Embedded Error Compensation for Energy Efficient DSP Systems

Sai Zhang *Student Member, IEEE* and Naresh R. Shanbhag, *Fellow, IEEE*

***Abstract*—Algorithmic noise-tolerance (ANT) is an effective statistical error compensation (SEC) technique for designing energy-efficient digital signal processing systems. A conventional ANT system employs an explicit estimator block to compensate for the large magnitude errors in the main block. The estimator presents area and power overheads, as large as 40% of the main block, to the system. In this paper, we propose ARCH-ANT, an architectural level embedded algorithmic noise-tolerance technique. ARCH-ANT achieves the same error compensation functionality as the conventional ANT by embedding the estimator block into the main block. Such embedding eliminates the estimator block and thus improves the system energy efficiency. A general optimization framework is proposed to design ARCH-ANT systems. Simulation results show that when applied to a multiply-accumulate (MAC) unit, 15.9%~69.4% and 59.2%~72.75% energy savings can be achieved by an 8×8 and 16×16 ARCH-ANT system, which is 5%~21.6% more than that of conventional ANT system, with no increase in mean square error (MSE).**

I. Introduction

The emerging applications in wireless sensor network, cloud computing, and big data services, require the design of intelligent and portable platforms. The limited energy supply in such platforms makes energy per operation E_{op} a critical design metric to be minimized. Additionally, there is a reliability challenge, caused by process, voltage and temperature (PVT) variations, leakage, soft errors and noise in sub-45 nm process technologies [1]. The result is an energy and robustness challenge in the design of nanoscale system-on-a-chips (SOCs). On the other hand, many of the above applications have relaxed precision requirements or employ statistical system level performance metrics such as signal to noise ratio (SNR) and detection rate [2]. Such statistical metrics enable the use of error resiliency techniques to design energy-efficient and robust SOCs.

Various error resiliency techniques have been proposed at the logic or circuit level. RAZOR [3] and error detection sequence (EDS) [4] employ a shadow latch to detect late arrived signals for error detection followed by a roll back scheme for precise error correction. Path-delay shaping [5] has been proposed to engineer the critical path of a DSP block such that timing errors are bounded to the least significant bits (LSBs). This type of error detection also requires RAZOR flip-flops. Other techniques such as digital stochastic computing [6] and Markov Random Field based circuits [7] have also been proposed to enhance the noise immunity of circuits.

At system level, fault-tolerance techniques such as N-modular redundancy (NMR) incurs large area and power overhead. Statistical error compensation (SEC) techniques are a class of techniques that utilize the signal and error statistics of DSP systems for enhancing robustness. Algorithmic noise-tolerance (ANT) [8] employs an explicit estimator block to compensate for the most significant bit (MSB) first errors in the main DSP block. ANT has been shown to provide up to 65% energy saving with little loss of performance. Soft NMR [9] makes explicit use of error probability mass functions (PMFs) to provide up to $10\times$ improvement in robustness with 35% energy saving. Likelihood processing [10] utilizes error statistics to perform inference at bit level and has been shown to provide up to $14\times$ improvement in robustness with 25% percent energy savings.

In general, circuit level error resiliency techniques operate close to point of first failure (PoFF) or in the low error rate ($< 0.1\%$) regime. In comparison, system level error resiliency techniques such as SEC can operate in the high error rate (>10%) regime. Previous studies have shown that a reduced precision replica ANT (RPR ANT) protected ECG processor [11] and MRF stereo matching block [12] can be fully functional at error rate of 58% and 21.3%, respectively. However, the improved robustness in RPR ANT comes at the price of 30% [11] to 40% [12] complexity overhead due to the explicit use of estimator blocks. This estimator overhead decreases the energy efficiency of the system and poses a concern for the application of SEC to more complex DSP systems.

