Trends in Energy-Efficiency and Robustness Using Stochastic Sensor Network-on-a-Chip

Girish V. Varatkar, Sriram Narayanan, Naresh R. Shanbhag and Douglas L. Jones
Coordinated Science Laboratory/ECE Department, University of Illinois at Urbana-Champaign
1308 W Main St., Urbana, IL, USA, 61801.
[varatkar, spnaraya, shanbhag, dl-jones]@uiuc.edu

ABSTRACT

The stochastic sensor network-on-chip (SSNOC) was recently proposed as an effective computational paradigm for jointly achieving energy-efficiency and robustness in nanoscale processes. In this paper, we study the trends in energy-efficiency and robustness exhibited by an SSNOC architecture as the feature size scales from 130nm to 32nm for a PN-code acquisition application. The conventional architecture exhibits a 3 orders-of-magnitude loss in detection probability P_{det} due to process variations in the 130nm and smaller technology nodes. At the 130nm and 90nm nodes, the proposed SSNOC architecture recovers from this performance loss, and exhibits a 2 orders-of-magnitude smaller variation in P_{det} compared to the conventional architecture. However, for the 65nm and 45nm technology nodes, the SSNOC architecture with assistance from circuit level techniques such as adaptive body bias (ABB) and adaptive supply voltage (ASV) shows a 2-3 order-of-magnitude better detection performance. In addition, the SSNOC architecture with ABB/ASV achieves 22% to 31% energy savings. For the 32nm node, the current version of SSNOC with ABB/ASV is not robust enough and thus motivates the need to explore even more powerful versions of SSNOC.

Categories and Subject Descriptors: B.7.3 Reliability and Testing: Redundant design.

General Terms: Reliability.

Keywords: sensor network-on-chip, robust design.

1. INTRODUCTION

The nanometer process technologies in the sub-100 nm regime suffer from non-idealities such as process variations, voltage or temperature induced noise and soft errors [1]. Previous schemes to *avoid* errors due to timing violations have relied on adaptive body biasing (ABB) to modulate the transistor threshold voltage V_t and adaptive supply voltage (ASV) [2]. However a worst-case design that *avoids* errors has much higher power consumption than a nominal-case design in the presence of such increased variations. The nominal-case design, even though it is energy-efficient, will exhibit intermittent errors. Therefore, error-resilient architectures and implementations which trade-off power with reliability are of great interest [3].

Recently, communications-inspired techniques such as the algorithmic noise-tolerance (ANT) [4] have shown enormous promise in jointly optimizing energy-efficiency and robustness of nanometer systems-on-a-chip (SOCs). The recently proposed stochastic sensor network-on-chip (SSNOC) [5] computational paradigm extends ANT into the networking realm by the use of distributed computational units (or sensors) that compute and communicate with each other in order to provide robustness and energy-efficiency. The computational sensor outputs are combined using robust statistical signal processing techniques to generate a reliable system output.

Preliminary results [5] showed the benefits of the SSNOC architecture in correcting voltage-overscaling (VOS) errors [4] in 130nm process for a PN-code acquisition application. The robustness of the SSNOC architecture to process variation errors for 130nm process node were studied in [6].

In this paper, we study the impact of technology scaling from the 130nm node to 32nm, on the effectiveness of the SSNOC architecture in correcting process variation as well as VOS errors and determine the achievable power savings. We employ a PN-code acquisition filter as an application of SSNOC. Simulations using Predictive Technology Models (PTMs) [7] and Carbon Nanotube FET (CNFET) model [8] are employed.

2. PRELIMINARIES

2.1 SSNOC for PN-code Acquisition

In code division multiple access (CDMA2000) standards, the transmitted signal is spread by a pseudo-random noise (PN-code) sequence. Matched filters are employed at the receiver to calculate the cross-correlation between received signal (transmitted signal corrupted by additive white gaussian noise (AWGN)) and the locally generated PN-code. The matched filters operate at a high frequency and consume significant portion of the total receiver power. Figure 1(a) shows the conventional centralized direct form implementation of the matched filter. Multiply-accumulate (MAC) units are commonly employed to compute the correlation of the received signal with the PN-code [**?**]. In this architecture, the MACs are designed and operated at a critical supply voltage $V_{dd-crit}$, such that the worst-case critical

GLSVLSI'08, May 4–6, 2008, Orlando, Florida, USA.

