Probabilistic CMOS Technology For Cognitive Information Processing*

Bilge E. S. Akgul Lakshmi N. Chakrapani Krishna V. Palem†

Center for Research on Embedded Systems and Technology
Georgia Institute of Technology
Atlanta, Georgia, USA 30332.

1. Introduction

Over the next decade the DoD will be challenged to deploy cognitive information processing systems that will embody learning and reasoning using semantically rich knowledge representations in applications such as autonomous vehicles, intelligence analysts, and electronic surveillance. In addition to the significant computational demands, cognitive applications must deal with uncertainty and inexact information and as a result probabilistic models have been central to many widely used cognitive kernels. Examples include Bayesian inference [7], probabilistic cellular automata, and randomized neural networks [4].

Further, as device scaling moves into the nanometer regime two significant technology challenges faced by embedded cognitive systems are the impact of noise and the significance of lower energy consumption, especially for mobile and autonomous embedded devices. These issues present significant challenges to the ability to sustain the performance benefits of Moore's Law over the next decade [8, 10, 5]. We can expect the innovation of novel computing architectures that are able to synergistically combine advances in device technology, algorithms, and new architectural primitives to deliver cognitive information processing systems that operate within the required size, weight, energy, and execution time constraints. In our current work we have innovated the concept of a probabilistic device or switch, whose output is guaranteed to be correct with a probability p, $1/2 < p \leq 1$ (where p is considered to be unity in the context of all conventional computing switches, in that the device is deemed to compute correctly and hence without error). We have shown how such devices can trade-off energy consumption with the probability of correctness [6]. Such devices can be fabricated using conventional CMOS technology and are referred to as probabilistic CMOS or PCMOS devices. We discuss how PCMOS devices can be used to build architectural solutions, probabilistic system-on-a-chip PSOC, to provide ultra-low energy architectures that are naturally matched to probabilistic components of cognitive applications. We discuss the impact of this technology in the cognitive domain and address future trends.

2. Probabilistic CMOS and Probabilistic-System-on-a-Chip Architectures

Any cognitive application consists of a deterministic component and a probabilistic component. As illustrated in Figure 1, the probabilistic component could be implemented on a *probabilistic* co-processor designed using PCMOS switches [3]. Such switches can trade-off energy for probability of correctness [6]. Such architectures—with a deterministic low energy host processor and a PCMOS based co-processor—will be referred to as probabilistic-system-on-a-chip (PSOC) architectures [2].

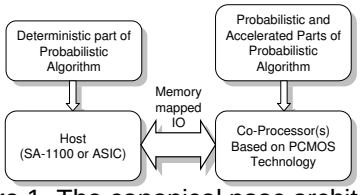

Figure 1. The canonical psoc architecture

To highlight and to analyze the benefits of PCMOS technology, the chief metric of interest is the **Energy performance product** or EPP for short. Given the EPP of two alternate realizations, they can be compared by computing the *energy performance product gain*. Energy performance product gain, $\Gamma_{\mathcal{I}}$, is the ratio of the EPP of the baseline denoted by β to the EPP of a particular architectural implementation $\mathcal{I}$. $\Gamma_{\mathcal{I}}$ is calculated as follows:

$$\Gamma_{\mathcal{I}} = \frac{Energy_{\beta} \times Time_{\beta}}{Energy_{\mathcal{I}} \times Time_{\mathcal{I}}} \quad (1)$$

Table 1 summarizes the application scenarios, and the application level EPP gains of PSOC over the baseline for the *Bayesian Inference* (BN) and *Randomized Neural Network* (RNN) application. The baseline implementation for BN and RNN applications is the StrongARM SA-1100 computing the deterministic as well as the probabilistic content and $\mathcal{I}$ is a PSOC executing an identical algorithm.

Algorithm	Applications	$\Gamma_{\mathcal{I}}$	
		Min	Max
BN	SPAM Filters, Battlefield Planning [9], Windows printer trouble shooting, Hospital Patient Management [1]	3	7.43
RNN	Image and pattern classification, Optimization of NP-hard problems	226.5	300

Table 1. Maximum and minimum EPP gains of PCMOS over the baseline implementation where the implementation $\mathcal{I}$ has a StrongARM SA-1100 host and a PCMOS based co-processor

As seen from Table 1, the application level gains of each of the application vary. In the subsequent sections, we explain this variation and in addition, analyze the factors affecting gains in a systematic way.

* This work is supported in part by DARPA under seedling contract #F30602-02-2-0124, by the DARPA ACIP program under contract #FA8650-04-C-7126 through a subcontract from USC-ISI and by an award from Intel Corporation.

† US Citizen and Corresponding Author

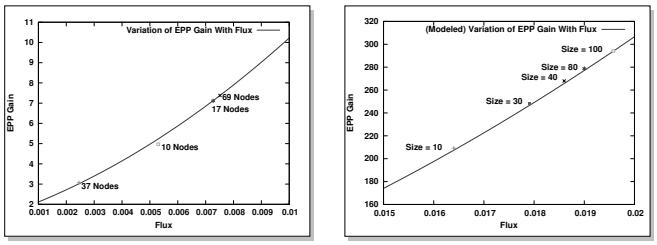

Figure 2. Variation of Gain With Respect to Flux for (a) Bayesian Network and (b) Randomized Neural Network

Application	gain over SA-1100	gain over CMOS
BN	9.99×10^7	2.71×10^6
RNN	1.25×10^6	2.32×10^4

Table 2. The EPP gain of PCMOS over SA-1100 and over CMOS for the core probabilistic step

3. PSOC Based Gains and an Analysis of Gains

Intuitively, the application level gain in energy and performance depend on three factors: (i) the "amount of opportunity" in the application to leverage the PCMOS based co-processor and (ii) the amount of gains afforded "per unit of opportunity" and (iii) its application-level impact. Broadly, the factors which influence gain can be studied under two categories *Implementation independent* characteristics and *implementation dependent* characteristics.