In this paper, we propose embedded error compensation (EEC), where the estimator is embedded into the main block via proper architecture and algorithm level transforms, resulting in a low overhead architecture with the same error compensation functionality. Various EEC techniques can be derived from existing SEC techniques. In this paper, we present ARCH-ANT, an EEC technique derived from ANT. ARCH-ANT achieves the same error compensation functionality as the conventional ANT inspite of embedding the estimator block into the main block. Such embedding is achieved by decomposing the main block into MSB and LSB components and employing the MSB component as the estimator output. As a result, the ARCH-ANT eliminates the need for an explicit estimator and achieves improved energy saving. To find the optimum ARCH-ANT architecture, we also propose a general optimization framework that integrates circuit, architecture and system level simulations. To illustrate the benefits of ARCH-ANT, a multiply-accumulate (MAC) unit is designed via ARCH-ANT and conventional ANT, and simulated in a commercial 45nm CMOS process. Simulation results show 15.9%~69.4% and 59.2%~72.75% energy savings can be

978-1-4799-7088-9/14/$31.00 ©2014 IEEE

achieved by 8×8 and 16×16 ARCH-ANT systems, which is 5%~21.6% more than that of the conventional ANT system, with no increase in MSE.

The rest of the paper is organized as follows. Section II describes the principle of ARCH-ANT. Section III illustrates the optimization of the ARCH-ANT technique in the context of a MAC unit. Section IV presents simulation results of the energy optimized ARCH-ANT MAC. Conclusion is presented in Section V.

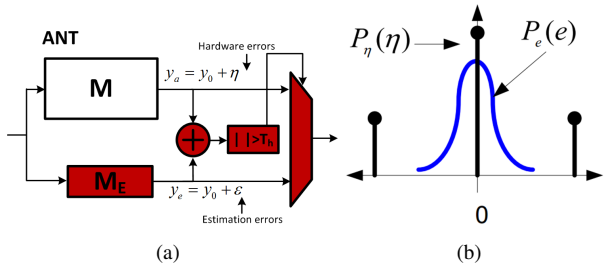

Figure 1. Algorithmic noise-tolerance (ANT): a) architecture and b) the error statistics in main and estimator block.

II. The Proposed EEC Technique: ARCH-ANT

A. Conventional ANT

Conventional ANT incorporates a main block and an estimator generating statistically similar results to the main block (see Fig. 1(a)). In RPR ANT [8], the estimator is obtained by reducing the precision of the main block. The main block is subject to large magnitude errors η, while the estimator is subject to small magnitude quantization errors ϵ (see. Fig. 1(b)), i.e.:

$$y_a = y_o + \eta$$
$$y_e = y_o + \epsilon$$

where y_o, y_a, y_e is the error free output, the main block output and the estimator output, respectively. ANT exploits the different error statistics of η and ϵ to obtain the final output $\hat{y}$, as follows:

$$\hat{y} = \begin{cases} y_a \text{ if } |y_a - y_e| \leq T_h \\ y_e \quad \text{otherwise} \end{cases} \tag{1}$$

where T_h is an application dependent parameter to maximize the performance of ANT.

B. Proposed ARCH-ANT

In RPR ANT, the main block and the estimator process the same data with different precision. This redundancy in data processing can be exploited to embed the estimator into the main block. This can be done by decomposing the main block into MSB and LSB blocks, and use the MSB block output as the estimator output y_e. By ensuring that the critical path of the MSB block is always shorter than that of the main block, the requirements of the error statistics on the main and estimator blocks are met.

Let $y_a = f(x)$ denote the main block functionality, where x and y_a are the input and output of the main block, respectively. A B_x-bit input x ($x = \{x_0, x_1 ..., x_{B_x-1}\}$) can be written in 2's complement form:

$$\begin{aligned} x &= -x_0 + \sum_{i=1}^{B_x-1} x_i 2^{-i} \\ &= x_M + x_L 2^{-(B_{x,M}-1)} \end{aligned}$$

where x_M is the value of B_{xM} MSB bits, and x_L is the value of $B_x - B_{xM}$ LSB bits, as follows:

$$x_M = -x_0 + \sum_{i=1}^{B_{xM}-1} x_i 2^{-i}$$
$$x_L = \sum_{i=B_{xM}}^{B_x-1} x_i 2^{-(i-B_{xM}+1)}$$

