Ten Years of Building Broken Chips: The Physics and Engineering of Inexact Computing

KRISHNA PALEM, Rice University and Nanyang Technological University
AVINASH LINGAMNENI, Rice University and CSEM SA

Well over a decade ago, many believed that an engine of growth driving the semiconductor and computing industries—captured nicely by Gordon Moore's remarkable prophecy (Moore's law)—was speeding towards a dangerous cliff-edge. Ranging from expressions of concern to doomsday scenarios, the exact time when serious hurdles would beset us varied quite a bit—some of the more optimistic warnings giving Moore's law until. Needless to say, a lot of people have spent time and effort with great success to find ways for substantially extending the time when we would encounter the dreaded cliff-edge, if not avoiding it altogether. Faced with this issue, we started approaching this in a decidedly different manner—one which suggested falling off the metaphorical cliff as a design choice, but in a controlled way. This resulted in devices that could switch and produce bits that are correct, namely of having the intended value, only with a probabilistic guarantee. As a result, the results could in fact be incorrect. Such devices and associated circuits and computing structures are now broadly referred to as inexact designs, circuits, and architectures. In this article, we will crystallize the essence of inexactness dating back to 2002 through two key principles that we developed: (i) that of admitting error in a design in return for resource savings, and subsequently (ii) making resource investments in the elements of a hardware platform proportional to the value of information they compute. We will also give a broad overview of a range of inexact designs and hardware concepts that our group and other groups around the world have been developing since, based on these two principles. Despite not being deterministically precise, inexact designs can be significantly more efficient in the energy they consume, their speed of execution, and their area needs, which makes them attractive in application contexts that are resilient to error. Significantly, our development of inexactness will be contrasted against the rich backdrop of traditional approaches aimed at realizing reliable computing from unreliable elements, starting with von Neumann's influential lectures and further developed by Shannon-Weaver and others.

Categories and Subject Descriptors: B.8.0 [**Performance and Reliability**]: General

General Terms: Reliability, Algorithms

Additional Key Words and Phrases: Co-design, EDA, energy-accuracy trade-off, Moore's law, inexact circuit design, probabilistic CMOS, VLSI design, low power/energy

---

This article is a full version of our previously published conference paper [Palem and Lingamneni 2012]. Early work was enabled in part by the U.S. Defense Advanced Research Projects Agency (DARPA) under seedling contract number F30602-02-2-0124 and further developed under the DARPA ACIP program under contract FA8650-04-C-7126 through a subcontract from USC-ISI and an award from Intel Corporation. The Moore distinguished faculty fellow program at the California Institute of Technology, the Canon distinguished professorship program of the Nanyang Technological University (NTU) at Singapore, the NTU-Rice Institute for Sustainable and Applied Infodynamics, and the Centre Suisse d' Electronique et de Microtechnique (CSEM SA) at Neuchatel, Switzerland, also supported various elements of our work.
Authors' addresses: K. Palem, NTU-Rice Institute of Sustainable and Applied Infodynamics, Nanyang Technological University, Singapore; and Department of CS, ECE & Statistics, Rice University, Houston, TX; A. Lingamneni, Department of ECE, Rice University, Houston, TX; and Wireless & Integrated Systems, CSEM SA, Neuchatel, Switzerland; email: avinash.l@rice.edu.

DOI: http://dx.doi.org/10.1145/2465787.2465789

ACM Reference Format:
Palem, K. and Lingamneni, A. 2013. Ten years of building broken chips: The physics and engineering of inexact computing. *ACM Trans. Embed. Comput. Syst.* 12, 2s, Article 87 (May 2013), 23 pages.
DOI: http://dx.doi.org/10.1145/2465787.2465789

1. RELIABLE COMPUTING IN THE BEGINNING

Computers process *bits* of information. A bit can take a value of 0 or 1, and computers process these bits through some physical mechanism. In the early days of electronic computers, this was done by electromechanical relays [Zuse 1993] which were soon replaced by vacuum tubes [Mauchly and Eckert 1947]. From the very beginning, these devices and the computers they were used to build were affected by concerns of reliability. For example, in a recently published interview with Presper Eckert [Randall V 2006] who co-designed ENIAC widely believed to be the first electronic computer built, he notes, "We had a tube fail about every two days, and we could locate the problem within 15 minutes."

