],\quad\forall 0\le k_L<L,0\le k_N<N.\tag{33}$$

The LHS of (33), which represents samples of the joint characteristic function of A and $z[n]$ is then expressed in terms of the characteristic functions of the dither samples and the impulse response of $H(z)$, namely $h[r]$. Once an expression for the LHS of (33) is obtained in terms of $h[r]$, conditions 1–3 listed in the theorem statement are substituted and equations (26)–(28) are proved. The expressions for the LHS of (33) depend on whether $A = s[n–p]$, $d[n–p]$, or $e[n–p]$. These specific details are presented next.

Convergence to signal independence

Substituting $A = C = s[n–p]$, $B = e[n]$, $D = \tilde{e}[n]$, and $L = S$ in (29) reduces it to equation (26). It follows from (33) that it is sufficient to prove that:

$$\Phi_{s[n-p],z[n]}\left(\frac{2\pi k_S}{S},\frac{-2\pi k_N}{N}\right)\xrightarrow{n_0\to-\infty}\Phi_{s[n-p]}\left(\frac{2\pi k_S}{S}\right)\cdot\delta[k_N],\ \forall 0\le k_S<S,0\le k_N<N. \quad (34)$$

Substituting $w_1 = 2\pi k_s/S$, and $w_2 = -2\pi k_N/N$ in Lemma B3, it follows that

$$\Phi_{s[n-p],z[n]}\left(\frac{2\pi k_s}{S},\frac{-2\pi k_N}{N}\right)=\Phi_{s[n-p],s_a[n]}\left(\frac{2\pi k_s}{S},\frac{-2\pi k_N}{N}\right)\cdot\prod_{r=0}^{n-n_0}\Phi_d\left(\frac{-2\pi k_N}{N}h[r]\right), \quad (35)$$
$$\forall 0\le k_N<N,0\le k_s<S,$$

where $\Phi_{s[n-p],s_a[n]}(w_1,w_2)$ is the joint characteristic function of $s_a[n]$, $s[n–p]$, and $\Phi_d(w)$ is the characteristic function of each of the *iid* random variables $d[n]$. Substituting (35) in equation (34) implies that to prove (26) true, it is sufficient to prove that

$$\Phi_{s[n-p],s_a[n]}\left(\frac{2\pi k_S}{S},\frac{-2\pi k_N}{N}\right)\cdot\prod_{r=0}^{n-n_0}\Phi_d\left(\frac{-2\pi k_N}{N}h[r]\right)\xrightarrow{n_0\to-\infty}\Phi_{s[n-p]}\left(\frac{2\pi k_S}{S}\right)\cdot\delta[k_N], \quad (36)$$
$$\forall 0\le k_N<N,0\le k_S<S.$$

This equation can be proved true for $k_N = 0$ by noting that $\Phi_{s[n-p],s_a[n]}(w,0)=\Phi_{s[n-p]}(w)$. The following equation is sufficient to prove (36) for $k_N \neq 0$, and hence (26):

$$\prod_{r=0}^{n-n_0}\Phi_d\left(\frac{-2\pi k_N}{N}h[r]\right)\xrightarrow{n_0\to-\infty}0,\qquad \forall 0<k_N<N, \quad (37)$$

Now conditions 1–3 are substituted to prove (37). First, consider those k_N for which the pair (k_N, 0) satisfies condition 1 listed in the theorem statement. It follows from setting $g[r]=k_N h[r]$ and invoking Lemma B6 that (37) is true for all such k_N. Then, consider those k_N for which the pair (k_N, 0) satisfies condition 2. Therefore, for each of these k_N, a positive integer $r_{N,0}$ exists such that $k_N h[r_{N,0}] \bmod N = N/2$. It follows

from setting $g[r] = k_N h[r_{N,0}]$ and invoking Lemma B7 that at least one of the terms in the LHS of (37) is zero, and hence (37) is true for all such k_N. The above two cases cover all k_N such that $0 < k_N < N$, because by hypothesis at least one of conditions 1–3 is satisfied for every pair (k_1, k_2) and $(k_N, 0)$ does not satisfy condition 3. Hence equation (26) is proved.

Convergence to dither independence

Substituting $A = C = d[n-p]$, $B = e[n]$, $D = \tilde{e}[n]$, and $L = 2$ in (29) reduces it to equation (27). It follows from (33) that it is sufficient to prove that:

$$\Phi_{d[n-p],z[n]}\left(\frac{2\pi k_2}{2}, \frac{-2\pi k_N}{N}\right) \xrightarrow{n_0 \to -\infty} \Phi_{d[n-p]}\left(\frac{2\pi k_2}{2}\right) \cdot \delta[k_N],\ \forall 0 \le k_2 < 2, 0 \le k_N < N. \quad (38)$$

Substituting $w_1 = 2\pi k_2/2$, and $w_2 = -2\pi k_N/N$ in Lemma B4, it follows that

$$\Phi_{d[n-p],z[n]}\left(\frac{2\pi k_2}{2}, \frac{-2\pi k_N}{N}\right) = \Phi_{s_a[n]}\left(\frac{-2\pi k_N}{N}\right) \cdot \Phi_d\left(\frac{-2\pi k_N}{N} h[p] + \frac{2\pi k_2}{2}\right) \cdot \prod_{\substack{r=0, \\ r \ne p}}^{n-n_0} \Phi_d\left(\frac{-2\pi k_N}{N} h[r]\right), \quad (39)$$

$$\forall 0 \le k_N < N, k_2 \in \{0,1\}.$$

Substituting (39) in (38) implies that, to prove (27) it is sufficient to prove that

$$\Phi_{s_a}\left(\frac{-2\pi k_N}{N}\right) \cdot \Phi_d\left(\frac{-2\pi k_N}{N} h[p] + \frac{2\pi k_2}{2}\right) \cdot \prod_{\substack{r=0, \\ r \ne p}}^{n-n_0} \Phi_d\left(\frac{-2\pi k_N}{N} h[r]\right) \xrightarrow{n_0 \to -\infty} \Phi_d\left(\frac{2\pi k_2}{2}\right) \cdot \delta[k_N],\ \forall 0 \le k_N < N, 0 \le k_2 < 2. \quad (40)$$

This equation can be proved true for $k_N = 0$ by noting that $\Phi_{s_a[n]}(0) = \Phi_d(0) = 1$. The following equation is sufficient to prove (40) for $k_N \neq 0$, and hence (27):

$$\prod_{\substack{r=0,\\ r\neq p}}^{n-n_0} \Phi_d\left(\frac{-2\pi k_N}{N} h[r]\right) \xrightarrow{n_0\to-\infty} 0, \qquad \forall 0 < k_N < N. \tag{41}$$

Now conditions 1–3 are substituted to prove (41). The proof is identical to the proof of (37) except for one detail. Consider those k_N for which either (k_N, 0) satisfies condition 2. Just as in the proof of (37), for each of these k_N, a positive integer $r_{N,0}$ exists such that $k_N h[r_{N,0}] \bmod N = N/2$. Unlike before, proving (41) by setting $g[r] = k_N h[r_{N,0}]$ and invoking Lemma B7 requires that $r_{N,0} \neq p$, which is guaranteed by condition 2. Hence (41) and (27) are proved.