Therefore, the main block output is expressed as:

$$y_a = f(x) = f(x_M + x_L 2^{-(B_{xM}-1)})$$

In ARCH-ANT, we decompose $f(x)$ as follows:

$$y_a = f(x) = f(x_M + x_L 2^{-(B_{x,M}-1)}) = g(f_M(x_M), f_L(x)) \tag{2}$$

where $f_M(x_M)$ and $f_L(x)$ are sub-functions that are combined by operator $g(\cdot)$ to generate the final output y_a. Such decomposition will exist when $f_M(x_M)$ is an approximation of y_a. If we ensure that the critical path of $f_M(x_M)$ is shorter than that of $g(f_M(x_M), f_L(x))$, then $f_M(x_M)$ can be directly employed as the estimator output y_e. The operation of ARCH-ANT can thus be summarized as follows:

$$\begin{aligned} y_a &= g(f_M(x_M), f_L(x)) \\ y_e &= f_M(x_M) \\ y &= \begin{cases} y_a \text{ if } |y_a - y_e| \leq T_h \\ y_e \quad \text{otherwise} \end{cases} \end{aligned}$$

where T_h is the error detection threshold as in (1).

Fig. 2 shows an 8x8 MAC unit ($y[n] = y[n-1] + x[n] \times w[n]$) transformed into ARCH-ANT MAC (the numbers in the parenthesis are (bit length, fraction length)). According to (2), the transformation can be written as follows:

$$\begin{aligned} y_a[n] &= y_e[n] + f_L(\mathbf{u}[n]) 2^{-(B_{msb}-1)} \\ y_e[n] &= f_M(\mathbf{u}_M[n]) = x_M[n] w_M[n] + y_M[n-1] \\ f_L(\mathbf{u}[n]) &= x_M[n] w_L[n] + x_L[n] w[n] \\ &\quad + y_L[n-1] 2^{-(B_{msb}-1)} \end{aligned} \tag{3}$$

where $\mathbf{u}_M[n] = [x_M[n], w_M[n], y_M[n-1]]^T$, $\mathbf{u}[n] = [x[n], w[n], y[n-1]]^T$ and $B_{xM} = B_{wM} = B_{msb}$.

From Fig. 2, we conclude that the critical path of the estimator is shorter than the critical path of the main block

due to the presence of the final adder A_f. Therefore, large magnitude errors will occur in the main block output, and we can directly use $y_e[n]$ as the estimator output.

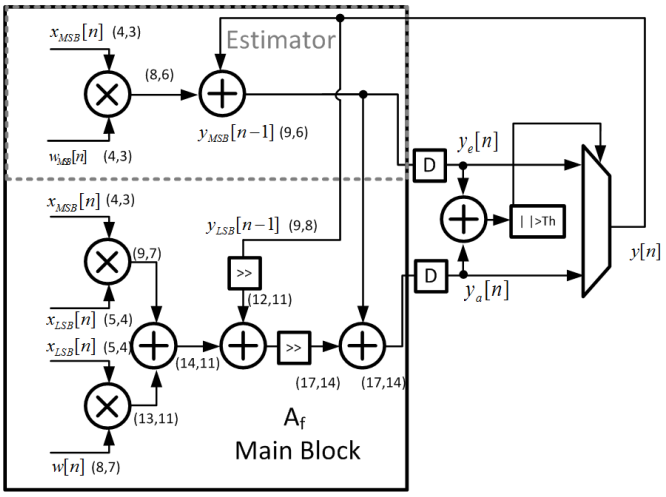

Figure 2. ARCH-ANT MAC architecture.

III. Energy-optimized ARCH-ANT

A. Energy Minimization Methodology

Fig. 3(a) shows the design methodology to optimize the energy consumption of ARCH-ANT subject to performance constraints. This methodology integrates circuit, architecture, and system level design variables, as indicated below:

1) At circuit level, HSPICE simulation is performed for basic blocks such as AND, XOR gates and the full adder using a commercial $45\,\text{nm}$ CMOS process to characterize delay and energy models under different supply voltages V_{dd}.

2) At architecture level, a structural HDL model for the ARCH-ANT kernel is developed in Verilog. To simulate the input dependent errors at different error rate, we apply voltage overscaling (VOS) [2], where the supply voltage $V_{dd} = K_{vos}V_{dd,crit}$, and $V_{dd,crit}$ is the minimum voltage necessary for error free operation. The delay model characterized at the circuit level is employed to obtain delays for the ARCH-ANT kernel at different V_{dd}s. Through HDL simulations, the error statistics under different error rate p_η (and thus voltage overscaling factor K_{vos}) are characterized.