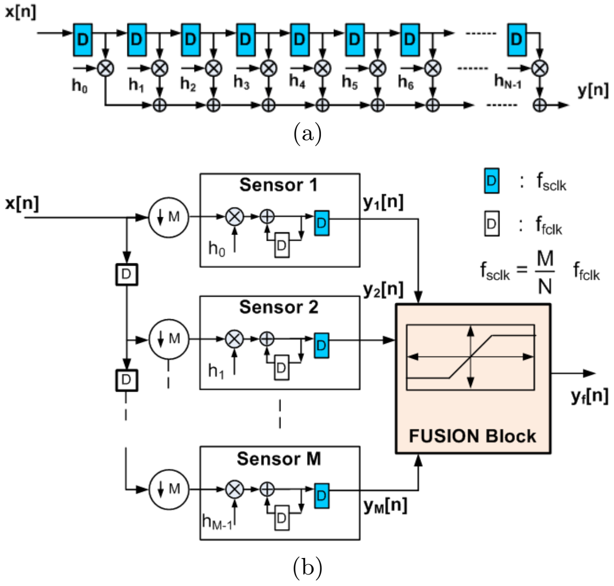

Figure 1: Matched filter for PN-acquisition: (a) traditional, and (b) an SSNOC-based matched filter.

path (with respect to process/voltage/temperature corner and input combination) is less than the clock period. The centralized nature of this computation makes it vulnerable to localized sources of non-idealities such as particle hits, hot-spots, and across die process variations and hence result in hardware errors if it is not operated at the worst-case corner.

The SSNOC computational paradigm is based on two key principles: 1) it employs the concept of *statistically similar decomposition* to decompose a centralized computation into a network of sensors, and 2) it employs *robust statistics theory* to construct/fuse the final output from the outputs of the sensor units. In the SSNOC architecture, statistically similar decomposition is achieved via poly-phase decomposition as shown in Fig. 1(b). The M sensor outputs exhibit an estimation error $\eta_{e,i}[n] = y[n] - My_i[n]$ due to input subsampling even under error-free hardware operation. Additionally, we expect the sensors to generate computational errors by designing them at the nominal process corner (instead of the traditional worst-case process corner design) or by operating them at sub-critical supply voltage, $V_{dd} = K_{vos}V_{dd-crit}$, where $0 < K_{vos} \leq 1$ and $V_{dd-crit}$ is the supply voltage below which timing violations occur. For a fair comparison, the conventional matched filter is also polyphase decomposed as in Fig. 1(b) but with an M-operand adder in place of the fusion block. The fusion block is implemented using principles of *Robust Statistics*.

Table 1: Characteristics of delay distributions due to process variations for $3\sigma_g$ slow die.

Gate		130nm IBM	90nm PTM	65nm PTM	45nm PTM	32nm PTM	32nm CNFET
Xor	μ_l	142	87	71	55	42	8.8
	σ_l	16	13	13	14	16	1.9
	$\frac{\sigma_l}{\mu_l}$	0.11	0.15	0.18	0.25	0.37	0.20
FA Carry	μ_l	223	130	106	83	62	16.36
	σ_l	19	12	11	12	12	3.48
	$\frac{\sigma_l}{\mu_l}$	0.08	0.09	0.10	0.15	0.20	0.21
FA Sum	μ_l	320	173	142	110	83	25.3
	σ_l	20	9	8	9	9	4.8
	$\frac{\sigma_l}{\mu_l}$	0.06	0.07	0.06	0.08	0.11	0.19

2.2 Robust Statistics

Sensor output errors can be modeled as random variables drawn from a distribution that is Gaussian (estimation error) with probability $(1-\epsilon)$ and some unknown distribution with probability ϵ (due to process/voltage non-idealities) for some $0 < \epsilon < 1$, i.e., an ϵ-contaminated distribution. Huber [9] shows that the following class of estimators, known as *M-Estimators*, are optimal in a certain robust sense:

$$\sum_{k=1}^{M} \psi[y_k - \theta] = 0 \tag{1}$$

where ψ is a general odd-symmetric function known as the influence function, and for ϵ-contaminated $\mathcal{N}(0,1)$ distributions, ψ is given by

$$\psi(y) = \begin{cases} y, & \text{if } |y| \leq k_{table} \\ k_{table}\ \mathrm{sgn}(y), & \text{else.} \end{cases} \tag{2}$$

where k_{table} is a constant that depends only on ϵ and the nominal distribution, $\mathcal{N}(0,1)$ [9]. The *One-step Huber* algorithm [9] can be employed to compute the parameters of the estimator, as shown below:

