

Supply Voltage Scalable System Design Using Self-Timed Circuits

W. Kuang, J.S. Yuan* and A. Ejnoui

Chip design and Reliability Laboratory

School of Electrical Engineering and Computer Science

University of Central Florida, Orlando, FL 32816

*Contact author email: yuanj@mail.ucf.edu

Abstract

Supply voltage scalable system design for low power is investigated using self-timed circuits in this paper. Two architectures are proposed to achieve supply voltage scalability, for preserved quality and energy-quality tradeoff respectively. In the first architecture, the supply-voltage automatically tracks the input data rate of the data path so that the supply-voltage can be kept as small as possible while maintaining the speed requirement and processing quality. In the second one, further energy saving is achieved at the cost of signal-noise-ratio loss in digital signal processing when an ultra-low supply voltage is applied. Cadence simulation shows the effectiveness for both architectures. More than 40% to 70% power can be saved by introducing -15dB to -10 dB error in a case study: speech signal processing.

1. Introduction

In recent years, very large-scale integrated (VLSI) system design for low power is of great interest given the proliferation of portable devices, the need to extend the lifetime of a battery, and reduce cooling cost. Many low power techniques have been developed at different design levels, such as the algorithmic level [1], architecture level [2], logic level [3], circuit level [4], and device level [5]. However, it is possible that the available energy is not sufficient to support a computing mission with a required quality even if all of possible low power techniques have been applied. It would be highly desirable to tradeoff some quality and extend the operation time. This tradeoff is referred to as energy scalability design [6].

The notion of energy -scalable system design was introduced by A. Chandrakasan in [6]. An algorithmic transformation-based energy scalability design method has been applied to digital signal processing (DSP) applications such as finite impulse response (FIR) filtering, discrete cosine transform (DCT), and classification. Based on the characteristics of data and the most-significant-first principle, most significant sub-missions are performed while less significant ones are

not, so that the energy consumed is reduced while the proportional hit in quality is minimal.

This paper presents the implementation of energy scalable system at circuit level by scaling supply voltage and using self-timed circuits. The contributions of this paper are: 1) a novel and simple adaptive supply-voltage scheme applied to self-timed circuits without quality loss, and 2) the implementation and thorough analysis of the energy-quality tradeoff in Soft DSP using self-timed circuits.

2. Previous work

Supply voltage scalable circuit design is another effective approach to energy scalable system design at circuit level. This method trades speed by reducing supply voltage for low power. Although reducing supply voltage V_{DD} leads to an increase in circuit delays, the increased delays are allowed as long as the circuit still meets the speed requirements. Significant power and energy savings are possible if the supply voltage is to scale down to the smallest possible while maintaining the specific speed requirements. A variable supply-voltage scheme [7] has been developed for synchronous system, where special attention should be paid on the increased delays due to the complicated clock distribution. This scheme includes a large logic control circuit, which consumes the area of a chip and significant power. So it is not efficient when applied to smaller circuits. A technique that combines self-timed circuitry with adaptive scaling of supply voltage was proposed by L.S.Nielsen et al [8]. However, this technique needs FIFO buffers, and the minimal size of the buffers depends on the data rate. The power dissipated by the buffers degrades the advantage of adaptive scaling.

When the available energy is not enough to support a given mission, in order to cover the lifetime of the mission the supply voltage can be reduced beyond the requirement imposed by the critical path delay. As a result, the computing quality will be lost in some degree. For DSP applications, an equivalent signal-to-noise ratio (SNR) will be introduced. This tradeoff was implemented by synchronous DSP architectures in [9], referred to as

soft digital signal processing. However, the application of this technique is limited by the following factors: 1) the applied circuit must have a characteristic of delay data-dependency; 2) the probability of the erroneous outputs depends not only on the supply-voltage but also on the distribution of inputs, and a high probability of the erroneous outputs may degrade the accuracy of the final outputs to an unacceptable degree.