3) At system level, the parameter space of ARCH-ANT design, such as $p_\eta(K_{vos})$, estimator bit width B_{msb}, is explored by employing the error statistics to inject errors and the energy model to estimate the E_{op}.

The optimization routine will output the optimum configuration that satisfies the performance requirements with minimum E_{op}.

B. Energy Consumption Model

We adopt a unified energy model which accounts for both dynamic energy and leakage energy [13], as follows:

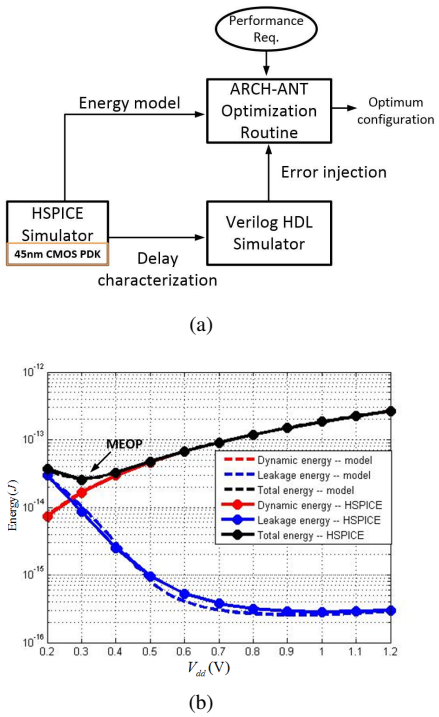

Figure 3. Design methodology: a) simulation setup, and b) HSPICE and model results for a 20 stage ripple carry adder in a $45\,\text{nm}$ CMOS process.

$$E_{core} = C_{core}V_{dd}^2 + V_{dd}I_{leak}(V_{dd})\frac{1}{f_{clk}} \tag{4}$$

with

$$I_{leak}(V_{dd}) = \mu C_{ox}\frac{W}{L}(m-1)V_T^2 e^{\frac{-V_t}{mV_T}} e^{\frac{-\eta_d V_{dd}}{mV_T}}(1-e^{\frac{-V_{dd}}{V_T}}) \tag{5}$$

where C_{core} is the effective load capacitance, V_{dd} is the supply voltage, V_t is the threshold voltage, V_T is the thermal voltage, μ is the carrier mobility, C_{ox} is the gate capacitance per unit W/L, m is a constant related to the sub-threshold slope factor, and η_d is the drain induced barrier lowering (DIBL) coefficient. Fig. 3(b) shows the output of HSPICE simulation and modeling results of a 20 stage ripple carry adder to show the accuracy of the unified model (within 5% for $0.2\,\text{V} \le V_{dd} \le 1.2\,\text{V}$). The unified model is able to correctly predict the minimum energy operating point (MEOP, see Fig. 3(b)) and will be employed to estimate the energy consumption of various architectures.

C. Energy Optimization Algorithm

We optimize the ARCH-ANT MAC proposed in Fig. 2 employing the methodology in Fig. 3(a). Since we adopt VOS to obtain different error rate, the parameters to be optimized are K_{vos} and estimator bit width B_{msb}, where we assume that $B_{x,msb} = B_{w,msb} = B_{msb}$. The optimization framework is general enough to consider the case when $B_{x,msb} \neq B_{w,msb}$. A grid search algorithm shown below is employed to systematically determine the optimum setting K_{vos}^* and B_{msb}^*, as shown in Algorithm 1. The B_{max} in Algorithm 1 is determined

via K_{vos}. It is the maximum estimator length under which the estimator does not make errors. The optimization routine gives the optimum ARCH-ANT configuration, including K^*_{vos}, B^*_{msb} and minimum energy E^*_{op}, at the output.

Algorithm 1 Energy Optimization algorithm for ARCH-ANT

1. Initialize K^*_{vos} = 1, B^*_{msb} = 0, E^*_{op} = energy of conventional MAC.
2. $K_{vos} = K_{vos} - \Delta, B_{msb} = 0$. Obtain maximum estimator precision B_{max} to ensure error free estimator operation.
3. $B_{msb} = B_{msb} + 1$, if $B_{msb} > B_{max}$, exit.
4. If MSE_{req} is satisfied, calculate energy $E(K_{vos})$, else go to step 3
5. If $E^*_{op} > E(K_{vos})$, $E^*_{op} = E(K_{vos})$, $B^*_{msb} = B_{msb}$
6. Go to step 2