Convergence to pair-wise independence

Substituting $A = e[n–p]$, $B = e[n]$, $C = \tilde{e}[n-p]$, $D = \tilde{e}[n]$, and $L = N$ in (29) reduces it to equation (28). It follows from (33) that it is sufficient to prove that:

$$\Phi_{e[n-p],z[n]}\left(\frac{2\pi k_{Np}}{N}, \frac{-2\pi k_N}{N}\right) \xrightarrow{n_0\to-\infty} \Phi_{\tilde{e}[n-p]}\left(\frac{2\pi k_{Np}}{N}\right) \cdot \delta[k_N], \forall 0 \le k_{Np}, k_N < N. \tag{42}$$

Since $e[n–p]$ is related to $z[n–p]$ through the fractional operator, $\langle x \rangle = x - \lfloor x \rfloor$, as given by (20), it follows from Lemma B2 that,

$$\Phi_{e[n-p],e[n]}\left(\frac{2\pi k_{Np}}{N}, \frac{-2\pi k_N}{N}\right) = \Phi_{z[n-p],z[n]}\left(\frac{-2\pi k_{Np}}{N}, \frac{-2\pi k_N}{N}\right) \quad \forall 0 \le k_{Np}, k_N < N. \tag{43}$$

Substituting equation (43), and the first equation of (31) into equation (42) implies that to prove (28), it is sufficient to prove that:

$$\Phi_{z[n-p],z[n]}\left(\frac{-2\pi k_{Np}}{N}, \frac{-2\pi k_N}{N}\right) \xrightarrow{n_0\to-\infty} \delta[k_{Np}] \cdot \delta[k_N], \forall 0 \le k_{Np}, k_N < N. \tag{44}$$

Substituting $A_1(z) = A(z)$, $A_2(z) = 0$, $H_1(z) = H(z)$, $H_2(z) = 0$, $N_1 = N_2 = N$,

$w_1 = -2\pi k_{Np}/N$, and $w_2 = -2\pi k_N/N$ in Lemma B5, results in the following:

$$\Phi_{z[n-p],z[n]}\left(\frac{-2\pi k_{Np}}{N},\frac{-2\pi k_N}{N}\right) = \Phi_{s_a[n-p],s_a[n]}\left(\frac{-2\pi k_{Np}}{N},\frac{-2\pi k_N}{N}\right) \cdot\prod_{r=0}^{n-n_0}\Phi_d\left(\frac{-2\pi}{N}\left\{k_{Np}h[r]+k_N h[r+p]\right\}\right) \cdot\prod_{r=0}^{p-1}\Phi_d\left(\frac{-2\pi}{N}k_N h[r]\right),\ \forall 0\le k_N,k_{Np}<N. \tag{45}$$

where $\Phi_{s_a[n-p],s_a[n]}(w_1,w_2)$ is the joint characteristic function of $s_a[n-p]$ and $s_a[n]$. Since $\Phi_{s_a[n-p],s_a[n]}(0,0)=1$, substituting (45) in (44) implies that, to prove (28) true it is sufficient to prove that:

$$\prod_{r=0}^{n-n_0}\Phi_d\left(\frac{-2\pi}{N}\left\{k_{Np}h[r]+k_N h[r+p]\right\}\right)\cdot\prod_{r=0}^{p-1}\Phi_d\left(\frac{-2\pi}{N}k_N h[r]\right) \xrightarrow{n_0\to-\infty} 0, \quad \forall 0\le k_N,k_{Np}<N, k_N+k_{Np}\neq 0. \tag{46}$$

Now conditions 1–3 are substituted to prove (46). First, consider those pairs $(k_{Np},k_N)\neq(0,0)$ which satisfy condition 1 listed in the theorem statement. It follows from setting $g[r]=k_{Np}h[r]+k_N h[r+p]$ and invoking Lemma B6 that the first product term in the LHS of (46) converges to zero, and hence (46) is true for all such pairs. Then, consider those pairs which satisfy condition 2. Therefore, for each of these pairs, a positive integer $r_{1,2}$ exists such that $(k_{Np}h[r_{1,2}]+k_N h[r_{1,2}+p]) \bmod N = N/2$. It follows from setting $g[r]=k_{Np}h[r_{1,2}]+k_N h[r_{1,2}+p]$ and invoking Lemma B7 that the first product term in the LHS of (46) is zero, and hence (46) is true for all such pairs. The remaining pairs satisfy condition 3. Therefore, for each of these remaining pairs, a positive integer r_2 exists such that $k_N h[r_2] \bmod N = N/2$. It follows from setting

$g[r] = k_N h[r_N]$ and invoking Lemma B7 that the second product term in the LHS of (46) is zero, and hence (46) is true for all the remaining pairs. Hence equation (28) is proved.

■

Corollary 1: For any given $p > 0$, suppose that the first condition of Theorem A1 is true for all integers k_1, k_2 such that $k_1 + k_2 \neq 0$, and $0 \leq k_1, k_2 \leq N-1$. Then, $e[n]$ converges in distribution to $\hat{e}[n]$ which is also independent of $s[n]$ and $d[n]$.

Proof: By hypothesis, condition 1 is true for integers $k_1 = k_N$, $k_2 = 0$. It follows that $k_N h[r] \bmod N$ does not converge to zero as $r \to \infty$. The result follows by setting $p = 0$ in equations (34)–(37) and (38)–(41) in the proof of Theorem A1.

■

Corollary 2: Suppose that the starting time index, n_0, is fixed and the conditions of Theorem A1 are satisfied for a given $p > 0$. Then, as $n \to \infty$, $e[n]$ and $e[n–p]$ converge in distribution respectively to uniformly distributed random variables, X_0 and X_p, such that X_0 is independent of $s[n–p]$, $d[n–p]$, and X_p.

Proof: The proof is identical to the proof of Theorem A1 except that the convergence is in terms of $n \to \infty$ instead of $n_0 \to -\infty$, and X_0 and X_p replace $\tilde{e}[n]$ and $\tilde{e}[n-p]$ respectively. By proceeding just as in the derivation of equations (37), (41), and (46) but with $n_0 \to -\infty$ replaced with $n \to \infty$, and $\tilde{e}[n]$ and $\tilde{e}[n-p]$ replaced with X_0 and X_p, it follows that to prove the required convergence of $e[n]$ and $e[n–p]$ it is sufficient to prove that equations (37), (41), and (46) are true as $n \to \infty$ instead of $n_0 \to -\infty$. It

is also evident from these equations that the conditions of Theorem A1 imply these three equations as long as $n - n_0 \to \infty$. Since, fixing n_0 and letting $n \to \infty$ achieves this purpose, the corollary is proved.