1. Compute scale estimate (Median Absolute Deviation):

$$\begin{aligned} T_0 &= \text{median}\{y_i\} \\ S_0 &= 1.4826 * \text{median}\{|y_i - T_0|\} \end{aligned}$$

2. Compute location estimate:

$$T_1 = T_0 + \frac{\frac{1}{M}\sum_i \psi(y_i - T_0, S_0 \cdot k_{table})}{0.5}$$

where 0.5 is used to approximate $\frac{1}{M}\sum_i \psi^{'}(\frac{y_i - T_0}{S_0})$ [9].

The One-step Huber architecture is shown in [5]. The value of k_{table} is pre-characterized and stored in a ROM. It needs to be read only once by the ψ block depending upon ϵ, the probability of hardware error due to non-idealities of process/voltage. The fusion block is implemented using T_1 in One-step Huber, and the median in T_0.

3. SIMULATION SETUP

3.1 Trend in Gate Delays with VOS

We characterized the gate delays for basic gates such as the full adder, xor and others with respect to supply voltage, body bias voltage combinations (denoted as (V_{dd}, V_b)) using HSPICE for an IBM 130nm CMOS process and using the PTMs from 90nm to 32nm. We note from simulations that the gate delay drops by approximately a factor 1.4x across the technology generations. The increase in full adder gate delay with respect to supply voltage is also noted for different process technologies. The gate delay increases slightly more rapidly as the technology scales down because the scaled supply voltages are closer to the respective threshold voltages.

We simulated the conventional and the SSNOC-based architectures ($M = 8$) at the gate-level using an HDL simulator using the pre-characterized delay values to obtain the sensor outputs. The throughputs were determined from Table **??**. The fusion algorithms were implemented in MATLAB to post-process the HDL output. The simulation was performed 1000 times to compute the receiver operating characteristic (ROC), the probability of detection (P_{Det}) versus probability of false-alarm (P_F). A threshold detector was used to evaluate the ROC by sweeping the threshold. A detection event is defined as correct detection of a PN-code in the input stream by the threshold detector. A false-alarm event occurs when the threshold detector declares the presence of a PN-code in the input incorrectly due to the AWGN channel noise or due to the computation errors originating from slow process/VOS. For our experiments, we first obtained the ROC by sweeping the threshold. For power vs. performance comparisons at different supply voltages, we calculated the P_{Det} at a $P_F = 5\%$ [10].

3.2 Trends in Delay Variations with Process

Next, we describe the simulation setup employed for simulating the effect of process variation induced timing errors on the performance of PN-code acquisition. The standard deviation of the gate delay due to die-to-die (D2D) process variations is denoted as σ_g. Within-die(WID) variations result in differences in device parameters for two instances of the same device on the same chip. The mean and the standard deviation of the gate delay due to WID process variations are denoted as μ_l and σ_l respectively. We first characterized the delay distribution of basic gates such as an xor, and a full adder due to WID variations for $90nm$ to $32nm$ process technologies using Monte-Carlo simulations with PTMs [**?**] and using CNFET models [11]. The characteristics of the distributions are tabulated in Table 1. Note that the relative variations (σ_l/μ_l) in the gate delays increase as the technology scales down.

Next, we sample the distributions in order to obtain the gate delays of a gate level implementation of the sensor MACs at: 1) the $3\sigma_g$ slow process corner operating at nominal supply voltage, 2) the $3\sigma_g$ slow process corner operating under ABB/ASV. This process is repeated 30 times in order to obtain 30 instances of the conventional and SSNOC architectures. We assume error-free operation by the fusion blocks and the detector, and simulate each of the 30 architectural instances with 100 independent data streams whose results are averaged to obtain the ROCs and the probabilities of detection.