IV. Simulation Results

Algorithm 1 is employed to optimize ARCH-ANT MAC with various precision and MSE requirements. The results are shown in Fig. 4. Fig. 4(a) and Fig. 4(c) show the optimization results for an 8×8 and 16×16 ARCH-ANT MAC, respectively. The dashed line shows that the maximum estimator precision B_{max} decreases as p_η increases, indicating that to ensure error free estimator operation, the estimator bit width is upper bounded. The solid line shows the optimum B_{msb} configuration for each p_η at different MSE requirements, with the green triangle marker indicating the (B^*_{msb}, p^*_η) pair achieving the MSE requirements with minimum E_{op}. Fig. 4(b) and Fig. 4(d) show the resulting energy comparison of the uncompensated (non-ANT) MAC, conventional ANT MAC and ARCH-ANT MAC, at different MSE levels. For the 8×8 MAC unit, the ARCH-ANT achieves energy savings between 15.9% and 69.4% for the MSE requirements of $10^{-2} \sim 10^{-5}$, while as the conventional ANT fails to achieve energy savings at the tight MSE requirement of 10^{-5} due to large estimator overheads. For the 16×16 MAC unit, both the ARCH-ANT and the conventional ANT achieve energy savings for all the MSE requirements, with the ARCH-ANT achieving 59.2%~72.75% energy savings compared with the non-ANT MAC.

Fig. 5 shows the energy savings achieved by the ARCH-ANT MAC as a function of MSE requirements and bit precision B_x. From Fig. 5 we can see that for fixed B_x, energy savings increase as MSE increases. This is because a larger MSE requirement allows the MAC to operate at higher p_η and reduces the estimator overheads. This is also confirmed in Fig. 4(b) and 4(d). Additionally, for fixed MSE, energy savings increase as B_x increases. This is because large B_x tends to tolerant more LSB errors, thus enabling the MAC to operate at higher p_η.

V. Conclusions

In this paper, we propose ARCH-ANT, an EEC technique derived from RPR ANT. ARCH-ANT performs the same error detection and correction functionality as the conventional RPR ANT by embedding the estimator block into the main block,

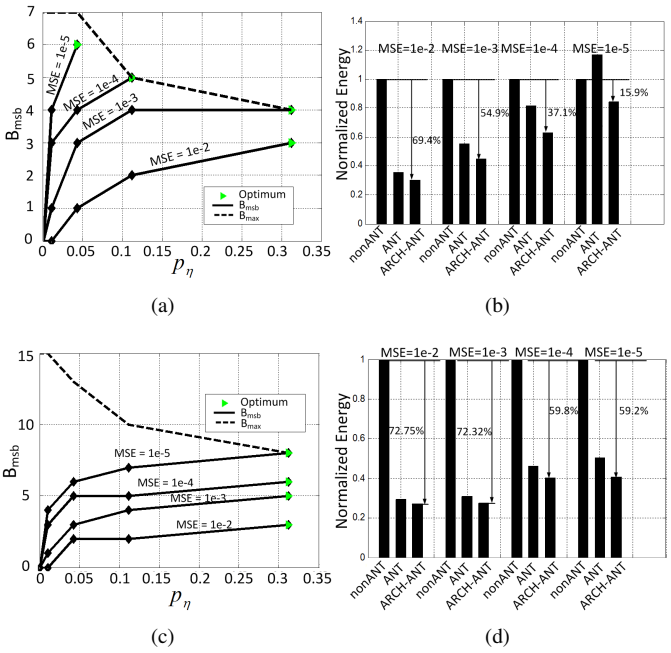

Figure 4. Optimization of ARCH-ANT MAC: a) optimization results of an 8×8 ARCH-ANT MAC for different MSE requirements, b) normalized energy of an 8×8 non-ANT MAC, conventional ANT MAC and ARCH-ANT MAC, c) optimization results of a 16×16 ARCH-ANT MAC for different MSE requirements, and d) normalized energy of a 16×6 non-ANT MAC, conventional ANT MAC, and ARCH-ANT MAC.

