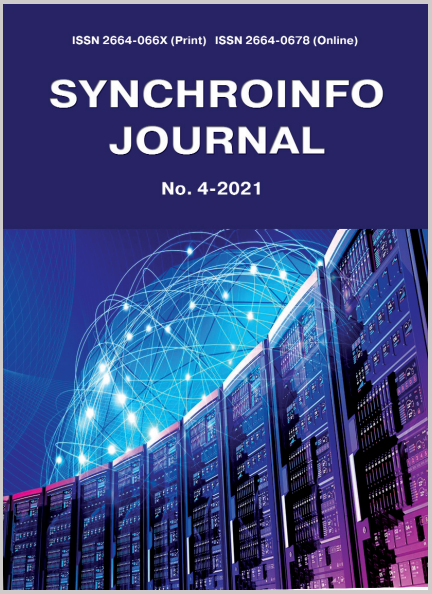
CONTENT
Vol. 7. No. 4-2021



Published bi-monthly since 2015.

ISSN 2664-066X (Print)
ISSN 2664-0678 (Online)

Publisher
Institute of Radio and Information Systems (IRIS), Vienna, Austria

Address:
1010 Wien, Austria,
Ebendorferstrasse 10/6b
media-publisher.eu/
synchroinfo-journal

AN EXPERIMENTAL VERIFICATION OF NEW NON-QUASI-STATIC SMALL-SIGNAL MOSFET MODEL

Tomasz Stefanski,
R&D Marine Technology Centre, Gdynia, Poland
stafan@ctm.gdynia.pl

Wieslaw J. Kordalski,
Gdansk University of Technology, Gdansk, Poland
kord@ue.eti.pg.gda.pl

Hans Hauer,
Fraunhofer Institut Integrierte Schaltungen, Am Wolfsmantel 33, 91058 Erlangen, Germany
hauer@iis.fraunhofer.de

DOI: 10.36724/2664-066X-2021-7-4-2-6

ABSTRACT

This article presents the results of an experimental verification of the New Non-Quasi-Static (NQS) Small-Signal MOSFET Model proposed in [1,2]. This model is valid in all operating modes, from weak to strong inversion and from nonsaturation to saturation. For the purpose of verification test transistors with de-embedding, dummy structures in 0.35 m technology were designed. The procedure of de-embedding was based on the open-short [3] method, optimised for RF measurement up to 30 GHz. The results obtained have confirmed aplicability of the model for the small-signal MOSFET simulation.

KEYWORDS: *Semiconductor device modeling, RF IC Design, CMOS and BiCMOS circuit simulations, circuits for communications.*

The article is reworked from unpublished 2nd IEEE International Conference on Circuits and Systems for Communications (ICCSC) materials.

I. INTRODUCTION

Small-signal MOSFET model and clear procedure of parameters extraction are needed to successfully design analog radio frequency integrated circuits (RFIC). When the device operates near or above the cut-off frequency, the NQS effects are strongly important. However, most models available in SPICE use the quasi-static (QS) formulation [4] which can result in unpredictable behavior of high-frequency circuits. QS approximation assumes the movable carriers in the channel of the transistor respond instantaneously to perturbations induced by a time-varying external signal, thus the channel charge achieve equilibrium once bias is applied.

As a result, serious inconsistencies arise when the QS approach is used to RF MOSFET modeling. For instance, according to the models presented in [5] magnitudes of transadmittances of voltage-controlled current sources tend to infinity as frequency increases. It is worth mentioning that any charge–based model, such as BSIM3v3, MOS Model 9 or EKV is inherently composed of such current sources as those in [5].

Also, for frequencies much smaller than the transistor cut-off frequency, NQS effects can occur. When an inductive load tunes out capacitance at some node, transistor behavior QS prediction can fail. Especially, PMOS devices can suffer from NQS effect because of lower holes mobility. Sometimes, it is possible to find an extremely long channel device, like 100µm with cut-off frequency below 1MHz where NQS effects can occur for low-frequencies [6].

The purpose of this work is to briefly present the results of an experimental verification in 0.35µm technology of the new NQS model, which was presented during 1st IEEE International Conference on Circuits and Systems for Communications in St. Petersburg, Russia [2]. We investigated admittance parameters of the measured test transistors and compared them to model simulation results. The experiment design and procedure of model parameters extraction are presented here.

II. NQS MODEL OF THE MOS TRANSISTOR [1-2]

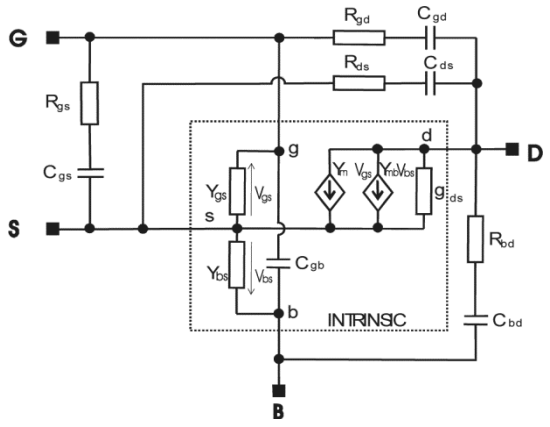


Fig. 1. Small-signal equivalent circuit of the MOSFET, with subcircuit enveloped in dotted line being the intrinsic part of the transistor

The equivalent circuit presented in the picture is simplified as compared to the one showed in [2] due to application of de-embedding strategy that allow us to reject the impact of pad parasitics. The set of equations establishing the model is as follows:

$$\mu(E_0)=\frac{\mu_0}{[1+(E_0/E_C)^{\beta}]^{1/\beta}},\left(E_0=\frac{|V_{DS}|}{L}\right), \quad (1)$$

$$C_{gb}=\frac{g_m\eta L}{(1+\eta)\mu(E_0)E_0},\left(\eta=\frac{g_{mb}}{g_m}\right), \quad (2)$$

$$Y_{mg}=g_m\exp(\gamma L),\ Y_{mb}=g_{mb}\exp(\gamma L), \quad (3)$$

$$Y_{gs}=j\omega\frac{C_{gb}D_C[\exp(\gamma L)-1]}{\eta\gamma L(1-\gamma V_T/E_0)}, \quad (4)$$

$$Y_{bs}=\eta^2 Y_{gs}, \quad (5)$$

$$\gamma=\frac{1}{2}\left(\frac{E_0}{V_T}-\sqrt{\frac{\sqrt{a^2+b^2}+a}{2}}-j\sqrt{\frac{\sqrt{a^2+b^2}-a}{2}}\right), \quad (6)$$

$$a=\left(\frac{E_0}{V_T}\right)^2,\ b=\frac{4\omega(1+D_C)}{V_T\mu(E_0)},\ V_T=\frac{kT}{q}. \quad (7)$$

where μ_0, EC and E0 are, respectively, the low-field mobility, the characteristic field, and the static longitudinal electric field at the Q-point. gds, gm and gmb are the QS drain-source conductance, gate transconductance and bulk transconductance.

For intrinsic part of the transistor we have $y_{12} = 0$. The intrinsic MOSFET without parasitics has to be unilateral ($y_{12} = 0$) for the reason that charge carriers (electrons or holes) are only injected through the source-channel barrier potential. When frequency goes to infinity, magnitude of transadmittance y_{21} goes to zero, and phase is linearly delayed. Transadmittance y_{21} predicted by the model behaves similarly to complex attenuated sinusoid, which was verified by experiment.

III. EXPERIMENT

To verify the model designed were eight test transistors in 0.35µm technology with geometries as follows (L denotes the channel length in µm, W – the channel width in µm, F – the number of fingers): (1) L=0.35 W=50 F=5, (2) L=0.5 W=50 F=5, (3) L=0.7

W=50 F=5, (4) L=1.4 W=50 F=5, (5) L=0.35 W=100 F=10, (6) L=0.5 W=100 F=10, (7) L=0.7 W=100 F=10, (8) L=0.35 W=200 F=20. The transistor geometries were optimized for DC measurement of the drain current with swept gate-source VGS, drain-source VDS and bulk-source VBS voltages and RF scattering parameters measurement of the transistor in common source configuration through Air Coplanar Probes. The strategy of de-embedding was based on the widely used open-short method [3].

Although transistor layouts provided possibility of drain current measurement with swept bulk-source voltage, short between bulk and source for RF measurement was obtained using ground of probes. Based on RF measurement of the short dummy structure, there were extrapolated values of the bulk and source inductance Lpar and resistance Rpar to ground in common source configuration. These elements cannot be removed from the device under test through de-embedding and had to be included in simulations of the model.

To obtain reliable and repeatable measurements four samples with described above eight transistors were measured. RF measurements were performed based on two different calibration techniques of the network vector analyzer (NWA). Two samples were measured making use of SOLT calibration method [3] in the 65MHz – 25GHz frequency range and other samples were measured making use of LRM method [3] in the 65MHz – 30GHz frequency range. The impedance substrate standard (ISS) was used to calibrate the NWA. Transmission lines on ISS were used to verify the quality of calibration. For all samples, it was obtained almost the same results, which eliminates risk of one-time wrong measurement. The procedure of transistor de-embedding was verified positively for all samples by measurement and de-embedding of known thru lines. Also, the characteristics obtained from the magnitude of hybrid h21-parameter let us treat measurements as reliable in the investigated frequency range. Transistors were measured at twenty four Q-points, V_{DS}[V]=0, 0.3, 0.6, 0.9, 2.1, 3,3 and V_{GS}[V]=0.9, 1.0, 1.1, 1.2.

In Fig. 2 and 3 shown are representative results, respectivelly, y12 and y21 transadmittances for 50µm-channel width transistors at Q-point VDS=3.3V and VGS=1.2V. gds, gm and gmb parameters were extracted through DC measurement and EC and µ values were taken from IC manufacturer data. Other values of the model parameters were obtained by manual curve-fitting of admittance experimental and simulated frequency characteristics. At first, real(y21) was fitted, which was aimed to obtain similar characteristic slope (L[µm]=0.35, 0.5, 0.7) or period for long channel transitor (L[µm]=1.4). Value of coupling parameter DC was extracted in this step. Afterwards, imag(y21) and imag(y12) were fitted by Cgd capacitance tuning. At VDS[V]=0.6, 0.9, 2.1, 3,3 Q-points, real(y12) was near to zero which agree with y12=0 thesis for intrinsic transistor. At these Q-points, extracted Rgd-resistance value was equal to zero. An increase in real(y12) for higher frequencies is a result of described above bulk-source short through ground of the measurement system. At VDS[V]=0, 0.3 Q-points it was necessary to tune Rgd to value different than zero, because of large decreasing tendency of real(y12) with frequency. In the next step, y11 characteristics were fitted, accurate fitting was obtained by Cgs and Rgs parameters tuning. Afterwards, it was possible to perfectly fit y22 characteristics through Cds, Rds and Cbd, Rbd tuning. In Fig 4 shown are representative results, respectivelly, y11 and y22 admittance parameters for transistor with L=0.35µm, W=50µm at Q-point VDS=3.3V, VGS=1.2V. Extracted parameter values for described transistors at Q-point VDS=3.3V and VGS=1.2V summarized are in Table.1.

IV. CONCLUSIONS

We presented the results of an experimental verification of the MOSFET model , we had proposed previously. The experimental results show that y21 MOSFET transadmittance vs. frequency behaves in fact like an attenuated complex sinusoid for long channel device (L=1.4µm), see Fig.3, which was theoretically predicted by the model. Of course, investigated frequency range is much wider than presented transistor cut-off frequency (fT=2.2GHz). The experiments also confirm real(y12)=0 at Q-points VDS[V]=0.6, 0.9, 2.1, 3,3 which agrees with thesis on y12=0 for intrinsic transistor. A simple parasitics circuit proposed in Fig.1 provides close fitting between measured and simulated y11 and y22 admittance characteristics. The model is in good agreement with the experimental results and does not have physical inconsistencies unlike other models.

Table 1

Extracted small-signal model parameters of 50µm-width transistors at Q-point V_{DS}=3.3V and V_{GS}=1.2V, other parameters: W=50um, F=5, β=2, E_C=2.812·106V/m, μ_0=475.8cm2/V/s, T=27°C, R_{gd}=0Ω, R_{par}=411mΩ, L_{par}=29.18pH; the cut-off frequencies are shown in the last row

	L=0.35um	*L=0.5um*	*L=0.7um*	*L=1.4um*
g_m*[mS]*	*9.99*	*7.684*	*5.822*	*3.236*
g_{mb}*[mS]*	*2.206*	*2.085*	*1.644*	*0.9488*
g_{ds}*[uS]*	*532*	*250*	*152*	*48*
D_C*[10^{-3}]*	*886.6*	*477.5*	*322.6*	*468.1*
C_{gd}*[fF]*	*8.433*	*9.57*	*9.863*	*14.16*
C_{gs}*[fF]*	*33.72*	*62.63*	*96.65*	*148.6*
C_{ds}*[fF]*	*22.18*	*20.8*	*20.8*	*20.8*
C_{bd}*[fF]*	*20.75*	*18.58*	*18.58*	*23.17*
R_{gs}*[Ω]*	*111.5*	*54.24*	*43.85*	*49.09*
R_{ds}*[Ω]*	*17.54*	*23.17*	*6.531*	*38.31*
R_{bd}*[Ω]*	*580.8*	*724.4*	*981.7*	*3020*
f_T*[GHz]*	*25.8*	*14.3*	*8*	*2.2*

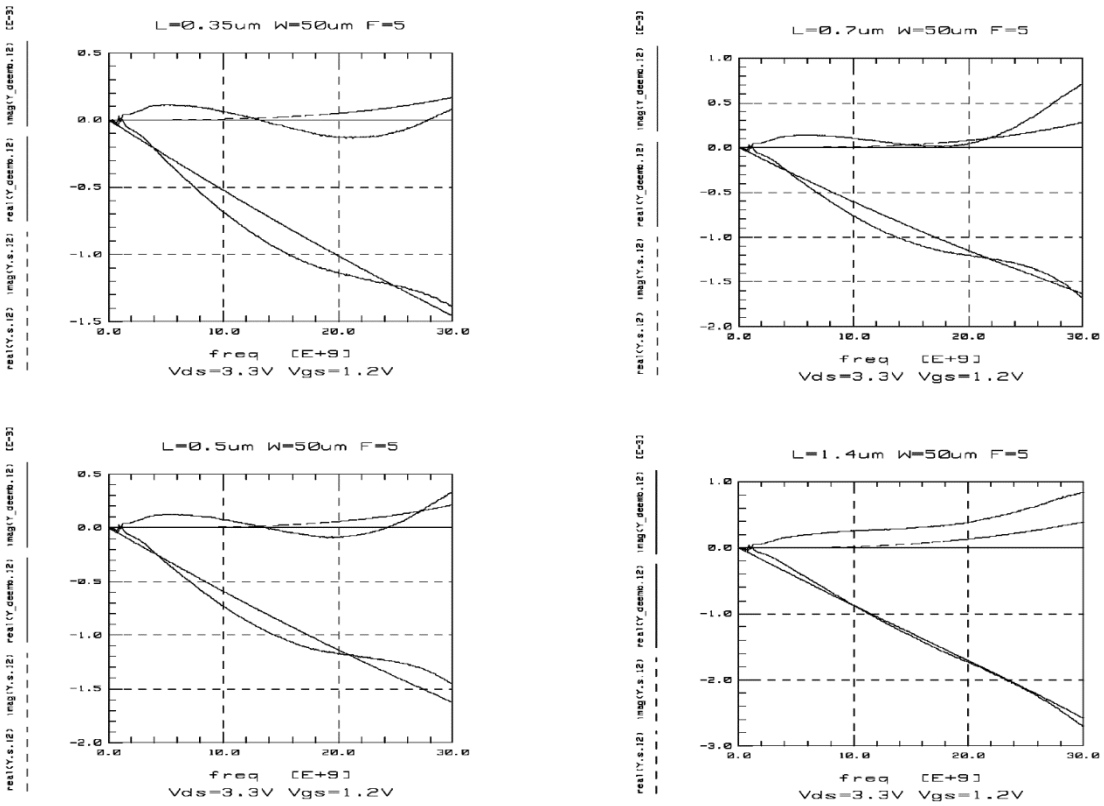


Fig. 2. Comparison between measured and simulated y12 parameters for transistors with L[µm]=0.35, 0.5, 0.7, 1.4, W=50µm (V_{DS}=3.3V , V_{GS}=1.2V)

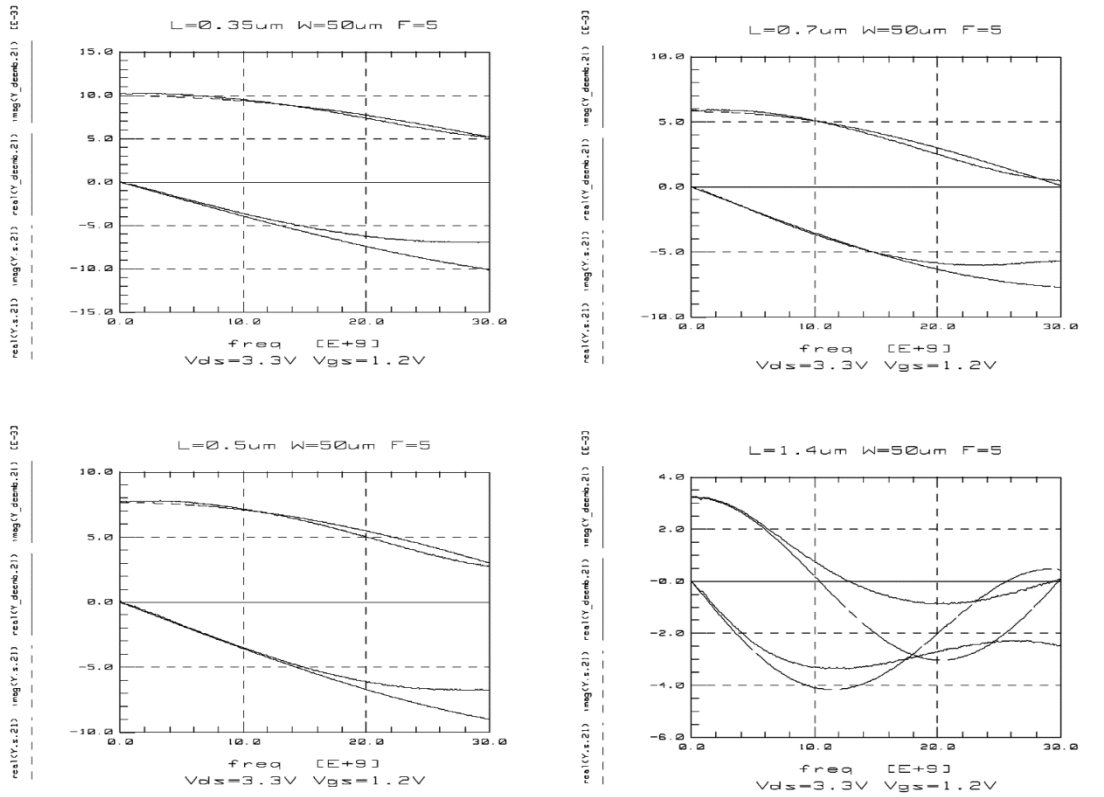


Fig. 3. Comparison between measured and simulated y21 parameters for transistors with L[µm]=0.35, 0.5, 0.7, 1.4, W=50µm (V_{DS}=3.3V , V_{GS}=1.2V)

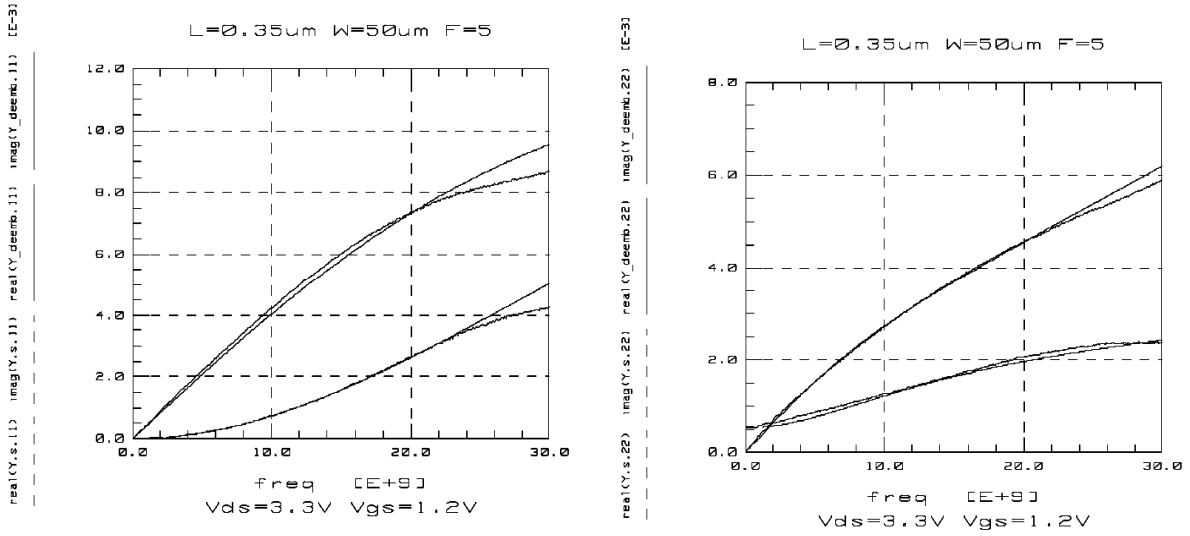


Fig.4. Comparison between measured and simulated y11 and y22 parameters for transistor with L=0.35µm, W=50µm (V_{DS}=3.3V , V_{GS}=1.2V)

REFERENCES

[1] W. Kordalski, "An Injection Non-Quasi-Static Small-Signal MOSFET Model", *International Conference on Signals and Electronic Systems*, 17-20 October 2000, Ustroń, Poland.