The pioneer John von Neumann who worked with Eckert and his colleague John Mauchly clearly understood the importance of reliability [von Neumann 1956]. It is worth spending some time reviewing this relatively old history and the issues from this period, both as a historical curiosity and for the important lessons to be learned from von Neumann's landmark lectures delivered at Caltech. Back in 1952, von Neumann notes, "The subject matter, as the title suggests, is error in logics, or in the physical implementation of logics—in automata synthesis. Error is viewed, therefore, not as an extraneous and misdirected or a misdirecting accident, but as an essential part of the process under consideration . . . " [von Neumann 1956]. von Neumann is concerned with errors that occur intrinsically in the physical implementations of logics, and today, these would occur in VLSI and ULSI circuits built out of CMOS transistors. Surprisingly, sixty years after von Neumann's lectures, we are again grappling with the same issue in the modern era, since many believe that the exponentially improving benefits of Moore's law will end in the next 10 to 20 years. So, six decades after von Neumann gave his lectures, we have returned to the same concerns in this article.

2. A WALK THROUGH HISTORY STARTING WITH VON NEUMANN'S LECTURES

At the dawn of the modern computing era, the issue of failures and reliability reared its ugly head. This motivated von Neumann and his collaborators to use probability to model error or failures through abstract models of hardware [von Neumann 1956]. Since these models were based on Turing's [1936] now classical paper and were presented in the McCulloch-Pitts style [1943], they had a cybernetic flavor but nevertheless captured the essence of a state machine or automaton widely used to abstractly capture hardware behaviors today (see [Minsky 1967] for an equivalence). We must remember that modern automata theory was nascent at this time, so some of von Neumann's constructions might seem cumbersome in retrospect. Nevertheless, we think the insights were incisive and continue to be relevant to us. For example, the model used in his lectures is essentially an abstract form of a modern computer, represented by a combinational logic component and a communication component corresponding to wires or interconnect, including a mechanism for encoding the state of a computation. In this model and with error in the logic components, he described ways for realizing reliable computational systems given that individual elements are prone to failure.

Four years later, Moore and Shannon [1953a, 1956b] in a sense reworked and extended the work presented in these lectures by introducing a model that is based on *switches*, which are the ubiquitous building blocks of computing systems even today. The particular model they used is based on Shannon's celebrated masters thesis,

"A Symbolic Analysis of Relay and Switching Circuits" [1937]. In this thesis, Shannon used the switch to abstractly represent an electromechanical relay and used it as a basis for building digital circuits. Probabilities were used to model the correct or incorrect functioning of a switch. Once again, focus was on the question of correcting errors introduced by potentially faulty switches and on determining the cost of such correction in realizing digital circuits. Since then, close to fifty years passed without much of an incident, until, spurred by the spectacular success of the transistor and its integration at a very large scale leading to VLSI, the revolutionary invention of one of the earliest microprocessors [Faggin and Hoff 1972] occurred, fueled by the historically unprecedented march of ever-decreasing transistor feature sizes prophesied by Moore [1965].

As physicist Gell-Mann notes in a recent interview Gell-Mann [1997] where he also describes his role with K. A. Bruckner in the work leading to von Neumann's celebrated lectures, concerns of reliability became insignificant with the remarkable reliability achieved with transistors replacing vacuum tubes as switches. However, by the 1990s, the call for approaches to help sustain Moore's law as CMOS transistor feature sizes approached nanoscale dimensions was becoming increasingly strident. Scholarly articles started appearing with daunting titles such as "End of Moore's Law: Thermal (Noise) Death of Integration in Micro and Nano Electronics" [Kish 2002]. So, in keeping with what seems to be history's penchant for repeating itself, the need to realize deterministically reliable computing architectures from unreliable switches resurfaced once again, this time in the modern context of CMOS transistor-based switching devices [Nikolic et al. 2001].