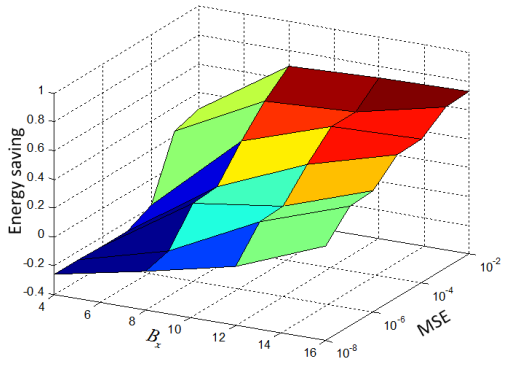

Figure 5. Energy savings vs. input precision and MSE.

and achieves improved energy efficiency. The design optimization of the ARCH-ANT system is formulated and solved using a general optimization flow that integrates circuit, architectural, and system level modeling. Simulation results using a commercial 45 nm CMOS process show that 15.9%~69.4% and 59.2%~72.75% energy savings can be achieved by an 8×8 and 16×16 ARCH-ANT MAC, which is 5%~21.6% more than that of the conventional ANT MAC while ensuring MSE requirements of $10^{-2} \sim 10^{-5}$.

Acknowledgment

This work was supported in part by Systems on Nanoscale Information fabriCs (SONIC), one of the six SRC STARnet Centers, sponsored by MARCO and DARPA.

REFERENCES

[1] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," in *Design Automation Conference, 2003. Proceedings*, June 2003, pp. 338–342.

[2] N. Shanbhag, R. Abdallah, R. Kumar, and D. Jones, "Stochastic computation," in *Design Automation Conference (DAC), 2010 47th ACM/IEEE*, June 2010, pp. 859–864.

[3] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, and T. Mudge, "Razor: a low-power pipeline based on circuit-level timing speculation," in *Microarchitecture, 2003. MICRO-36. Proceedings. 36th Annual IEEE/ACM International Symposium on*, Dec 2003, pp. 7–18.

[4] J. Tschanz, K. Bowman, S.-L. Lu, P. Aseron, M. Khellah, A. Raychowdhury, B. Geuskens, C. Tokunaga, C. Wilkerson, T. Karnik, and V. De, "A 45nm resilient and adaptive microprocessor core for dynamic variation tolerance," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, Feb 2010, pp. 282–283.

[5] P. Whatmough, S. Das, D. Bull, and I. Darwazeh, "Error-resilient low-power dsp via path-delay shaping," in *Design Automation Conference (DAC), 2011 48th ACM/EDAC/IEEE*, June 2011, pp. 1008–1013.

[6] W. Qian, X. Li, M. Riedel, K. Bazargan, and D. Lilja, "An architecture for fault-tolerant computation with stochastic logic," *Computers, IEEE Transactions on*, vol. 60, no. 1, pp. 93–105, Jan 2011.

[7] K. Nepal, R. I. Bahar, J. Mundy, W. Patterson, and A. Zaslavsky, "Designing logic circuits for probabilistic computation in the presence of noise," in *Design Automation Conference, 2005. Proceedings. 42nd*, June 2005, pp. 485–490.

[8] B. Shim, S. Sridhara, and N. Shanbhag, "Reliable low-power digital signal processing via reduced precision redundancy," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 12, no. 5, pp. 497–510, May 2004.

[9] E. Kim and N. Shanbhag, "Soft nmr: Analysis and application to dsp systems," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 1494–1497.

[10] R. Abdallah and N. Shanbhag, "Robust and energy-efficient dsp systems via output probability processing," in *Computer Design (ICCD), 2010 IEEE International Conference on*, Oct 2010, pp. 38–44.

[11] ——, "An energy-efficient ecg processor in 45-nm cmos using statistical error compensation," *Solid-State Circuits, IEEE Journal of*, vol. 48, no. 11, pp. 2882–2893, Nov 2013.

[12] J. Choi, E. Kim, R. Rutenbar, and N. Shanbhag, "Error resilient mrf message passing architecture for stereo matching," in *Signal Processing Systems (SiPS), 2013 IEEE Workshop on*, Oct 2013, pp. 348–353.

[13] R. Abdallah, P. Shenoy, N. Shanbhag, and P. Krein, "System energy minimization via joint optimization of the DC-DC converter and the core," in *Low Power Electronics and Design (ISLPED) 2011 International Symposium on*, Aug. 2011, pp. 97 –102.