[2] W. Kordalski, T. Stefański "A Non-Quasi-Static Small-Signal Model of the Four-Terminal MOSFET for Radio and Microwave Frequencies", *1st IEEE International Conference on Circuits and Systems for Communications*, 26-28 June 2002, 2002, St.Petersburg, Russia.

[3] F. Sischka, Agilent Technologies, "IC-CAP Characterization & Modeling Handbook" http://eesof.tm.agilent.com/docs/iccap2002/iccap_mdl_handbook.html.

[4] M. Chan, K.Y. Hui, Ch. Hu, P.K. Ko, "A Robust and Physical BSIM3 Non-Quasi-Static Transient and AC Small-Signal Model for Circuit Simulation", *IEEE Trans. on Electron Devices*. Vol.45. No.4. 1998, pp. 834-841.

[5] C.C. Enz, Y.Cheng, "MOS Transistor Modelling for RF IC Design", *IEEE Trans. on Solid-State Circuits*. Vol.35. No.2. 2000, pp. 186-201.

[6] A.S. Porret, J.M. Sallese, C.C. Enz, "A Compact Non-Quasi-Static Extension of a Charge-Based MOS Model", *IEEE Trans. on Electron Devices*. Vol.48. No.8. 2001, pp. 1647-1654.

A 36 MW, 13 B, 2.1 MS/S MULTI-BIT ΔΣ ADC IN 0.18 M DIGITAL CMOS PROCESS USING AN EFFICIENT TOP-DOWN DESIGN METHODOLOGY

Mona Safi-Harb,
Microelectronics & Computer Systems Laboratory, McGill University, Canada,
mona@macs.ece.mcgill.ca

Gordon W. Roberts,
Microelectronics & Computer Systems Laboratory, McGill University, Canada,
roberts@macs.ece.mcgill.ca

DOI: 10.36724/2664-066X-2021-7-4-7-11

ABSTRACT

A systematic method to design a switched-capacitor (SC) multi-bit ΔΣ ADC integrated circuit is presented. The modulator consists of a fourth-order, multi-stage (2-1-1) architecture, with a 3-bit fash ADC in the last stage only. The modulator building blocks specifcations were designed using a systematic top-down methodology. Trade-offs between circuit building block specifcations, optimization time and computing resources are derived. When sampled at 50 MHz, measured performance reveals an 81.3 dB dynamic range for an output Nyquist rate of 2.1 MS/s while using a single 1.8 V supply and dissipating 36 mW of power.

KEYWORDS: *switched-capacitor, multi-bit fash ADC, optimization time, computing resources, digital CMOS technology.*

The article is reworked from unpublished 2nd IEEE International Conference on Circuits and Systems for Communications (ICCSC) materials.

I. INTRODUCTION

Delta-sigma modulators (ΔΣMs) constitute an essential block in mixed-signal designs. The increasingly stringent requirements of today's communication systems and portable devices are however rendering the design of those ΔΣMs more challenging. Extending the input frequency range to exceed the 1 MHz range, while maintaining a feasible sampling frequency is further rendered more complicated by the ever shrinking transistor dimension, and in turn, the supply voltage. The above two challenges are causing ΔΣMs to suffer from a long design cycle increasing therefore its time-to-market.

Synthesizing the modulator's main building block specifications on the system level is an efficient design method that has been suggested in the literature [1]-[3]. In this paper, an observation of the weak correlation among the parameters to be synthesized allows for the deduction of a fast synthesis method using Matlab/Simulink [4]. Parameters dictating the final performance of the modulator were separated reducing the complexity of the algorithm from exponential to linear dependency on the variables. In other words, the complexity of the algorithm used to deduce the optimum parameters, assuming n parameters exist, is reduced from On() to ·nO1(). Long optimization time could therefore be avoided and simple sweeps on the key parameters were enough to determine blocks' specifications on the system level.

This method will be demonstrated efficient in the implementation of a well-performing modulator. This paper is organized as follows: A detailed description of the system-level design of the modulator follows in section 2. The circuit implementation and experimental results from the fabricated IC are presented in sections 3 and 4 respectively. Concluding remarks are given in the last section of this paper.

II. SYSTEM-LEVEL DESIGN

2.1. Architecture

The selection of the architecture is the first step in the system-level design. In order to achieve a large dynamic range, a detailed study for the choice of the appropriate topology and proper topology parameters was outlined in [5][6]. Therefore, the assumption for the remainder of this paper is that the architecture is already selected. The focus will be on the impact of the circuit non-idealities on the system performance.

In order to meet the goal specifications: minimum of 14 bits of resolution for ~ 1 MHz input bandwidth, a 4th order, single-bit, cascade 2-1-1 architecture with a 3-bit quantizer in the last stage and 50 MHz sampling frequency, fs were chosen.

2.2. Non-Idealities Considered

The complete system-level representation of the non-ideal model of the integrator was achieved in Matlab/Simulink. The inputs to this block include but are not limited to the following non-idealities: operational transconductance amplifier (OTA) DC gain, unity-gain bandwidth (BW), slew rate (SR), input thermal noise density (S_n), switch on-resistance (R), and sampling capacitor (Cs). Other constants include the system oversampling ratio, OSR, Boltzman's constant (K) and the temperature (T). The comparator offset is another parameter that was included in the non-ideal characterisation of the modulator, even though noise-shaping alleviates its degradative effect on the overall system performance.

A description of the effect of each one of those non-idealities on the overall system performance is presented next. In this paper, the choice was made to carry out the optimization for parameters such as OTA DC gain, BW, SR, etc., rather than parameters such as OTA differential pair transconductance, output impedance, and biasing current, as was done in [2]. The latter method assumes an op amp topology constraining therefore the mapping of the obtained results to any other topology. The effect of each circuit non-ideality is explained next.

The OTA thermal noise is computed using a 2-pole system approximation. The root-mean-squared, RMS input-referred thermal noise is then given by

$$\sqrt{(S_n \cdot BW)/(2A)},$$

where A represents the finite DC gain of the OTA.

The switched- capacitor (also knows as KT/C) noise is computed according to

$$\sqrt{(K \cdot T)/(C_s \cdot OSR)}.$$

The distortion due to the front-end switches were deduced separately. The switches of the front-end sampling network introduce distortion due to the dependency of their on-resistance on the input voltage. The worst-case distortion due to the variation in the resistance of the input switches was quantified in [2] and is given by:

$$THD = e^{[-4R \cdot C_s \cdot f_s]^{-1}} / \left(1 - e^{[-4R \cdot C_s \cdot f_s]^{-1}}\right), \qquad (1)$$

where THD stands for the total harmonic distortion.

A plot of the magnitude of the THD,|THD|, as a function of the witch on-resistance reveals that in the case where C_s = 1 pF, f_s = 50 MHz and a maximum allowable distortion of 80 dB, a maximum switch on-resistance, R_{max} of 540 Ω is allowed.

All the previously mentioned inputs were then added together, with an input saturation block to represent the limited input swing of the OTA, to constitute the input to a modified delayed-integrator transfer function. The finite DC gain of the OTA, A, as well as capacitor ratio mismatch denoted by *a*, will transform the transfer function from the ideal $1/(z-1)$ transfer function for a delayed integrator to:

$$1LSB = (V_{FS+} - V_{FS-})/(2^N - 1). \quad (3)$$

The limited OTA SR and BW were also modeled using a Matlab function that computes the current output given a current input, previous output and three cases which are determined by a purely slewing behavior, purely exponential behavior or a combination of both. A Matlab function is dedicated to compute the output value after checking for each one of the three cases.

Finally the OTA output limited swing is modeled by a saturation block at the output of the modulator.

Accounting for all the non-idealities discussed above will transform the transfer function of the integrator from a single-input, single-output transfer function given by $1/(z-1)$ to a more complex model which was generated in Simulink and Matlab as mentioned earlier.

In order to study the simultaneous effects of some of the circuit non-idealities on the system's performance, 3-D simulation sweeps on the key non-idealities were carried out. Observations of the results shown in Figure 1 reveal that little correlation exists between the swept variables. The absence of maxima in the plots reveals that no one optimum solution exists, rather, the plots can be used to find the minimum specification needed for each one of the swept variables to meet the desired SNR. As an example, a plot of the SNR (where noise throughout includes harmonic bins) as a function of the OTA thermal noise density, S_n and BW is shown in Figure 1 (b). The monotonically increasing plots imply that these parameters could be considered uncorrelated over a practically large range of OTA BW and S_n. Similar conclusions could be drawn when observing the other plots in Figure 1, hence the validation of the separation of variable method, with the exception of BW and SR which both determine the settling behavior of the OTA and need to be swept simultaneously.

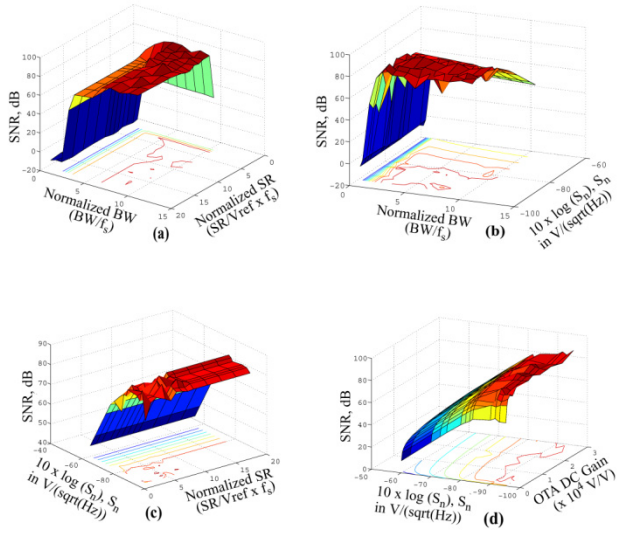


Fig. 1. Peak SNR versus: (a) BW and SR, (b) BW and S_n, (c) SR and S_n and (d) DC Gain and S_n

The above observation eliminates the need for a time-consuming, difficult-to-setup multi-dimensional optimization.

The difficulties associated with setting up multi-dimensional optimization and selecting a set of good initial conditions for convergence purposes could also be avoided. Instead, 1-D sweeps on each circuit non-ideality separately is enough for all practical purposes to deduce a set of specifications that will eventually be mapped into a transistor-level system.

In [7], the method presented above was used to design a single-bit ΔΣ ADC. The aim of the current work presented in this paper is to extend those results to include multi-bit modulators for higher achievable resolution. A complete section is therefore devoted for discussing the DAC non-linearities and their effect on the system performance, and is presented next.

2.3. DAC Non-Idealities in Multi-bit Quantization

As mentioned earlier, with multi-bit quantization, the DAC linearity becomes a concern when high resolution ΔΣ ADCs are required. If the DAC exhibits non-linearities that exceed the overall system linearity requirement, then the converter will be limited to the DAC accuracy. Different calibration techniques were proposed in the literature to improve the performance of the internal DAC through the use of individual level averaging, data weighted averaging, and other modified versions of the previous methods. Each one of those proposed methods has advantages and disadvantages which won't be discussed here.

One common feature to all these correction techniques is that they improve the DAC linearity at the expense of additional power dissipation. For the design under consideration, and in order to keep the power dissipation to a minimum, system-level simulations presented next will show that the required DAC Integral-Non-Linearity, INL is believed to be met with a simple resistor string and careful resistance choice and layout techniques. Note that multi-bit quantization in the last stage only was used as explained in great details in [6].

The effect of the DAC INL on the overall dynamic range of the modulator was simulated in Matlab/Simulink. The reference voltages in both the ADC and DAC of the last stage of the modulator were deviated from their ideal levels. To do so, an additive noise source having a Gaussian distribution, with variance (noise power) swept between 0 LSB (ideal DAC curve) and 1 LSB, where LSB = Least Significant Bit, and defined as:

$$1LSB = (V_{FS+} - V_{FS-})/(2^N - 1). \quad (3)$$

The DAC INL was then calculated using the best-fit line method. The results of the selected INL curves and the corresponding SNDR as a function of input power of the modulator are shown in Figure 2.

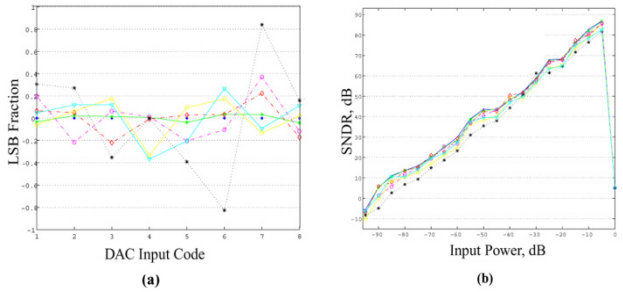


Fig. 2. (a) INL for different non-ideal DAC transfer curves and (b) corresponding SNDR

From the plots, it can be seen that provided that the maximum DAC INL is limited to 0.15 · LSB, the achievable resolution is ~ 14 bits.

2.4. Building Block Specifications Synthesis

Based on the observation of the little correlation existing between the variables, specifications on the building blocks were deduced using 1-D sweeps only. One variable representing one circuit non-ideality was swept at a time, while the other variables were made ideal.

Plots of the SNR versus different circuit non-idealities were all found to be monotonic, as expected. The minimum OTA DC gain, BW, SR, input thermal noise density, comparator offset as well as the input switching network time constant could therefore be deduced. The DAC maximum allowable non-linearity could be deduced from Figure 2.

III. SC CIRCUIT IMPLEMENTATION

3.1. Operational Transconductance Amplifier

A single-stage folded-cascode topology was used for the implementation of the OTA due to its excellent frequency characteristics. A complementary differential pair at the input was used in order to allow for a larger input swing. Gain boosting were added to the output stage in order to increase the OTA DC gain. Pole-zero modeling of the OTA was performed in order to choose the optimum biasing and transistor sizing while meeting the stringent requirements using a single 1.8 V supply and minimum power dissipation. A switched-capacitor common-mode feedback circuit was used due to its low power dissipation. The optimized OTA achieves 100 dB DC gain. When loaded with 1.5 pF capacitive load, the OTA achieves 410 MHz unity-gain bandwidth, a phase margin of 58°, and 236, 246 V/μs rising, falling slew rates respectively.

3.2. Comparator & Multi-bit Quantizer

A CMOS dynamic comparator and a clocked RS latch were used to implement the 1-bit quantizer. In the case of the multi-bit (3b) quantizer in the last stage, a flash ADC comprising of 7 dynamic CMOS comparators was used. The reference voltages were generated using a resistor string, and a switched-capacitor arrangement to carry out the subtraction between the reference levels and the integrator outputs [8]. The resistor values had to be chosen small enough to avoid long settling time, but large enough to minimize power dissipation. Values chosen for the resistors also have an impact on the matching which directly affects the resolution of the multi-bit ADC/DAC.

As explained earlier, the INL of the DAC had to be kept below 0.15 · LSB. Assuming a 3-bit DAC with full swing range of about 0.5 V, and using (3), we get that the noise on the DAC reference voltages due to mismatch in the resistor values generating those levels (as well as some other noise sources such as substrate noise) have to be kept below 10.7 mV. Since it is only matching between resistors and not absolute values that matters, with proper layout techniques, the matching between any two or more resistors could be made as small as 0.01% (1%being an achievable mismatch ratio with moderate layout techniques) [9].

Even with a moderate matching of 1% between the various resistors, the INL requirement of 10.7 mV is achievable with a simple resistor string.

It is worth mentioning that the choice of the resistor layout dimensions (width and length) also affects the matching percentage between two or more resistors due to process variations. All of the above considerations were taken into account when the resistors were laid out in order to minimize the process variation effects controllable by proper layout techniques. The final resistance choice was set to 200 Ω.

No pre-amplification stage was used for the comparator due to the its relaxed offset requirements. The implemented comparator achieves a resolution equivalent to 11 bits, capable therefore of resolving a differential signal of ~880 μV.

3.3. Switches

The switches were implemented using regular transmission gates. Despite the low voltage used to clock the gates of the switches, sizing the NMOS and PMOS according to (1) will limit the distortion of the front-end sampling network to an allowable level.

3.4. Other Blocks

The four-phase non-overlapping clocks were generated on-chip, from a single off-chip clock. Delayed versions of the clocks (together with fully differential implementation) were used to minimize charge injection. Both clocks and their complementary signals were generated for both the NMOS and PMOS transistors of the switches. Buffers appropriately sized for driving the clock line capacitive load with fast rise/fall times were inserted at the clock outputs.

The overall SC circuit is shown in Figure 3, and its mapping into a 0.18 μm, single-poly, 6-metal CMOS process IC is shown in Figure 4. Post-layout simulations were performed on the IC and it is worth mentioning that no transistor-level iterations were needed since no performance degradation in the SNR between the non-ideal system (Simulink) and transistor-level system (Cadence) were observed.

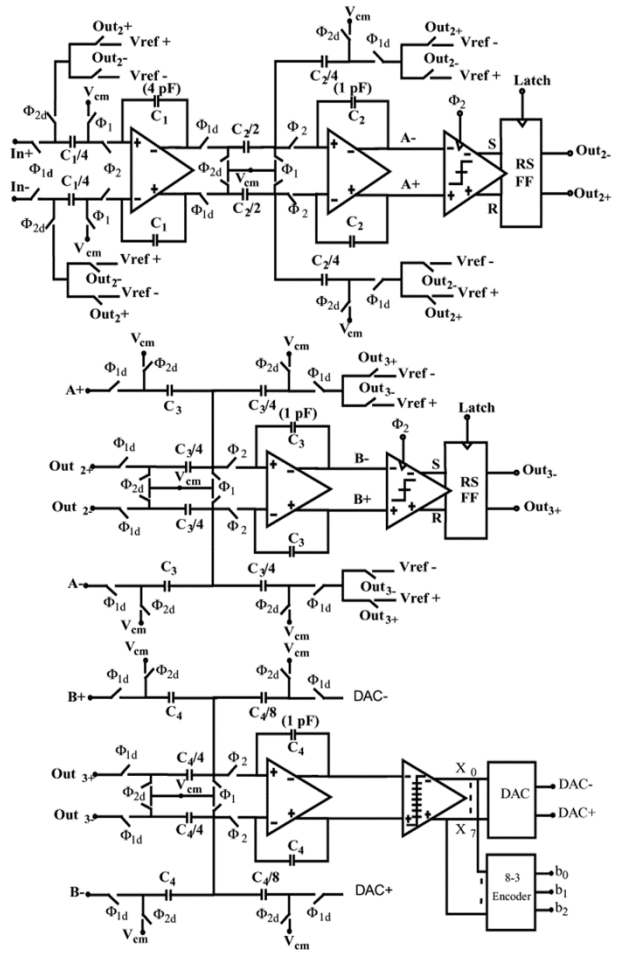


Fig. 3. SC implementation of the multi-bit modulator

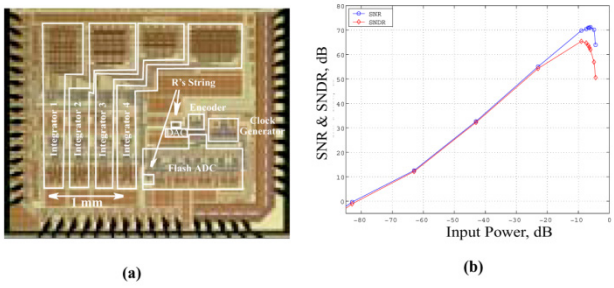


Fig. 4. IC (a) microphotograph and (b) measured DR

IV. EXPERIMENTAL RESULTS

A four-layer printed circuit board was used to test the IC. Separation of analog and digital planes, de-coupling capacitors and voltage regulation for the supply voltage lines were included. The differential input was generated using a Teradyne A567 mixed-signal tester and the clock using a Hewlett Packard high-quality pulse generator model number 81130A. The achieved dynamic range, DR shown in Figure 4 (b) is 13.2 bits of resolution. Table 1 summarizes the measured specifications of the IC.

V. CONCLUSIONS

We have presented the implementation and experimental results of a performing multi-bit ΔΣ ADC in 0.18 μm digital CMOS technology. Careful system-level modeling and synthesis of circuit specifications were used to minimize the power dissipation. It was shown that assuming independency among the circuit non-idealities and optimizing for each separately can provide a practical means for achieving a target specifications in the presence of complex and unpredictable interactions between the non-idealities. Experimental results agreed to some extent with the system-level modeling. The expected dynamic range was 14 bits while experimental results showed a resolution of 13.2 bits. Verification of the system-level modeling method was also conducted on a single-bit modulator, revealing even closer matching between expected and experimental results, justifying further the usefulness of the proposed method.

Table 1

Summary of measured specifications

Technology	0.18μm CMOS	***Output Rate***	~ 2 MS/s
Supply	1.8 V	***Ref. Volt.***	1.8 V, 0 V
Core Area	2.55x2.3 mm^2	***Input Range***	1.5 V_{p-p}
Power	36 mW	***DR***	81.3 dB
f_s	50 MHz	***Peak SNR***	72.7 dB
OSR	24	***Peak SNDR***	65.7 dB

ACKNOWLEDGMENTS

This work was supported by Micronet, a Canadian network of centres of excellence dealing with microelectronic devices, circuits and systems and the Canadian Microelectronics Corporation.

REFERENCES

[1] S. Brigati, F. Francesconi, P. Malcovati, D. Tonietto, A. Baschirotto, and F. Maloberti. Modeling Sigma-Delta Modulator Non-Idealities in Simulink, *Proc. IEEE ISCAS*, Vol. 2, pp. 384-387, 1999.

[2] N. Chandra. A Top-Down Approach to Delta-Sigma Modulator Design, *M. Eng. Thesis*, McGill University, 2001.

[3] K. Francken, P. Vancorenland, and G. Gielen. Dedicated System-Level Simulation of ΔΣ Modulators, *International Conference on Computer Aided Design*, pp. 188-192, 2000.

