COMPUTATION AS ESTIMATION: ESTIMATION-THEORETIC IC DESIGN IMPROVES ROBUSTNESS AND REDUCES POWER CONSUMPTION

Sriram Narayanan, Girish V. Varatkar, Douglas L. Jones and Naresh R. Shanbhag

Department of Electrical and Computer Engineering, UIUC
1308 West Main St., Urbana, IL 61801

ABSTRACT

Modern Integrated Circuits (ICs) are designed as massively parallel systems as a consequence of diminishing silicon feature sizes. This has adversely impacted reliability because of increased errors due to process and environmental variations, and particle hits. Viewing hardware errors as analogous to measurement or system noise allows us to borrow results from estimation theory and extend Moore's law. The estimation-theoretic framework provides a design optimization formalization that enables power/reliability trade-off in broad classes of applications. Two applications described here show that specific instantiations of the framework yield significant power savings and system reliability.

Index Terms— Estimation-theoretic computation, Sensor Networks-on-Chip, Low-power IC design, Robust gigascale systems

1. INTRODUCTION

Driven by Moore's law, technology scaling has continued to increase the density of IC systems. Complex ICs are now designed as systems comprising many smaller subcomponents. One may view circuit elements as collaborating nodes that collectively produce the final output expected from the system. The notion of on-chip networks is becoming increasingly commonplace in the literature [1].

As a side-effect of diminishing dimensions, lower supply voltages, and increased clock frequencies, modern ICs are plagued by numerous nonidealities [2]. Nanometer ICs are affected by variations in the manufacturing process. Soft errors caused by neutron and alpha particle hits adversely impact reliability. Designing for the worst-case, typically through overprovisioning supply voltage or hardware, may avoid these errors but is often unacceptable due to increased system power consumption and area overhead.

This research is supported by the Gigascale System Research Center (GSRC), one of five research centers funded under the Focus Center Research Program (FCRP), a Semiconductor Research Corporation program, and Texas Instruments, Inc. Special thanks to Chhay Kong for hardware characterizations.

The reliability issue in ICs has been traditionally addressed by introducing hardware redundancy (*e.g.*, N-modular redundancy), which adds to the system cost, power, and area. More recently proposed Algorithmic Noise Tolerance (ANT) techniques [3] use a lower-complexity estimator to recover from any errors made by the main DSP block. Recent research efforts such as the Better Than Worst-Case Design [4] also attempt to avoid power over-provisioning. Here, the main block is aggressively optimized for power or performance and is paired with a checker block that detects and corrects any faults made by the main block. In allowing hardware errors, these approaches make a power/reliability trade-off that is fundamental to the extension of Moore's law.

The problem of maintaining system robustness in the presence of subcomponent failures occurs in other contexts. Sensor networks have traditionally been used to achieve robust estimation of physical phenomena even when some of the constituent nodes may fail. By allowing node-level failures, sensor networks have proven to be energy-efficient [5]. Estimation algorithms that tolerate noisy measurements at the sensor nodes are the enabling technology for sensor networks. Because modern IC design also strives to achieve robustness and energy-efficiency using unreliable hardware, treating errors as computational noise may enable us to borrow results from estimation theory.

Identifying estimation problems in computation will allow an optimal power-reliability trade-off that is essential for sustaining IC system development. The proposed estimation-theoretic framework provides the necessary formalization for practical design optimization.

2. ESTIMATION THEORY FOR COMPUTATION

When viewing computation as a special case of estimation, hardware errors due to process variations, voltage or frequency scaling, and soft errors represent measurement or system noise. This analogy enables us to leverage many already known algorithms in estimation theory. But designing IC systems presents a set of unique issues that need to be addressed.

Cost and technology considerations may limit the IC system designer to a finite number of architectures. For a given architecture, the choices for the various operating parameters

such as supply voltage, clock frequency, and register word-lengths may also be constrained. Because these parameters have a direct impact on both the system power consumption and hardware error rate, it is important to optimally choose them. Therefore, we need a general estimation-theoretic framework that optimizes system performance or power consumption while accounting for design constraints.