3. Basic self-timed architecture

In the data path of the self-timed circuit, each bit of the data is encoded by dual rails, shown in Table 1. The state DATA 0 ($D0=1, D1=0$) corresponds to a Boolean logic 0. The DATA 1 ($D0=0, D1=1$) corresponds to a Boolean logic 1. Spacer ($D0=0, D1=0$) corresponds to the empty set meaning that value of the bit is not yet available. The state ($D0=1, D1=1$) is forbidden.

Table 1 Dual-rail encoding scheme

Bit value	Rail logic value	
	D0	D1
DATA 1	0	1
DATA 0	1	0
Spacer	0	0
Invalid	1	1

In order to achieve speed-independence, the data path must work under Seitz's weak condition [10]. Some techniques, such as differential cascode voltage switch logic (DCVSL) [11] and NULL convention logic (NCL) [12], can be exploited to design the dual-rail data path to meet Seitz's weak condition. Martin's delay-insensitive full adder [13] can be used in the data path. Registers need to be dedicatedly designed to manage the handshake protocol. The structure of a 2-bit register is composed of C-elements and OR gates, as shown in Fig. 1. If the request signal from the post-stage is high to request data, then data are allowed to pass through the register, and when each bit is datum, the acknowledge signal will become low to request spacer from the pre-stage, which means the computation is finished and the circuit needs to be reset. Similarly, if the request signal from the post-stage is low to request spacer, then spacer is allowed to pass through the register, and when all of bits are spacers, the acknowledge signal will become high to request another data from the pre-stage, which means the reset is finished and the circuit can start another computation.

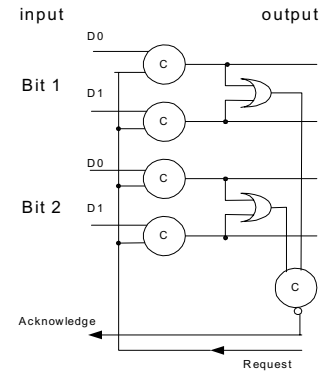


Fig.1 A 2-bit register

In many applications, the input data of the self-timed come from a synchronous system, such as an A/D converter, and the data rate is constant and independent of the delay of the self-timed circuit, as shown in Fig. 2 (a). However, the allowed maximum input data rate is limited by the speed of the self-timed circuit. The timing constraint is illustrated in Fig. 2 (b), where T_{data} is the input data cycle, D_{data} is the propagation delay of data from register 1 to register 2, which includes the delays of two registers and the data path, similarly D_{spacer} is the propagation delay of spacer from register 1 to register 2. The sum of D_{data} and D_{spacer} must be no more than T_{data} , i.e., a complete set of data (or spacer) must arrive after the corresponding request signal. Otherwise, the self-timed circuit will miss some input data. Therefore, the allowed maximum input data rate is given by

$$f_{max} = \frac{1}{D_{data} + D_{spacer}} \quad (1)$$

Usually a speed margin is needed to guarantee that the self-timed circuit works correctly.