[4] The Math Works Inc., *SIMULINK and MATLAB User's Guides*, The Math Works Inc., 1997.

[5] A. Marques, V. Peluso, M. Steyaert, and W. Sansen. Optimal Parameters for ΔΣ Modulator Topologies, *IEEE TCAS II*, Vol. 45, No. 9, pp. 1232-1241, 1998.

[6] R. del Rio, F. Medeiro, J. M. de la Rosa, B. Pérez-Verdu, and A. Rodríguez-Vasquez. A 2.5-V ΣΔ modulator in 0.25-μm CMOS for ADSL, *Proc. IEEE ISCAS*, Vol. 3, pp. 301-304, 2002.

[7] M. Safi-Harb and G. W. Roberts. Design Methodology for Broad-band Delta-Sigma Analog-to-Digital Converters, *Proc. IEEE MWSCAS*, Vol. 2, pp. 231-234, 2002.

[8] B. Brandt and B. A. Wooley. A 50 MHz Multibit Sigma-Delta Modulator for 12-b 2-MHz A/D Conversion, *IEEE JSSC,* Vol. 26, No. 12, pp. 1746-1756, 1991.

[9] A. Hastings, *The Art of Analog Layout,* Prentice Hall, New Jersey, 2001.

COMPARATIVE ANALYSIS OF DIFFERENT RECEIVE ALGORITHMS FOR BLAST ARCHITECTURE IN MOBILE COMMUNICATION SYSTEMS

Salem Elabed
Institut Superieur des Etudes Technologiques en Communications de Tunis, Ariana, Tunis

DOI: 10.36724/2664-066X-2021-7-4-12-15

ABSTRACT

The Layered Space-Time Processing approach to STC was first introduced by Lucent's Bell Labs, with their BLAST family of STC structures. The information bits are demultiplexed into individual streams, which are then fed into individual encoders. These coders may be binary convolutional coders, or even no coding at all. The outputs of the coders are modulated and fed to the separate antennas, from which they are transmitted, using the same carrier-frequency/symbol waveform (TDMA) or Walsh code (CDMA). At the receiver, a spatial beamforming/nulling (zero-forcing) process is used at the front end in order to separate the individual coded streams, and feed them to their individual decoders. The outputs of the decoders are multiplexed back to reconstruct the estimate of the original information bitstream. Various receivers for communication system with Space-Time Coding in MIMO Channel are considered. Performance analysis was provided by statistical simulation for various parameters of communication system.

KEYWORDS: *Multiple-Input Multiple Output, BLAST, Space-Time Coding, Zero-Forcing, MMSE, V-BLAST detector*

The article is reworked from unpublished 2nd IEEE International Conference on Circuits and Systems for Communications (ICCSC) materials.

INTRODUCTION

The Layered Space-Time Processing approach to STC was first introduced by Lucent's Bell Labs, with their BLAST family of STC structures [1]. The basic concept behind layered STC is illustrated in Figure 1. The information bits are demultiplexed into individual streams, which are then fed into individual encoders. These coders may be binary convolutional coders, or even no coding at all. The outputs of the coders are modulated and fed to the separate antennas, from which they are transmitted, using the same carrier-frequency/symbol waveform (TDMA) or Walsh code (CDMA). At the receiver, a spatial beamforming/nulling (zero-forcing) process is used at the front end in order to separate the individual coded streams, and feed them to their individual decoders. The outputs of the decoders are multiplexed back to reconstruct the estimate of the original information bitstream.

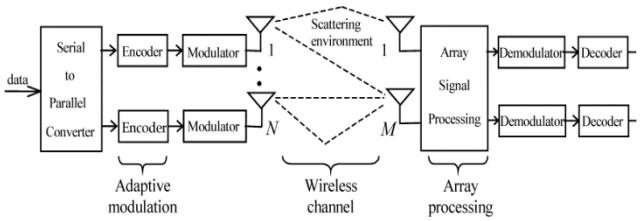


Fig. 1. Model of digital communication system with multiple transmitting and receiving antennas

Several types of BLAST structures were proposed [1]-[6]: Horizontal BLAST (H- BLAST), Diagonal BLAST (D-BLAST) and Vertical BLAST (V-BLAST). The system model with V-BLAST architecture is shown by the Fig.1. This model may be considered as a system with spatial multiplexing.

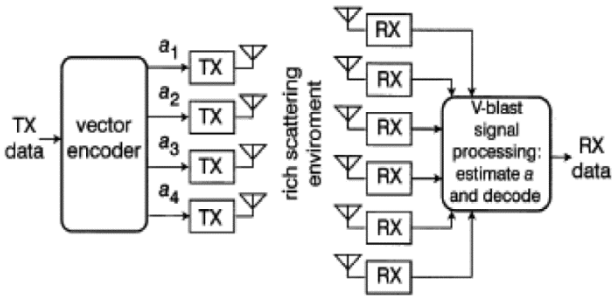


Fig. 2. V-BLAST high-level system diagram

The system model with V-BLAST can be described by the following observation equation:

$$Y = \sqrt{\frac{E_s}{N}}\mathbf{H}A + \eta\,,$$

where A – is the complex QAM symbol vector with dimension N; Y – is the complex observations vector with dimension M; H – is an ($M \times N$) dimensioned channel matrix, whose each element is complex Gaussian random value with zero mean and unit variance; η – is the complex Gaussian random vector with zero means and σ_n^2 variances; E_s – is energy of radiated signal.

MIMO RECEIVER ARCHITECTURES

In this section, we shall discuss receiver architectures for spatial multiplexing (N= M). Hence, receiver techniques (that have been studied in detail) such as zero-forcing (ZF), minimum-mean square error estimation (MMSE) and (optimal) maximum-likelihood sequence estimation (MLSE) can be applied directly.

The problem faced by a receiver for spatial multiplexing is the presence of multi-stream interference (MSI), since the signals launched from the different transmit antennas interfere with each other (recall that in spatial multiplexing the different data streams are transmitted co-channel and hence occupy the same resources in time and frequency). For the sake of simplicity we restrict our attention to the case (N= M).

A. Maximum-likelihood (ML) receiver. The ML receiver performs vector decoding and is optimal in the sense of minimizing the error probability. Assuming equally likely, temporally uncoded vector symbols, the ML receiver forms its estimate of the transmitted signal vector according to

$$A = \arg\min_{A}\left(\left\|Y - \sqrt{\frac{E_s}{N}}\mathbf{H}A\right\|^2\right)$$

where the minimization is performed over all possible transmit vector symbols A. Denoting the alphabet size of the scalar constellation transmitted from each antenna by Q, a brute force implementation requires an exhaustive search over a total of vector Q^N symbols rendering the decoding complexity of this receiver exponential in the number of transmit antennas. However, the recent development of fast algorithms [7], [8], [9] for sphere decoding techniques [10] offers promise to reduce computational complexity signi_cantly (at least for lattice codes). As already pointed out above, the ML receiver realizes N-th order diversity for Horizontal Encoding (HE) and (full) NM-th order diversity for Vertical Encoding (VE) and Diagonal Encoding (DE).

B. Linear receivers. We can reduce the decoding complexity of the ML receiver significantly by employing linear receiver front-ends to separate the transmitted data streams, and then independently decode each of the streams. We discuss the zero-forcing (ZF) and minimum mean squared error (MMSE) linear front-ends below.

ZF receiver: The ZF front-end is given by

$$G_{ZF} = \sqrt{\frac{M}{E_s}}\left(H^H H\right)^{-1} H^H$$

The output of the ZF receiver is obtained as

$$Z = A + G_{ZF}\eta,$$

which shows that the ZF front-end decouples the matrix channel into MT parallel scalar channels with additive spatially-colored noise. Each scalar channel is then decoded independently ignoring noise correlation across the processed streams. The ZF receiver converts the joint decoding problem into N single stream decoding problems (i.e., it eliminates MSI) thereby signi_cantly reducing receiver complexity. This complexity reduction comes, however, at the expense of noise enhancement which in general results in a signi_cant performance degradation (compared to the ML decoder). The diversity order achieved by each of the individual data streams equals (M-N+1) [11], [12].

MMSE Receiver: The MMSE receiver front-end balances MSI mitigation with noise enhancement and is given by

$$G_{MMSE} = \sqrt{\frac{N}{E_s}}\left(H^H H + \frac{\sigma_\eta^2 N}{E_s} I_N\right)^{-1} H^H.$$

In the low-SNR regime $E_s / \sigma_\eta^2 >> 1$, the MMSE receiver approaches the matched-filter receiver given by

$$G_{MMSE} = \frac{1}{\sigma_\eta^2}\sqrt{\frac{E_s}{N}} H^H,$$

and outperforms the ZF front-end (that continues to enhance noise). At high SNR $E_s / \sigma_\eta^2 >> 1$,

$$G_{MMSE} = G_{ZF}$$

i.e., the MMSE receiver approaches the ZFreceiver and therefore realizes (M-N+1)-th order diversity for each data stream.

Successive cancellation receivers. The key idea in a successive cancellation (SUC) receiver is layer peeling where the individual data streams are successively decoded and stripped away layer-by-layer. The algorithm starts by detecting an arbitrarily chosen data symbol (using ZF or MMSE) assuming that the other symbols are interference. Upon detection of the chosen symbol, its contribution from the received signal vector is subtracted and the procedure is repeated until all symbols are detected. In the absence of error propagation SUC converts the MIMO channel into a set of parallel SISO channels with increasing diversity order at each successive stage [2], [13]. In practice, error propagation will be encountered, especially so if there is inadequate temporal coding for each layer.

The error rate performance will therefore be dominated by the first stream decoded by the receiver (which is also the stream experiencing the smallest diversity order).

Ordered successive cancellation receivers. An improved SUC receiver is obtained by selecting the stream with the highest SINR at each decoding stage. Such receivers are known as ordered successive cancellation (OSUC) receivers or in the MIMO literature as V-BLAST [3], [4].

OSUC receivers reduce the probability of error propagation by realizing a selection diversity gain at each decoding step. The OSUC algorithm requires slightly higher complexity than the SUC algorithm resulting from the need to compute and compare the SINRs of the remaining streams at each stage.

SIMULATION

BER versus SNR curves are shown by the Figures 3-6 for ZF receiver, MMSE receiver V-BLAST receiver with ZF algorithm, V-BLAST receiver with MMSE algorithm and optimal ML receiver.

QPSK case and N=M=4 is shown by the Fig.3. Spectral efficiency of such system is 8 bps/Hz. 16QAM case and N=M=4 is shown by the Figure 4. Spectral efficiency of such system is 16 bps/Hz. QPSK case and N=M=8 is shown by the Figure5.

Spectral efficiency of such system is 16 bps/Hz. 16QAM case and N=M=8 is shown by the Figure 6. Spectral efficiency of such system is 32 bps/Hz. Channel with independent Rayleigh fadings is considered in all mentioned cases.

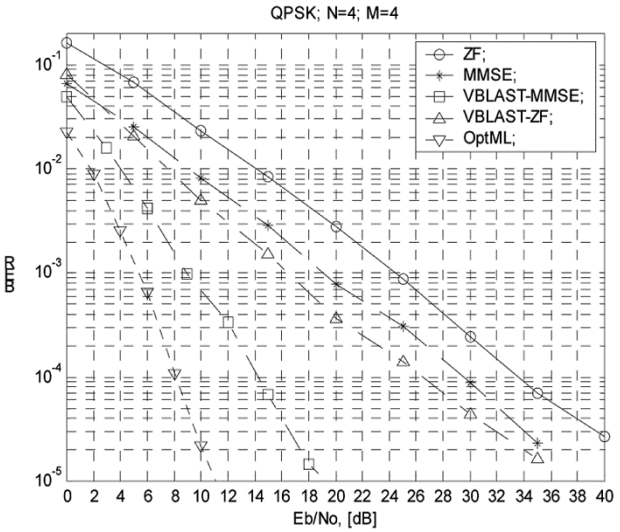


Fig. 3. BER versus SNR for N=M=4 and QPSK modulation

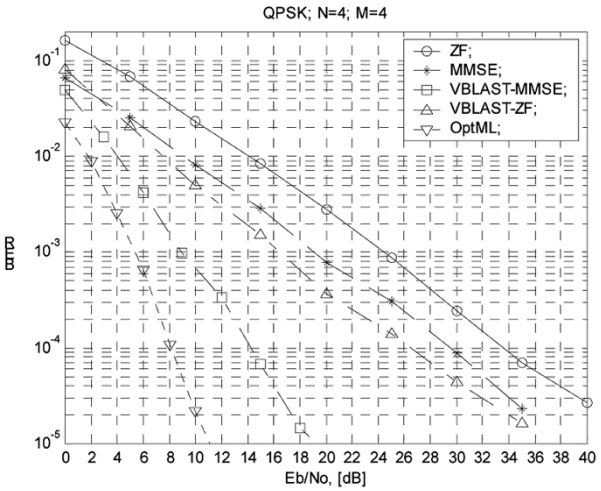


Fig. 4. BER versus SNR for N=M=4 and 16QAM modulation

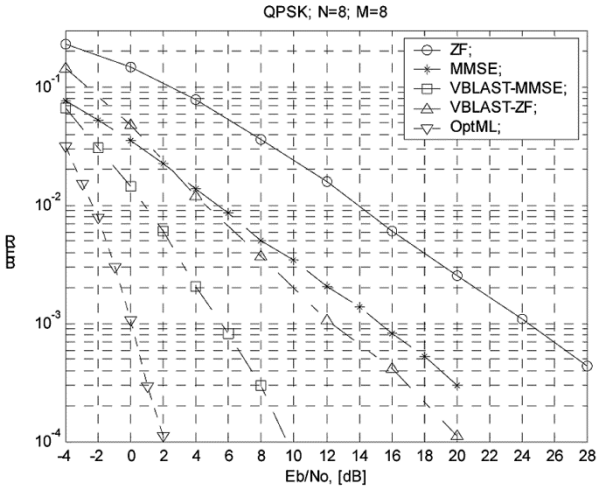


Fig. 5. BER versus SNR for N=M=8 and QPSK modulation

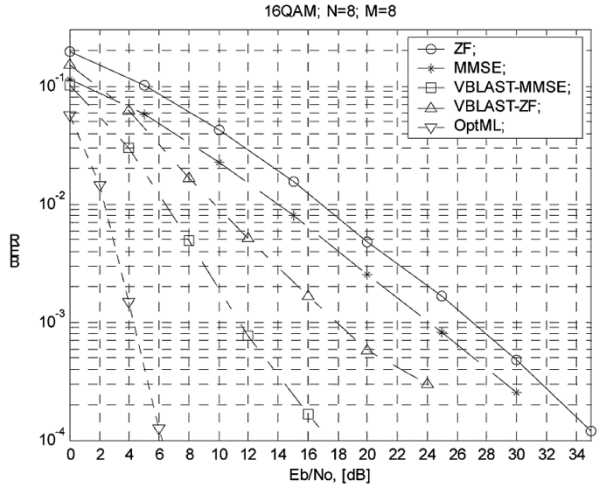


Fig. 6. BER versus SNR N=M=8 and 16QAM modulation

Looking at these plots, it is possible to se, that V-BLAST receiver with MMSE algorithm in each iteration has best performance among suboptimal reception algorithms. This receiver has about 4…9 dB losses at BER=0.001, versus optimal receiver. Note, that efficiency of such algorithm drops for high order modulation.

REFERENCES

[1] G.J. Foschini. Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas, *Bell Labs Technical Journal*, Vol. 1, No. 2, Autumn 1996, pp. 41-59.

[2] G.J. Foschini. and M.J. Gans. On Limits of Wireless Communications in Fading Environment when using Multiple Antennas, *Wireless Personal Communications*, Vol. 6, 1998, Kluwer Academic Publishers, pp. 311-335.

[3] P. Wolniansky, G. Foschini, G. Golden, and R. Valenzuela, V-BLAST: An Architecture for Realizing Very High Data Rates Over the Rich- Scattering Wireless Channel. *ISSSE '98*, oct 1998.

[4] G.D. Golden, C.J. Foschini, R.A. Valenzuela and P.W. Wolniansky. Detection algorithm and initial laboratory results using V-BLAST space-time communication architecture, *Electronics Letters,* 7th January 1999, Vol. 35 No. 1.

[5] Da-shan Shiu and Joseph M. Kahn. Layered Space-Time Codes for Wireless Communications Using Multiple Transmit Antennas, 1999 IEEE.

[6] D. Shiu and J.M. Kahn. Layered space-time codes for wireless communications using multiple transmit antennas, *ICC 99*, Vancouver, Canada, June 1999.

[7] E. Viterbo and J. Buotros. A universal lattice code decoder for fading channels, *IEEE Trans. Inf. Theory*, vol. 45, pp. 161-163, May 2000.

[8] O. Damen, A. Chkeif, and J. C. Belfiore. Lattice code decoder for space-time codes, IEEE Comm. *Letters*, vol. 4, no. 5, pp. 161-163, May 2000.

[9] B. Hassibi and H. Vikalo. On the expected complexity of sphere decoding, *Proc. Asilomar Conf. on Signals, Systems and Computers,* vol. 2, Nov. 2001, pp. 1051-1055.

[10] U. Fincke and M. Pohst. Improved methods for calculating vectors of short length in a lattice, including a complexity analysis, *Mathematics of Computation*, vol. 44, pp. 463-471, April 1985.

[11] J. H. Winters, J. Salz, and R. D. Gitlin. The impact of antenna diversity on the capacity of wireless communications systems, *IEEE Trans. Comm*., vol. 42, no. 2, pp. 1740-1751, Feb. 1994.

[12] D. Gore, R. W. Heath, and A. Paulraj. On performance of the zero forcing receiver in presence of transmit correlation, *Proc. IEEE ISIT, Lausanne*, Switzerland, July 2002, p. 159.

[13] S. Loyka and F. Gagnon. Performance analysis of the V-BLAST algorithm: An analytical approach, *Proc. International Zurich Seminar on Broadband Communications*, Feb. 2002, pp. 5.1-5.6.

DESIGN OF A LOW LATENCY ROUTER FOR ON-CHIP NETWORKS

Sumant Sathe, Daniel Wiklund, Dake Liu,
Dept. of Electrical Engineering, Linköping University, Sweden

DOI: 10.36724/2664-066X-2021-7-4-16-20

ABSTRACT

A problem with long on-chip wires is the resulting delays, and repeater insertion becomes essential to mitigate this problem. Point-to-point wiring leads to an increase in the area and repeater insertion leads to an increase in power consumption. The complexity of System-on-Chip (SoC) designs con- tinues to increase, and traditional bus-based interconnects will not be sufficient to manage the communication requirements of future billion transistor chips. The properties of our OCN are now reviewed. The OCN provides reliable communication. This is achieved by ensuring that data dropping is not allowed. On-Chip Network's (OCN's) provide a scalable alternative to existing on-chip interconnects. The key element of the OCN is the router. We present a prototype design of a 5-input, 5-output, scalable router. The router is constructed from a collection of parameterizable and reusable hardware blocks and is a basic building block of the OCN. The router supports wormhole routing, and is characterized by an area of 0.3 mm2 in 0.18 micron CMOS technology.

KEYWORDS: *System-on-Chip, On-Chip Network's, On-Chip Network's.*

The article is reworked from unpublished 2nd IEEE International Conference on Circuits and Systems for Communications (ICCSC) materials.

INTRODUCTION

With the increasing complexity of SoC's, there is an increasing demand for high performance communication between the IP's (computation blocks) [1-4, 6-8]. The current solution for implementing SoC's with multiple processors, memories, etc. is to use a time-division multiplexed (TDM) bus, e.g. AMBA from ARM. All the communication is done over a shared transmission medium, so only a single master can drive the medium at any given instant. An arbitration mechanism has to be used to allow only one master to drive the shared medium at a time. This does not enable an extremely high-performance or a highly scalable solution.

A problem with long on-chip wires is the resulting delays, and repeater insertion becomes essential to mitigate this problem. Point-to-point wiring leads to an increase in the area and repeater insertion leads to an increase in power consumption.

OCN's provide an elegant solution to these problems because [1-4] they (a) structure and manage global wires in new deep submicron technologies, (b) share wires, lowering their number and increasing their utilization, (c) can be energy-efficient and reliable, (d) are scalable when compared to traditional buses, and (e) allow multiple simultaneous transactions.

The properties of our OCN are now reviewed. The OCN provides reliable communication. This is achieved by ensuring that data dropping is not allowed. The OCN is free FROM deadlock. This can be ensured if no resource is allowed to be locked indefinitely while waiting for another resource. Deadlock can be avoided without dropping data by introducing constraints either in the topology or routing. The OCN conforms to data ordering to eliminate the need for reordering modules. This minimizes the buffer space. The overall latency and hardware overhead of the OCN is minimal. This is achieved by the use of wormhole routing.

In this paper, we describe our OCN architecture and explain how transactions are handled. We present the design of a 5-input, 5-output, scalable router suitable for our OCN.

OCN ARCHITECTURE

The OCN has been implemented as a 2-dimensional mesh as seen in Figure 1 [5]. It uses 5-port routers that allocate four ports to connect to other routers and one port to connect to the local IP block interface (port).

The local IP port is connected through a wrapper to the router. The wrappers handle IP and network port differences such as transaction handling, port width, endianness, etc. The wrappers act as an interface between the IP block clock domain and the interconnect (OCN) clock domain.