3. CROSSING OVER TO THE DARK SIDE BY USING UNRELIABLE ELEMENTS UNRELIABLY

Around 2002, Krishna Palem asked the following question. What if we consider building unreliable yet useful, circuits and computing blocks from unreliable hardware elements, rather than striving to build reliable switches, circuits, and computing hardware from potentially unreliable components? We could potentially have a richer domain of switches to draw upon and therefore be much less constrained if we were to strive for reliability all the time. Thus, half a century after von Neumann's lectures, by not pursuing the goal of reliable computing advocated there, realizing useful and yet unreliable computing devices by design was shown to be viable using a variety of models [Palem 2003a], including a *random access machine* [Palem 2003b] and models representing circuits built from switches whose probabilities quantify the error [Palem 2005].

Why would anyone want to build unreliable computing elements knowing that they do not compute correctly? While this is counterintuitive, it turns out that there is an entirely different relationship between the resource cost associated with the physical implementations of switches and their probabilistic or erroneous behaviors. Specifically, there is a relationship between the correctness or erroneous behavior and the energy consumed by the physical implementation of a switch. The work of Palem [2003b, 2005] characterized the relationship of the amount of energy savings in the physical implementation of switches as we increase the associated correctness or error, for the first time, as far as we can determine. Thus, erroneous hardware switches and circuits built from them that are not deterministic were shown to be potentially beneficial since their accuracy or error can be traded for energy savings [Palem et al. 2005].

Exploiting this trade-off became interesting since by the late 1990s there was a rapidly increasing concern about the amount of energy consumed by computing systems [Mudge 2001]. Therefore, implementing probabilistic switches designed to be erroneous as a basis for realizing energy savings and building computing systems from

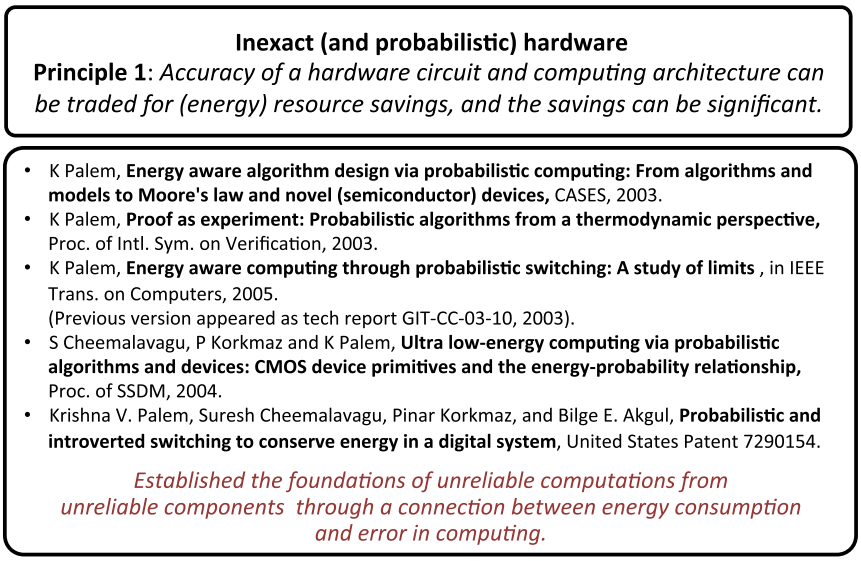

Fig. 1. The first principle in the domain of inexact design and the papers and patent that established it.

them for doing useful work, seemed to offer an attractive alternative. Palem became a "heretic" [Jonietz 2008] as a result of starting an active project and building a group to work on this idea, around 2001–2002. This effort received significant stimulation with support from DARPA through a seedling grant in 2002. A summary of our work over the decade to follow, which represents a break from the historical legacy of always seeking deterministically reliable computing from error-prone computing elements, is the subject of this article. Since then and based on the establishment of the *first principle* of our work as shown in Figure 1, that of *designing unreliable hardware and computing systems that are useful from unreliable computing elements while garnering resource savings in return*—the field has been growing steadily. We will summarize the current state of the field which we refer to as *inexact design* in Section 5.