■

Theorem A2: Suppose that the conditions of Theorem A1 are satisfied for all positive integers p. Then, the requantizer error $e[n]$ has the property of *whiteness* with statistical mean and auto-covariance,

$$M_e = \frac{1}{2},$$
$$C_{ee}(p) = \frac{N^2 - 1}{12}\delta[p].$$

Proof: If the conditions of Theorem A1 are satisfied for all $p > 0$, then $\tilde{e}[n]$ is independent of $\tilde{e}[n-p]$ for all $p > 0$. Since Theorem A1 assumes nothing about the index n except that $n > n_0$, we can replace n by $n + p$. As $n_0 \to -\infty$, this can be done for all $p > 0$. Consequently, $\tilde{e}[n+p]$ is independent of $\tilde{e}[n]$ and vice versa. Hence, $\tilde{e}[n]$ is independent of (and hence, uncorrelated with) $\tilde{e}[n+p]$ for all $p \neq 0$.

The mean and auto-covariance of the requantization error sequence are defined as:

$$M_e[n] \triangleq \lim_{n_0 \to -\infty} E\big[e[n]\big],$$
$$C_{e[n]e[n-p]}[n,p] \triangleq \lim_{n_0 \to -\infty} E\Big[\big(e[n] - M_{e[n]}\big)\big(e[n-p] - M_{e[n-p]}\big)\Big].$$

Since $e[n]$ converges in distribution to $\tilde{e}[n]$ and since $e[n]$ is bounded and has finite

support $M_e[n]$ is equal to $E[\tilde{e}[n]]$ and hence is independent of n [6][5]. Consequently, the index n can be dropped and it follows that,

$$M_e = \sum_{m=-N/2+1}^{N/2} \frac{m}{N} = \frac{1}{2}.$$

By analogous reasoning it follows that

$$C_{ee}[n,p] = E\left[\left(\tilde{e}[n]-M_e\right)\left(\tilde{e}[n-p]-M_e\right)\right]$$

and hence is independent of n. Consequently, the index n can be dropped and it follows that,

$$C_{ee}[p] = \begin{cases} \left(\sum_{m=-N/2+1}^{N/2} \frac{m^2}{N}\right) - M_e^2 = \frac{N^2-1}{12}, & p=0 \\ 0, & p \neq 0 \end{cases}.$$

The independence of M_e and $C_{ee}[p]$ from n along with the uncorrelated-ness proves the result.

■

Corollary: Suppose that the starting time index, n_0, is fixed and the conditions of Theorem A1 are satisfied for all positive integers $p > 0$. Then,

$$E[e[n]] \xrightarrow{n\to\infty} M_e,$$
$$E\left[\left(e[n]-M_e\right)\left(e[n-p]-M_e\right)\right] \xrightarrow{n\to\infty} C_{ee}[p].$$

Proof: The proof is similar to the proof of Theorem A2 except that Corollary 2 of Theorem A1 is used instead of Theorem A1, and random variables X_0 and X_p are used instead of $\tilde{e}[n]$ and $\tilde{e}[n-p]$.

---

[5] See Corollary 8.3.1, pg. 261.

■

The theorems presented so far are concerned with the ensemble statistics of the requantization error *i.e.*, the average behavior of $e[n]$ over numerous realizations of the random dither sequences. However, these results would be of practical use only if $e[n]$ is proven to exhibit identical average behavior over time for any arbitrary realization of the dither sequence.

Theorem A3: Suppose that the conditions of Theorem A1 are satisfied. Then, the following are true:

1. The ensemble averages M_e, $C_{ee}[p]$ are equal to the corresponding time averages *i.e.*,

$$\begin{aligned} &\frac{1}{L}\sum_{n=n_0}^{L+n_0-1} e[n] \xrightarrow{L\to\infty} M_e, \\ &\frac{1}{L}\sum_{n=n_0+p}^{L+n_0+p-1} \big(e[n]-M_e\big)\big(e[n-p]-M_e\big) \xrightarrow{L\to\infty} C_{ee}[p], \quad \forall p \geq 0, \end{aligned} \tag{47}$$

   where the convergence is in probability.

2. There is no average time correlation between $e[n]$, and the desired signal, $s[l]$, for any $n,l\in\mathbb{Z}$ and $(n-l)$ finite *i.e.*,

$$\frac{1}{L}\sum_{n=n_0}^{L+n_0-1} \big(e[n]-M_e\big)s[l] \xrightarrow{L\to\infty} 0. \tag{48}$$

   where the convergence is in probability.

Proof: Convergence in probability means "the probability that the LHS of one of the equations in (47) or (48) is unequal to its RHS *i.e.*, $P(LHS \neq RHS)$, converges to zero

as $L \to \infty$". The proofs for all three cases are similar. Only the second equation in (47) is proved here. Define the random process,

$$X_k \triangleq \big(e[k]-M_e\big)\big(e[k-p]-M_e\big)-C_{ee}[p].$$

It follows from the corollary to Theorem A2 that

$$E\{X_k\} \xrightarrow{k\to\infty} 0. \tag{49}$$

By setting $n_0 = 0$, and defining $\eta_k \triangleq d[k]$ and $\mu_k \triangleq s[k]$ equation (20) can be rewritten as:

$$e[n] = \frac{N}{2} - N\left\langle \frac{z[n]}{N} + \frac{1}{2} \right\rangle$$

$$\text{where} \quad z[n] = \sum_{m=0}^{n} \mu_m a[n-m] + \sum_{m=0}^{n} \eta_m h[n-m].$$

Consequently, for each $k \in \mathbb{Z}^+$, X_k is a bounded, measurable function $X_k = f\left(\eta_0,...,\eta_k,\mu_0,...,\mu_k\right)$. Therefore, it follows from (49) that

$$E\left\{X_k \mid \eta_0,...,\eta_j,\mu_0,...,\mu_j,...\right\} \to 0$$

in probability as $k-j \to \infty$ with $k > j \geqslant 0$. It follows from Lemma A2 in [7] that:

$$\frac{1}{L}\sum_{k=p}^{L+p-1} X_k \to 0.$$

Hence the theorem is proved.