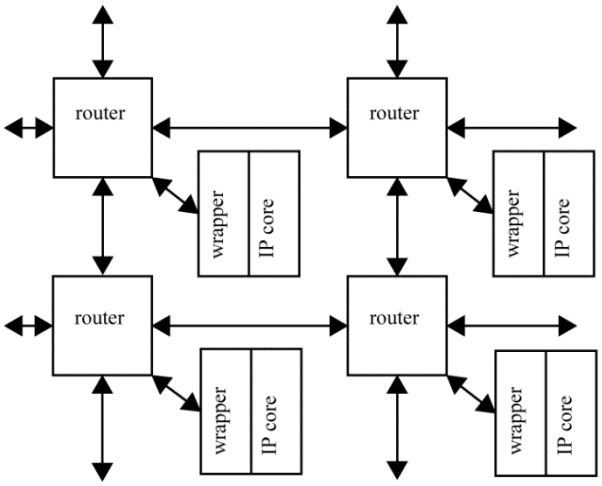


Fig. 1. A 2 x 2 OCN with routers, wrappers and IP blocks

NETWORK SWITCHING OPTIONS

A network using circuit switching has low complexity routers because their main function is to connect an incoming link to an outgoing link. Deadlocks are easily avoided since the circuit setup can either succeed or fail, but it cannot stall somewhere in this process.

Packet switching leads to more complex routers as the router has to buffer every packet before routing it, or the router has to use several virtual channels, or the router has to restrict the possible paths, to avoid deadlocks. Packet switching also suffers from latency problems, wherein the packet delay through the OCN can be several hundred or even several thousand cycles, depending on the routing algorithm and router implementation. Even for a wormhole routing network there is a possibility that the packets will be stalled for a long time due to other traffic.

This is because there is always a statistical distribution of packet delays in a packet switched network, which could also lead to the out-of-order arrival of packets at the destination. Circuit switching has an advantage over packet switching since the data transfer latency is only dependent on the distance, and there is no dependency on other factors, viz. other traffic in the network. The only dependency on the traffic situation in a circuit switched network is when setting up a route. All data is also guaranteed to arrive in the same order it is sent.

NETWORK TRANSACTION HANDLING

Route Setup Flow

The network transactions consist of four to six phases depending on whether the first routing try is successful or not.

A successful transaction has four phases. (I) First a request is sent from the source to the network. As this request finds its way through the network, the route is temporarily locked, and cannot be used for any other transaction. (II) The second phase starts when the request reaches its destination. An "acknowledge" is sent back along the route to the source. (III) When the "acknowledge" has returned to the source, the third phase starts. The actual transfer of the data payload is done during this phase. (IV) Finally after the data has been transferred, a 'cancel' request is sent that releases all resources as it follows the route. Figure 2 shows a successful transaction.

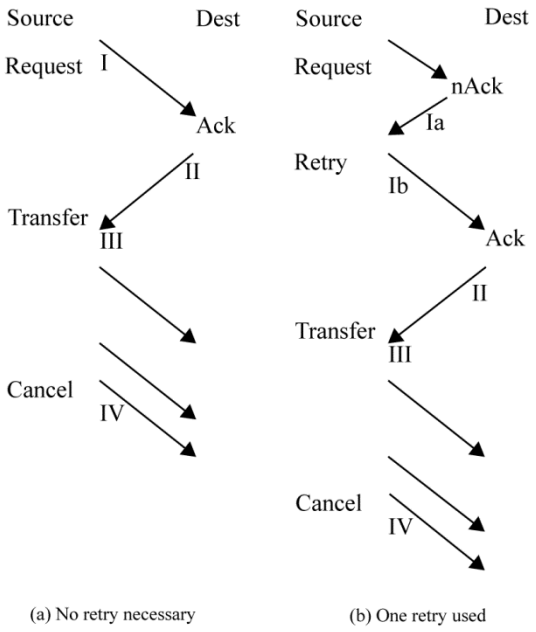


Fig. 2. Two successful circuit setups

PCC: Packet Connected Circuit

We refer to the novel hybrid circuit switching with packet-based setup introduced in section *Route Setup Flow* as "packet connected circuit" or PCC for short. PCC has the following nice properties:

(a) PCC is deadlock free since no resources are locked while waiting (indefinitely) for other resources.

(b) The routing hardware in the router becomes very simple since no special cases like stalls or virtual channels must be considered.

(c) There is no inherent limit on route selection algorithms in the PCC scheme.

(d) No central controller is required, and hence it is scalable.

Routing

A minimum path-length algorithm has been selected. Since the network does not change after the chip has been manufactured, this knowledge is static and decided at the time of high level synthesis of the network. The routing decisions are simply based on the destination address and the known direction. If there is more than one direction that leads to the destination, one is selected. If the primary selection is occupied, the second choice will be used. If there are no free outputs that lead to the destination, routing is not possible and the router will send a "negative acknowledgement" to the source.

ROUTER DESIGN

The interface to the router is shown in Figure3. In total 19 wires are used in each direction. 16 wires carry forward going data and routing request packets, 1 wire is used for forward control, and 2 wires are used for reverse control. The routing request packet is 16 bits wide, and comprises of the 8 bit destination address and 8 auxiliary bits.

The forward control handles the framing of transmissions and clock information to allow for easy retiming of the transfers when using mesochronous clocking (i.e. same frequency, but unknown phase). The reverse control carries the positive and negative acknowledgements. The diagonal line in the figure represents a similar interface to the local IP wrapper.

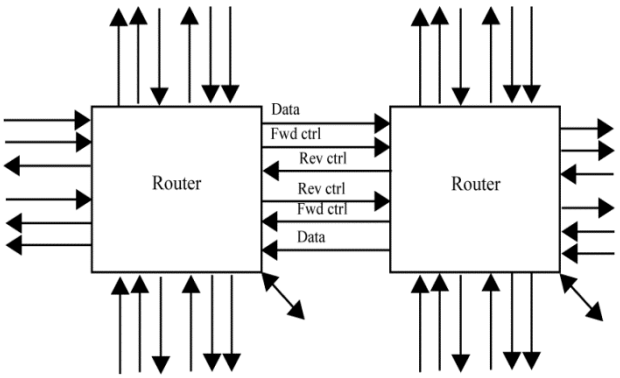


Fig. 3. Interface to the router

Clocking Methodology

Considering the high clock rate and the distributed nature of an on-chip network, the wire delays between the components become a serious problem in a traditional synchronous design methodology. In order to allow for wire delay and skew, we propose the use of mesochronous clocking with signal retiming in the OCN [9].

Router Block Diagram

Fig. 4 is the block diagram of the router. There is one input fsm, one fifo, one address decode module, and one output fsm for each port of the router. This implementation allows for easy scalability.

The number of IP cores connected to the router can be increased by instantiating the additional number of input fsm's, fifo's, address decode modules, and output fsm's. The priority encoder and the arbiter can be easily scaled using parameterizable Verilog modules.

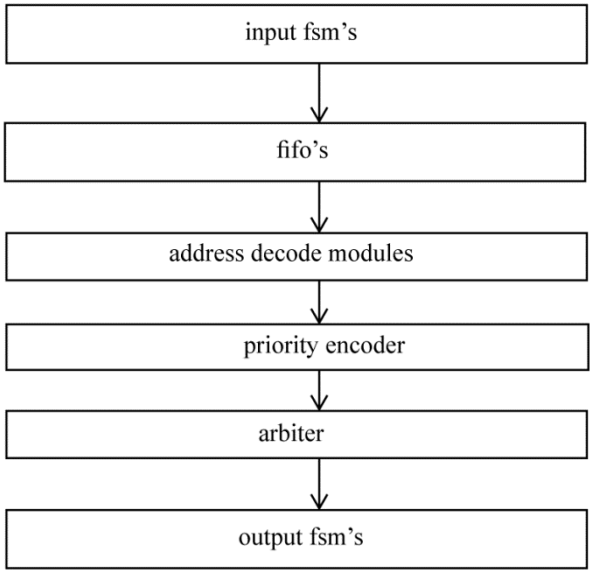


Fig. 4. 5-input 5-output router

Address Decode Module

In Figure 5, the destination address field, which is a part of the routing packet, is 8 bits wide. The address decode unit compares the higher 4 bits and the lower 4 bits of the router address respectively to determine the routing direction for the input routing packet.

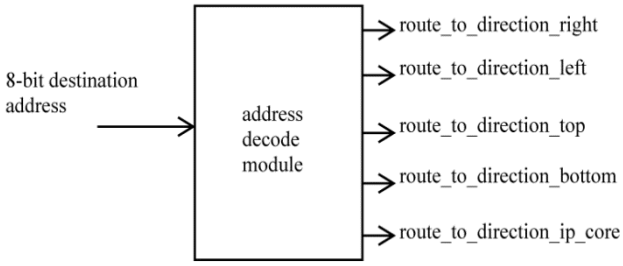


Fig. 5. Address Decode module

Priority Encoder

The 5 ports of the router have been assigned fixed priorities. The priority encoder performs the input-side arbitration for the case when multiple routing packets arrive at the different inputs of the router simultaneously. Although this fixed-priority scheme is not fair to all inputs, and contributes to congestion in the network, this scheme has been adopted to save area.

Arbiter

The arbiter does the output-side arbitration. An output port being used by an input port, is unavailable for use by the remaining input ports. The arbiter routes the routing packet to the appropriate output port, if the output port is available, i.e. not locked by another input. If no output port is available, the arbiter indicates its inability to establish a route by sending a 'negative acknowledgement'.

LATENCY ANALYSIS

There are two primary types of latency in the network. One is related to the route setup time and the other is related to the payload transfer. Both latencies are linearly dependent on the distance between the source and the destination. The route setup latency consists of first the request handling latency, which is 4 network clock cycles per router for request buffering and route selection. The second part of the route setup latency is the acknowledgement latency, which is 1 network clock cycle per router. Since the network is circuit switched, the data transfer latency is just 1 network clock cycle per router to allow for retiming.

During the 4 clock cycles that it takes for a router to forward the routing packet to the next router, the incoming data is stored in the FIFO in the router. In the previous implementation of the router [10], this was not possible. In the previous version of the router, the data transfer from the source could not begin until the connexion was setup.

As soon as a connection is set up, this data (stored in the FIFO) is then streamed out to the destination. If a connexion cannot be set up, the data in the FIFO is cleared, and a negative acknowledgement is sent to the source. Since it takes 4 clock cycles for the router to forward the routing packet, the FIFO has to store 4 data packets in the best case. In the worst case, if all ports receive routing packets simultaneously, the FIFO in the lowest priority input has to store 8 packets.

The overall worst case FIFO size is thus equal to the worst case FIFO size per router multiplied by the worst case number of hops between the longest path through the OCN. On account of the high network speed desired, the FIFO's have been implemented using flip-flop's, and not dual-port RAM's. The FIFO area (16*5 slots) is a significant percentage of the router area, but this trade-off was made with the goal of minimizing the latency through the OCN.

Due to the high internal clock rate in the network, the latency will appear lower from the IP block perspective. With a targeted network clock frequency of 1.2 GHz and a targeted route setup latency of 6 clock cycles, and a typical IP block clock frequency of 300 MHz, the apparent route setup latency will be only 1.75 cycles. The data transfer latency will appear as 0.25 cycles per router.

The wrappers will incur some setup and data transfer latency of their own. A discussion about the latency due to the wrappers has been omitted, since the wrappers have yet to be implemented.

We emphasize that our OCN architecture provides guaranteed throughput and latency, after a route setup has been successful. While a successful route setup cannot be guaranteed, this problem can be solved by proper scheduling at the software level. This trade-off has enabled us to minimize the complexity of the router.

CONCLUSIONS

An area-efficient design and architecture of a router for future OCN applications has been presented. The router is ideal for applications requiring high throughput and low latency. The router has been synthesized using Cadence PKS in 0.18 micron CMOS technology, and has an area of 0.3 mm^2.

FUTURE WORK

Wrappers for commonly used IP cores such as ARM, etc. will be designed. A demonstrator system is planned, which will feature the complete OCN solution consisting of the routers, wrappers, and IP cores.

REFERENCES

[1] P. Guerrier, A. Greiner. A generic architecture for on-chip packet interconnections, *in DATE* 2000.

[2] W. Dally, B. Towles. Route packets, not wires: On-chip interconnection networks, *in DAC* 2001.

[3] L. Benini, G. De Micheli. Networks on Chips: A New SoC Paradigm, *IEEE Computer*, 35(1):70-78, 2002.

[4] K. Goossens, et. al. Networks on silicon: Combining best-effort and guaranteed services, *in DATE* 2002.

[5] D. Wiklund, D. Liu. SoCBUS: Switched Network on Chip for Hard Real-Time Embedded Systems, *in IPDPS* 2003.

[6] E. Rijpkema, et. al. Trade Offs in the Design of a Router with both Guaranteed and Best-Effort Services for Networks on Chips, *in DATE* 2003.

[7] I. Saastamoinen, et. al. Interconnect IP Node for Future System-on-Chip Designs, *IEEE Internal Workshop on Electronic Design, Test and Applications*, 2002.

[8] P. Pande, et. al. Design of a Switch for Networks on Chip Applications, *in ISCAS* 2003.

[9] F. Mu, C. Svensson. Self-Tested Self-Synchronization Circuit for Mesochronous Clocking, *IEEE TCAS-II: Analog and Digital Signal Processing,* vol. 48, no. 2, Feb. 2001.

[10] S. Sathe, et. al. Design of a Switching Node (Router) for On-Chip Networks, *2003 5th International Conference on ASIC (ASICON2003).*

A SERIES VOLTAGE REGULATOR INTEGRATED IN CMOS TECHNOLOGY

Helene Tap-Beteille, Marc Lescure,
Laboratoire d'electronique INP-ENSEEIHT, Toulouse, France,
tap@len7.enseeiht.fr

DOI: 10.36724/2664-066X-2021-7-4-21-25

ABSTRACT

Since the regulation of voltage supply is one of the most critical requirement of the electronic system design, the monolithic voltage regulator has become one of the most important building block of both analog and digital systems. This importance has been recently increased with the emergent low voltage technologies, encouraging industrials and researchers to work on new regulator structures. The voltage regulator presented here, has been first calculated and simulated through PSpice. The paper deals with the conception of a series voltage regulator integrated in 0.6Bμm CMOS technology. This type of regulators has become one of the most important building block of both analog and digital systems. It is constituted by a bandgap voltage reference and an error amplifier associated with a ballast element. The circuit obtained is unconditionally stable with good performances: a power-supply-rejection-ratio < 2%, an output resistance < 0.3 Ω and a temperature coefficient < 10 ppm. It is integrated on an ASIC for on board applications where low volume and low power consumption are key elements.

KEYWORDS: *Analog design, Bandgap circuit, CMOS technology, Integrated circuits, Voltage regulator.*

The article is reworked from unpublished 2nd IEEE International Conference on Circuits and Systems for Communications (ICCSC) materials.

INTRODUCTION

Since the regulation of voltage supply is one of the most critical requirement of the electronic system design, the monolithic voltage regulator has become one of the most important building block of both analog and digital systems. This importance has been recently increased with the emergent low voltage technologies, encouraging industrials and researchers to work on new regulator structures. The voltage regulator presented here, has been first calculated and simulated through PSpice. Then, the layout of the circuit has been drawn through Cadence Virtuoso and sent to foundry (0.6 µm CMOS technology).

We will first present the voltage regulators (section II). Then, we will present the bandgap voltage reference (section III) and the CMOS AOP circuit (section IV). To finish with, in section V, we will give the performances obtained.

THE VOLTAGE REGULATOR

The function of a voltage regulator is to provide a specified and constant output voltage from a fluctuating input voltage. The output voltage of this regulator (independent from the changing load conditions) would then be used to supply other circuits. Today, there are two different types of IC voltage regulators: the series regulators and the switching regulators [1]. The series regulator are connected in series with the load and the unregulated input voltage. They consist of three main elements as shown on figure 1. The voltage reference source generates a reference voltage Vr, independent of the unregulated input voltage variations or the temperature changes. The error amplifier compares Vr with a fraction of the output voltage Vo and generates a corrective error signal to regulate the voltage drop across the ballast element. Vo is derived from the actual output voltage by means of the sampling resistances R1 and R2.

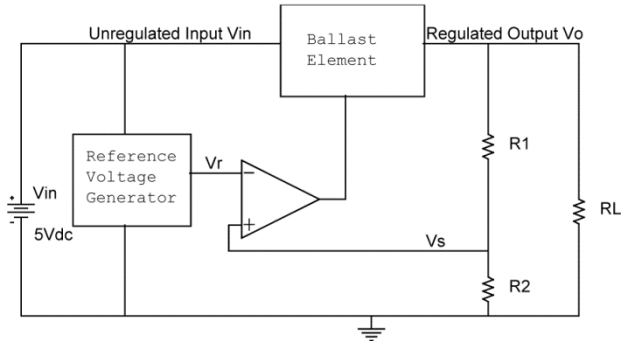


Fig. 1. Block diagram of series regulator

If the error amplifier gain Ad is sufficiently high, the voltage drop across the ballast element will vary with the fluctuations of the unregulated input voltage to maintain the output voltage V_0 constant and equal to:

$$V_o = \frac{A_d}{1+\alpha A_d} V_r \quad (1)$$

where α is the feedback factor, determined by the sampling resistances, that is:

$$\alpha = \frac{V_r}{V_0} = \frac{R_2}{R_1 + R_2} \quad (2)$$

Thus, the circuit produces an output voltage which is, to first order, independent of the input voltage and proportional to V_r. For most applications [2], the open-loop voltage gain A_d is on the order of 60-70 dB, which is usually obtained from two gain stages (a source-coupled pair and a common-source stage). That's why it has to be frequency-compensated to ensure stability under all operating conditions.

In CMOS technology, both MOS and/or bipolar transistor can be applied to generate the basic signals for voltage reference [4-6]. In case of MOS transistors, the basic signals are derived from the threshold voltage and the mobility. In the bipolar transistors, the base-emitter voltage and the saturation current are used for the extraction of the basic signals. It appears that the base-emitter voltage and saturation current of the bipolar transistors show better temperature characteristics than the threshold voltage and mobility of the MOS transistors. The low accuracy of the CMOS reference circuits is due to mismatching of components, drift, temperature effects, 1/f noise and mechanical stress. Thus, most of the voltage reference circuits apply bipolar transistors as the basic components [7]. That's why, we have integrated a bandgap reference, presented in next section.

THE BANDGAP REFERENCE

Figure 2 shows the implementation of the band-gap reference concept. The operational amplifier is in a feedback loop so the input differential voltage has to be a zero [3].

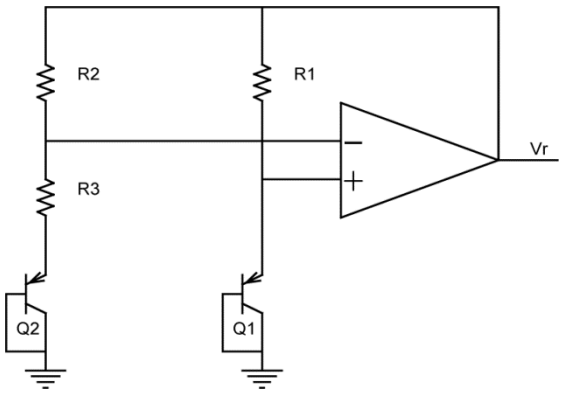


Fig. 2. Band-gap voltage reference circuit

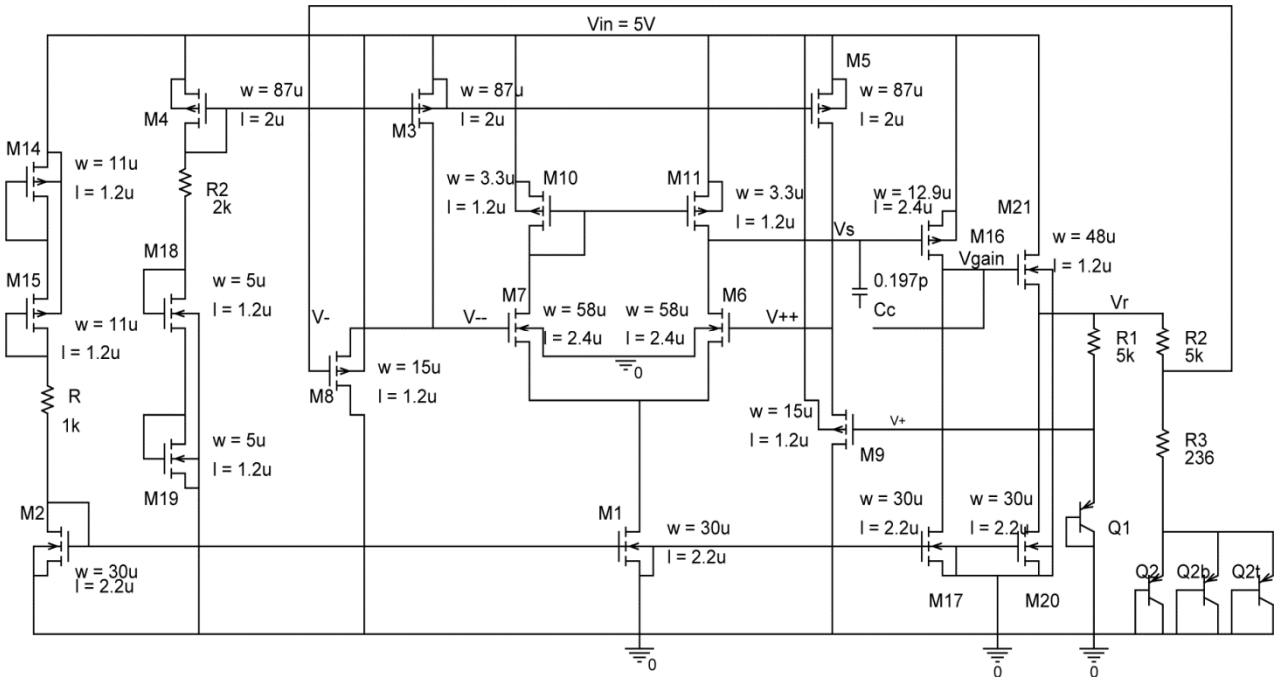


Fig. 3. Bandgap circuit

That's why I_1 and I_2 are forced to be equal to the ratio:

$$\frac{I_1}{I_2} = \frac{R_2}{R_1} \tag{3}$$

If Q_1 and Q_2 are well-matched and neglecting the base currents, the difference between their base-emitter voltages are:

$$V_{be1} - V_{be2} = \Delta V_{be} = U_T Ln\left(\frac{I_1 I_{s2}}{I_2 I_{s1}}\right) = U_T Ln\left(\frac{R_2 A_{E2}}{R_1 A_{E1}}\right) \tag{4}$$

With I_S = saturation current of the transistor, A_E = emitter surface of the transistor This differential voltage ΔV_{be} appears directly across resistance R_3, that is:

$$\Delta V_{be} = I_2 R_3 = \frac{I_1 R_1 R_3}{R_2} \tag{5}$$

That's why the reference voltage V_r is:

$$V_r = V_{be1} + R_1 I_1 = V_{be1} + U_T \frac{R_2}{R_3} Ln\frac{R_2 A_{E2}}{R_1 A_{E1}} = V_{be1} + K U_T \tag{6}$$

knowing that $\frac{\Delta V_{be}}{\Delta T} = -2\ mV/°C$ and $\frac{\Delta U_T}{\Delta T} = +0.086\ mV/°C$, $K = -\frac{2}{0.086} = 23.3$

So V_r comes about +1.25V which is very nearly equal to the band-gap voltage of silicon.