4. SIMULATION RESULTS

4.1 Power Savings using VOS

Isodelay curves are obtained by varying (V_{dd},V_b) via HSPICE using PTMs. We simulate the schematic netlists of conventional architecture and the SSNOC architecture using HSPICE for a few random input vectors to obtain the power consumption for both the architectures at the (V_{dd},V_b) combinations obtained from the isodelay curves. The power-optimum (V_{dd},V_b) combination for the conventional and SSNOC architectures from these simulations are tabulated in Table 2. The maximum achievable power savings (for $P_{Det} \approx 0.6$ at a $P_F = 5\%$) obtained using SSNOC architecture for each process technology is also shown in Table 2. Note that the power optimum (V_{dd},V_b) combination changes from forward body-bias (increased leakage) to reverse body-bias (decreased leakage) when we use smaller process technologies with intrinsically high leakage currents. The maximum power savings decrease with technology scaling from 46% to 36% as we move from $90nm$ to $32nm$. The CNFET models are only voltage overscaled and they are not body-biased because CNFETs have both the NFET and PFET body terminal connected together [11].

Table 2: Power savings using SSNOC under VOS for equal P_{Det} (at a $P_F = 5\%$).

	130nm IBM	90 nm PTM	65 nm PTM	45 nm PTM	32 nm PTM	32 nm CNFET
Conv. (V_{dd}, V_b) (V)	1.2, 0.2	1.2, 0.1	1.1, 0	1.0, 0	0.9, -0.1	0.9, 0
SSNOC (V_{dd}, V_b) (V)	0.9, 0.2	0.8, 0.1	0.75, 0	0.72, 0	0.66, -0.1	0.6, 0
Power sav. (%)	36	46	40	38	36	46

Table 3: Power savings using SSNOC with ABB/ASV under process errors.

	130nm IBM	90 nm PTM	65 nm PTM	45 nm PTM
Conventional (V_{dd}, V_b) (V)	1.35, 0.2	1.3, 0.1	1.2, 0	1.1, 0
SSNOC (median) (V_{dd}, V_b) (V)	1.15, 0.2	1.2, 0.1	1.1, 0	1.0, 0
Power savings (%)	31	25	23	22

4.2 Impact of Process Variations on P_{Det}

Recall that we obtained 30 instances of the conventional and the SSNOC architectures under process variations. Each of the 30 instances results in a different P_{Det}. Thus, the P_{Det} is a random variable and hence will have a distribution. The means μ_i and the standard deviations σ_i of the P_{Det} corresponding to these distributions are tabulated in Table 4 for all the process technologies.

We note from Table 4 that the mean P_{Det} drops severely by approximately 3 orders of magnitude for the conventional architecture at the $3\sigma_g$ slow process corner operating at nominal supply voltage. In $130nm$ and $90nm$ process technologies, the SSNOC architecture improves P_{Det} by close to 3 orders-of-magnitude over the conventional architecture under identical process and voltage conditions. In addition, the SSNOC improves the variation in P_{Det} (σ/μ) by 2 orders-of-magnitude. As the technology scales down

Table 4: Characteristics of P_{Det} distributions (at a $P_F = 5\%$) of the three architectures due to process.

Process		Process			Process and ABB/ASV		
		Conventional	SSNOC (1-step)	SSNOC (median)	Conventional	SSNOC (1-step)	SSNOC (median)
130 nm	μ_i	0.001	0.78	0.80	0.84	0.98	0.97
	σ_i	0.003	0.015	0.008	0.24	0.005	0.003
	σ_i/μ_i	3	0.02	0.01	0.3	0.005	0.003
90 nm	μ_i	0.009	0.02	0.51	0.002	0.79	0.82
	σ_i	0.009	0.016	0.06	0.004	0.03	0.02
	σ_i/μ_i	1.0	0.75	0.13	1.8	0.034	0.025
65 nm	μ_i	0.004	0.005	0.008	0.009	0.59	0.77
	σ_i	0.007	0.006	0.02	0.008	0.1	0.02
	σ_i/μ_i	1.6	1.2	2.4	0.89	0.17	0.03
45 nm	μ_i	0.001	0.0027	0.003	0.006	0.053	0.60
	σ_i	0.004	0.005	0.006	0.008	0.05	0.07
	σ_i/μ_i	2.2	1.9	1.9	1.3	0.9	0.12
32 nm	μ_i	0.008	0.01	0.01	0.007	0.006	0.006
	σ_i	0.01	0.008	0.008	0.007	0.007	0.007
	σ_i/μ_i	1.1	0.8	0.8	1	1.2	1.2

beyond $90nm$, we note that the SSNOC architecture alone is unable to increase the mean P_{Det} or reduce its variation because of increased process variations. In that case, we use the SSNOC architecture in combination with conventional circuit level techniques such as ABB/ASV.