Example. *The application developer has available a system consisting of a set of N identical processing elements. (This may be a common scenario since such highly parallel designs have recently become rather inexpensive.) The different processing elements may be operated at supply voltages chosen from a set $V = \{V_i\}$, at clock frequencies chosen from the set $F = \{f_i\}$, the word-length of the registers may be chosen from $H = \mathbb{N}$. The set of architectural choices, $\mathbb{A}$, is then given by $\mathbb{A} = V \times F \times H$.*

The estimation-theoretic computational system design optimization may be stated in two canonical problems. The first problem seeks to minimize power consumed in arriving at a computational result (analogous to estimate), while constraining the average system accuracy (analogous to average risk). Let θ be the result being computed, and let $\hat{\theta}$ be the estimator that operates on the input set, Y. We can state this problem as follows:

$$\begin{aligned} &\hat{\theta}(Y) = \arg\left\{ \min_{\theta\in\Lambda, A\in\mathbb{A}} Power(\hat{\theta}(Y)) \right\} \\ &\text{subject to} \\ &E_\theta\{C[\hat{\theta}(Y),\theta]\} \le C_{Target} \\ &\mathbb{A} = \{A_1(\bar{\lambda}), A_2(\bar{\lambda}), \ldots\} \end{aligned} \tag{1}$$

where $\mathbb{A}$ is the set of architectural choices, and the vector $\bar{\lambda}$ defines the parameters of an architecture (*e.g.*, supply voltage, clock frequency).

Different computation problems may call for different risk functions. Squared-error risk functions may be appropriate for signal processing applications in which it is common to minimize mean squared error (MSE), $C[\hat{\theta},\theta] = (\hat{\theta}-\theta)^2$. General-purpose computing systems may follow a model in which only errors up to some limit, Δ, may be tolerated. For such systems, we suggest the following risk function:

$$C[\hat{\theta},\theta] = \begin{cases} 1, & |\hat{\theta}-\theta| \ge \Delta \\ 0, & |\hat{\theta}-\theta| < \Delta \end{cases},$$

Another example of a risk function is the absolute error, $C[\hat{\theta},\theta] = |\hat{\theta}-\theta|$.

Problem (2) seeks to minimize the average risk incurred in misestimating the result while constraining the power consumed to be within budget. For a battery-operated system, the amount of stored energy may be used to arrive at a power budget. Cooling and packaging costs may define this budget for mains-powered systems.

$$\begin{aligned} &\hat{\theta}(Y) = \arg\left\{ \min_{\theta\in\Lambda, A\in\mathbb{A}} E_\theta\{C[\hat{\theta}(y),\theta]\} \right\} \\ &\text{subject to} \\ &Power(\hat{\theta}(Y)) \le P_{Budget} \\ &\mathbb{A} = \{A_1(\bar{\lambda}), A_2(\bar{\lambda}), \ldots\} \end{aligned} \tag{2}$$

3. APPLICATIONS OF THE ESTIMATION-THEORETIC FRAMEWORK

The canonical problems of the estimation-theoretic framework are very general and different applications may lend themselves to very different specific estimation problems. The value of the abstraction presented through the canonical problems is that they allow the system designer to quickly recognize relevant estimation theory results to solve many important VLSI design problems. We show two specific problems that illustrate the usefulness of the estimation problems resulting from the general framework.

3.1. FIR Filtering

FIR filters arise in many circuit designs and often consume the majority of chip area and power. They are commonly implemented in a parallel fashion using a polyphase decomposition. The outputs of the filter banks are summed to yield the desired result. For sufficiently low-frequency inputs, the outputs of these parallel filter banks serve as *estimates* of the mean of the overall computation. By reducing the supply voltage of the filter banks to subcritical levels, we can expect squared gains in the overall system power consumption. But these savings often come at the cost of increased errors because the resulting slower computations may not always meet the timing requirements of the system clock. Since computations are typically performed in an LSB-first fashion, Voltage Overscaling (VOS) errors tend to be large in magnitude. The output of the estimators is therefore contaminated by a mixture of the noise already present in the input and the large-magnitude VOS errors. This noise can be modeled as random variables drawn from a class of distributions that is Gaussian with probability $(1-\epsilon)$ and some unknown distribution with probability ϵ for some $0 < \epsilon < 1$, as shown below

$$P_\epsilon = \{F | F = (1-\epsilon)\Phi + \epsilon H, H \in \mathcal{R}\} \tag{3}$$

where Φ is the class of standard normal distributions and H is the class of arbitrary densities with zero mean and finite but unbounded variance. Studies performed in [6] link the voltage overscaling factor, K_{VOS}, and ϵ, the probability of hardware error. Because such hardware errors are an artifact of voltage scaling and the input may be modeled as a uniform random process, we assume that they are uncorrelated with the input.

