# A Fundamental Basis for Power-Reduction in VLSI Circuits

Naresh R. Shanbhag

Coordinated Science Laboratory/ECE Dept.
University of Illinois at Urbana-Champaign,
Urbana, IL 61801. *shanbhag@uivlsi.csl.uiuc.edu*

### Abstract

Presented in this paper is a fundamental mathematical basis for power-reduction in VLSI systems. This basis is employed to 1.) derive lower bounds on the power dissipation in digital systems and 2.) unify existing power-reduction techniques under a common framework. The proposed basis is derived from information-theoretic arguments. In particular, a digital signal processing algorithm is viewed as a process of information transfer with an inherent information transfer rate requirement of $R$ bits/sec. Different architectures implementing a given algorithm are equivalent to different communication networks each with a certain capacity $C$ (also in bits/sec). The absolute lower bound on the power dissipation for any given architecture is then obtained by *minimizing* the signal power such that its channel capacity $C$ is equal to the desired information transfer rate $R$. Numerical calculations for a simple static CMOS circuit and fundamental basis for the power-reduction capabilities of parallel processing and pipelining are presented.

## 1 Introduction

Numerous applications in the area of signal processing and communications have emerged in recent years, which require an implementation of highly complex algorithms along with stringent requirements on the power dissipation. These applications include signal compression, mass data storage, high bit-rate digital subscriber loops etc.. Therefore, development of power-reduction techniques is currently of great interest. Numerous power-reduction techniques [1] have been presented in the past. Some work has been done in determining the lower bounds on the achievable power dissipation [2-3].

In this paper, we employ an information-theoretic approach to develop a mathematical basis for power-reduction in VLSI systems. The proposed basis has two advantages: 1.) it allows us to derive lower bounds on the power dissipation in digital systems and 2.) it enables us to unify existing power-reduction techniques under a common framework. The utility of the proposed theory is demonstrated via numerical calculations for a simple static CMOS circuit. Furthermore, we also analyze the power-reduction capabilities of parallel processing and pipelining techniques. Thus, the proposed theory can be universally applied at different levels of the design hierarchy.

## 2 Preliminaries

In order to provide the necessary background, we will review some basic information-theoretic concepts in this section.

### 2.1 Entropy and Mutual Information

Consider a discrete source generating symbols from the set $S_X = X_0, X_1, \ldots X_{L-1}$ according to a probability distribution $Pr(X)$. A measure of the information content of this source is given by its *entropy* $H(X)$, which is defined as follows

$$H(X) = -\sum_{i=0}^{L-1} P_i log_2(P_i), \tag{2.1}$$

where $P_i \stackrel{\text{def}}{=} Pr(X = X_i)$ for $i = 0, \ldots, L-1$ and $H(X)$ is in bits. The *mutual information* $I(X;Y)$ between the input $X$ and the output $Y$ (of a channel) is defined as

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X), \tag{2.2}$$

where $H(X|Y)$ is the *conditional entropy* of $X$ conditioned on $Y$. The mutual information $I(X;Y)$ can be viewed as the *reduction in uncertainty* in $X$ due to the knowledge of $Y$.

### 2.2 Information Transfer Rate

The reduction in uncertainty (by an amount $I(X;Y)$) in $X$ is due to the information transferred from the input of the channel to its output. Thus, the *information transfer rate* $R$ is defined as

$$R = f_{op} I(X;Y), \tag{2.3}$$

where $f_{op}$ is the rate at which the symbols are generated by the source.

0-7803-3073-0/96/$5.00 

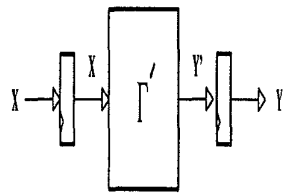


Figure 1: A noisy transformation.

### 2.3 Channel Capacity

In his seminal work [4], Shannon showed that the capacity ($C$) of a channel band-limited to frequency $W$, is given by

$$C = \int_0^W log_2[1 + SNR(f)]df, \qquad (2.4)$$

where $C$ is in bps. From (2.4), it is clear that the capacity $C$ depends upon the $SNR$ and the transmission bandwidth $W$.

## 3 A Fundamental Basis for Power-Reduction

In this section, we will first show that all logic transformations have a inherent information transfer rate requirements $R$ associated with them. Next, we will present a theorem which allows one to determine the lower bounds on the power dissipation for a given architectural implementation.