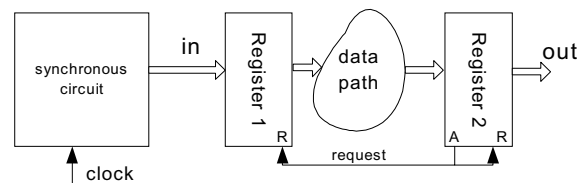


Fig. 2 (a) A self-timed circuit receives data from a synchronous circuit

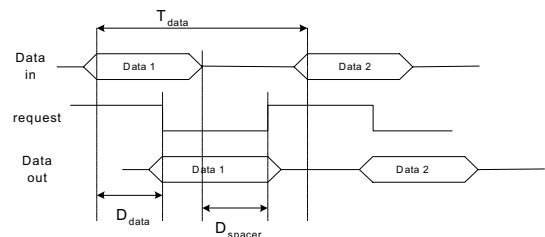


Fig.2 (b) Timing constraint of self-timed circuit

4. Supply voltage scalability design in self-timed circuits

4.1 Adaptive supply-voltage scheme without quality loss

Under the assumption that the data path works faster at a fixed V_{DD} than the speed required by the input data rate, a feedback circuit can be designed to provide the data path with a lower supply voltage, as long as the delay of the data path meets the timing constraint in Fig. 2(b). The feedback circuit is implemented as shown in Fig. 3. It consists of a completion detector, a D-flip-flop and a DC-DC buck converter [14]. The completion detector can be constructed by C-elements. The output of the detector is high when a complete set of spacers arrives. The output is low when a complete set of data arrives. Otherwise, the output doesn't change. The high level output of the D-flip-flop implies that the data path is waiting for data input and that the data path works faster than required, and therefore that the supply voltage is allowed to decrease, and vice versa.

To observe the waveforms during simulation, the data path is implemented by a chain of 26 inverters, which has a significant delay while the circuit is relatively simple for reasonable simulation time. The typical waveforms in the close-loop voltage control are shown in Fig.4. Initially, the maximum voltage V_{DD} is applied to the self-timed data path by set M_1 on and M_2 off. Data and spacers are input to register 1 alternatively at a constant rate below the maximum allowed data rate.

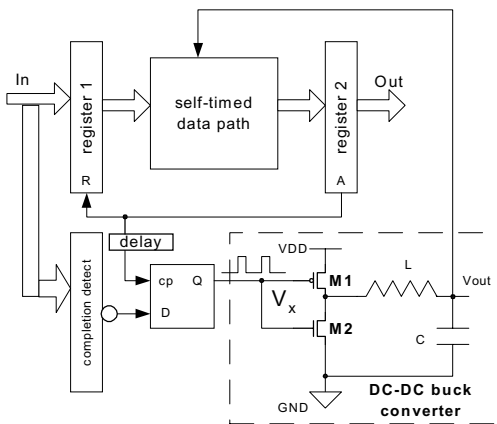


Fig. 3 The proposed adaptive supply-voltage scheme

After one data cycle, the output of the D-flip-flop becomes high until A (in Fig. 4) because the request signal arrives at register 1 earlier than corresponding data does, which means that the data path operates too fast, and that the supply-voltage needs to decrease for saving power. However, V_x becomes low after A, which means that the data path operates too slowly, and that the supply-

voltage needs to increase for speed. Actually, even when V_x is low, V_{out} still decreases from A to B due to the continuity of current in inductor, instead of increasing. To guarantee a safe speed, two methods can be used to compensate for the amount of voltage drop from A to B. One is to design some buffers before register 1 to store some data temporally in case the data path is too slow from A to B. Another one is to put a delay element on the clock input of the D-flip-flop to make the falling edge of V_x occur earlier a little bit, thus to increase V_{out} by a proper amount.

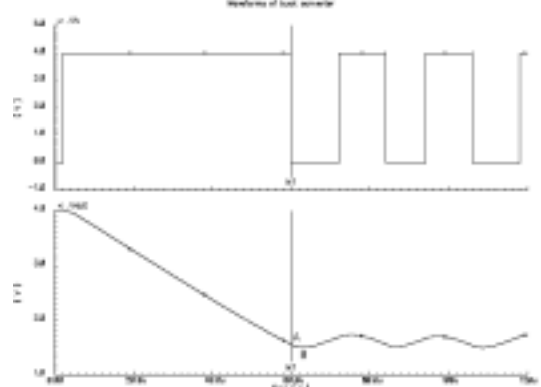


Fig.4 Waveforms V_x and V_{out} from Cadence simulation

When the circuit is stable, the waveform of V_x should be a pulse signal with a duty cycle P associated with an input data rate. The output of the buck converter is a rough DC voltage with a small ripple, and the DC component is given by

$$V_{out} = (1 - P) \cdot V_{DD} \quad (2)$$

The frequency of V_x mainly depends on the input data rate and the delay characterization of data path. To achieve a reasonable ripple on the output voltage, the values of L and C are chosen so that the LC frequency constant $f_0 = \frac{1}{2\pi\sqrt{LC}}$

of V_x . On the other hand, the increase of L and/or C will result in a longer time for the circuit to track the input data rate. Thus, choosing the values of L and C requires making tradeoff between ripple and transient performance.