3.1. The Energy-Error Relationship from a Thermodynamic Perspective

That the energy consumed by a switch and its associated error are related is not an entirely surprising fact. von Neumann came close to this connection when he remarks in his lectures that "our present treatment of error is unsatisfactory and ad-hoc. It is the author's conviction, voiced over many years, that error should be treated by thermodynamical methods and be the subject of a thermodynamical theory, as information has been, by the work of L. Szilard and C. E. Shannon (Cf. 5.2) [von Neumann 1956]." This comment made in passing is actually related to the genesis of our idea and the subject matter of this article, in particular to Principle 1 (Figure 1).

We will now take a short digression to understand its full import. First, we note that physical implementations of switches using electrical means, whether they are built using vacuum tubes or transistors, consume energy when they perform the switching function. Thus, they are governed by the laws of classical physics in general and thermodynamics in particular. By the time von Neumann delivered his lectures, classical thermodynamics had a clear statistical foundation and interpretation following the seminal work of Boltzmann and Brush [1995] and Gibbs [1902]. In particular,

Szilard's work Szilard [1929] to which von Neumann refers is an important part of this development.

As background, one of the more significant debates in classical physics which spanned a sixty-year period has to do with the validity of the celebrated second law of thermodynamics on which Szilard's work played a central role. Through a clever device based on Maxwell's construction of reasoning about the celebrated second law [Maxwell 1871], he created an object which has since come to be known as Szilard's engine. Loosely speaking, Szilard's engine is a physical structure which we can think of as a cylinder delineated into two halves separated by a (weightless) trap door. In its simplistic form, we can think of a single molecule (of some gas) is in motion in the cylinder. From Maxwell's Gedanken experiment, an external agent or "demon" can, by raising and lowering the door, trap the molecule in either half of the cylinder. For helpful comments about the validity of the simplification involving a single molecule and the use of ensemble averages instead of averaging a system with multiple molecules over time, please see Feynman's lectures.

While it is not widely known, this construct had and continues to have a powerful influence on the way we reason about information and its relationship to physical implementations, especially as they relate to issues of energy consumption. It is the first device that we know of which can be in one of two *states* and which can be switched to induce a change of state by raising and lowering the trapdoor. It is, therefore, the earliest instance that we have found of an abstract representation of a physical implementation of a *bistable* switch. In this sense, in our opinion, it is also the first known link between a physical object and a method for producing abstract information through switching. In this instance, information can be created through the actions of the external agent by manipulating the trap door. Thus, each act of raising and lowering the trapdoor represents a switching step which produces a bit of information that is recorded, thereby encoding the state of the switch. With this interpretation, Szilard's construct can be viewed as a switching engine for producing information. Since the state of the engine is determined by the location of the gas molecule, the laws of statistical thermodynamics can apply. Therefore, it provides a very natural and to use von Neumann's term, an "intrinsic" statistical basis for relating the energy consumed to the information being produced or computed by physical switches.

Returning to von Neumann's comment, the "unsatisfactory" aspect, we believe, has to do with the fact that in the approach he described, the probability of error is modeled synthetically in a manner that is extrinsic to the implementation. The probability of a switch performing its activity correctly is simply a numerical value which is not derived from the inherent thermodynamics associated with the physical implementation of a switch. This has an advantage in that it allows reasoning in purely abstract terms without being encumbered with the physical details. This detachment was exploited by Moore and Shannon subsequently [1956a]. As a result, the relationship between the (probabilistic) error as it varies and the associated thermodynamics was never a part of the classical treatment of probabilistic switching.

An exception of course is the remarkable work of Landauer [1961]. He explored the thermodynamics of a bistable switch which led to his historic insight on the minimum amount of energy needed to perform the act of recording a bit of information by an agent during a single switching step. However, it is very important to note that Landauer concerned himself with producing a bit correctly, and therefore, he did not characterize the energy associated with probabilistic switches with varying probabilities of correctness or error. In this sense, Principle 1 is not something that Landauer's work explored, other than at a single point where the probability of error tends to zero. So, to summarize, models of probabilistically correct switching which do not have a relationship with energy consumption is what Moore and Shannon had developed