### 3.1 Logic Transformation

We can represent any noisy logic transformation as shown in Fig. 1, where noise could have many sources including the implementation media itself. Without any loss of generality, we assume that the inputs and the outputs are latched synchronously.

The definition of $\Gamma'$ is shown in Fig. 2, where the input space $S_X$ is mapped onto the output space $S_{Y'}$. The dark dots in the set $S_X$ represent the discrete values that the input $X$ can assume, while the ones in the set $S_Y$ denote the values that the output can assume if the noise power were zero.

Assuming that the noise probability density function is identical for all possible noiseless outputs $Y$, then we can represent the system in Fig. 1 as shown in Fig. 3 with the corresponding mapping for $\Gamma$ as shown in Fig. 4. In this figure, all the noise in $\Gamma'$ has been referred to the output and we now have a noiseless transformation $\Gamma$ mapping the input space $S_X$ to a noiseless output space $S_Y$. In this paper, we will assume that the Fig. 3 can be employed to represent any noisy transformation $\Gamma'$. Note that the computation model in Fig. 3 is quite similar to a generic digital communications system. The input and output latches can be viewed as transceivers (transmitter-receiver) and the clock as providing the correct sampling epoch (output of a timing-recovery block). In a digital communications system, the transformation $\Gamma$ is usually an identity transformation.

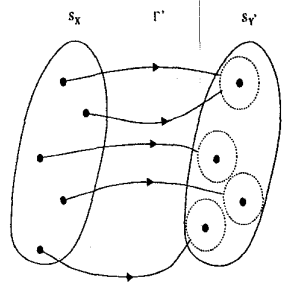


Figure 2: Input-output mapping for $\Gamma'$ in Figure 1.

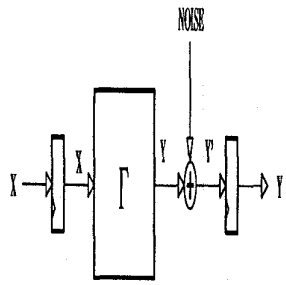


Figure 3: Another representation of a noisy transformation with noise referred to the output.

*Theorem 1: If* $\Gamma: \mathbf{Z}^L \rightarrow \mathbf{Z}^M$ *is a deterministic mapping, where* $\mathbf{Z}$ *is the set of integers, then the information transfer rate is given by*

$$R = f_{op} H(Y). \qquad (3.1)$$

Theorem 1 can be easily proved by observing that in a noiseless case $I(X;Y') = I(X;Y) = H(Y)$ and substituting the result in (2.3). As an example, for a two-input AND gate with equiprobable inputs, (2.2) implies that $I(X;Y) = 0.8112$ bits. If such a system operates at 100 Mhz then $R = 81.12$ Mb/s.

Thus, all digital transformations, in particular linear finite-precision digital signal processing have an inherent information transfer rate requirement $R$ given by (3.1). This requirement is an inherent property of

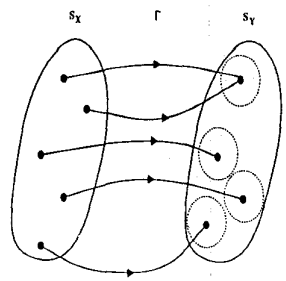


Figure 4: Input-output mapping for $\Gamma$ in Figure 3.

the transformation and is *independent of the implementation media or the architecture.*

### 3.2 Lower Bounds on Power Dissipation

It is well-known that there can be many different digital architectures which achieve the same functionality. In the present context, we view each of these architectures as a communication network with a certain capacity $C$. Clearly, from [4], the capacity $C$ should be greater than or equal to $R$ for a meaningful computation to take place. This result can be exploited formally to present the following theorem.