The bandgap circuit is shown on figure 3. The gain of the differential stage is given by:

$$A_{diff} = \frac{V_s}{V^+ - V^-} = -\frac{g_{m6}}{g_{ds11} + g_{ds6}} = -158 = 44dB \tag{7}$$

with $g_m = transconductance = \sqrt{\frac{2 K_P W I_D}{L}}$ (8)

and $g_{ds} = \lambda I_D$ (9)

The gain of the second stage is given by:

$$A_{M16} = -\frac{g_{m16}}{g_{ds16} + g_{ds17}} = -47.9 = 34dB \tag{10}$$

So, the gain of the whole circuit is A_0 = 78dB. This two-stages amplifier is only marginally stable when used in a feedback circuit. So, we have to introduce a pole-splitting capacitance C_c to make the circuit always stable.

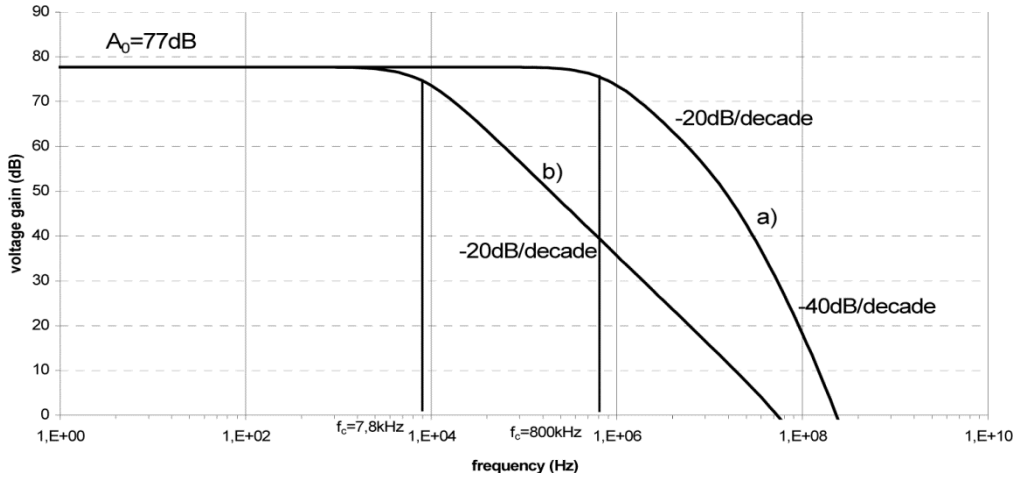


Fig. 4. a. Gain response without C_c; b. Gain response with C_c

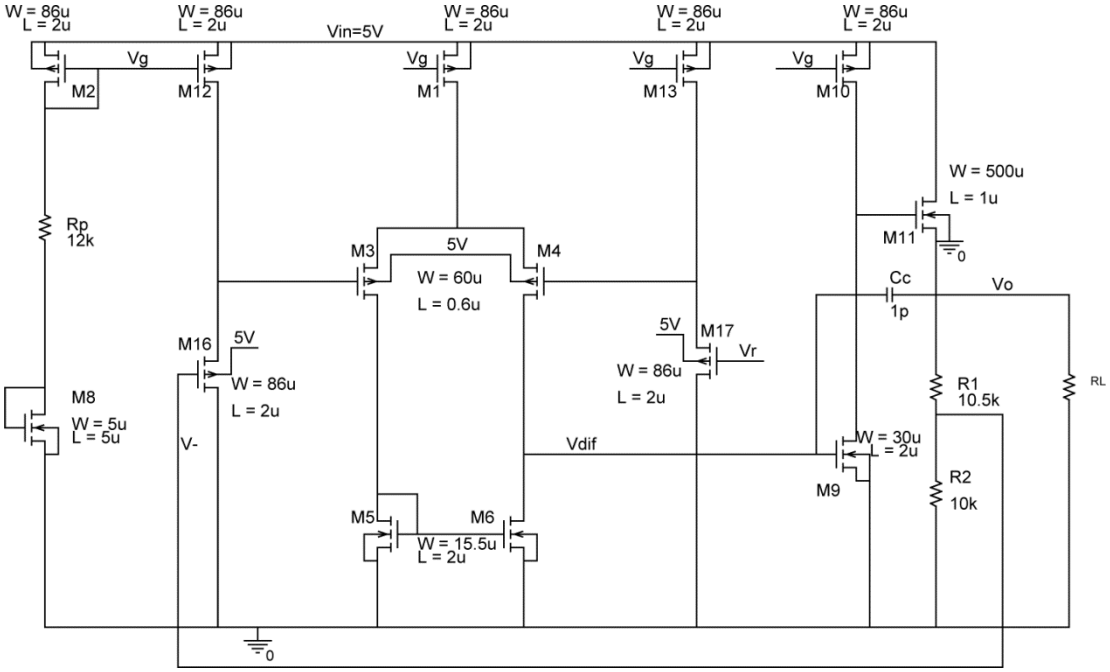


Fig. 5. Error amplifier and ballast element circuit

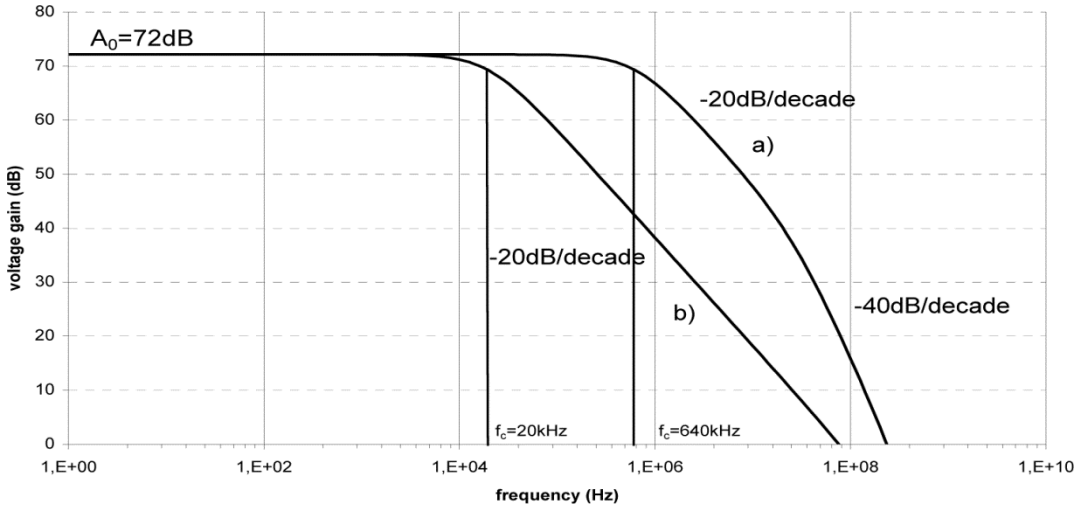


Fig. 6. a. Gain response without C_c; b. Gain response with C_c

The gain frequency response obtained with or without the compensation capacitance is shown on figure 4.

THE AOP AND THE BALLAST ELEMENT

The error amplifier and ballast element circuit is shown of figure 5.

The gain of the differential stage is given by:

$$A_{diff} = \frac{V_{diff}}{V^+ - V^-} = -\frac{g_{m4}}{g_{ds6} + g_{ds4}} = -35 = 31dB \quad (11)$$

The gain of the second stage is given by:

$$A_{M9} = -\frac{g_{m9}}{g_{ds9} + g_{ds10}} = -112 = 41\,dB \quad (12)$$

Neglecting the voltage gain of the source follower stages, the gain of the whole circuit is $A_0 = 72$ dB.

A pole-splitting capacitance C_c is used to make the circuit always stable.

The gain frequency response obtained with or without the compensation capacitance is shown on figure 6.

PERFORMANCES AND CONCLUSIONS

The temperature coefficient given by PSpice simulations is:

$$k_T = \frac{1}{V_0}\frac{\Delta V_0}{\Delta T} = 8\ ppm/^\circ C \quad (13)$$

The total power dissipation is 9mW. The global consumption without an external resistance load RL is 1.9mA. The power-supply-rejection-ratio is 1.4% and the output resistance is 0.025Ω.

So this paper shows that it is interesting to integrate a voltage regulator in a ASIC in order to regulate the input voltage of both analog and digital circuits, because the performances obtained are quite good for a reduced place. The circuit has been realised in CMOS 0.6µm technology.

This ASIC will be tested to compare the measured performances with the expected performances. If the measured performances are satisfying, the ASIC will be used in on board applications where low volume and low power consumption are key elements.

REFERENCES

[1] A.B. Grebene. Bipolar and MOS analog integrated circuit design, John Wiley and Sons, 1983.

[2] R.L. Geiger, P.E. Allen, N.R. Strader. VLSI techniques for analog and digital circuits, Mc Graw Hill, 1990.

[3] R. Gregorian. Introduction to CMOS Op-Amps and comparators, John Wiley and Sons, 1999.

[4] R.A. Blauschild, P.A. Tucci, R.S. Muller, R.G. Meyer. A new NMOS temperature-stable voltage reference*, IEEE J. Solid –State Circuits*, vol.SC-13, pp. 767-774, Dec.1978.

[5] H.J. Song, C.K. Kim. A temperature-stabilized SOI voltage reference based on threshold voltage difference between enhancement and depletion NMOSFET's, *IEEE J. Solid-State Circuits*, vol. SC-28, pp. 671-677, June 1993.

[6] Y.P. Tsivids, R.W. Ulmer. "CMOS voltage reference, *IEEE, J. Solid-State Circuits*, vol. SC13, pp. 774-778, Dec 1978.

[7] G.C.M. Meijer, G. Wang, F. Fruett. Temperature sensors and voltage references implemented in CMOS technology, *IEEE Sensors Journal,* Vol.1, No.3, pp. 225-234, October 2001.

A MIXED TIME-FREQUENCY DOMAIN APPROACH FOR THE QUALITATIVE ANALYSIS OF AN HYSTERETIC OSCILLATOR

Chris Taillefer,M. Bonnin, M. Gilli and P. P. Civalleri,
Department of Electronics, Politecnico di Torino, Torino, Italy

DOI: 10.36724/2664-066X-2021-7-4-26-29

ABSTRACT

Frequency domain techniques, like harmonic balance and describing function, are classical methods for studying and designing electronic oscillators and nonlinear microwave circuits. In most applications spectral techniques have been used for determining the steady-state behavior of nonlinear circuits that exhibit a single periodic attractor. On the other hand, the global dynamics of nonlinear networks and systems is usually investigated through time-domain techniques, that require to introduce rather complex and sophisticated concepts. Recently some HB based techniques have been proposed for investigating bifurcation processes in nonlinear circuits that present several attractors (the authors have considered systems that admits of a Lur'e representation). Their approach presents the advantages of providing a simple and qualitative description of the system dynamics, that can be effectively exploited for design purposes. In this manuscript we will examine a third order hysteretic oscillator, that cannot be described in the classical Lur'e form and we will show that its dynamics can be investigated through the joint application of the describing function technique and of a suitable time-domain method for estimating Floquet's multipliers.

KEYWORDS: *Hysteretic Oscillator, Frequency domain techniques, Floquet's multipliers.*

The article is reworked from unpublished 2nd IEEE International Conference on Circuits and Systems for Communications (ICCSC) materials.

Frequency domain techniques, like harmonic balance and describing function, are classical methods for studying and designing electronic oscillators and nonlinear microwave circuits [1], [2], [3]. In most applications spectral techniques have been used for determining the steady-state behavior of nonlinear circuits that exhibit a single periodic attractor. On the other hand, the global dynamics of nonlinear networks and systems is usually investigated through time-domain techniques, that require to introduce rather complex and sophisticated concepts [4].

Recently some HB based techniques have been proposed for investigating bifurcation processes in nonlinear circuits that present several attractors [5]-[11]. In [12], [13] the authors have considered systems that admits of a Lur'e representation. They have shown that the describing function technique (i.e., HB with a single harmonic) is able to predict the occurrence of chaotic behavior and several bifurcation phenomena. Their approach presents the advantages of providing a simple and qualitative description of the system dynamics, that can be effectively exploited for design purposes.

In this manuscript we will examine a third order hysteretic oscillator, that cannot be described in the classical Lur'e form and we will show that its dynamics can be investigated through the joint application of the describing function technique and of a suitable time-domain method for estimating Floquet's multipliers.

We consider the third order hysteretic oscillator shown in Figure 1 of [14] and described by the following set of normalized state equations:

$$\dot{x} = -(x+y) \quad (1)$$

$$\dot{y} = \alpha(x+y) - y - \beta z \quad (2)$$

$$\dot{z} = \gamma f(x-z) - \delta \sinh(z) \quad (3)$$

where the dot denotes the time-derivative and α, β δ and γ depend on circuit parameters. According to [14] f(·) is defined as the R → R function, obtained by finding for each w the unique solution of the following transcendental equation.

$$f = \delta \sinh[w - f] \quad (4)$$

As a preliminary step we show that the above set of equations can be reduced to a scalar Lur'e like system. The first two equations (1) and (2) allow one to derive a linear relationship between $x(t)$ and $z(t)$:

$$z(t) = L(D)x(t) \quad (5)$$

where $D = d/dt$ denotes the first order differential operator and

$$L(D) = \frac{D^2 + (2-\alpha)D + 1}{\beta} \quad (6)$$

By substituting expression (5) in equation (3), we obtain the following Lur'e like model, in term of the scalar variable x(t):

$$DL(D)x(t) = \gamma f[x(t) - L(D)x(t)] - \delta \sinh[L(D)x(t)] \quad (7)$$

As pointed out above, we will show that the most significant dynamic properties of the hysteretic oscillator under study, can be qualitatively revealed through the joint application of the describing function technique and of a suitable time-domain method for computing limit cycle Floquet's multipliers.

We assume that any periodic signal of period $T = 2\pi/\omega$ can be approximated by a bias term and a single harmonic. We have

$$x(t) \approx \tilde{x}(t) = A + B\sin(\omega t) \quad (8)$$

where A denotes the bias term and B is the amplitude of the first order harmonic.

The following expressions can be readily computed:

$$L(D)\tilde{x}(t) = L(0)\,A + B\,\mathbf{Re}[L(j\omega)]\,\sin(\omega t) + B\,\mathbf{Im}[L(j\omega)]\,\cos(\omega t) \quad (9)$$

$$DL(D)\tilde{x}(t) = B\,\omega\,\mathbf{Re}[L(j\omega)]\,\cos(\omega t) - B\,\omega\,\mathbf{Im}[L(j\omega)]\,\sin(\omega t) \quad (10)$$

$$L(0) = \frac{1}{\beta},\ \mathbf{Re}[L(j\omega)] = \frac{1-\omega^2}{\beta},\ \mathbf{Im}[L(j\omega)] = \frac{(2-\alpha)\omega}{\beta} \quad (11)$$

By exploiting (11), it is also derived that expression $f[\tilde{x}(t) - L(D)\tilde{x}(t)]$ can be considered as a function (hereafter named $\tilde{F}(\cdot)$) of the describing function parameters A, B and ω and of time:

$$f[\tilde{x}(t) - L(D)\tilde{x}(t)] = f\left[\left(1 - \frac{1}{\beta}\right) A + B\left(1 - \frac{1-\omega^2}{\beta}\right) \times \right.$$
$$\left. \times \sin(\omega t) - B\,\frac{(2-\alpha)\omega}{\beta}\cos(\omega t)\right] = \tilde{F}(A, B, \omega, t) \quad (12)$$

According to the describing function technique, the following first harmonic approximation of function $f[\tilde{x}(t) - L(D)\tilde{x}(t)]$ holds:

$$f[\tilde{x}(t) - L(D)\tilde{x}(t) \approx F^A(A,B,\omega) + \\ + F^B(A,B,\omega)\sin(\omega t) + F^C(A,B,\omega)\cos(\omega t) \quad (13)$$

where

$$F^A(A,B,\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \tilde{F}(A,B,\omega,t)\,d\omega t$$

$$F^B(A,B,\omega) = \frac{1}{\pi}\int_{-\pi}^{\pi} \tilde{F}(A,B,\omega,t)\sin(\omega t)\,d\omega t$$

$$F^C(A,B,\omega) = \frac{1}{\pi}\int_{-\pi}^{\pi} \tilde{F}(A,B,\omega,t)\cos(\omega t)\,d\omega t \quad (14)$$

Since the explicit expression of function $f(\cdot)$ is not known (because the latter is found as the solution of the tran-scendental equation (4)), the integrals above do not admit analytical expressions.

However they can be numerically computed to any desired accuracy, for any set of parameters *A, B* and ω.

By following a similar procedure, the first harmonic approximation of sin*h*[*L*(*D*)*x*(*t*)] (the other nonlinear function appearing in the scalar model (7)) can be found. We have:

$$\sinh[L(D)\tilde{x}(t)] \approx G^A(A,B,\omega) + \\ + G^B(A,B,\omega)\sin(\omega t) + G^C(A,B,\omega)\cos(\omega t) \tag{15}$$

where

$$G^A(A,B,\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\sinh[L(D)\tilde{x}(t)]d\omega t$$

$$G^B(A,B,\omega) = \frac{1}{\pi}\int_{-\pi}^{\pi}\sinh[L(D)\tilde{x}(t)]\sin(\omega t)d\omega t$$

$$G^C(A,B,\omega) = \frac{1}{\pi}\int_{-\pi}^{\pi}\sinh[L(D)\tilde{x}(t)]\cos(\omega t)d\omega t \tag{16}$$

The integral expressions (16) can be analytically computed. In fact, after some algebraic manipulations, the following fundamental integrals can be derived (for any set of real coefficients *R P, Q*).

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\sinh[R+P\sin(\theta)+Q\cos(\theta)]\,d\theta = \\ = I_0(\sqrt{P^2+Q^2})\,\sinh(R) \tag{17}$$

$$\frac{1}{\pi}\int_{-\pi}^{\pi}\sinh[R+P\sin(\theta)+Q\cos(\theta)]\,\sin(\theta)\,d\theta = \\ = 2\frac{P}{\sqrt{P^2+Q^2}}\,I_1(\sqrt{P^2+Q^2})\,\cosh(R) \tag{18}$$

$$\frac{1}{\pi}\int_{-\pi}^{\pi}\sinh[R+P\sin(\theta)+Q\cos(\theta)]\,\cos(\theta)\,d\theta = \\ = 2\frac{Q}{\sqrt{P^2+Q^2}}\,I_1(\sqrt{P^2+Q^2})\,\cosh(R) \tag{19}$$

where I_0 and I_1 denotes the modified Bessel function of the first kind of order zero and one respectively.

Since, according to (9) and (11) the following expression stands for $L(D)\tilde{x}(t)$

$$L(D)\tilde{x}(t) = \frac{1}{\beta}A + \frac{1-\omega^2}{\beta}B\,\sin(\omega t) + \\ + \frac{(2-\alpha)\omega}{\beta}B\,\cos(\omega t) \tag{20}$$

the analytical close form for the integrals (16) can be readily derived by replacing *R, P* and *Q* in (17) - (19) with the expressions shown below:

$$R = \frac{1}{\beta}A, \quad P = \frac{1-\omega^2}{\beta}B, \; Q = \frac{(2-\alpha)\omega}{\beta}B \tag{21}$$

By substituting in (7) the first order harmonic approximations of the nonlinear functions $f[\tilde{x}(t)-L(D)\tilde{x}(t)]$ and $\sinh[L(D)\tilde{x}(t)]$ derived in (13) and (15), and of $DL(D)\tilde{x}(t)$ derived in (10), we obtain a non-di˙erential system of three equations with three unknowns *A, B* and ω (hereafter named describing function system):

$$\gamma F^A(A,B,\omega) - \delta G^A(A,B,\omega) = 0$$
$$B\omega\,\mathbf{Im}[L(j\omega)] + \gamma F^B(A,B,\omega) - \delta G^B(A,B,\omega) = 0$$
$$B\omega\,\mathbf{Re}[L(j\omega)] - \gamma F^C(A,B,\omega) + \delta G^C(A,B,\omega) = 0 \tag{22}$$

where Re[*L*(*j*ω)] and Im[*L*(*j*ω)] are given by (11) and the coefficient $F^{A,B,C}$ and $G^{A,B,C}$ have the expressions reported in (14) and (16) respectively.