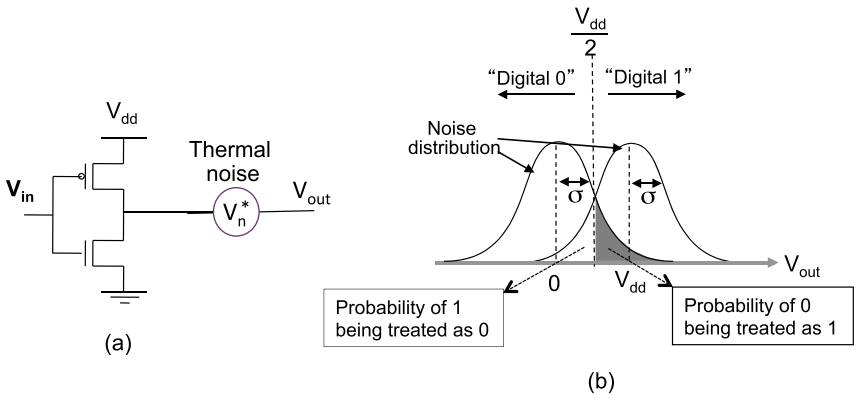

Fig. 2. (a) An inverter as an example of a PCMOS switch characterized by coupling a noise source V_n^*, (b) The digital value 0 (and 1) corresponding to the noisy output (input) voltage of the probabilistic inverter is represented by a Gaussian distribution with a mean value of 0 (or Vdd) and a standard deviation σ which is the rms value of the noise [Korkmaz et al. 2006].

on the one hand. On the other hand, through Landauer's work, we have a connection between the energy cost as it relates to a correct switching step for producing one bit of information and recording it without error.

3.2. Our Path to Connecting Switches Back to Their Physical Reality

It turned out that having both of the elements—the probabilistic error and the associated energy consumption—in a single framework was the key to unearthing Principle 1 of trading error for energy (and subsequently hardware resource) savings. To do this, we went back to Szilard's roots, and rather than focus on his understanding the nature of the second law of thermodynamics, instead looked at his engine as a technical and perhaps even as a technological construct. By doing this, we were able to extend his engine to a model of a switch [Palem 2005], which can be switching correctly with some probability of correctness p and relate this probability very naturally to the associated energy consumption. We now have a switching device which can produce a bit of information through some physical medium wherein the probability of it being correct is the parameter p, and the greater its value, the greater the energy consumed. The first attribute of our switch is the probability of correctness p associated with each bit of information being produced by it, whereas the second attribute is the associated energy cost. Both of these attributes are explicit and can be related to each other.

In 2003, we did this through a probabilistic switch that captures these two attributes simultaneously through the work referred to earlier [Palem 2005]. In spirit and philosophy, it went against the direction of abstracting away the physical attributes through clean models as von Neumann and Moore-Shannon set out to do. Our switches connected probabilistic error during switching back to the physics and the energy consumed. In our experience, this formulation of a probabilistic energy-aware switch has proven to be a useful foundational construct to understanding and reasoning about potentially unreliable energy-efficient hardware.

However, in order to really use this idea in current reality, we had to consider CMOS implementations. As outlined in Figure 2, by 2004, we could show that *probabilistically* correct CMOS switches, since referred to as PCMOS (switches), did exhibit this trade-off (see Cheemalavagu et al. [2004, 2005]). In fact, the trade-off between p, its probability of correctness, and the associated energy was modeled mathematically and also measured physically. It turned out to be much more favorable than we had

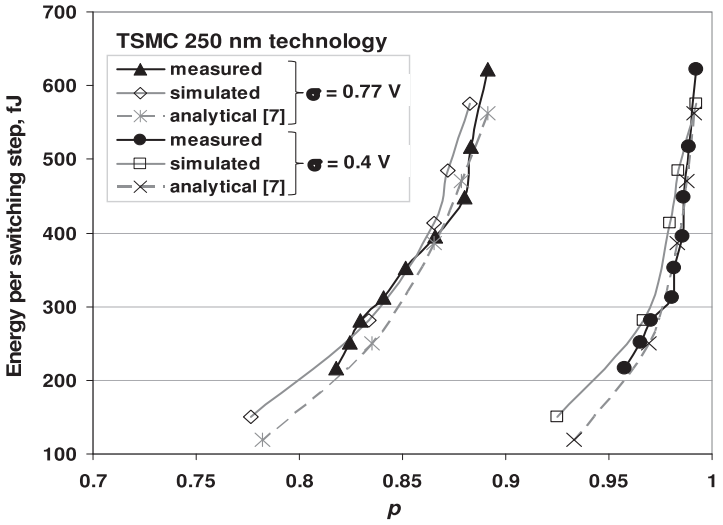

Fig. 3. Measurement, simulation, and analytical results for the E-p relationship of PCMOS inverters based on TSMC 0.25μm technology with a noise magnitude of 0.77V RMS and 0.4V RMS (from [Korkmaz 2007] and [Palem et al. 2009b]).