*Theorem 2: For a given channel with the following properties*

- *bandlimited to $W$ Hz.*
- *with a noise power spectral density of $S_{NN}(f)$.*
- *a desired information transfer rate $R$ bps.*
- *power dissipation at frequency $f$ ($P_D(f)$) being related to the signal power spectrum $S_{XX}(f)$ as*

$$P_D(f) = F[S_{XX}(f)], \tag{3.2}$$

*where $F()$ is a linear monotonically increasing function of its argument.*

*The lower bound on the power dissipation for such a channel is given by*

$$P_D > F[\int_0^W (\nu_{min} - S_{NN}(f))^+ df], \tag{3.3}$$

*where $(arg)^+$ denotes the positive part of arg and $\nu_{min}$ is a unique constant which can be obtained as a solution to the following equation*

$$R = \int_0^W log_2[1 + \frac{(\nu_{min} - S_{NN}(f))^+}{S_{NN}(f)}]df. \tag{3.4}$$

The proof of this Theorem 2 follows along similar lines as that of [4] and is omitted here for the sake of brevity.

Thus, for a given signal processing transformation $\Gamma$, Theorem 1 allows us to calculate the information transfer rate $R$ and Theorem 2 enables us to determine the lower bound on the power dissipation of *a particular architectural implementation of* $\Gamma$. It should be clear that the results obtained from the application of Theorem 1 and Theorem 2 would be a function of the algorithm ($\Gamma$), the technology ($S_{NN}(f)$ and $F(.)$) and the architecture ($C$).

Note that Theorem 2 does not provide us with the technique to achieve the lower bound. This is not surprising given the fact that Theorem 2 is derived from Shannon's joint source-channel coding theorem [4], which in turn provides a proof of achievability but not the method.

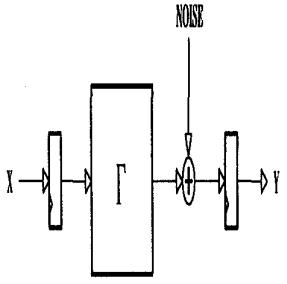


Figure 5: Serial architecture.

## 4 Lower Bound Calculations

In this section, we present a two examples of the application of Theorem 1 and Theorem 2. In the first example, we compare the power-reduction capabilities of parallel and pipelined architectures. In the second example, we determine numerically the lower bounds on power dissipation in a simple static CMOS circuit. The total noise power in digital systems is a function of signal power, temperature, semiconductor properties, frequency of operation, etc.. However, for conventional digital systems the noise is mainly due to the phenomenon of ground bounce. Just for the sake of demonstration, and without any loss of generality, we assume that the implementation technology is CMOS with a flat noise spectrum with average power $\sigma_N^2 = 10^{-2}$ V$^2$ over a usable bandwidth of $W = 100$ MHz. The bandwidth is a function of the parasitic capacitances and resistances in the critical path of the architecture under consideration. The noise power is consistent with the value of ground bounce noise in a typical sub-micron CMOS technology. Furthermore, without any loss of generality, we assume that a $V_{dd}$ pulse corresponds to a '1' and a zero pulse corresponds to a '0'. The signal power $\sigma_X^2$ (or the variance) is therefore equal to $V_{dd}^2/4$. The function $F$ for static CMOS is defined as follows

$$F = C_L V_{dd}^2 2W, \tag{4.1}$$

where we assume a maximum of $2W$ channel uses per second.

### 4.1 Parallel and Pipelined Processing

Let $R$ be the required information transfer rate (for all architectures). Furthermore, let $W$ be the channel bandwidth, $\sigma_N^2$ be the noise power and $C$ be the channel capacity for the serial architecture in Fig. 5.

The lower bound on the supply voltage for a serial architecture $V_{dd1,min,ser}$ is calculated as follows

$$C = R$$

$$W log_2[1 + \frac{V_{dd1,min,ser}^2}{4\sigma_N^2}] = R$$

$$V_{dd1,min,ser} = [(2^{\frac{R}{W}} - 1)4\sigma_N^2]^{\frac{1}{2}}, \tag{4 . 2}$$

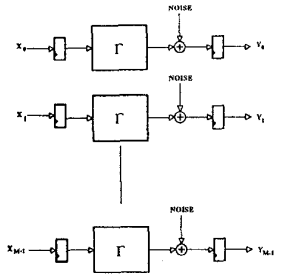

Figure 6: Parallel architecture.

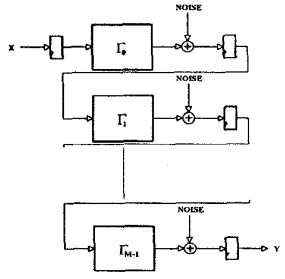

Figure 7: Pipelined architecture.