■

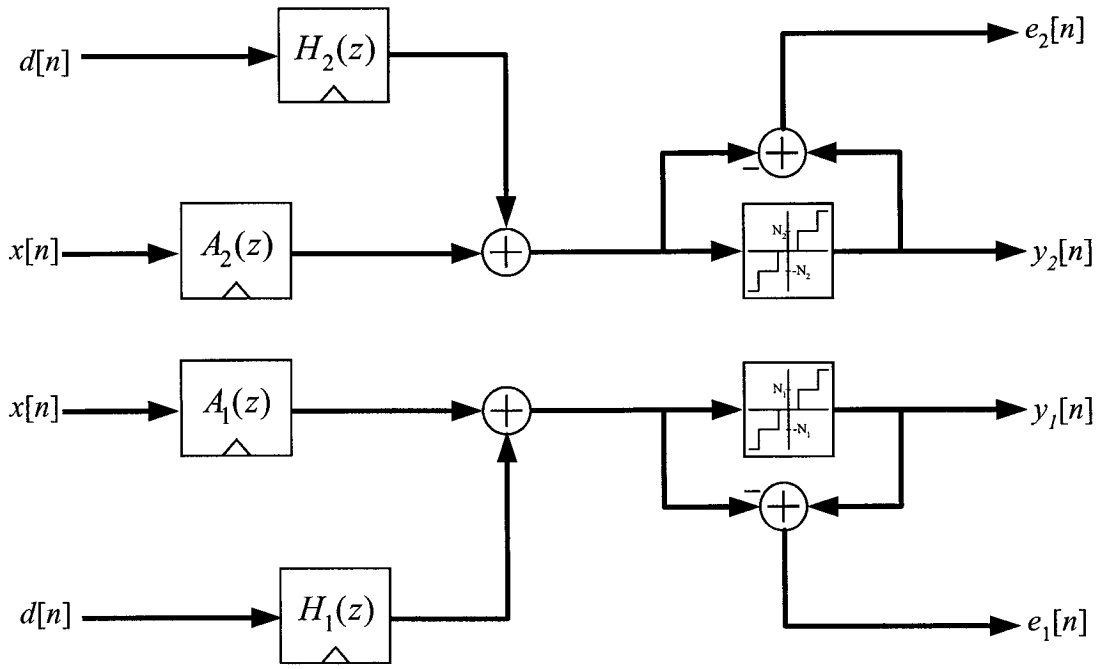

Fig. 3.11: Framework for the theoretical analysis of multi-stage dithered quantization.

The following theorem refers to the system shown in Fig. 3.11.

Theorem A4: Suppose that for a given positive integer p and any k_1, k_2 both not zero, $0 \le k_1 \le N_1 - 1$, and $0 \le k_2 \le N_2 - 1$, at least one of the following is true:

1. The following sequence does not converge to zero as $r \to \infty$

$$\left(k_1 h_1[r] + \frac{N_1}{N_2} k_2 h_2[r+p] \right) \bmod N_1$$

2. A non-negative integer $r_{1,2} \neq p$ exists such that

$$\left(k_1 h_1[r_{1,2}] + \frac{N_1}{N_2} k_2 h_2[r_{1,2}+p] \right) \bmod N_1 = N_1/2$$

3. A non-negative integer $r_2 < p$ exists such that $\left(k_2 h_2[r_2]\right) \bmod N_2 = N_2/2$

Then for i = 1, 2, the requantization error $e_i[n]$ converges to $\tilde{e}_i[n]$ such that $\tilde{e}_1[n]$ is

independent of $\tilde{e}_2[n+p]$.

Proof: The proof is analogous to that of pair-wise independence (equation (28)) in Theorem A1 using $\Phi_{e_1[n-p],e_2[n]}(k_1,k_2)$ instead of $\Phi_{e[n-p],e[n]}(k_1,k_2)$.

■

APPENDIX B

Lemma B1: Suppose A, B, C, D are discrete random variables, which are functions of an arbitrary integer n_0, such that $A,C \in \{-L/2+1,...,0,...,L/2\}$, and $B,D \in \{-N/2+1,...,0,...,N/2\}$, where L, N are even positive integers. Suppose $\Phi_{A,B}(w_1, w_2)$, $\Phi_{C,D}(w_1, w_2)$ are the joint characteristic functions of A and B, and C and D, respectively. Then to prove that

$$P(A=i,B=j) \xrightarrow{n_0 \to -\infty} P(C=i,D=j), \quad \forall -L/2 < i \le L/2, -N/2 < j \le N/2,$$

it is sufficient to prove that

$$\Phi_{A,B}\left(\frac{2\pi k_L}{L},\frac{2\pi k_N}{N}\right) \xrightarrow{n_0 \to -\infty} \Phi_{C,D}\left(\frac{2\pi k_L}{L},\frac{2\pi k_N}{N}\right), \ \forall 0 \le k_L < L, 0 \le k_N < N.$$

Proof: The result follows if $P(A=i,B=j)$ and $P(C=i,D=j)$ are uniquely determined by the samples of $\Phi_{A,B}(w_1, w_2)$, $\Phi_{C,D}(w_1, w_2)$ used in the above expression. The uniqueness is proved by expressing $P(A=i,B=j)$, $P(C=i,D=j)$ respectively as functions of the $L*N$ samples of $\Phi_{A,B}(w_1, w_2)$ and $\Phi_{C,D}(w_1, w_2)$.

Consider $Q(A=i,B=j)$ defined for integers i,j such that $-L/2 < i \le L/2$, and $-N/2 < j \le N/2$:

$$Q(A=i,B=j) \triangleq \frac{1}{NL}\sum_{k_L=0}^{L-1}\sum_{k_N=0}^{N-1} e^{-j\frac{2\pi i k_L}{L}} e^{-j\frac{2\pi j k_N}{N}} \Phi_{A,B}\left(\frac{2\pi k_L}{L}, \frac{2\pi k_N}{N}\right).$$

It is shown below that $Q(A=i,B=j) = P(A=i,B=j)$, thereby proving that $P(A=i,B=j)$ is uniquely determined by the samples of $\Phi_{A,B}(w_1, w_2)$. Substituting the joint characteristic function of A and B, which is defined as

$$\Phi_{A,B}(w_1,w_2) = \sum_{m_L=-L/2+1}^{L/2}\sum_{m_N=-N/2+1}^{N/2} e^{jw_1 m_L} e^{jw_2 m_N} P(A=m_L,B=m_N), \qquad \forall w_1, w_2 \in [0,2\pi),$$

in the definition of $Q(A = i, B = j)$ results in the following:

$$Q(A=i,B=j) = \frac{1}{NL}\sum_{k_L=0}^{L-1}\sum_{k_N=0}^{N-1} e^{-j\frac{2\pi i k_L}{L}} e^{-j\frac{2\pi j k_N}{N}} \sum_{m_L=-L/2+1}^{L/2}\sum_{m_N=-N/2+1}^{N/2} e^{j\frac{2\pi k_L m_L}{L}} e^{j\frac{2\pi k_N m_N}{N}} P(A=m_L,B=m_N).$$

Interchanging the order of the summations and simplification results in the following:

$$Q(A=i,B=j) = \sum_{m_L=-L/2+1}^{L/2}\sum_{m_N=-N/2+1}^{N/2} P(A=m_L,B=m_N)\frac{1}{L}\sum_{k_L=0}^{L-1} e^{j\frac{2\pi k_L (m_L-i)}{L}} \frac{1}{N}\sum_{k_N=0}^{N-1} e^{j\frac{2\pi k_N (m_N-j)}{N}}.$$

Using the relation

$$\frac{1}{N}\sum_{k_1=0}^{N-1} e^{j\frac{2\pi k_1 (l_1-m_1)}{N}} = \sum_{r=-\infty}^{\infty} \delta[l_1 - m_1 - rN],$$

in the above expression for $Q(A=i,B=j)$ simplifies it to the following:

$$Q(A=i,B=j) = \sum_{m_L=-L/2+1}^{L/2}\sum_{m_N=-N/2+1}^{N/2} P(A=m_L,B=m_N)\sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty} \delta[m_L - i - rL]\cdot\delta[m_N - j - sN].$$

For $-L/2 < i \le L/2$, and $-N/2 < j \le N/2$ therefore $Q(A=i,B=j) = P(A=i,B=j)$,

thereby proving that the sample of the joint characteristic function uniquely determine the joint probability distribution function. An analogous result is obtained for the random variables C, D, which completes the proof.