The describing function system can be solved in a very efficient way, by exploiting standard numerical methods. If for a given set of parameters α, β, γ and δ a solution (*A, B* and ω) exists, the latter is called predicted limit cycle. Since the describing function technique is approximate in nature, in general one cannot guarantee that there is a one to one correspondence between the describing function solutions (predicted limit cycles) and the actual limit cycles of the system. There are however two possible approaches for discussing the accuracy of the describing function prediction. The first one consists in the computation of the distortion index Δ, that for the system under study can be defined by adapting the well known definition given in [13] for classical Lur'e systems:

$$\mathbf{\Delta}(A,B,\omega) = \frac{||\overline{x}(t)-\tilde{x}(t)||_2}{||\tilde{x}(t)||_2} \tag{23}$$

where

$$\overline{x}(t) = [DL(D)]^{-1}\{\gamma f[\tilde{x}(t) - \\ - L(D)\tilde{x}(t)] - \delta\sinh[L(D)\tilde{x}(t)]\} \tag{24}$$

If the distortion index is small enough, i.e. $\Delta < 0.01$, then the describing function prediction can be considered reliable.

The second approach requires to verify a set of rather complex conditions, under which one can guarantee the existence of an actual limit cycle in a computable neighborhood of the predicted limit cycle [3].

As we have already pointed out, several describing function based methods have been developed for investigating limit cycle stability properties and their most significant bifurcation phenomena [13]. Such methods mainly rely on the Loeb's stability criterion and are based on the investigation of the homogeneous linearized system equation, obtained by perturbing the original solution in a way that depends on the bifurcation under study.

In this manuscript we will show that an accurate investigation of limit cycle stability and bifurcations, can be carried out by linearizing the system equation (1)-(3) along the solution predicted by the describing function technique and then by computing the Floquet's multipli-

ers of the corresponding variational equation. By denoting with $\tilde{w}(t) = [\tilde{x}(t), \tilde{y}(t), \tilde{z}(t)]'$ the describing function solution and with $w_p(t)$ a generic perturbation of $\tilde{w}(t)$, the variational equation is readily derived:

$$\dot{w}_p(t) = A(t) w_p(t) \tag{25}$$

$$A(t) = \begin{pmatrix} -1 & -1 & 0 \\ \alpha & \alpha - 1 & -\beta \\ \gamma f'(\tilde{x} - \tilde{z}) & 0 & -\gamma f'(\tilde{x} - \tilde{z}) - \delta \cosh(\tilde{z}) \end{pmatrix} \tag{26}$$

where $\tilde{z}(t)$ can be computed as $L(D)[\tilde{x}(t)]$ by exploiting (5) and (9); $f'(\cdot)$ denotes the derivative of function $f(\cdot)$ with respect to its argument, whose expression can be obtained from (4):

$$f'(w) = \frac{\delta \cosh[w - f(w)]}{1 + \delta \cosh[w - f(w)]} \tag{27}$$

Once matrix $A(t)$ is known, the Floquet's multipliers can be effectively computed by using the time-domain numerical algorithm described in [15].

The application of the mixed time-frequency domain technique described above, allows one to identify and characterize the main dynamic features of the hysteretic oscillator under study. We will show in the final version of the manuscript that the most significant bifurcation curves (pitchfork and flip bifurcations) can be detected and that the parameter regions in which more attractors coexist can be identified with a good accuracy (see Fig. 2 of [14], where a bifurcation brute-force analysis was carried out, for a synthetic description of the oscillator dynamic behavior). The accuracy of the technique has been checked by comparing the results with those obtained by applying the spectral method described in [17], based on the approximation of the state through a large number of harmonics.

As a conclusion we remark the main characteristics of our approach, in comparison with previous works on spectral techniques [13] and on hysteretic oscillators [14].

Remark 1: The proposed techniques applies to an hysteretic oscillator, that cannot be described as a classical Lur'e system. Then it represents a sort of extension of the general results presented in [13].

Remark 2: In comparison with the spectral technique described in [13] the proposed method represents an alternative and simpler way for studying bifurcations, that in some cases gives rise to more accurate results (see for example [16], where bifurcations in a Colpitts' oscillator were investigated). Moreover in all cases that we have considered the predictions yielded by our method are never less accurate and precise than those provided by applying the technique developed in [13].

Remark 3: With respect to the brute force method presented in [14], the proposed technique allows one to predict the existence of unstable limit cycles, that play an important role for understanding bifurcation phenomena. Hence it provides a more complete knowledge of the global dynamics of the hysteretic oscillators.

ACKNOWLEDGMENT

This work was supported in part by Ministero dell'Istruzione, dell'Università e della Ricerca, under the FIRB project no. RBAU01LRKJ.

REFERENCES

[1] K. S. Kundert, A. Sangiovanni-Vincentelli. Simulation of nonlinear circuits in the frequency domain, *IEEE Transactions on Computer-Aided Design*, pp. 521-535, 1986.

[2] A. Ushida, T. Adachi, and L. O. Chua. Steady-state analysis of nonlinear circuits, based on hybrid methods, *IEEE Transactions on Circuits and Systems*: I, vol. 39, pp. 649-661, 1992.

[3] A. I. Mees. Dynamics of feedback systems, John Wiley, New York, 1981.

[4] Y. A. Kuznetov. Elements of applied bifurcation theory, New York: Springer-Verlag, 1995.

[5] C. Piccardi. Bifurcations of limit cycles in periodically forced nonlinear systems, *IEEE Transactions on Circuits and Systems:* I, vol. 41, pp. 315-320, 1994.

[6] C. Piccardi, "Bifurcation analysis via harmonic balance in periodic systems with feedback structure," International Journal of Control, vol. 62, pp. 1507–1515, 1995.

[7] C. Piccardi. Harmonic balance analysis of codimension-2 bifurcations in periodic systems, *IEEE Transactions on Circuits and Systems:* I, vol. 43, pp. 1015-1018, 1996.

[8] V. Rizzoli, A. Neri. State of the art and present trends in nonlinear microwave techniques, *IEEE Transactions on Microwave Theory and Techniques*, pp. 343-365, 1988.

[9] V. Rizzoli, A. Neri, D. Masotti. Local stability analysis of microwave oscillators based on Nyquist's theorem, *IEEE Microwave Guided Wave Letters,* vol. 7, pp. 341-343, Oct. 1998.

[10] A. Suarez, J. Morales, R. Quere. Syncronization analysis of autonomous microwave circuits using new global-stability analysis tools, *IEEE Transactions on Microwave Theory and Techniques*, vol. 46, pp. 494-504, May 1998.

[11] D. W. Berns, J. L. Moiola, and G. Chen. Predicting period-doubling bifurcations and multiple oscillations in nonlinear time-delayed feedback systems, *IEEE Transactions on Circuits and Systems:* I, vol. 45, pp. 759-763, 1998.

[12] R. Genesio, A. Tesi. A Harmonic Balance Approach for Chaos Prediction: Chua's circuit, *International Journal of Bifurcation and Chaos,* vol. 2, no. 1, pp. 61-79, 1992.

[13] M. Basso, R. Genesio, A. Tesi. A frequency method for predicting limit cycle bifurcations, *Nonlinear Dynamics*, vol. 13, pp. 339-360, 1997.

[14] F. Bizzarri and M. Storace. Coexistence of attractors in an oscillator based on hysteresis, Proceedings of the 2002 *IEEE International Symposium on Circuits and Systems (ISCAS'2002),* Scottsdale, Arizona, May 2002, paper 1713.

[15] M. Farkaj, Periodic motions, Springer Verlag, 1994, pp. 58-59.

[16] M. Gilli, G. M. Maggio, and P. Kennedy. An approximate analytical approach for predicting period doubling in the Colpitts oscillator, *IEEE International Symposium on Circuits and Systems*, Monterey (CA-USA), 1998.

[17] F. Bonani and M. Gilli. Analysis of stability and bifurcations of limit cycles in Chua's circuit through the harmonic balance approach, *IEEE Transactions on Circuits and Systems:* Part I., vol. 46, no. 8, pp. 881-890, August 1999.

LOW POWER DIGITAL CMOS VLSI CIRCUITS DESIGN WITH DIFFERENT HEURISTIC ALGORITHMS

Wladyslaw Szczesniak,
Faculty of Electronics, Telecommunications, and Informatics, Gdansk University of Technology, Gdansk, Poland
wlad@ue.eti.pg.gda.pl

Piotr Szczesniak,
Faculty of Electronics, Telecommunications, and Informatics, Gdansk University of Technology, Gdansk, Poland;
R&D Marine Technology Centre, Gdynia, Poland
piotr@ue.eti.pg.gda.pl

DOI: 10.36724/2664-066X-2021-7-4-30-34

ABSTRACT

The growing demand for portable computing devices leads to new electronic systems fulfilling the requirements for the low power dissipation in the chip. Although, reduction of supply voltage is one of the most effective techniques of decreasing the power consumption in digital CMOS VLSI circuits it results in chip throughput degradation. This paper presents three versions of the Inserting Idle Operation with Interchanging heuristic algorithm, namely simple IIOI, MAximal RELativity (MAREL) and UNIform LOad (UNILO). They are applied to the high-level synthesis of CMOS VLSI circuits with power reduction. Comparison of the obtained results for the chosen set of benchmarks show the different levels of power reduction obtained by different algorithms applied. For different benchmarks the power reduction reaches up to 15%, and up to 68% with extending latency by 50%.

KEYWORDS: *Low power design, VLSI digital circuits, high level synthesis and low power design, heuristic algorithms.*

The article is reworked from unpublished 2nd IEEE International Conference on Circuits and Systems for Communications (ICCSC) materials.

INTRODUCTION

The growing demand for portable computing devices leads to new electronic systems fulfilling the requirements for the low power dissipation in the chip. Although, reduction of supply voltage is one of the most effective techniques of decreasing the power consumption in digital CMOS VLSI circuits it results in chip throughput degradation [1], [2], [3].

The main part of the power dissipated in the CMOS circuit (single functional unit fu_i) is dynamic power represented by:

$$P_{d_i} = \frac{a_i}{2} C_{Li} V_{dd_i}^2 f_{clk} \tag{1}$$

where ai is the activity factor (i.e. the probability of a power consuming transition) of the functional unit fu_i, C_{Li} is the load capacitance, V_{ddi} is the supply voltage, and f_{clk} is the clock frequency.

By decreasing the voltage swing for a given load capacitance of the chosen functional units, we can reduce power consumption quadratically as in (1) but their toggling time (T_{ti}) will be longer and is defined as [1], [3]:

$$T_{ti} \cong \frac{C_L V_{ddi}}{K\left(V_{ddi} - V_t\right)^{\alpha}} \tag{2}$$

where K is a technology constant, V_t is a threshold voltage and α is a constant depending on device technology ($1<\alpha<2$). The toggling time (T_{ti}) defines the delay time (tdi) of the functional unit f_{ui}.

Figure 1. shows the dependency of supply voltage (V_{dd}) on normalized delay times for an inverter as a functional unit.

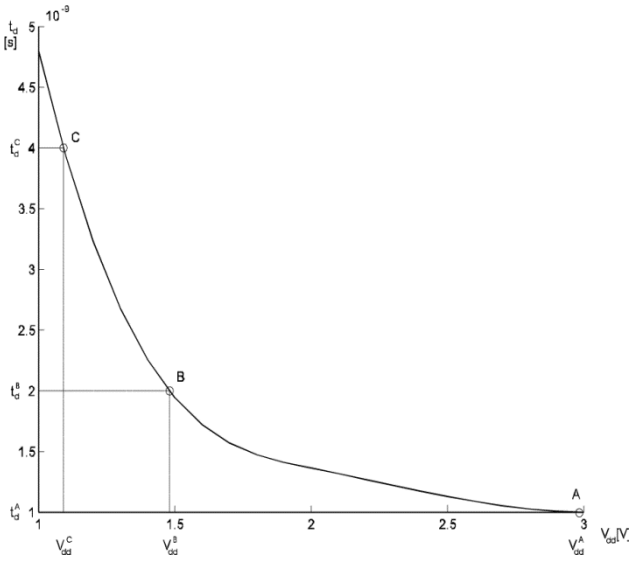


Fig. 1. Influence of the supply voltage (V_{dd}) on the nominal delay (t_d) of CMOS inverter with a load of 0.1pF for the Alcatel Mietec 0.35μm technology (V_t = 0.66V) [3]

For this example the relation between the nominal delays t_d (see Fig. 1) for points A ($V_{dd}^A = 3.0V$) and B ($V_{dd}^B = 1,48V$) is given by:

$$t_d^A = \frac{1}{2} t_d^B \tag{3}$$

In fact, to reduce the power dissipated in an electronic CMOS circuit, the supply (V_{dd}) and the threshold (V_t) voltages should be tuned according to the system activity requirements [1], so the maximal toggling time of the circuit is expressed by (2).

It leads to a lower value of the supply voltage Vdd that results in reducing the dynamic power consumption ($P_{di} \sim V^2_{ddi}$).

This paper presents the comparison of the results of three versions of Inserting Idle Operations with Interchanging (IIOI) heuristic algorithm, namely raw IIOI, MAximal RELativity (MAREL) and UNIform LOad (UNILO) applied to the high-level synthesis (HLS) of CMOS VLSI circuits with power reduction.

All of them utilize as soon as possible (ASAP) algorithm. All three presented algorithms lead to decreasing the supply voltage of the chosen functional units (from the constrained set of resources) without degradation of the chip throughput.

PROBLEM FORMULATION

Our task is to minimise the dynamic power (P_d) of CMOS circuits/systems during the HLS using three heuristic algorithms.

In HLS the behavioural specifications of digital systems are mapped to structural designs at the register transfer level (RTL). The behavioural specifications consist of algorithms which describe the behaviour of circuit outputs in terms of circuit inputs and timing constraints [1], [2], [3].

A computational task to be scheduled on the set of functional units {fu_i} (resources) can be modelled by a data flow graph DFG(V,E) where a set of vertices V ($|V| = n$) represents operations of the program and a set of edges E represents the data (variables and constants) [2], [4], [6].

For the given DFG the result of the HLS can be represented by the scheduling graph (SG) in which each vertex of DFG is assigned to the appropriate element of the set of resources {fu_i} and a control step(s) (CS) in such a way that all data dependencies are satisfied. Each column of SG corresponds to one element fu_i of the set of resources and each row to the successive control step CS.

The HLS task is constrained by a limited set of resources ({fui}) e.g. multipliers (MULT), adders (ADD), subtractors (SUB), gates, etc., and an acceptable throughput (described by the maximal number of control steps CSMAX) [3].

If it is possible to introduce slacks for chosen resources without decreasing the throughput of the whole system, then they can be supplied with a lower V_{dd}, which results in reducing power consumption. Note that in some applications, especially for systems that have different working modes, CSMAX can be extended. In the experimental part of the paper we have considered CSMAX extension by 10, 20 and 50 percent.

According to the above discussion, we can formulate the problem to be solved as follows: for given DFG assign the operations to functional units, introduce the slacks for chosen resources and construct the scheduling graph so that the dynamic power consumption of CMOS circuit/system in question is minimised without deteriorating its throughput (measured by the number of control steps CS).

The slack of the functional unit is introduced by reduction of its supply voltage *Vdd*. This results in increasing the delay of the unit and appropriate reduction of the power consumption as explained in the introduction (in this paper we consider delay by 2, 4 or 8 times). The problem of assigning the operations and distributing the functional unit slacks is solved in two stages. First, the ASAP-like base algorithm is used to construct the non-delayed scheduling graph. Then it is modified with the IIOI [3] scheduling algorithm, and its two extensions, namely MAREL and UNILO described in the next section.

SCHEDULING ALGORITHMS DESCRIPTION

Three scheduling algorithms, namely IIOI, MAREL and UNILO, used for power reduction in digital CMOS circuits during high-level synthesis are presented. All the algorithms are based on the same two-stage core, described below.

The First Stage of the Algorithms

The first stage of the algorithms is a modified ASAP scheduling process. Modifications include resource constraints and multi-cycle operations.

The pseudo-code of the ASAP stage of the algorithms is presented below, with function explanations following.

assign_p_labels ( ) (line 1)

The function assigns the p-labels to each operation.

P-label of an operation is equal to the number of the control steps needed to perform all child operations in the DFG. Operations without any child-operations (i.e. having system outputs only) have p-label equal 0;

V_r = find_ready_operations ( ) (line 5)

```
1  assign_p_labels( )
2  cstep ← 0
3  WHILE ( V ≠ ∅ )
4  {
5  Vr ←find_ready_operations(cstep)
6     WHILE ( Vr ≠ ∅ )
7     {
8    vs ←vi ∈ Vr: p_label( vs )=min
9        fus ← find_fu( vs )
10       IF ( fus ≠ NULL )
11       {
12          assign( vs , fus )
13          V ← V \ vs
14       }
15       Vr ← Vr \ vs
16    }
17    cstep ← cstep + 1
18 }
```

Fig. 2. The first (ASAP) stage of the IIOI/MAREL/UNILO algorithms

This function returns the Vr set of operations ready to be scheduled, i.e. assigned to the chosen functional unit, in the current control step. Operation is ready to be scheduled if all input values required are calculated, i.e. all parent-operations are finished.

f_{us} ← find_fu(v_s) (line 9)

This operation finds the functional unit for the chosen operation that is ready to be scheduled at the current control step. If there are no free resources at the moment, function returns NULL, and selected operation scheduling is delayed.

If there is only one functional unit of the proper type available, then it is returned.

However the differences between IIOI, MAREL and UNILO algorithms depict the situation when there is more than one functional unit of the proper type available.

The IIOI algorithm, simply chooses the first available functional unit.

The MAREL algorithm calculates the relation distance for all available functional units.

The UNILO algorithm selects the functional unit that have the smallest number of operation assigned. This ensures that load is uniformly balanced over all of the functional units.

assign(v_s , f_{us}) (line 12)

The functions assigns the operation to the chosen functional unit at the current control step. If the operation being assigned is a multi-step one then the appropriate number of following control steps is also reserved.

The Second Stage of the Algorithms

The second stage (Fig. 3.) is the same for all the algorithms. It consists of delaying of the initial SG created in the first stage. The C_{su} set includes all columns suitable for delaying, and C_{ms} indicates the most suitable column selected out of this set. The chosen C_{ms} column actually undergoes the process of delaying.

Functions used in the pseudo-code are explained below.

```
1  Csu ← SG
2  WHILE ( |Csu| > 0 )
3  {
4   FOREACH ( Ck ∈ Csu )
5  IF ( NOT fsc_fulfiled( Ck ) )
6     Csu ← Csu \ Ck
7   IF ( |Csu| == 0 )
8    GOTO WHILE_END
9   IF ( |Csu| > 1 )
10  {
11   FOREACH ( Ck ∈ Csu )
12  IF ( there_is_same_fu( Ck ))
13     {
14    Cl←column_of_the_same_fu(Ck)
15      FOREACH ( vi ∈ Ck )
16       IF (fvi(vj)>fvi( vi ) )
17        interchange( vi , vj )
18     }
19  Cms ← Ck ∈ Csu: fck( Ck )=max
20  } ELSE
21  Cms = HEAD( Csu )
22  SGbackup ← SG
23  FOREACH ( vi ∈ Cms )
24  {
25   insert_idle_operations(vi)
26  IF(NOT delay_all_successors(vi)
27   {
28    SG ← SGbackup
29    Csu ← Csu \ Cms
30    GOTO WHILE_END
31   }
32  }
33  :WHILE_END
34 }
```

Fig. 3. The second stage of the IIOI/MAREL/UNILO algorithms

fsc_fulfiled(C_k) (line 5)

This function performs the check of the free space condition (f_{sc}), defined by the formula:

$$n_i \cdot l_k \le r_o \tag{4}$$

where n_i is the number of cycles needed to perform the operation, l_k is the number of operations assigned to the C_k functional unit column, r_o is the number of free operation slots after the first occurrence of an operation in the C_k functional unit column.

The condition (4) checks if there are enough free *csteps* for the idle operations identifying the longer processing time of a fu_i. Every Ck selected for the frequency lowering has to fulfil the condition (4). Despite the fact, that cascading operations from the other C_k's are not taken into consideration while calculating f_{sc} , it is sufficient for quick pre-rejection of some C_k from the C_{su} set, before starting the time consuming delaying process.

there_is_same_fu(C_k) (line 12)

This functions simply indicates, whether there is another *fu* of exactly the same type as the one assigned to C_k, i.e. being capable of performing the same type of operation in the same time.

C_l =column_of_the_same_fu(C_k) (line 14)

This function seeks for a column containing the same operations as Ck and sets the Cl pointer to it.

fvi(v_i) (line 16)

The $fv_i(v_i)$ function calculates fv_i factor for the v_i operation with the following formula:

$$f_{vi} = \frac{i_{vi} + o_{vi} + (f_{avi} - s_{vi} \cdot n_i)}{p_i + 1} \tag{5}$$

where i_{vi} is the number of independent inputs of the operation v_i, o_{vi} is the number of system outputs of the operation v_i , $f_{avi} = cs_M - (cs_e + n_i)$, cs_M is the maximal number of csteps admissible, and cs_e is the number of *cstep*, which v_i is assigned to, s_{vi} is the number of operations of the same type as operation v_i in the path of DFG below the operation v_i, n_i is the number of cycles needed to perform the operation, p_i is the p – label of the operation v_i, the minimal p label of an operation equals 0, hence addition of 1 in the denominator is necessary to avoid dividing by 0.