3.1. Implementation Independent Characteristics Influencing PSOC Gains

In the PSOC based realization of Cognitive applications, the *core probabilistic step* of each application is implemented in the PCMOS based co-processor and one core probabilistic step will be regarded as one "unit of opportunity". The "amount of opportunity" is formalized through the notion of *Flux* $\mathcal{F}$ (or flux for short) where $\mathcal{F}$ of an algorithm is defined as the *ratio of the core probabilistic steps to the total number of operations of an algorithm during a typical execution of the algorithm*. Figure 2 shows how $\Gamma_{\mathcal{I}}$ varies with the flux illustrating the results of both analytical as well as simulation models.

3.2. Implementation Dependent Characteristics Influencing PSOC Gains

The application level gains not only depends on the flux of an application but on the energy and performance gains afforded per "unit of opportunity". Table 2 presents the EPP gain of PCMOS based co-processor for the core probabilistic step of each of the applications of interest. The second column in the table corresponds to the case where β is the SA-1100 host without any co-processor and the third column corresponds to the case where β is a SA-1100 host coupled to a conventional CMOS based co-processor.

For a given flux, the application level gain would increase with increase in the energy as well as performance gain per unit flux. To illustrate this, let us revisit the Bayesian Network application (in particular a network of 37 nodes with a flux of 0.25), and the gain $\Gamma_{\mathcal{I}}$ where $\mathcal{I}$ is a PSOC and the baseline is a StrongARM SA-1100 host without a co-processor. As illustrated in Figure 3, higher the energy and time *saved* per invocation of the core probabilistic step, higher is the gain afforded by the PSOC implementation. The point where the surface intersects the z axis, presents the performance and energy consumption per unit flux which corresponds to a gain of 3 which correlates with the simulation results.

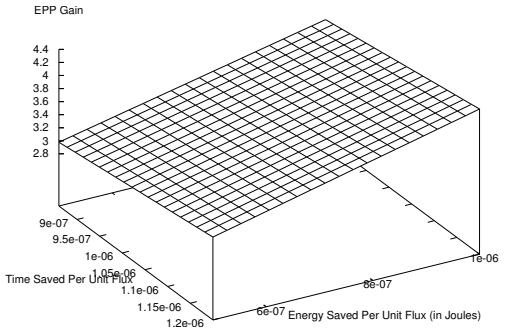

Figure 3. For a fixed Flux, variation of gain with respect to energy saved per unit Flux and time saved per unit Flux by using PCMOS

4. Concluding Remarks

Cognitive applications place stringent requirements on computational resources, which are unlikely to be met using the current strategy of performance increases purely due to Moore's law. In addition, we have seen how the probabilistic steps of these applications could be realized *directly* at the architectural and device level using future PCMOS devices. Such realizations would yield orders of magnitude improvement in energy and performance due to three reasons (i) Abundant opportunity for "acceleration" in cognitive application, (ii) Extremely efficient implementation of key steps of cognitive application using PCMOS technology and (iii) Favorable trends which indicates that custom solutions further increase gains.

References

[1] I. Beinlich, G. Suermondt, R. Chavez, and G. Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proceedings of the Second European Conference on AI and Medicine*, pages 247–256, 1989.

[2] L. N. Chakrapani, B. E. S. Akgul, S. Cheemalavagu, P. Korkmaz, K. V. Palem, and B. Seshasayee. Ultra efficient embedded soc architectures based on probabilistic cmos technology. In *Proceedings of The 9th Design Automation and Test in Europe (DATE)*, Mar. 2006.

[3] S. Cheemalavagu, P. Korkmaz, K. V. Palem, B. E. S. Akgul, and L. N. Chakrapani. A probabilistic CMOS switch and its realization by exploiting noise. In *Proceedings of The IFIP International Conference on Very Large Scale Integration*, 2005.

[4] E. Gelenbe. Random neural networks with negative and positive signals and product form solution. *Neural Computation*, 1(4):502–511, 1989.

[5] L. B. Kish. End of Moore's law: thermal (noise) death of integration in micro and nano electronics. *Physics Letters A*, 305:144–149, 2002.

[6] P. Korkmaz, B. E. S. Akgul, L. N. Chakrapani, and K. V. Palem. Advocating noise as an agent for ultra low-energy computing: Probabilistic cmos devices and their characteristics. *Special Issue of Japanese Journal of Applied Physics (JJAP)*, 45(4B):3307–3316, Apr. 2006.

[7] D. MacKay. Bayesian interpolation. *Neural Computation*, 4(3), 1992.

[8] K. Natori and N. Sano. Scaling limit of digital circuits due to thermal noise. *Journal of Applied Physics*, 83:5019–5024, 1998.

[9] A. Pfeffer. *Probabilistic Reasoning for Complex Systems*. PhD thesis, Stanford Univeristy, 2000.

[10] N. Sano. Increasing importance of electronic thermal noise in sub-0.1mm Si-MOSFETs. *The IEICE Transactions on Electronics*, E83-C:1203–1211, 2000.