■

Lemma B2: Suppose X, Y, and Z are discrete random variables and N is an even positive integer such that

$$Y = \frac{N}{2} - N\left\langle \frac{Z}{N} + \frac{1}{2} \right\rangle, \quad \text{where } \langle x \rangle = x - \lfloor x \rfloor.$$

Then, the following is true about their joint characteristic functions, $\Phi_{X,Y}(w_1, w_2)$ and $\Phi_{X,z}(w_1, w_2)$:

$$\Phi_{X,Y}\left(w_1, \frac{2\pi k}{N}\right) = \Phi_{X,Z}\left(w_1, \frac{-2\pi k}{N}\right) \quad \forall 0 \le k < N.$$

Proof: Random variable Y takes on a value $j \in \{-N/2+1, ..., 0, ..., N/2\}$ whenever random variable $Z = rN - j$, where r is an integer. Therefore, the joint probability distribution of X and Y is related to that of X and Z as follows:

$$P(X = i, Y = j) = \sum_{r=-\infty}^{\infty} P(X = i, Z = rN - j) \quad \forall i \in \mathbb{Z}, j \in \{-N/2+1, ..., 0, ..., N/2\}. \quad (50)$$

The joint characteristic function of X and Z is defined as

$$\Phi_{X,Z}(w_1, w_2) = \sum_{m_1=-\infty}^{\infty} \sum_{m_2=-\infty}^{\infty} e^{jw_1 m} e^{jw_2 m_2} P(X = m_1, Z = m_2).$$

The second summation in the above expression can be split into a summation over blocks of length N,

$$\Phi_{X,Z}(w_1,w_2)=\sum_{m_1=-\infty}^{\infty}\sum_{r=-\infty}^{\infty}\sum_{l=-N/2+1}^{N/2}e^{jw_1m}e^{jw_2(rN-l)}P(X=m_1,Z=rN-l),$$

where the substitution $m_2 = rN - l$ is made. On substituting $w_2 = 2\pi k/N$, noting that $e^{j2\pi kr} = 1$ the above expression further simplifies to,

$$\Phi_{X,Z}\left(w_1,\frac{2\pi k}{N}\right)=\sum_{m_1=-\infty}^{\infty}\sum_{r=-\infty}^{\infty}\sum_{l=-N/2+1}^{N/2}e^{jw_1m}e^{-j\frac{2\pi kl}{N}}P(X=m_1,Z=rN-l).$$

Substituting (50) reduces the above expression to

$$\begin{aligned}\Phi_{X,Z}\left(w_1,\frac{2\pi k}{N}\right)&=\sum_{m_1=-\infty}^{\infty}\sum_{l=-N/2+1}^{N/2}e^{jw_1m_1}e^{j\left(\frac{-2\pi k}{N}\right)l}P(X=m_1,Y=l)\\&=\Phi_{X,Y}\left(w_1,\frac{-2\pi k}{N}\right).\end{aligned}$$

Substituting k for $-k$ implies the result.

■

Lemmas B3, and B4 suppose the following:

$$z[n]=\underbrace{\sum_{m=n_0}^{n}s[m]a[n-m]}_{s_a[n]}+\sum_{m=n_0}^{n}d[m]h[n-m]$$

where $d[n]$ is a sequence of *iid* random variables each independent of $s[n]$ for every $n \geq n_0$. All variables are assumed to be zero for $n < n_0$. The characteristic function of each of the identical random variables $d[n]$ is denoted by $\Phi_d(w)$.

Lemma B3: For any given finite, positive integer p,

$$\Phi_{s[n-p],z[n]}(w_1,w_2)=\Phi_{s[n-p],s_a[n]}(w_1,w_2)\cdot\prod_{r=0}^{n-n_0}\Phi_d(w_2h[r])\quad\forall w_1,w_2\in[0,2\pi),$$

where $\Phi_{s[n-p],z[n]}(w_1,w_2)$ is the joint characteristic function of $z[n]$ and $s[n-p]$ and

$\Phi_{s[n-p],s_a[n]}(w_1, w_2)$ is the joint characteristic function of $s_a[n]$ and $s[n\text{-}p]$.

Proof: The joint characteristic function of $z[n]$ and $s[n\text{-}p]$ is,

$$\Phi_{s[n-p],z[n]}(w_1, w_2) = E\left\{e^{jw_1 s[n-p]} e^{jw_2 z[n]}\right\} = E\left\{e^{jw_1 s[n-p]} e^{jw_2 s_a[n]} \prod_{m=n_0}^{n} e^{jw_2 d[m]h[n-m]}\right\}.$$

Since the samples of $d[n]$ are independent of themselves and $s[l]$ for all $l \geqslant n_0$, the above expression reduces to,

$$\Phi_{s[n-p],z[n]}(w_1, w_2) = E\left\{e^{j(w_1 s[n-p]+w_2 s_a[n])}\right\} \prod_{m=n_0}^{n} E\left\{e^{j(w_2 d[m]h[n-m])}\right\}.$$

Note that the expectations in the above expression are the characteristic functions, $\Phi_{s[n-p],s_a[n]}(w_1, w_2)$, and $\Phi_d(w)$, with the appropriate arguments. Therefore, the above expression reduces to

$$\Phi_{s[n-p],z[n]}(w_1, w_2) = \Phi_{s[n-p],s_a[n]}(w_1, w_2) \cdot \prod_{m=n_0}^{n} \Phi_d(w_2 h[n-m]).$$

Substituting $r = n - m$ in the product proves the lemma.

■

Lemma B4: For any given finite, positive integer p,

$$\Phi_{d[n-p],z[n]}(w_1, w_2) = \Phi_{s_a[n]}(w_2) \cdot \Phi_d(w_2 h[p] + w_1) \cdot \prod_{\substack{r=0,\\ r\neq p}}^{n-n_0} \Phi_d(w_2 h[r]) \quad \forall w_1, w_2 \in [0, 2\pi),$$

where $\Phi_{s_a[n]}(w)$ is the characteristic function of $s_a[n]$.