A 4x4-bit NCL Baugh-Wooley multiplier [15] is designed as the data path to demonstrate the energy saving effectiveness of the proposed scheme. The simulation is based on 0.18 μm CMOS technology. The dependences of supply voltage V_{out} and energy consumed by the multiplier on data rate are plotted in Fig. 5 (a) and Fig. 5 (b) respectively. When the maximal supply voltage $V_{DD} = 3.3\text{V}$ is applied to the data path, the maximal allowed data rate is approximately 180 MHz. The supply voltage will adaptively decrease whenever the input data rate goes down. Fig. 5 (b) shows that significant energy can be saved when the rate of input sample is relatively

small. Note that the energy loss of the DC-DC converter is not included in Fig. 5 (b).

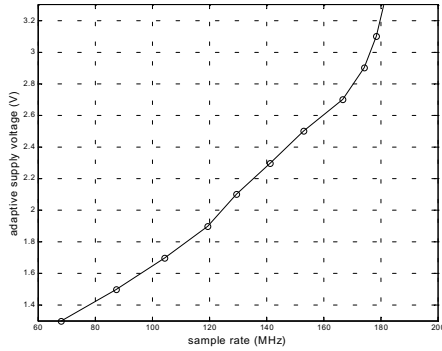


Fig. 5 (a) Adaptive supply voltage vs. data rate

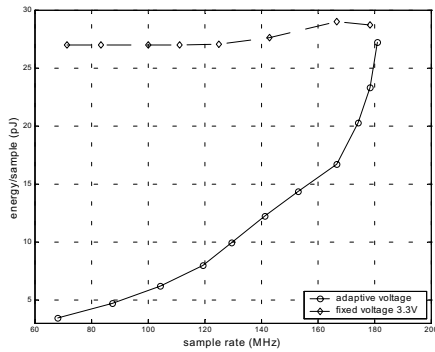


Fig.5 (b) Energy vs. data rate

4.2 Soft DSP applications

A supply voltage is called ultra-low supply voltage (ULSV) if the supply voltage is so small that the circuit can't work fast enough to process the input data stream. When an ULSV is applied to a self-timed circuit, further energy saving can be achieved while some errors are introduced because the constraint defined by (1) is violated. The following analysis shows that under the condition of ULSV the self-timed circuit would miss some input samples (DATA or SPACER), and the outputs corresponding to the inputs not missed are always correct. In other words, an output is either lost or delivered correctly. This ultra-low supply voltage scaling is particularly useful in systems with highly sequential algorithms that perform a large number of computation steps per data sample [8]. In DSP systems, the missing introduces an equivalent SNR loss, leading to a soft DSP.

For the sake of simplicity, we make the following assumptions: 1) The data rate of input is fixed, and the duration of DATA is equal to that of SPACER, i.e., $T_{data} = T_{spacer} = 0.5T_{data+spacer}$. 2) The delay of DATA is the same as the delay of SPACER, i.e., $D_{data} = D_{spacer} = D$. This assumption requires that the rising time of the circuit is equal to its falling time. 3) $D < T_{data+spacer}$ so that no two

consecutive samples are missed. This assumption makes sure of a miss rate no more than 50%.

Figure 6 shows the implementation of soft DSP using self-timed circuits. Since no consecutive DATA samples are missed under assumption (3), one bit flag DATA0 and DATA1 can be attached to two consecutive DATA samples respectively for miss detection. This flag bit passes from the input register to the output one without processing. If two consecutive outputs have the same flag DATA0 (or DATA1), there must be an output missed between these two consecutive outputs. The missed output is estimated by the interpolation based on the outputs delivered by the self-timed circuit. A linear interpolation method is adopted in this paper. The average of two consecutive outputs with the same flag is the estimation of the missed output.