The *fvi* value of an operation indicates its suitability for being slowed down. It is used when there is more than one column of the same type, in order to create least interconnected column by interchanging operations.

interchange(v_i , v_j) (line 17)

This function swaps the v_i and v_j operations, so that the v_i is located in operation slots formerly occupied by v_j and vice versa.

fck(C_k) (line 19)

The value of the function is given by:

$$f_{Ck} = P^n_{dC_k} \cdot l_{Ck} \tag{6}$$

Table I

Power reduction obtained by IIOI (I), MAREL (M) and UNILO (U) for HLS for chosen DFG benchmarks

Benchmarks		CS_{MAX} extension (ΔCS)											
circuit	gates	0%			10%			20%			50%		
		I	M	U	I	M	U	I	M	U	I	M	U
c1355	514	4.1	3.5	7.7	12.6	12.9	13.1	12.6	13.1	13.1	13.1	13.2	13.2
s208	104	14.5	14.5	14.5	22.0	22.0	22.8	22.0	31.8	34.2	31.8	46.5	42.9
s5378	2779	8.2	6.6	15.3	17.0	20.4	44.0	17.0	44.8	51.8	41.0	47.1	66.8
s9234	5597	6.4	6.7	3.6	32.2	36.1	51.6	33.0	32.3	65.4	60.7	66.3	67.6

Where P^n_{dCk} is the normalised dynamic power dissipated in C_k functional unit column (*fu*-assigned to the C_k column, P^n_{dCk} is normalised to the fu_i having the lowest value of P_{di}), l_{Ck} is a number of operations in the C_k functional unit column.

The *fck* function is responsible for selecting the most suitable column (C_{ms}) for inserting idle operations, from the C_{su} set. It chooses the column assigned to the fu_i that has the highest power demand, hence gives the highest power demand reduction when slowed down.

insert_idle_operations(vi) (line 25)

This function simply adds new operation slots with idle operations after the v_i operation. If there is an empty operation slot after the last cstep occupied by v_i, then an idle operation is added there. However, when there is no empty room for a new idle operation, then the next operation in the column of v_i is delayed. Next the data interconnections between v_i and its successor operations must be checked. This is done by the delay_all_successors function described below.

delay_all_successors(v_i) (line 26)

This function checks if all the data needed to perform successor operation (si) of v_i are available on time, by checking the condition:

END_CSTEP(VI) ≤ START_STEP(SI) (7)

If it is not fulfilled, then the successor is delayed as many cycles as needed (so that start_step(s_i)=end_cstep(v_i)). Such delay implies the need for checking all the data interconnections between the successors of s_i. If the delay is not possible due to the cs_M constraint, the frequency lowering of v_i (and the functional unit it is assigned to) fails. In such a case the column containing v_i, i.e. C_{ms} is removed from the C_{su} set F, and the process starts from the beginning.

EXPERIMENTAL RESULTS

In our experiments, we also have considered CS_{MAX} extension by 10, 20 and 50 percents. The results obtained for chosen benchmarks [4] are presented in Table 1.

CONCLUSIONS

The level of power reduction obtained by IIOI, MAREL and UNILO algorithms are very promising, yet they strongly depend on the benchmark structure. The main advantage of the presented algorithms is their simplicity which makes them very robust. Further comparison of the results obtained by IIOI, MAREL and UNILO algorithms with the evolutionary one [2] show that for some benchmarks heuristic algorithms give better results.

Moreover, the heuristic algorithms run in significantly shorter time than evolutionary ones and in many cases lead to the acceptable results.

REFERENCES

[1] L. Benini, A. Bogliolo, and G. De Micheli. A survey of design techniques for system-level dynamic power management, *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 8 3, pp. 299-316, 2000.

[2] S. Koziel, W. Szczesniak. Application of adaptive evolutionary algorithm for low power design of CMOS digital circuits, *Proc. of Ninth IEEE ICECS'2002*, Dubrovnik, pp. 685-688.

[3] W. Szczesniak, B. Voss, M. Theisen, J. Becker and M. Glesner. Influence of high-level synthesis on average and peak temperatures of CMOS circuits, *Microelectronics Journal*, vol. 32, pp. 855-862, 2001.

[4] Collaborative BenchmarkingLaboratory, ftp.cbl.ncsu.edu.

HIGH POWER AMPLIFIER PREDISTORTER ASIC IN STANDARD DIGITAL CMOS TECHNOLOGY

H. Tap-Beteille,
LEN7/ENSEEIHT, Toulouse, France
tap@len7.enseeiht.fr

D. Roviras,
TeSA/IRIT/ENSEEIHT, Toulouse, France
daniel.roviras@tesa.prd.fr

M. Lescure,
LEN7/ENSEEIHT, Toulouse, France

A. Mallet,
CNES, Toulouse, France
alain.mallet@cnes.fr

DOI: 10.36724/2664-066X-2021-7-4-35-39

ABSTRACT

Satellite communications offer a wide coverage for global communication systems. In order to increase spectral efficiency, non constant modulus constellations are very attractive compared to the classical BPSK and QPSK schemes. A 16-QAM modulation could offer a significant increase in spectral efficiency for satellite communications. Because the available power on board the satellite is strongly limited, High Power Amplifiers (HPA) like Travelling Wave Tubes (TWT) or Solid State Power Amplifiers (SSPA) are generally operated near the saturation point with a low back off. When operated with such low back off, HPA are highly non linear amplifiers. So, the signal to amplify being strongly distorted, a predistorter has been developed. A high power amplifier predistorter has been implemented in 0.6 m CMOS technology. First, the predistorter is briefly described. Then, the implementation of the predistorter is shown. The circuits designed for the neuron first layer are described, as well as the simulation results obtained.

KEYWORDS: *High power amplifier, MLP neural network predistorter, Gilbert cell, Kirshoff adder, analog circuits, 0.6 m CMOS technology.*

The article is reworked from unpublished 2nd IEEE International Conference on Circuits and Systems for Communications (ICCSC) materials.

PREDISTORTER DESCRIPTION

The predistorter must fight against two impairments given by the non linear amplifier, the AM/AM and AM/PM conversions. In order to do that, a mimic structure has been adopted [1], [2]. The predistorter is composed of two Neural Networks (NN), one NN computes the phase shift correction and the other one computes the gain correction. The input of both NN is the square modulus, $\rho^2 = I^2 + Q^2$, of the input signal I+jQ. Data I and Q having a 25MHz bandwidth, the input of the NN, ρ^2, is a 50MHz signal. All details concerning the structure of the predistorter can be found in [4] for example. Each NN is a MLP network with 10 neurons in the hidden layer. The architecture of one of the NN is shown on figure 1.

W1k and b1k are respectively the input weight and bias (or offset) of the first layer of neuron number k. W2k is the output weight (or gain) of neuron k, while f(.) is a non linear function like a hyperbolic tangent function. b2 is the output bias of layer 2. The weights and biases are updated following a gradient based algorithm using back propagation [3].

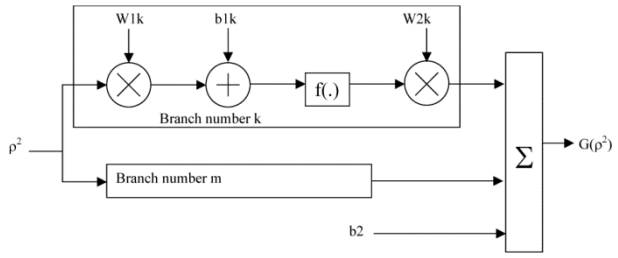


Fig. 1. Architecture of the MLP Neural Network

PREDISTORTER IMPLEMENTATION

The NN core is implemented in an analog ASIC whereas the adaptive algorithm is implemented in an external DSP. For the digitization of signals, 8 bits DAC converters are sufficient. This technique is particularly attractive because of the high integration density for the NN core; the good accuracy and the easy memorization for the algorithm part. Moreover, analog implementation permits to control high frequency signals unlike digital techniques; and implementing the algorithm on a DSP permits the future evolution of the adaptive algorithm.

The NN core is realized with a submicronic CMOS technology because CMOS offers a good reliability as well as reproducibility. The maximal bandwidth and the power supply are determined by the minimum length of the transistors channel. So, we have first verified that the circuits bandwidth was sufficiently high in 0.6 µm technology (> 50MHz) [4]. This technology permits to use a 5V power supply instead of 3.3V for a 0.35 µm technology. Having a higher voltage permits to have a higher dynamic range of the neuronal functions. The choice of the packaging is made by computing the number of inputs/outputs, considering the neural network, one test neuron, test elementary cells and more than 10 ground pins evenly distributed. This balance sheet, reported on table, imposes to use a JLCC84 case.

Table 1

Balance sheet of the 84 pins ASIC

function	Pin names	mode	quantity
square modulus ρ^2	I and Q	input	8
square modulus ρ^2	ρ^2	output	1
Neuron first layer	ρ^2	input	1
Neuron first layer	First weight W1	input	10
Neuron first layer	First bias b1	input	10
Hyperbolic tangent function	Tanh output	output	10
Neuron second layer	Second weightW2	input	10
Neuron second layer	Second bias b2	input	1
Neuron second layer	Predistorter output	output	1
Test branch	/	/	7
Test cells	/	/	13
Electrical ground	/	/	12
			Total 84

To implement the neuronal functions, three types of analog cells are necessary [5]: adders, four-quadrant multipliers and a non linear function. Many possibilities have been explored to realize these functions. At last, all the functions implemented on the ASIC are Gilbert cells, Kirshoff adders and sources-coupled differential cell to realize the hyperbolic tangent function. Each cell has been individually optimized depending on its specifications and place in the predistorter.

THE SQUARE MODULUS FUNCTION ρ^2

The implementation of the square modulus function is shown on figure 2. It is composed of 2 Gilbert cells configured like frequency doublers, to perform I^2 and Q^2 respectively. A Kirshoff adder is used to perform $I^2 + Q^2$.

The maximum dynamics of the input signals(I(t) and $Q(t)$) has been evaluated, in order to reduce noise while preserving a signal distorsion < 1% at the output of the function. The maximum amplitude of the differential input voltage $\Delta V_{d\ Max}$ allowed for I and Q signals in order to avoid the saturation of the output differential current is [6]:

$$\Delta V_{d\,Max} < \sqrt{\frac{I_{pol}}{K_P (W/L)}} \cong 250\,mV \qquad (1)$$

Ipol = 500 µA +/- 20%, the current supplied by Q7 or Q'7, K_p = µon C_{ox} = 120e^{-6} A^2/V +/-20%, W/L = 40µm / 0.6µm the channel size of the sources-coupled pair.

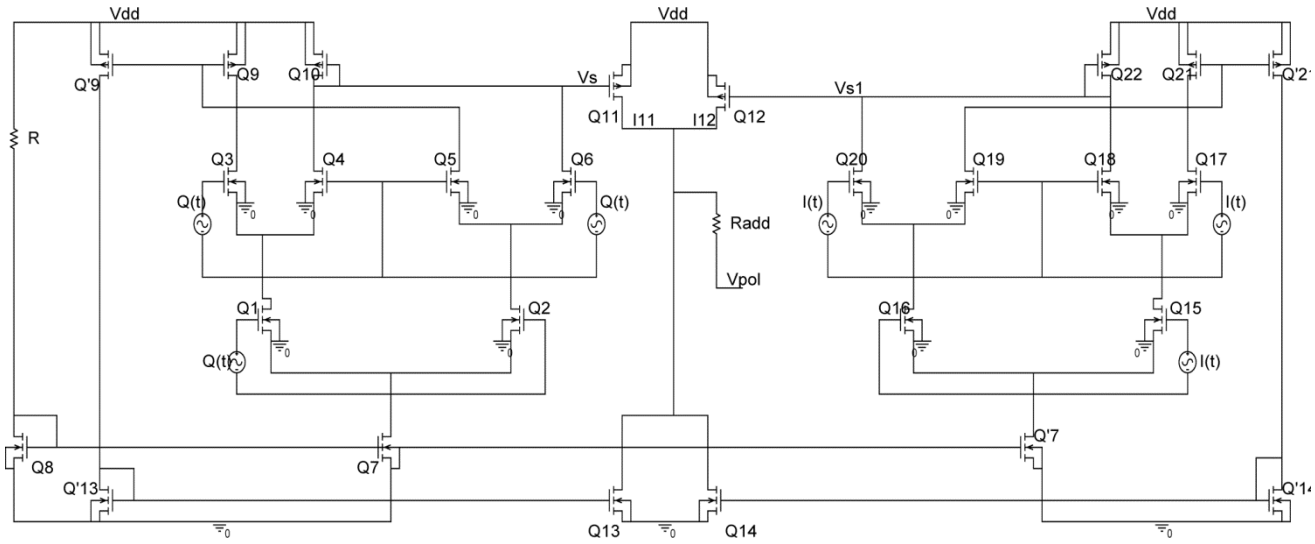


Fig. 2. The square modulus function is composed of 2 Gilbert cells and a Kirshoff adder

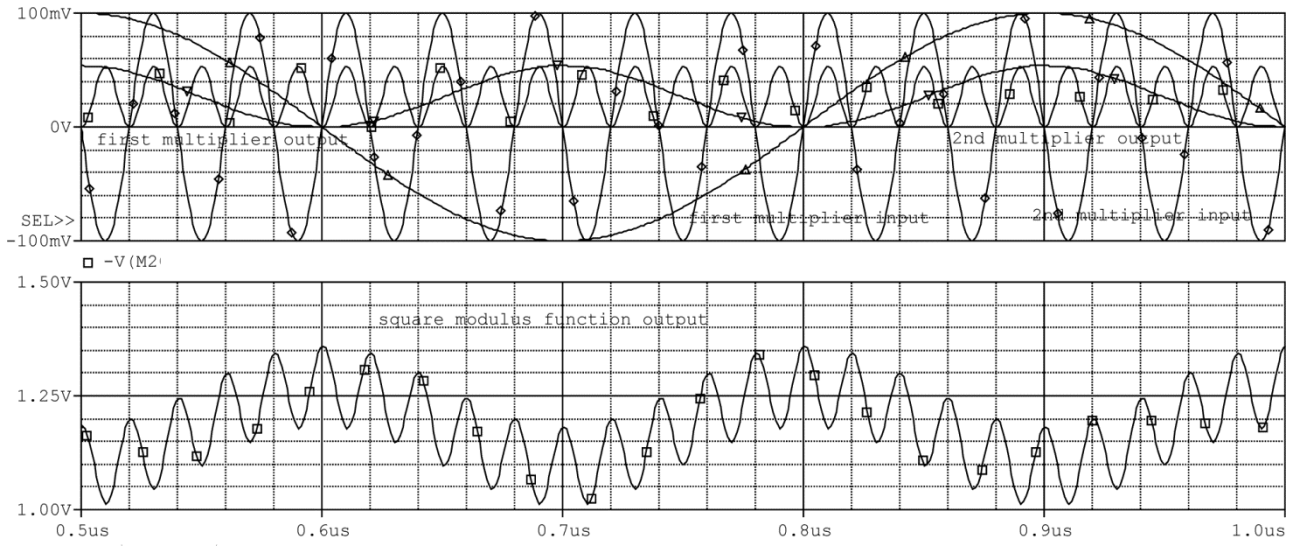


Fig. 3. a) ◇ first multiplier input, □ first multiplier output, Δ second multiplier input, ▽ second multiplier output;
b) □ square modulus function output

I_{pol} is not very well known because of the absolute tolerance on the sheet resistances, as well as K_p. That is why we have chosen I and Q range around 200mV to stay in the linear region of the transfert of the differential pairs as well as being strongly superior to the voltage noise. Simulation results showing the multipliers outputs and the function output are reported on figure 3.

THE NEURON FIRST LAYER

In the neuron first layer (fig. 1), there is to multiply the square modulus with a weight (to control the amplitude and the polarity of the signal) and to add it a bias (to control the d.c.value). Both functions have been implemented together as shown on figure 4.

The output of the square modulus function is applied on the gate of Q3 and Q4 through the use of an external capacitor in order to filter the d.c.value of ρ2. Matlab simulations have shown that the high-pass filter obtained must have a low cutoff frequency less than or equal to 1MHz to keep a Signal to Error Ratio of 36dB [7]. SER being the ratio between the HPA output power and the power of the identification error. Knowing that the input resistance R = 10kΩ, the external capacitor C must be equal to or higher than 16pF.

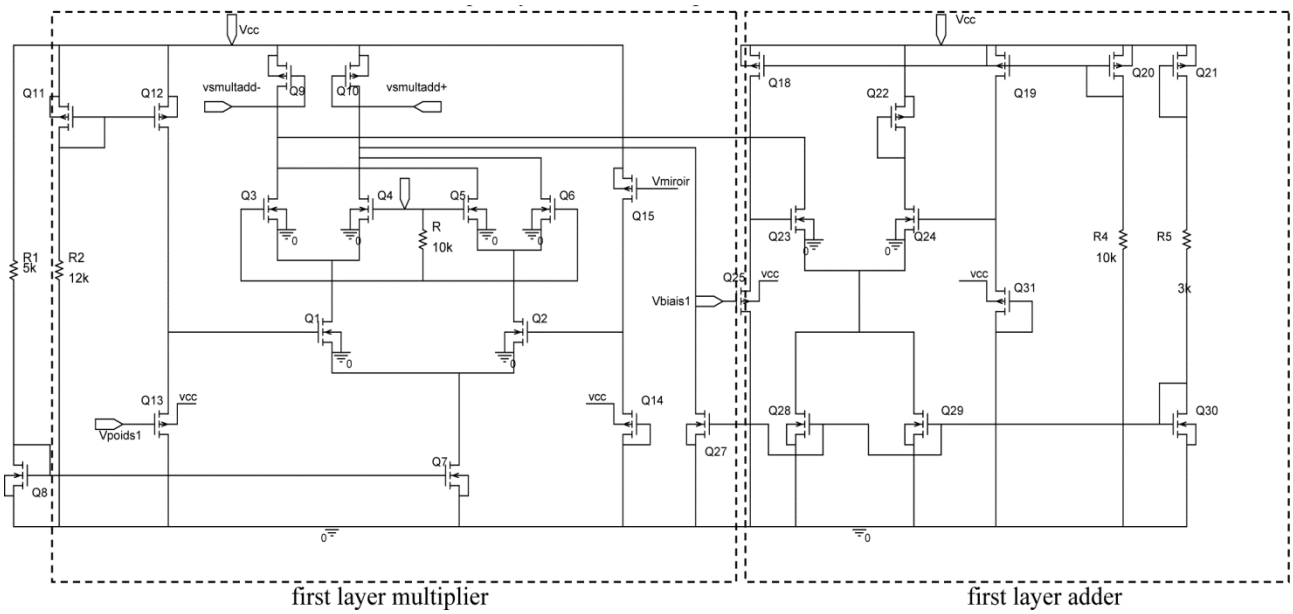


Fig. 4. Neuron first layer: the square modulus function output is multiplied with a weight called "vpoids1" and is added with an offset called "vbiais1"

This value is too high to be integrated in the ASIC, that is why the connection between the output of the square modulus function and the input of the neuron first layer is external. The multiplier used is a Gilbert cell where the signals to multiply are the output of the square modulus function and the weight called Vpoids1.

The principle of the adder part is to create a current I in Q30 which is mirrored in Q27, Q28, Q29. Q27 will always impose a current I to one of the cell outputs (Vsmultadd+). A current 2I must be provided by Q23 + Q24. Each of them will provide a current I when Vbiais1 = 0. So, Q29 will impose a current I to one of the outputs of the multiplier, as well as Q27 to the other output. In this case the d.c.value of the cell output (Vsmultadd+ - Vsmultadd-) is equal to zero.

If Vbiais1 increases, current in Q23 increases, so the d.c. value of (Vsmultadd+ - Vsmultadd-) increases ; and conversely. The output of the cell (Vsmultadd+ - Vsmultadd-) is presented on figure 5a with vbiais1 = 0V and Vpoids1 varying between –0.5V and +0.5V and on figure 5b with Vpoids1 = 0.1V and Vbiais1 varying between –0.5V and 0.5V. The input of the cell is a 50MHz sine.

Then, the output signal is multiplied with a tanh function (fig.1). The function has been realized with a source-coupled differential pair [4].

Figure 6 represents both drain currents versus the input differential voltage (curve a). The hyperbolic tangent function has also been plotted (curve b).

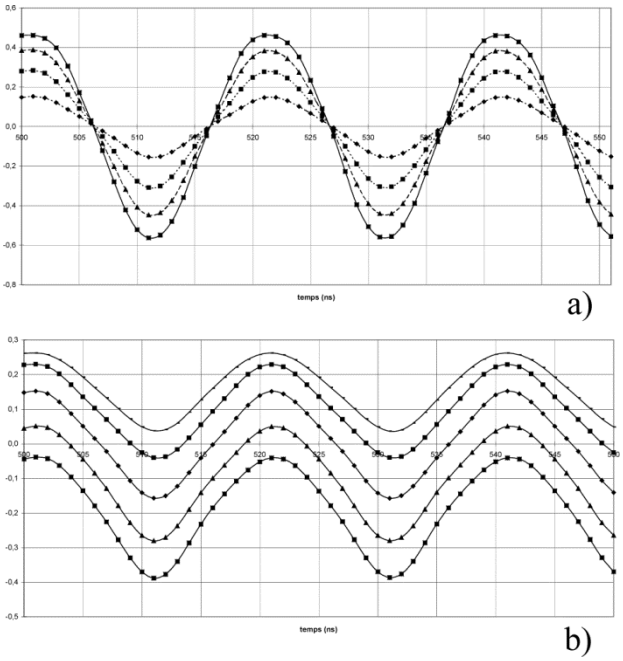


Fig. 5. Output of the first layer multiplier-adder (Vsmultadd+ - Vsmultadd-):
a) vbiais1 = 0V, -0.5V<Vpoids1<0.5V;
b) vpoids1 = 0 V, -0.5V<Vbiais1<0.5V

The neuron second layer, as well as the output adder, consists of the same functions as those constituting the first layer.