Substituting (4.2) into (4.1) gives us the strict lower bound on $P_{D1}$ for a serial architecture $P_{D1,min,ser}$ as shown below

$$P_{D1,min,ser} = C_L(2^{\frac{R}{W}} - 1)8\sigma_N^2 W. \qquad (4.3)$$

For the parallel architecture, we assume that each channel in Fig. 6 is identical to the serial channel in Fig. 5. Proceeding in an identical fashion as before, we get

$$V_{dd1,min,par} = [(2^{\frac{R}{MW}} - 1)4\sigma_N^2]^{\frac{1}{2}}. \qquad (4.4)$$

The lower bound on $P_{D1}$ is given by

$$P_{D1,min,par} = C_L(2^{\frac{R}{MW}} - 1)8\sigma_N^2 MW. \qquad (4.5)$$

The pipelined architecture in Fig. 7 achieves power reduction in a slightly different manner than a parallel architecture. Pipelining by $M$-levels results in an $M$-fold expansion of the transmission bandwidth per stage due to the reduction in the critical path length. Assuming identical pipeline stages, the noise power per stage over a bandwidth of $W$ is $\sigma_N^2/M$. This implies that the noise power per stage over a bandwidth of $MW$ is equal to the noise power of the serial architecture ($\sigma_N^2$). Hence, the lower bound on the $V_{dd}$ for the pipelined architecture is also given by (4.4). Note that all the stages in a pipelined architecture need to transmit $R$ bits/s of information. Assuming a maximum usage of $2MW$ uses/s, the lower bound on the $P_{D1}$ is

$$P_{D1,min,pip} = C_L(2^{\frac{R}{MW}} - 1)8\sigma_N^2 M^2 W. \qquad (4.6)$$

From (4.3) and (4.5), it can be shown that the parallel architecture will always have a smaller lower bound than the serial architecture for $M > 1$. Furthermore, for $M = 1$, the lower bounds for the serial and parallel architectures are identical. Interestingly, a comparison between (4.5) and (4.6) indicates that $P_{D1,min,par}$ is lower by a factor of $M$ as compared to $P_{D1,min,pip}$. This is counter to the well-prevalent notion that parallelization and pipelining have identical power-reduction capabilities. However, both architectures are equivalent if the *Area* $\times$ *Power* product is considered. This is due to the fact that a $M$-level parallel architecture requires $M$-times the area of a serial architecture. On the other hand, the area requirements of a pipelined architecture is of the same order as that of a serial architecture.

## 4.2 A Single Bit Two-Latch System

Consider a simple CMOS digital system where the output of a 1-bit latch is connected to the input of another. We consider $H(X) = H(Y) = 1$ bit with $f_{op} = 100$ Mhz. From Theorem 1, we get the desired information transfer rate as $R = 100$ Mb/s.

The lower bounds on the supply voltage and the power dissipation for the single bit case can be obtained by substituting $R = 100$ Mb/s, $W = 100$ MHz, $\sigma_N^2 = 10^{-2}$ V$^2$, and $C_L = 0.5$ pF into (4.2) and (4.3), respectively. These lower bounds are

$$V_{dd} > 0.2 \quad \text{V} \qquad (4.7)$$

$$P_D > 0.004 \quad \text{mW}. \qquad (4.8)$$

Note that the value of $V_{dd}$ obtained via the proposed theory is in the same range as the 20 MHz encoder-decoder circuit in [5], where $V_{dd} = 0.2$ V.

## References


[1] A. P. Chandrakasan, and R. W. Brodersen, "Minimizing power consumption in digital CMOS circuits", *Proceedings of IEEE*, vol. 83, no. 4, pp. 498-523, April 1995.

[2] E. A. Vittoz, "Low-power design: Ways to approach the limits", *ISSCC '94*, pp. 14-18, San Fransisco, CA.

[3] J. D. Meindl, "Low power microelectronics: Retrospect and prospect", *Proceedings of IEEE*, vol. 83, no. 4, pp. 619-635, April 1995.

[4] C. E. Shannon, "A mathematical theory of communications," *Bell System Technical Journal*, vol. 27, Part I, pp. 379-423, Part II, pp. 623-656.

[5] J. B. Burr and J. Schott, "A 200mv self-test encoder/decoder using Stanford ultra-low power CMOS", *ISSCC '94*, pp. 84-85, San Fransisco, CA.