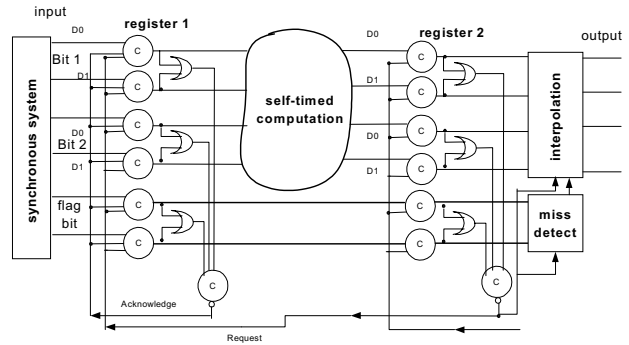


Fig. 6 The proposed architecture for soft DSP (Only two bits in data bus for simplicity)

It can be observed that the effect of the time difference Δt accumulates until a sample (DATA or SPACER) is missed, where

$$\Delta t = D - 0.5T_{data+spacer} \quad (3)$$

Furthermore, let n be defined by

$$n = \left\lfloor \frac{0.5T_{data+spacer}}{\Delta t} \right\rfloor \quad (4)$$

where $\lfloor x \rfloor$ is the floor function of x . Note that $n \geq 1$ due to assumption (3).

If there is a (DATA, SPACER) pair missed after average k pairs of (DATA, SPACER) are delivered at output, then the miss rate of (DATA, SPACER) pair is defined by

$$R_m = \frac{1}{k+1} \quad (5)$$

where k is given by

$$k = \frac{n+1}{2} \quad (6)$$

where n is defined by (4). Note that k is not necessarily an integer. Obviously, the miss rate is a two-dimensional function of input data rate and circuit delay (or supply

voltage). By defining the input data rate f as the reciprocal of the DATA-SPACER cycle $T_{data+spacers}$, replacing D in (3) by $D(V_{dd})$, and combining (3), (4), (5), and (6), the miss rate in (5) can be rewritten as

$$R_m(V_{dd}, f) = \frac{2}{n+3}, \quad n \leq \frac{1}{2f \cdot D(V_{dd}) - 1} < n+1, \quad n = 1, 2, 3, \dots \quad (7)$$

where

$$D(V_{dd}) = \frac{C_L V_{dd}}{\beta(V_{dd} - V_t)^\alpha} \quad (8)$$

C_L is the total node capacitance,
 β is gate transconductance,
 V_t is the device threshold voltage.

The circuit delay is estimated by (8) with a good accuracy [8]. Since a self-timed circuit has a characteristic of average-case delay, instead of worst-case delay, $D(V_{dd})$ in (7) is a average delay in real operation environments.

As an example, the miss rate $R_m(V_{dd}, f)$ for a chain of 8 full adders is plotted in Fig.7, where the plane (V_{dd}, f) is partitioned into different regions, and each region corresponds to a miss rate of (DATA,SPACER) pair. Given an input data rate, the supply voltage can be reduced significantly by allowing a tolerable miss rate. Similarly, given a supply voltage, the maximal input data rate can be increased by allowing a tolerable miss rate. The curve "critical V_{dd} " shows the minimal supply voltage for no-error operation.

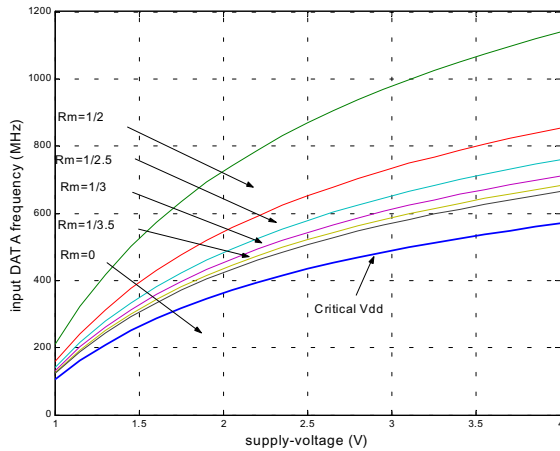


Fig.7 Miss rate as a function of data rate and supply-voltage

A typical speech signal $y(n)$ and its spectrum without missing are plotted in Fig. 8. The output $\hat{y}(n)$ from the interpolation includes the ideal output signal $y(n)$ and error signal $e(n)$, expressed by