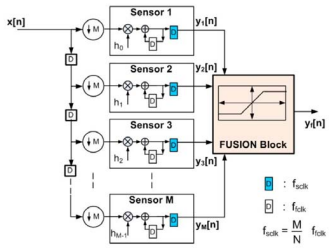

Fig. 1. Polyphase decomposition of the matched filter yields multiple statistically similar *estimates* which are *fused* to obtain the final result.

3.1.1. Robust Statistics

An inference method is said to be *robust* if it exhibits optimal or near-optimal performance when the assumed model is correct, the performance is only slightly worse when the deviation is mild, and large deviations from the assumed model do not cause drastic performance losses [7].

We seek an estimator that minimizes the worst-case estimate variance for errors drawn from probability distributions belonging to P_ϵ. The robust estimate, θ, is the solution to the following equation [7]:

$$\sum_{k=1}^{n} \psi[Y_k - \theta] = 0 \tag{4}$$

where ψ is a general odd-symmetric function known as the *influence function*, and Y_k are the measurements. To obtain the worst-case maximum likelihood estimate, we set f to be the probability distribution function of the least-informative distribution from the class defined in Eq. (3) and $\psi(x) = -f^{'}(x)/f(x)$. For the case of ϵ-contaminated $\mathcal{N}(0,1)$ distributions, the influence function, ψ, is given by

$$\psi(x) = \begin{cases} x, & \text{if } |x| \leq k \\ k \text{ sgn}(x), & \text{else.} \end{cases}$$

where k is a constant that depends only on ϵ and the nominal distribution, $\mathcal{N}(0,1)$ [7].

The computation problem in this application corresponds to Problem (2) of the estimation-theoretic framework. We seek to minimize the maximum asymptotic variance of the estimate while constraining the power consumption of the robust estimator to be some value less than that of a traditional FIR filter. The architectural space consists of N identical processing elements that may be operated at a continuously varying range of supply voltages.

3.1.2. PN Code Acquisition and Simulation Results

Spread spectrum communication systems commonly use pseudo-noise (PN) codes for user identification. The receiver correlates the noisy received signal with a local code and this output is then processed by a detector [8]. Matched filters with a locally generated code sequence as tap weights are commonly used for this purpose. The peaks in the output of the matched filters are used for detection and synchronization of PN sequences. This code acquisition is the computationally critical block in a spread spectrum communication receiver [8].

Prior work in [9] used an HDL simulation to show that the robust PN acquisition system can reduce power consumption by up to 36% when the median is used to approximate the optimal estimator. Figure 2 shows the power savings at different levels of voltage overscaling for a detector operating at a constant false-alarm rate of 5%.

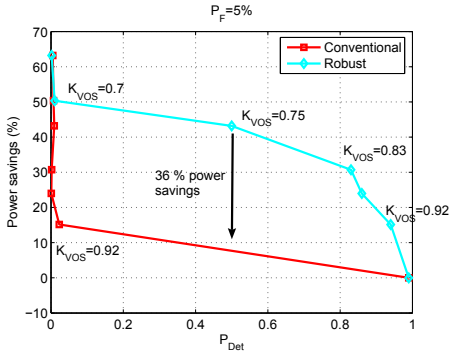

Fig. 2. Robust implementation offers marked reduction in power consumption at different levels of voltage scaling.

3.2. Word-length Optimized ANT Systems

Traditional Algorithmic Noise Tolerance (ANT) techniques have proven to be effective in improving system robustness and lowering power consumption. The Reduced Precision Redundancy (RPR) ANT [3] system pairs a main DSP block with a lower-precision estimator block. While voltage overscaling may cause the main DSP block to make sporadic errors, the lower-precision block remains immune to such errors (since VOS errors occur in the MSB of the result), thereby offering system power savings. The motivating idea behind this technique is that by reducing word-length of the registers, one may save storage power, but at the cost of lowered performance in terms of Signal-to-Quantization-Noise-Ratio (SQNR).

The RPR system uses the fact that the bound on system SQNR depends only on the word-length of the estimator to search over possible choices for its word-length. Using the estimation-theoretic framework, we can further reduce power consumption by jointly optimizing over the word-lengths of both the main block and the estimator while meeting application SQNR performance. Let the word-length of the main block and estimator block be B_1 and B_2 respectively. This

requires storage power of POW_{main} and POW_{RPR}, respectively. This problem may be formalized as follows:

$$[B_1, B_2, V_{DD,main}, V_{DD,RPR}] = \arg\{\min POW_{main}(B_1, V_{DD,main}) + POW_{RPR}(B_2, V_{DD,RPR})\}$$

subject to

$$P_{TE}(B_1, V_{DD,main}) SQNR_{RPR}(B_2) + \{1 - P_{TE}(B_1, V_{DD,main})\} SQNR_{main}(B_1) \geq SQNR_{avg}$$
$$V_{DD,main} \geq V_{DD,RPR} \quad \& \quad B_2 \leq B_1$$

where $V_{DD,main}$ and $V_{DD,RPR}$ are the supply voltages of the main block and the estimator block, respectively.