Proof: The joint characteristic function of $z[n]$ and $d[n\text{-}p]$ is,

$$\Phi_{d[n-p],z[n]}(w_1,w_2)=E\left\{e^{jw_1d[n-p]}e^{jw_2z[n]}\right\}=E\left\{e^{j(w_1+w_2h[p])d[n-p]}e^{jw_2s_a[n]}\prod_{\substack{m=n_0,\\ m\neq n-p}}^{n}e^{jw_2d[m]h[n-m]}\right\}.$$

Since the samples of $d[n]$ are independent of themselves and $s[l]$ for all $l \geqslant n_0$, the above expression reduces to,

$$\Phi_{d[n-p],z[n]}(w_1,w_2)=E\left\{e^{jw_2s_a[n]}\right\}\cdot E\left\{e^{j(w_1+w_2h[p])d[n-p]}\right\}\cdot\prod_{\substack{m=n_0,\\ m\neq n-p}}^{n}E\left\{e^{jw_2d[m]h[n-m]}\right\}.$$

Note that the expectations in the above expression are the characteristic functions, $\Phi_{s_a[n]}(w)$, and $\Phi_d(w)$, with the appropriate arguments. Therefore, the above expression reduces to

$$\Phi_{d[n-p],z[n]}(w_1,w_2)=\Phi_{s_a[n]}(w_2)\cdot\Phi_d(w_1+w_2h[p])\cdot\prod_{\substack{m=n_0,\\ m\neq n-p}}^{n}\Phi_d(w_2h[n-m]).$$

Substituting $r = n - m$ in the product proves the lemma.

■

Lemma B5 supposes the following:

$$z_i[n]=\underbrace{\sum_{m=n_0}^{n}s[m]a_i[n-m]}_{s_{ai}[n]}+\sum_{m=n_0}^{n}d[m]h_i[n-m]\quad\forall i=1,2$$

where $d[n]$ is a sequence of *iid* random variables each independent of $s[n]$ for every $n \geq n_0$. All variables are assumed to be zero for $n < n_0$. The characteristic function of each of the identical random variables $d[n]$ is denoted by $\Phi_d(w)$. The sequences $a_i[n]$, $h_i[n]$ are the impulse responses of the filters $A_i(z)$ and $H_i(z)$ respectively, for i = 1 or 2

in Fig. 3.11.

Lemma B5: For any given positive integer p, and $i, j = 1$ or 2,

$$\Phi_{z_i[n-p],z_j[n]}(w_1,w_2) = \Phi_{s_{ai}[n-p],s_{aj}[n]}(w_1,w_2)\cdot\prod_{r=0}^{n-n_0}\Phi_d\left(w_1h_i[r]+w_2h_j[r+p]\right)$$
$$\cdot\prod_{r=0}^{p-1}\Phi_d\left(w_2h_j[r]\right)\quad \forall w_1,w_2\in[0,2\pi),$$

where $\Phi_{z_i[n-p],z_j[n]}(w_1,w_2)$ the joint characteristic function of $z_i[n\text{-}p]$ and $z_j[n]$, and $\Phi_{s_{ai}[n-p],s_{aj}[n]}(w_1,w_2)$ is the joint characteristic function of $s_{a_i}[n-p]$ and $s_{a_j}[n]$.

Proof: The joint characteristic function of $z_i[n]$ and $z_j[n+p]$ can be expressed as,

$$\Phi_{z_i[n-p],z_j[n]}(w_1,w_2) = E\left\{e^{jw_1z_i[n-p]}e^{jw_2z_j[n]}\right\}$$
$$= E\left\{e^{j(w_1s_{ai}[n-p]+w_2s_{aj}[n])}\prod_{m=n_0}^{n-p}e^{jd[m]\left(w_1h_i[n-p-m]+w_2h_j[n-m]\right)}\prod_{m=n-p+1}^{n}e^{jd[m]w_2h_j[n-m]}\right\}.$$

Since the samples of $d[n]$ are independent of themselves and $s[l]$ for all $l \geqslant n_0$, the above expression reduces to,

$$\Phi_{z_i[n-p],z_j[n]}(w_1,w_2) = E\left\{e^{j(w_1s_{ai}[n-p]+w_2s_{aj}[n])}\right\}\cdot\prod_{m=n_0}^{n-p}E\left\{e^{jd[m]\left(w_1h_i[n-p-m]+w_2h_j[n-m]\right)}\right\}$$
$$\cdot\prod_{m=n-p+1}^{n}E\left\{e^{jd[m]w_2h_j[n-m]}\right\}$$
$$= \Phi_{s_{ai}[n-p],s_{aj}[n]}(w_1,w_2)\cdot\prod_{m=n_0}^{n-p}\Phi_d\left(w_1h[n-p-m]+w_2h[n-m]\right)$$
$$\cdot\prod_{m=n-p+1}^{n}\Phi_d\left(w_2h[n-m]\right).$$

Substituting $r = n - p - m$ in the first product, and $r = n - m$ in the second product reduces the above expression to

$$\Phi_{z_i[n-p],z_j[n]}(w_1,w_2)=\Phi_{s_{ai}[n-p],s_{aj}[n]}(w_1,w_2)\cdot\prod_{r=0}^{n-p-n_0}\Phi_d\left(w_1h[r]+w_2h[r+p]\right)\cdot\prod_{r=0}^{p-1}\Phi_d\left(w_2h[r]\right).$$

■

Lemma B6: Suppose that d is an integer random variable with a probability distribution,

$$P(d[n]=m)=\begin{cases}0.5, & m=0,\\ 0.5, & m=1.\end{cases}$$

and its characteristic function is $\Phi_d(w)$. Also suppose that N is a positive integer and that $g[r]$ is an integer sequence. If the sequence $g[r]\bmod N$ does not converge to zero as $r\to\infty$ then

$$\lim_{L\to\infty}\prod_{r=0}^{L}\Phi_d\left(\frac{-2\pi g[r]}{N}\right)=0. \tag{51}$$

Proof: The characteristic function of the random variable under question is

$$\Phi_d(w)=0.5\left(1+e^{jw}\right)=e^{jw/2}\cos\left(w/2\right). \tag{52}$$

Consequently,

$$\left|\Phi_d(-w)\right|<1 \qquad \forall w\neq 2m\pi, m\in\mathbb{Z}.$$

So, if $\left(k\cdot f[r]\right)\bmod N$ does not converge to zero as $r\to\infty$ then

$$\left|\Phi_d\left(\frac{-2\pi g[r]}{N}\right)\right|<1$$

for an infinite number of values of r. Since the characteristic function is bounded by unity, the infinite product in the LHS of (51) equals zero. This proves the lemma.

■

Lemma B7: Suppose that d is an integer random variable taking on consecutive integers 0 and 1 with equal probability and its characteristic function is $\Phi_d(w)$. Also suppose that N is a positive integer and g is any integer. If $g \bmod N = N/2$ then

$$\Phi_d\left(\frac{-2\pi g}{N}\right) = 0.$$

Proof: As shown in (52), the characteristic function of the random variable under question is

$$\Phi_d(w) = e^{jw/2}\cos(w/2). \tag{53}$$

Consequently,

$$\Phi_d(-w) = 0 \qquad \forall w = (2m+1)\pi, m \in \mathbb{Z}.$$

So, if $g \bmod N = N/2$ then

$$\Phi_d\left(\frac{-2\pi g}{N}\right) = 0.$$

This proves the lemma.