$$\hat{y}(n) = y(n) + e(n) \quad (9)$$

The magnitude of interpolation error, normalized to ideal output signal, is defined by

$$M_{error} = 20 \cdot \lg\left(\frac{\sigma_{error}}{\sigma_y}\right) \quad (10)$$

where σ_{error}^2 is the variance of error $e(n)$ and σ_y^2 is the variance of ideal output signal $y(n)$. Fig.9 shows the estimation error versus the reciprocal of miss rate for this speech signal. The error depends on the bandwidth of signal, interpolation method, and miss rate.

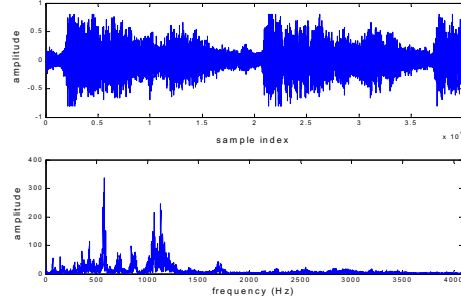


Fig. 8 A typical speech signal and its spectrum

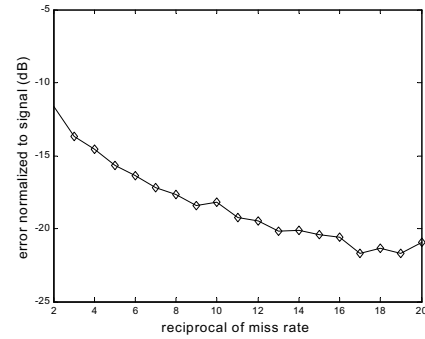


Fig.9 Error versus the reciprocal of miss rate

As a case study, the delay of the 4x4-bit NCL multiplier in section 4.1 is modeled by (8) where $\alpha=1.1967$, and $C_L / \beta = 0.899 \cdot 10^{-9} F \cdot V^2 / A$, $V_t=0.75V$, based on 0.18 μ m CMOS technology. If the above speech signal passes through the multiplier with a ULSV, the following simulation results are obtained. Given an input DATA frequency, the reciprocal of miss rate is plotted in Fig.10 as a function of supply voltage. The reduction in power dissipation is characterized by power savings (PS) defined as

$$PS = \frac{P_{critical} - P_{ULSV}}{P_{critical}} \quad (11)$$

where $P_{critical}$ is the power dissipation with $V_{dd}=V_{critical}$, and P_{ULSV} is the power dissipation with V_{dd} less than $V_{critical}$. Neglecting the power dissipated by the error compensation circuit, the curves of power savings due to ULSV are plotted in Fig.11 for input DATA rate 200 MHz, 400 MHz, 600 MHz respectively. More than 40% to 70 % power can be saved by introducing -15dB to -10 dB error, which is tolerable in many DSP applications.

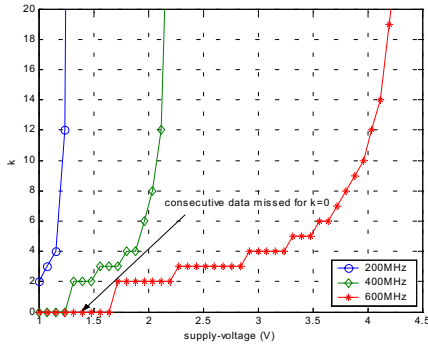


Fig.10 the reciprocal of miss rate versus supply voltage

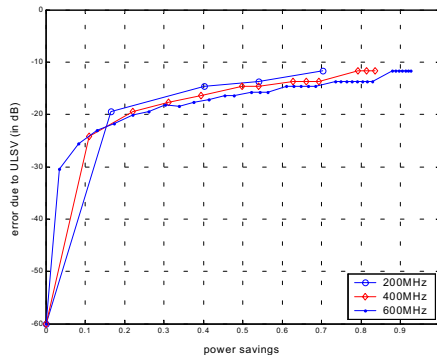


Fig.11 Error versus power savings

5. Summary

Low power operation can be achieved for data path by combining self-timed design with an adaptive supply-voltage scheme. The scheme proposed consists of a simple logic circuit and a buck converter. The handshake signals in self-timed pipeline are employed to track the input data rate automatically, and thus to keep the supply-voltage of the data path as small as possible. The cost of the overhead circuitry for logic control is negligible