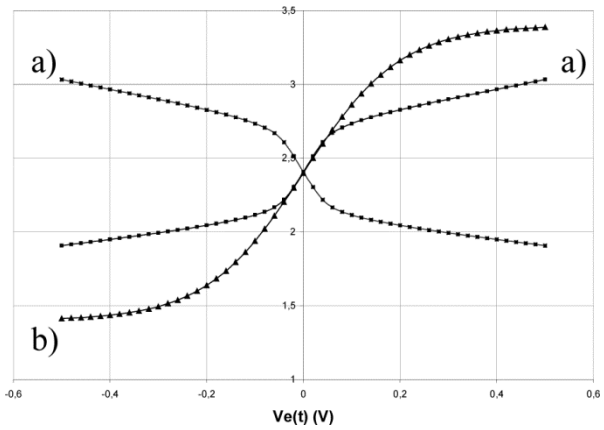


Fig. 6. a) drain currents of the source-coupled differential pair (simulation results); b) hyperbolic tangent function

CONCLUSIONS

We have simulated a Neural Network predistorter. This NN is based on a MLP structure. Because of the high speed of the transmitted signal (higher than 25MHz), we have adopted an analog implementation of the MLP Neural Network. The circuit has been implemented in 0.6µm CMOS technology.

REFERENCES

[1] F. Langlet, H. Abdulkader and D. Roviras. Predistorsion of non-linear satellite channels using neural networks : Architecture, Algorithm and Implementation, *EUSIPCO 2002*, Toulouse, France, Sept. 2002.

[2] F. Langlet, H. Abdulkader, D. Roviras and F. Castanie, A. Mallet, L. Lapierre. New predistorsion schemes for satellite power amplifiers, *Microwave Technology and Techniques Workshop*, ESTEC, ESA, Netherlands, October 2002.

[3] Simon Haykin. Neural Network: A Comprehensive Foundation, Prentice Hall, 1994.

[4] H. Tap-Beteille, D. Roviras, M. Lescure , F. Castanie, A. Mallet, Integration in CMOS Technology of a High Power Amplifier Predistorter, *SPSC2003*, Catania, Italy, 24-26 September 2003.

[5] C. Mead, Analog VLSI and neural systems, Addison-Wesley Publishing Company, 1989.

[6] R.L. Geiger, P.E. Allen, N.R. Strader. VLSI techniques for analog and digital circuits, McGraw Hill, 1990.

[7] D. Roviras, H. Abdulkader, H. Tap-Béteille, F. Castagne, M. Lescure, A. Mallet. Multi-layer preceptron neural network implementation and integration in CMOS technology, *International Conference on Information & Communications Technologies : from Theory to Applications (ICCTA'04),* Damascus, Syria, 19-23 April 2004.

FAST FRACTAL IMAGE ENCODER DESIGN

Yung-Gi Wu,
Department of Computer Science and Information Engineering Institute of Applied Information, Leader University, Tainan, Taiwan
wyg@mail.leader.edu.tw

DOI: 10.36724/2664-066X-2021-7-4-40-44

ABSTRACT

Fractal theory has been widely applied in the filed of image compression due to the advantage of resolution independence, fast decoding, and high compression ratio. However, it has a fatal shortcoming of intolerant encoding time because that every range block is need to find its corresponding best matched domain block in the full image. Therefore, it has not been widely applied as other coding schemes in the field of image compression. In this paper, an algorithm is proposed to improve this time-consuming encoding drawback by the adaptive searching window, partial distortion elimination and characteristic exclusion algorithms. Proposed can efficient decrease the encoding time significantly. In addition, the compression ratio is also raised due to the reduced searching window. Conventional fractal encoding for a 512 by 512 image need search 247009 domain blocks for every range block. Experimental results show that our proposed method only search 120 domain blocks which is only 0.04858% compared to conventional fractal encoder for every range block to encode Lena 512 by 512 8-bit gray image at the bit rate of 0.2706 bits per pixel (bpp) while maintaining almost the same decoded quality as conventional fractal encoder does. This paper contributes to the research of encoder of fast image communication system.

KEYWORDS: *Fast communication, fractal encoder, compression.*

The article is reworked from unpublished 2nd IEEE International Conference on Circuits and Systems for Communications (ICCSC) materials.

1. INTRODUCTION

A picture may be worth a thousand words, but it requires far more memory to store or bandwidth to transmit. With the successes of multimedia technology and the era of wideband network, peoples' desire to the high quality of multimedia still can not be satisfied. All the scholars in the universities and the engineers in the industry want to get the compromise between the limited network bandwidth and the unlimited human desire. Among all the digitized data that we people can touch every day, such as digital library, VCD, DVD, JPEG, etc, the kernel of the system or the standard that commercial products use is the compression technique.

There are many coding schemes that have been developed such as DCT, VQ, Wavelets, BTC, Fractal, etc. In general, the criteria to evaluate the performance of a compression system include 1) compression ratio 2) reconstructed quality 3) processing time. Fractal compression can achieve the highest compression ratio among all the existed coding schemes theoretically; however, its encoding time is terrible intolerant to practical industrial applications. If this shortcoming of long encoding time can be improved, the application of fractal will be a practical consideration.

The cause of fractal image coding with high compression is that the minority of blocks through rotations represent the majority of blocks. In a word, fractal encoding is based on Partitioned Iterated

Function System (PIFS). The detailed descriptions of PIFS can be found in [2], [3], [4]. To improve the drawback of time consuming in encoding process, we utilize classification and local search techniques to exclude those un-related searching domains to reduce the encoding time while decreasing the bit rate as well in this paper. The remaining sections are organized as follows: Section 2 will depict the basic fractal image coding. Proposed method is given in Section 3 and results will be shown in Section 4.

FRACTAL IMAGE CODING

Fractal image coding is based on partition iterated function system (PIFS). Let an original image be partitioned into non-overlapping regions called range blocks (R) and overlapping regions called domains blocks (D). The size of each domain block should be larger than that of the range block to satisfy the property of contraction. Let D' denotes the down sampled domain block of D and the D' size is equal to range blocks to match the contractive property.

The transformations are composed of a geometric transformation and a massic transformation. The geometric transformation consists of moving the domain block to the location of the range block and adjusting the size of domain block to match the size to size range block. The massic transformation adjusts the intensity and orientation of the pixels in the domain block after it has been operated on by the geometric transform. The massic transformation t_i can be depicted as follows:

For each range block, we must find the best matching domain block (D') by the affine transformations

$$t_i\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} a_i & b_i & 0 \\ c_i & d_i & 0 \\ 0 & 0 & s_i \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} e_i \\ f_i \\ o_i \end{bmatrix} \quad (1)$$

where si controls the contrast and oi controls the brightness. $Z = f(x, y)$ is the gray level value at (x, y) and a_i, b_i, c_i, d_i, e_i, f_i denote the eight-symmetry such as (1) Identity mapping (2) Rotation by 90 degrees (3) Rotation by 180 degrees (4) Rotation through -90 degrees (5) Reflection about mid-vertical axis (6) Reflection about mid-horizontal axis (7) Reflection about diagonal (8) Reflection about cross diagonal. The i in s_i and o_i denotes one of the above mentioned symmetries which means ($1 <= i <= 8$).

In practice, we compare a range block and down sampled domain blocks using RMS metric as follows

$$RMS = \sum_{k=1}^{n\times n}\left(s_i \times a_k + o_i - b_k\right)^2 \quad (2)$$

Where ak represents the pixel value of the domain blocks (D') after eight transformations and b_k represents the pixel value of the range blocks and the block size for both R and D' is n by n. This RMS metric allows easy computation for optimal values of s_i and o_i in equation (1). This will give us contrast and brightness settings that make the affinely transformed a_k values have the least squared distance from the b_k values. The minimum of RMS occurs when the partial derivatives with respect to s_i and o_i are zero, which occurs when

$$si = \frac{n\times n\left(\sum_{k=1}^{n\times n} a_k b_k\right) - \left(\sum_{k=1}^{n\times n} a_k\right)\left(\sum_{k=1}^{n\times n} b_k\right)}{n\times n\sum_{k=1}^{n\times n} a_k^2 - \left(\sum_{k=1}^{n\times n} a_k\right)^2} \quad (3)$$

$$Oi = \frac{\sum_{k=1}^{n\times n} b_k - si\sum_{k=1}^{n\times n} a_k}{n\times n} \quad (4)$$

There are many best matched criteria to choose. The root mean square (RMS) is usually used in fractal image coding and the minimal RMS is the better matching. We use equation (2) to find the optimal si and oi and then quantize them for storage or transmission.

In addition, the encoder must record the position of the best matched domain block (D') and its transformation for each range block so as to reconstruct the decoded block on the decoder side.

The following example depicts how this encoding can be done. Suppose the data to be dealt with is 512 x512 pixel image in which each pixel can be one of the 256 levels of gray(ranging from black to white). Let Ri be the 8x8 pixel non-overlapping range block (i=1,..,4096) and let D be the collection of all the 16x16 overlapped sub-squares of the image.

The collection of D contains 497x497=247009 squares. For each R_i, search through all of collection of D_i to find one which minimizes the RMS as equation (2); that is, find the part of the image that most looks like the image above R.

There are 8 ways to map one square onto another, so that this means comparing 8x247009=1976072 squares with each of the 4096 range blocks. In addition, we must fulfill down-sampling for each Di to get the same size of R to satisfy the property of contraction.

Choosing 1 from each 2x2 sub-square of D_i or averaging the 2x2 sub-square corresponding to each pixel of R can achieve the goal of down-sampling. From the above descriptions about the conventional fractal encoding, we know the huge computation overhead is obviously. The time to search the best matched domain block for every range block is intolerant a time consuming job in practical application.

Therefore, we develop a new encoding algorithm to reduce the time in this research. In addition, the bit rate is also be reduced by the smaller search window. A lot of people make efforts in fractal improvement. Some investigate region-based image coding methods in [5] and some combine fractal with other algorithm such as wavelet in [6], genetic algorithms in [7], discrete cosine transform in [8], [9]. In a word, they make use of classification methods to reduce time.

FAST ENCODING ALGORITHM

The major factor to cause the huge computation for fractal encoding is the searching number of domain blocks. Therefore, decreasing the number of it is an institute method to speed up the encoding time. The methods to achieve the goal in proposed method are by classifying range and domain blocks, respectively and searching the algebraic adjacent region for every range block. Following sub-sections depict the detailed descriptions of proposed method.

The classification approach has the advantage that it elegantly excludes the range block to search those domain blocks whose characteristics do not match the characteristic of range block.

Here, variance of a block is used to be the basic criterion of classification. The formulas to get the variance of block X are given as follows:

$$u(X) = \frac{1}{n \times n} \sum_{i=1}^{n \times n} X_i \tag{5}$$

$$\mathrm{var}(X) = \frac{1}{n \times n - 1} \sqrt{\sum_{i=1}^{n \times n} (X_i - u(X))^2} \tag{6}$$

nxn is the block size of X. There are four classes of c1, c2, c3 and c4 after the classification. c_1 is classified by variance merely once the condition var(X)<=T1 is met. Those range blocks belonged to c1 class are encoded and transmitted or storage by the mean value $u(X)$ only. The other range blocks whose classifications are not c1 will be encoded by fractal and those blocks are classified into c_2, c_3 and c_4. The criterion to classify those blocks into (c_2, $c_3$3, c_4) is according to the distortion after fractal encoding. Detailed algorithm is depicted as following:

Input: range blocks (R) whose var(R)>T1

Output: c_i ($i \in \{1,2,3\}$)

Step 1: Implement fractal encoding to find the best matched D within the searching window and record its distortion RMS by equation (2).

Step 2: If RMS>T2, output c_4;

else ((RMS>T3)&&(RMS<=T2)) output c3;

else output c_2;

We set three thresholds {T1, T2, T3} to get the classification.

Note that T1 is the variance of input range block for classification and {T2, T3} is the RMS distortion to classification. Refer to Figure 1 and Figure 2, which are the resulted regions after classification when the threshold set is to {0.5, 25000, 15000}.

Figure 1 is c1 class regions.

Fig. 1. Pure blocks after classification

There are three classes regions shown in Figure 2.

Fig. 2. Fractal encoding regions

The regions marked by light lattice denote the c_4 region, dark lattices represents c_3 region and the other regions without lattice belong to c_2. Here, the size of each range block is 8 by 8. c_1 blocks are compressed or transmitted by the mean value of itself and c_2 ~c_4 blocks are compressed or transmitted by fractal encoding. The following sections depict the reduction of searching window and the computation reduction method to find the best matched domain block for each $c_2 \sim c_4$ range blocks.

If the size of each overlapped domain block (D) is 16x16 and the we move D pixel by pixel to cover the whole image, there are 247009 domain blocks for a 512 x 512 image. A range block is encoded by projecting it tentatively on to the entire 247009 domain blocks in conventional fractal encoding. Therefore, its computation burden is terrible huge. Proposed method decreases the searched number of domain blocks to about 120 while losing a little fidelity. For each range blocks in $c_2 \sim c_4$ classes, the searching window (SW) is only constrained to the nearby regions instead of the entire domain pool and we do not move D pixel by pixel within the SW to reduce the computation burden furthermore.

The step size to move D within the searched window is set to 4 pixels in the later experiments. In addition, we exclude the domain blocks whose variances do not match the range block to be encoded within search window. Detail descriptions of the reduction of searching window for $c_2 \sim c_4$ range blocks are given as follows:

Input: Class of range block (cx) and position of range block (i,j), image_size=MxN.

Output: range of search window.

```
        switch (cx)
{
        case (c2):  SW=SW_2;  break;
        case (c3):  SW=SW_3; break;
        case (c4):  SW=SW_4; break;
}
search_range_of_i= (i-SW) ~ (i+SW);
search_range_of_j= (j-SW) ~ (j+SW);
if (i-SW)<0  search_range_of_i=0 ~ 2 x SW;
if(i+SW)>M search_range_of _i=(M-2 x SW) ~ M;
if (j-SW)<0  search_range_of_j=0 ~ 2 x SW;
if(j+SW)>N search_range_of _j=(N-2 x SW) ~ N;
```

If the size of the domain block is 16 by 16 and we move the domain block pixel by pixel then the total searched number of domain block is *(2xSW-15)x(2xSW-15)*. In this research, we move the range block four pixels every time so that the number of the searched domain block is

$$\frac{2\times SW-15}{4}\times\frac{2\times SW-15}{4} \tag{7}$$

Compared to the conventional fractal encoding algorithm whose searched numbers of domain block for every range block is (M-15)x(N-15) when the image size is M by N and the size of the domain block is 16 by 16, the economize searched number of domain block is given in (8)

$$\frac{\frac{(2\times SW-15)^2}{16}}{(M-15)\times(N-15)} \tag{8}$$

Let the image size be 512 by 512 and the SW be 32, it only searches about 144 domain blocks nearby the range block in the proposed method, which is only about 1/1715 compared to the conventional fractal encoding scheme. After selecting the searching window for every range block, the next step is to calculate as equation (2) to find the best matched one and its correspondent transformation within the window. Note that for every input range block whose class is not c1, we treat it as c2 class at first. If it meets the criterion of classification as described in the previous section, it will select larger searching window size to find better domain block.

The major factor to lead the fractal encoding become a time consuming technique is the vast computation cost on RMS. Previous section depicts the algorithm to select the nearby domain blocks which decreasing the number of RMS of computation instead of the whole image. This section uses another methodology to decrease the computation furthermore.

For every selected domain block, we must compute the error between the range block and the eight-transformed domain blocks to find the best matched one whose error is minimum. We use a method to reduce the computation of RMS metric. This algorithm is called as Partial Distance Elimination (PDE) which is devised by Bei and Gray in 1985 for the application of fast Vector Quantization encoding. The PDE algorithm discards all domain blocks whose partial distortion relative to an input vector exceeds the current available nearest distortion. The operation of the PDE algorithm is summarized as follows:

Input: Range block (R);

Selected sampled domain blocks D' (number is z);

Output: Best matched D' and its best isometric;

Step1: find si and oi by equation (3)(4) for first D';

i=0;

$$RMS=\sum_{k=1}^{n}\sum_{l=1}^{n}(D_1^{'}(k,l)\times s_i+o_i-R(k,l))^2 \tag{9}$$

```
        current_error =RMS;
        best_D'=1;
        best_isometric=0;
Step 2:
    for(p=1;p<=z; p++)  {
    find si and oi by equation (3)(4) for every  D' ;
    if(p==1)  a=1; else a=0;
// the RMS of first D' and its first isometric has calculated already
in Step 1;
            for(i=a; i<8;i++) {
            RMS=0;
            for(k=1; k<=n; k++)
             for(l=1; l<=n; l++)
              {
```

```
        RMS+ = (D_i'(k,l) × s_i + o_i − R(k,l))^2  (10)
          if (RMS>=current_error)  goto exit;
        }
        If(RMS<current_error)
        {
        current_error=RMS;
        best_D'=p;
        best_isometric=m;
        }
        exit: { };
          }
}
return (best_D', best_isometric);
```

There are z D_i' (i=1..z) to be compared to R and every D' has eight-symmetry forms; thus, the total RMS computation is zx8. Step 1 calculates a RMS error between R and the first isometric of D_1'. Step 2 calculates all the other RMS including the other seven isometrics of D_1' and the eight isometrics of all the D_2' ~ D_z'; meanwhile, it compares the distortion between the current minimum error and increasing RMS in (10). As long as the increasing RMS exceeds the current minimum error, it stops the RMS calculation and exits to the outer of the loop. Such a method can decrease the heavy computation of RMS effectively.

SIMULATION RESULTS AND CONCLUSION

In order to evaluate the performance of the proposed method, several images including Lenna, Lena are employed to test. of the test images are 512 by 512 with 8-bit gray level. The circumstance of our experiment was on a single PC with Pentium 4-2.4GHz CPU and 256 MB RAM. In our discussion of the results, we will focus on the visual quality, encoding time, compression ratio, and PSNR. Visual quality is an objective judgment and PSNR is an evaluation in mathematic sense.

The bit rate for a 512 by 512 image whose size for a range block is 8 by 8 is calculated by the following:

$$BR=\frac{\#c1\times 8+(\#c2\times(2\times sw\_c2-15)^2+\#c3\times(2\times sw\_c3-15)^2+\#c4\times(2\times sw\_c4-15)^2)/16+(\#c2+\#c3+\#c3)\times(15)}{512\times 512}+\frac{2}{64}\ \text{bpp} \quad (11)$$

#*cn* denotes the number of each class. *sw_cn* represents the searching window size of class *n*. Because the c1 class range blocks are all pure, only the mean value of each one is transmitted instead of the fractal encoding.

The second term (#c2× (2× sw _ c2 −15)2 +#c3× (2× sw _ c3 −15)2 +#c4× (2× sw _ c4 −15)2) /is the bits used to represent position of the searched D within the searching window for c2 ~ c4 classes. The size of each D is 16 by 16 and moves the D four pixels a time. (#c2+#c3+#c4)×15 is used to specify the 8 transformations (3-bit) and si (7-bit) and oi (5-bit) for each R in class c2~c4. The last term 2/64 is used for classification of each range block.

The resulted data of encoding time, searched domain blocks and decoded quality after running to all the four images by the proposed method and conventional fractal encoding are given in table 1.

Table 1

Simulation Results

	Lena	L
Bit_rate(bpp)	2706	0.2475
Compression Ratio	29.562	32.322
Encoding time (second)	15.89	12.546
PSNR(dB)	28.9641	30.0904

The thresholds of classification (T1, T2, T3) is (0.5, 25000, 15000) and the searching window size for class_2, class_3, class_3 is 8, 16 and 64, respectively. The table lists the number of each class. The blocks belongs to Class_1 is pure so that fractal encoding is not employed.

The average searched domain blocks for each range block with respect to Lena and Lenna are 129 and 120, respectively. Practical running time is 15.89 and 12.546 seconds. Due to the reduced search window for each range block, the bits used to the represent best matched domain block also decreased.

Refer to table 1, the bit rate is 0.515625 bpp for conventional fractal encoding and the proposed method decreases the bit rate to 0.2706 bpp and 0.2475 bpp for Lena and Lenna, respectively. Running time is 28685 and 28681 seconds for the two test images and bit rate for the two images are 0.515625 bpp by using the conventional fractal encoding. Because the numbers of searched domain blocks are restricted to the searching window, the decay of quality is indeed irreproachable. However, the visual difference is almost unnoticeable.

REFERENCES

[1] M. F. Barnsley. *Fractal everywhere*, Academic Press, New York, 1988.

[2] Y. Fisher. *Fractal Image Compression. Theory and Application*, New York: Springer-Verlag, 1994.

[3] A.E. Jacquin. Fractal image coding: a review, *Proceedings of the IEEE*, Vol.81, No.10, pp.1451-1456 Oct.1993.

[4] A.E. Jacquin. Image coding based on a fractal theory of iterated contractive image transformations, *IEEE Trans.Image Processing*, vol. 1, pp.18-30, Jan. 1992.

[5] B. Wohlberg and G. de Jager. A review of the fractal image coding literature, *IEEE Trans. Image Processing*, vol. 8, pp. 1716-1729, Dec. 1999.

[6] G.M. Davis, A wavelet-based analysis of fractal image compression," *IEEE Trans. Image Processing*, vol. 7, pp. 141-154, Feb. 1998.

[7] M. Takezawa, H. Honda, J. Miura, H. Haseyama and H. Kitajima. A genetic-algorithm based quantization method for fractal image coding, *IEEE Trans. Image Processing*, vol. 1, pp. 458-461, 1999.

[8] Y. Zhao and B. Yuan, "Image compression using fractals and discrete cosine transform," *Electron. Lett.*, vol. 30, no. 6, pp. 474-475, 1994.

[9] Y. Zhao and B. Yuan. A hybrid image compression scheme combining block-based fractal coding and DCT, *Image Communication.* Vol. 8, No. 2, pp. 73-78, Mar. 1996.

[10] C. Bei and R.M. Gray. An improvement of the minimum distortion encoding algorithm for vector quantization, *IEEE Trans. Commun*., vol. COM-33, pp. 1132 1133, Oct. 1985.