■

APPENDIX C

Theorem C1: Suppose $H(z) = z^{-1}(1-z^{-1})^{-1}$. Then none of the conditions of Theorem 1 are satisfied for one or more positive values of p.

Proof: The impulse response of $H(z)$ is $h[r] = u[r-1]$. Suppose $k_1 = k \neq N/2$ and k_2

$= N - k$. Then,

$$(k_1 h[r] + k_2 h[r+p]) \bmod N = \left(ku[r-1] + (N-k)u[r+p-1]\right) \bmod N$$
$$= \begin{cases} (N-k) & \forall r = 0 \\ 0 & \forall r \geq 1 \end{cases}.$$

This implies that there exist values of p for which conditions 1 and 2 of Theorem A2 are not satisfied. Moreover, for these values of p, $(k_2 h_1[r]) \bmod N = k \neq N/2 \quad \forall r \geq 0$. So, condition 3 of Theorem A2 is not satisfied either.

■

Theorem C2: Suppose $H(z) = z^{-2}(1 - z^{-1})^{-2}$. Then at least one of the three conditions of Theorem 1 are satisfied for all $p > 0$. Furthermore, there exists at least one value of $p > 0$ for which condition 1 of Theorem 1 is satisfied for all integers k_1, k_2, such that $k_1 + k_2 \neq 0$, and $0 \leqslant k_1, k_2 \leqslant N - 1$.

Proof: As proved below, for most positive values of p, condition 1 is satisfied. Therefore, this proof identifies situations in which condition 1 is not satisfied. Then it is proved that in such situations, either condition 2 or condition 3 is satisfied.

The impulse response of $H(z)$ is $h[r] = (r - 1)u[r - 2]$. Substituting this in the expression for condition 1 of Theorem 1 results in,

$$\left(k_1 h[r] + k_2 h[r+p]\right) \bmod N = \begin{cases} 0 & \forall r < 2 - p \\ \left(k_2[r+p-1]\right) \bmod N & \forall 2 - p \leq r < 2. \\ \left([k_1 + k_2][r-1] + k_2 p\right) \bmod N & \forall 2 \leq r \end{cases} \quad (54)$$

Condition 1 of Theorem 1 is not satisfied only if

$$[k_1 + k_2][r-1] + k_2 p = mN, \quad m \in \mathbb{Z}, \quad r \geq r_0 \geq 2. \quad (55)$$

To determine the cases where condition 1 is not satisfied, suppose that (55) is true. Then, there exist two consecutive integers $r_*, r_*+1 > r_0$, which satisfy (55):

$$\begin{aligned}[k_1+k_2][r_*-1]+k_2p &= m_1N,\\ [k_1+k_2][r_*]+k_2p &= m_2N, \quad m_1,m_2 \in \mathbb{Z}.\end{aligned} \tag{56}$$

By subtracting the first of the two equations in (56) from the second we get,

$$\begin{aligned}k_1+k_2 &= m_3N,\\ k_2p &= m_4N, \quad m_3,m_4 \in \mathbb{Z}.\end{aligned} \tag{57}$$

So, condition 1 of Theorem 2 is true for all triplets (k_1, k_2, p) except those specified by (57). As shown below, most of these triplets satisfy condition 3. The ones that do not are shown to satisfy condition 2.

Since $0 < k_2 < N = 2^M$, k_2 can be expressed as $k_2 = 2^s(2t + 1)$ where s, t are non-negative integers such that $s < M$. Therefore, the integer values of p which satisfy the second equation of (57) are $p = 2^{M-2}(2t+1)m_5$, where m_5 is a positive integer. The choice $r_2 = 1 + N/2^{s+1}$ results in $k_2h[r_2] \bmod N = N/2$ since,

$$k_2h[r_2] \bmod N = k_2(r_2-1) \bmod N = (2t+1)\frac{N}{2} \bmod N = \frac{N}{2}.$$

Except for the case where $s = M - 1$, $p = 2$ *i.e.*, ($k_2 = N/2$, $p = 2$), it is readily shown that $r_2 < p$:

$$r_2 = 1+\frac{N}{2^{s+1}} = 1+\frac{2^{M-s}}{2} < 2^{M-s}(2t+1)m_5.$$

Hence condition 3 of the theorem is satisfied for all triplets of (57) except ($N/2$, $N/2$, 2). For $p = 2$, choosing $r_{1,2} = 0 \neq p$ in equation (54) implies that condition 2 of the theorem is satisfied for this triplet. Hence, for all $p > 0$, at least one of conditions 1–3

is satisfied.

Moreover, choosing $p < N$ to be relatively prime to N ensures that no integer k_2 such that $0 \le k_2 \le N-1$ satisfies the second equation in (57). For example, if p = 3, and $3k_2 = m_4 N$ is not satisfied for any $0 < k_2 \le N-1$ and integer m_4. This proves the second part of the theorem.

■

Theorem C3: Suppose $H(z) = z^{-L}(1-z^{-1})^{-L}$ and $L > 3$. Then the conditions of Theorem 1 are satisfied for all $p > 0$.

Proof: Define $H_S(z) \triangleq z^{-S}(1-z^{-1})^{-S}$ for positive integers S and suppose $h_S[r]$ is its impulse response. This results in a difference equation between $h_S[r]$ and $h_{S-1}[r]$,

$$h_S[r] - h_S[r-1] = h_{S-1}[r] \tag{58}$$

Then it needs to be shown that $h[r] = h_L[r]$ satisfies conditions of Theorem 1 for all $p > 0$. The theorem will be proved by the principle of mathematical induction.

Part (i): It has already been shown in Section III that $h_3[r]$ satisfies condition 1 of Theorem 1 for all $p > 0$.

Part (ii): It needs to be shown that if $h_S[r]$ satisfies condition 1 of Theorem 1 for a given $p > 0$, so does $h_{S+1}[r]$.

To prove by contradiction, for a given $p > 0$, and some $0 \leqslant k_1, k_2 \leqslant N-1$ such that $k_1 + k_2 \neq 0$, suppose that $h_S[r]$ satisfies condition 1 of Theorem 1, but $f_{S+1}[r]$ does not. Then $\left(k_1 h_{S+1}[r] + k_2 h_{S+1}[r+p]\right) \bmod N$ does not converge to zero as $r \to \infty$ but,

$$k_1 h_{S+1}[r] + k_2 h_{S+1}[r+p] = mN, \quad m \in \mathbb{Z}, \quad \forall r \geq r_0. \tag{59}$$

For all pairs of consecutive integers $r_* + 1$ and $r_* + 2$ where $r_* \geq r_0$, (59) implies the following:

$$\begin{aligned} k_1 h_{S+1}[r_* + 1] + k_2 h_{S+1}[r_* + 1 + p] &= m_1 N, \\ k_1 h_{S+1}[r_* + 2] + k_2 h_{S+1}[r_* + 2 + p] &= m_2 N. \end{aligned} \tag{60}$$

Subtracting the first of the two equations in (60) from the seconds results in

$$k_1 \left(h_{S+1}[r_* + 1] - h_{S+1}[r_*] \right) + k_2 \left(h_{S+1}[r_* + p + 1] - h_{S+1}[r_* + p] \right) = (m_2 - m_1) N. \tag{61}$$

Substituting (58) in (61) results in

$$k_1 h_S[r_*] + k_2 h_S[r_* + p] = m_3 N, \quad m_3 \in \mathbb{Z}, \quad r_* \geq r_0. \tag{62}$$

This contradicts the hypothesis and hence proves the theorem.