Simulation in SPICE and VERILOG characterized the delay for full adders ranging from 4 to 16 bits wide at different supply-voltage settings using the IBM 130nm CMOS transistor model [10]. The above optimization problem was solved using a brute-force search. The running time of this search is not especially consequential since it is run once at design time.

SQNR	Main		Estimator		Power
(dB)	$V_{DD}(V)$	WL	$V_{DD}(V)$	WL	(μW)
30	0.75	5	0.75	4	62.50
50	0.75	9	0.75	4	90.33
70	0.75	13	0.75	4	118.17
90	0.80	16	0.75	4	156.14
95	0.90	16	0.75	6	205.13

Table 1. Given a specified SQNR requirement, an estimation-theoretic design avoids overprovisioning word-length and voltage in an ANT system.

Table 1 shows the optimal choice of word lengths for the main block and the RPR blocks along with their supply voltages for various values of the system SQNR. If an application's SQNR requirement is known, a design based on this optimization avoids overprovisioning of word-length and supply voltage.

4. CONCLUSIONS

In accordance with Moore's law, technology trends have resulted in extremely dense ICs. Consequently, it is common to design modern ICs as systems-on-chip. A side effect of technology scaling is increased hardware unreliability due to process variations, soft errors and voltage scaling. Often these errors may be avoided by designing with sufficient margins, but the power overhead of such worst-case design is typically unacceptable. This requires the system designer to make a power/reliability trade-off. System-level approaches that allow subcomponent unreliability may be a key to continued technology scaling.

Estimation theory has offered numerous algorithms to deal with noisy data in many signal processing applications. Treating hardware errors as a new source of noise allows us to borrow many ideas from estimation theory to develop next-generation IC systems. However, a set of engineering constraints unique to IC system design need to be accommodated. The estimation-theoretic framework presented in this work captures these constraints and allows us to turn design problems into an optimization formalization. In the first canonical problem, we seek to minimize power consumption while meeting any system accuracy specifications. The second problem aims at achieving the best system accuracy while operating under a rigid power budget. Stating design problems in the language of this framework helps us recognize classes of applications and points us to already existing solutions to established estimation problems.

Two specific applications, namely, FIR filtering and word-length-optimized ANT systems, were shown to gain substantial power savings and system robustness when viewed under the proposed framework. A robust implementation of PN code acquisition yielded around 36% savings in power consumption. By optimizing the word-lengths in a traditional ANT system, one may tailor a system that exactly meets requirements without overprovisioning hardware resources.

5. REFERENCES

[1] D. Bertozzi and L. Benini, "Xpipes: a network-on-chip architecture for gigascale systems-on-chip," *IEEE Circ. and Sys. Magazine*, vol. 4, no. 2, pp. 18–31, 2004.

[2] C. Constantinescu, "Trends and challenges in VLSI circuit reliability," *IEEE Micro*, vol. 23, no. 4, July 2003.

[3] B. Shim et al., "Reliable low-power digital signal processing via reduced precision redundancy," *IEEE Transactions on VLSI Systems*, vol. 12, no. 5, May 2004.

[4] Todd Austin et al., "Opportunities and challenges for better than worst-case design," in *Asian South Pacific Design Automation Conference*, January 2005.

[5] S.C. Zhang et al., "Feasibility analysis of stochastic sensor networks," in *IEEE SECON*, 2004.

[6] Byongho Shim, *Error-Tolerant Digital Signal Processing*, Ph.D. thesis, UIUC, 2005.

[7] P. Huber, *Robust Statistics*, John Wiley & Sons, 1981.

[8] David Tse and Pramod Viswanath, *Fundamentals of Wireless Communication*, Cambridge, 2005.

[9] G. Varatkar, S. Narayanan, N.R. Shanbhag, and D.L. Jones, "Sensor network-on-chip," in *International Symposium on SoC*, Nov. 2007, to appear.

[10] "IBM Process Design Manual," May 2004.