In $130nm$ process technology, using ABB/ASV, the conventional architecture recovers its loss in detection performance to achieve a $P_{Det} \approx 0.83$, which is comparable to that achieved by the SSNOC architecture. However, the SSNOC with ABB/ASV has a 2 orders-of-magnitude better (smaller) P_{Det} variation (σ/μ), and a smaller power as shown later. As the technology scales down to $45nm$, we note that the SSNOC (median) architecture with ABB/ASV is able to improve the mean P_{Det} by 2 orders of magnitude and reduce its variation (σ/μ) by 1 order of magnitude. For $32nm$ process technology, however, the SSNOC with ABB/ASV is unable to recover the P_{Det} because of high variations in gate delays.

4.3 Power Savings with Process Variations

Next, we present the power overhead of using SSNOC with a median fusion block. Since the performance of the SSNOC(median) is comparable to that of SSNOC(1-step), we choose SSNOC(median) in the following discussion of power-performance trade-off. We simulate the transistor-level netlists of sensors (MACs) in the conventional and the SSNOC architecture using HSPICE with a few random input vectors to obtain the power consumption of the sensors at different (V_{dd}, V_b) voltage levels. The power consumption of the fusion blocks (conventional M-operand adder and median) are obtained for $130nm$ IBM process technology using *Synopsys Design Analyzer*. From these fusion block power consumption numbers, we estimate their power consumption for other process technology nodes by scaling them by an appropriate factor. The factor is determined as the ratio of inverter power in that particular process technology divided by an identical inverter power in $130nm$ IBM process.

The maximum achievable power savings (for equal $P_{Det} \approx 0.6$ at a $P_F = 5\%$) obtained using SSNOC architecture for each process technology is shown in Table 3. Note that the power optimum (V_{dd},V_b) combination shows the same trend as VOS. The maximum achievable power savings decrease with technology scaling from 31% for $130nm$ process to 22% for $45nm$ process. Thus, SSNOC architecture is able to trade-off power and performance effectively with robust P_{Det} performance. The SSNOC with ABB/ASV is unable to recover the performance for $32nm$ process technologies. This motivates the need for more powerful versions of SSNOC using advanced techniques for robust fusion.

5. ACKNOWLEDGEMENT

The authors acknowledge the support of the Gigascale System Research Center (GSRC), one of five research centers funded under the Focus Center Research Program (FCRP), a Semiconductor Research Corporation program.

6. REFERENCES

[1] W. Zhao, and Y. Cao, "New Generation of Predictive Technology Model for Sub-45 nm Early Design Exploration," *IEEE Trans. on Electron Devices*, vol. 53, Page(s):2816 - 2823, Nov. 2006.

[2] T. Chen, and S. Naffziger, "Comparison of adaptive body bias (ABB) and adaptive supply voltage (ASV) for improving delay and leakage under the presence of process variation," *IEEE Trans. VLSI*, vol. 11, Oct. 2003.

[3] S. Borkar et. al., "Parameter variations and impact on circuits and microarchitecture," in *Proc. of DAC*, 2003.

[4] R. Hegde, and N. R. Shanbhag, "Soft digital signal processing," *IEEE Trans. on VLSI*, vol. 9 pp. 813-823, Dec. 2001.

[5] G. Varatkar et. al., "Sensor Network-On-Chip," in *Proc. of Int. Symp. on SOC*, Nov. 2007.

[6] G. Varatkar et. al., "Variation-Tolerant, Low-power PN-Code Acquisition using Stochastic Sensor NOC," in *Proc. of ISCAS*, May 2008.

[7] http://www.eas.asu.edu/∼ptm

[8] http://nano.stanford.edu/

[9] P. Huber, *Robust Statistics*, John Wiley and Sons, 1981.

[10] Gordon J. R. Povey, "Spread Spectrum PN Code Acquisition Using Hybrid Correlator Architectures," *Wireless Personal Communications: An Int. Journal*, September 1998.

[11] J. Deng, et. al., "Carbon Nanotube Transistor Circuits: Circuit-Level Performance Benchmarking and Design Options for Living with Imperfections," *in Proc. of ISSCC*,San Francisco, Feb. 2007.