■

Proof of Theorem 3: Since $g_i[n]$ is integer valued by assumption and $y_i[n]$ is an integer multiple of N_i, just as in equation (6), the i^{th} requantizer error can be shown to be

$$\begin{gathered} e_i[n] = \frac{N_i}{2} - N_i \left\langle \frac{e_{i-1}[n] * f_i[n]}{N_i} + \frac{1}{2} \right\rangle \\ \text{where} \quad e_0[n] \triangleq \left(s[n] + v[n] * d[n] \right) * f_1[n]. \end{gathered} \tag{63}$$

In other words,

$$e_{i-1}[n] * f_i[n] = m_i N_i - e_i[n] \quad \text{where} \quad m_i \in \mathbb{Z}.$$

Similarly,

$$e_{i-2}[n] * f_{i-1}[n] = m_{i-1} N_{i-1} - e_{i-1}[n] \quad \text{where} \quad m_{i-1} \in \mathbb{Z}. \tag{64}$$

Substituting (64) in (63) results in

$$
\begin{aligned}
e_i[n] &= \frac{N_i}{2} - N_i\left\langle \frac{m_{i-1}N_{i-1} * f_i[n]}{N_i} + \frac{-e_{i-2}[n] * f_{i-1}[n] * f_i[n]}{N_i} + \frac{1}{2} \right\rangle \\
&= \frac{N_i}{2} - N_i\left\langle \frac{-e_{i-2}[n] * f_{i-1}[n] * f_i[n]}{N_i} + \frac{1}{2} \right\rangle
\end{aligned} \tag{65}
$$

since N_{i-1}/N_i is a positive integer by assumption. By proceeding recursively it is readily shown that

$$
\begin{aligned}
& e_i[n] = \frac{N_i}{2} - N_i\left\langle \frac{z_i[n]}{N_i} + \frac{1}{2} \right\rangle, \\
& z_i[n] = \sum_{m=n_0}^{n} s[m]q_i[n-m] + \sum_{m=n_0}^{n} d[m]t_i[n-m] \quad \forall 1 \le i \le K
\end{aligned} \tag{66}
$$

where $q_i[n]$ is the impulse response of the fictitious filter

$$
Q_i(z) \triangleq (-1)^{i-1} \prod_{l=1}^{i} F_l(z) \quad \forall 1 \le i \le K \tag{67}
$$

and $h_i[n]$ is the impulse response of the fictitious filter

$$
\begin{aligned}
T_i(z) &\triangleq (-1)^{i-1} \prod_{l=1}^{i} F_l(z) \cdot V(z) \\
&= Q_i(z) \cdot V(z) \quad \forall 1 \le i \le K.
\end{aligned} \tag{68}
$$

Consequently the behavior of the K^{th} requantization error, $e_K[n]$, is governed by the filter $T_K(z)$ defined in the Theorem statement. Setting $A_1(z) = Q_K(z)$, $A_2(z) = 0$, $H_1(z) = T_K(z)$ and $H_2(z) = 0$ in Theorems A2, A4 and A5 in succession with $i = 1$ proves the result.

■

Proof of Theorem 4: Equations (66)-(68) derived in the proof of Theorem 3 in Appendix A can be readily used in proving Theorem 4.

Part (i): Setting $A(z) = Q_j(z)$, and $H(z) = T_j(z)$ and applying Theorem A1 and Corollary

1 of Theorem A1 for every $p > 0$, and then applying Theorem A2 proves that in terms of ensemble statistics, $e_j[n]$ has the properties of *uniformity*, *signal-independence*, *dither-independence*, and *pair-wise independence*. The application of Theorem A3 proves that the $e_j[n]$ has these properties in a time-averaged sense as well, along with the property of *whiteness*, with mean and auto-covariance as given by (8), with N replaced by N_j for a given j. Repeating for all $1 \leqslant j \leqslant K$ proves the result.

Part (ii): For any given $p \geq 0$, setting $A_1(z) = Q_i(z)$, $A_2(z) = Q_j(z)$, $H_1(z) = T_i(z)$ and $H_2(z) = T_j(z)$ in Theorem A4 proves the result for a given i, j where $i \neq j$. This can be repeated for all $p \geq 0$ and all i, j where $1 \leq i, j \leq K, i \neq j$. An ergodicity proof similar to Theorem A3 can be readily provided for this situation as well, but is not included for sake of brevity. The result can be proved for all $p < 0$ by interchanging the roles of $A_1(z)$ and $H_1(z)$ with those of $A_2(z)$ and $H_2(z)$.

■

CHAPTER ACKNOWLEDGEMENTS

The text of this chapter is to be submitted for review and publication as two regular papers in the *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*. The dissertation author is the primary researcher. Ian Galton supervised the research which forms the basis of this chapter. Jared Welz assisted with formulation of some of the theorems. The author is grateful to Ashok Swaminathan for their help with reviewing the material in this chapter.

REFERENCES

1. J. C. Candy, G. C. Temes, "Oversampling Methods for A/D and D/A Conversion," *Oversampling Delta-Sigma Data Converters Theory, Design and Simulation*, New York: IEEE Press, 1992, pp. 1-25.

2. S. R. Norsworthy, R. Schreier, G. C. Temes, *Delta-Sigma Data Converters Theory, Design, and Simulation*, IEEE Press, pp. 75-121, 1996.

3. M. H. Perrott, M. D. Trott, C. G. Sodini, "A Modeling Approach for Σ–Δ Fractional-N Frequency Synthesizers Allowing Straightforward Noise Analysis," *IEEE Journal of Solid State Circuits*, Vol. 37, No. 8, August 2002, pp. 1028-38.

4. J. C. Candy, A. N. Huynh, "Double interpolation for digital-to-analog conversion," *IEEE Transactions on Communications*, Vol. COMM-34, pp. 77-81, January 1986.

5. K. Uchimura et al., "Over-sampling A-to-D and D-to-A converters with multistage noise shaping modulators," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 21, December 1991, pp. 1746-1756.

6. Sidney I. Resnick, *A Proability Path*, Birkhauser, Boston, 1998.

7. Ian Galton, "Granular Quantization Noise in the First-Order Delta-Sigma Modulator," *IEEE Transactions on Information Theory*, Vol. 39, No. 6, November 1993, pp. 1944-1956.