We have also proposed an approach to low voltage low power design for DSP applications. This approach exploits the robustness of self-timed circuit to ULSV to achieve significant power saving. The effectiveness of this approach is demonstrated by miss rate analysis and a DSP case study. However, the bandwidth and SNR of input signal limit the accuracy of error correction. On the other hand, the accuracy can be improved by a smaller miss rate and/or an advanced interpolation method such as linear prediction based on multi-samples and data autocorrelation.

6. References

- [1] N. R. Shanbhag and M. Goel, "Low -power adaptive filter architectures and their application to 51.84 mb/s ATM-LAN," *IEEE Trans. Signal Processing*, vol. 45, pp. 1276-1290, May 1997.
- [2] P. E. Landman and J. M. Rabaey, " Architectural power analysis: The dual bit type method," *IEEE Trans. VLSI Syst.*, vol. 3, pp. 173-187, June 1995.
- [3] S. Iman and M. Pedram, "An approach for multilevel logic optimization targeting low power," *IEEE Trans. Comput.-Aided Design*, vol. 15, pp. 889-901, 1996.
- [4] R. K. Krishnamurthy and L. R. Carley, "Exploring the design space of mixed swing quadrail for low power digital circuits," *IEEE Trans. VLSI Syst.*, vol. 5, pp. 388-400, Dec. 1997.
- [5] R. Zhang and K. Roy, "Low-power high-performance double-gate fully depleted SOI circuit design," *IEEE Trans. Electron Devices*, vol. 49, no. 5, pp. 852-862, May 2002.
- [6] A. Sinha, A. Wang, and A. Chandrakasan, "Energy scalable system design," *IEEE Trans. VLSI Syst.*, vol. 10, no. 2, pp. 135-145, April 2002.
- [7] T. Kuroda, et al, "Variable supply-voltage scheme for low-power high-speed CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 33, no. 3, pp. 454-462, March 1998.
- [8] L. S. Nielsen, C. Niessen, J. Sparso, and K.V. Berkel, "Low-power operation using self-timed circuits and adaptive scaling of the supply voltage," *IEEE Trans. VLSI Syst.*, vol. 2, no.4, pp. 391-397, Dec. 1994.
- [9] R. Hegde and N. R. Shanbhag, "Soft digital signal processing," *IEEE Trans. VLSI Syst.*, vol. 9 no. 6, pp. 813-823, Dec. 2001.
- [10] C. Mead, L. Conway, *Introduction to VLSI systems*, Addison-Wesley Publishing Company, 1980.
- [11] P. Ng, P. T. Balsara, and D. Steiss, " Performance of CMOS differential circuits," *IEEE J. Solid-State Circuits*, vol. 31, no. 6, pp. 841-846, June 1996.
- [12] K. M. Fant and S. A. Brandt, "NULL Convention Logic: a complete and consistent logic for asynchronous digital circuit synthesis," *Proc. of international conf. on application specific systems, architecture, and processors*, pp. 261-273, 1996.
- [13] A. J. Martin, "Asynchronous datapaths and the design of an asynchronous adder," *Formal Methods in System Design*, vol.1, no.1, pp. 119-137, July 1992.
- [14] A. Stratakos, et al, "A low-voltage CMOS DC-DC converter for a portable battery-operated system," *IEEE Power Electronics Specialists Conference*, pp. 619- 626, 1994.
- [15] S. C. Smith, R. F. Demara, J. S. Yuan, M. Hagedorn, D. Ferguson, "Delay-insensitive gate-level pipelining," *Integration, the VLSI Journal*, 2001, 30, (2), pp. 103-131.