anticipated in the following sense. As the probability p of a PCMOS switch being correct approached 1, the energy consumed increased dramatically in the region where CMOS switches are operated correctly (very close to $p = 1$). In other words, a small decrease in the probability of correctness (below a value of $p = 1$) could be traded for a significant drop in the energy consumed. These PCMOS switches (Figure 2) embodied probabilistic behavior induced by noise, whose statistics were akin to those of "thermal" noise sources Cheemalavagu et al. [2004, 2005] and whose magnitude is succinctly represented by the root-mean-square (RMS) value σ of the associate probability distribution.

This relationship between p and the energy consumed is illustrated in Figure 3 [Akgul et al. 2006] for one switching step for an inverter. As we can see there, for a noise magnitude σ of 0.4V and a p of 0.99, the corresponding energy consumption is 600 fJ, whereas with a p of 0.95, this value drops to 200 fJ per switching step. This represents an indicating factor of 3 (or 300%) reduction in energy consumption for a mere $\sim$4% decrease in the probability of correctness!

Paraphrased from Akgul et al. [2006], this relationship between p and the energy consumption of a PCMOS switch can be restated as follows: "For any fixed technology generation determined by the transistor's feature size, which determines the capacitance C of a switch built using such transistors, and a fixed noise (RMS) magnitude σ, the switching energy $E_{C,\sigma}$ of a probabilistic switch grows with the probability of correctness p. Furthermore, the order of growth of $E_{C,\sigma}$ in p is asymptotically bounded below by an exponential in p." Over the course of the next two years, we fabricated working switches and demonstrated that, indeed, measured behaviors matching those predicted by the models and simulations could be realized [Korkmaz et al. 2006, 2007].

4. THE LOGIC OF ERRONEOUS HARDWARE DESIGN AND A PROBABILISTIC BOOLEAN LOGIC

The behavior of our probabilistic CMOS switches and the way they can be combined to design electrical circuits is based on Boole's strikingly contemporary formulation of

logic citeboole from the nineteenth century! In our current terminology, Boole's logic is always reliable, since it captures the behavior of gates that are deterministically correct. Since correctly functioning or exact circuits have always been the basis for designing computational building blocks, boolean logic has been the universally used framework during design. However, if we set out to design hardware that is imprecise and hence error-prone, as we have over the past decade, boolean logic has to be extended to admit incorrect outcomes.

To illustrate this, let us consider the simple act of adding two digital numbers. An adder circuit that could perform this addition is composed of a collection of boolean logic gates. In the engineering domain, its behavior is typically represented as a truth table, as shown in Figure 4(a) for the example of the hardware used to compute the carry bit. Using this example, we ask the question as to what type of logic is needed to design a full adder which can compute the sum and carry given two input bits and a carry-in that can admit probabilistic error. Since logic gates are structurally the atomic elements from which we build the adder circuit, we need to develop a framework for modeling erroneous gates and consequently extending Boole's rules.

Around 2007, motivated by the fact that significant energy gains could be achieved through probabilistic CMOS switches, we extended boolean logic to systematize their design. What does this mean? Initially, we thought of allowing probabilities of error to be associated with bits of information at the inputs and outputs of conventional boolean operators synthetically in the von Neumann-Moore-Shannon style, the operators or gates themselves remained deterministic. To understand this idea, let us consider the example boolean formula from Figure 4(a). Let us further suppose that the sole erroneous (operator) gate is in the clause $(y \wedge z)$ and that it is correct with a probability of 3/4. Now, we can inject error through a pseudorandom source whose output is combined with that of the deterministic gate implementing the clause $y \wedge z$. Since the gate and the input bits are deterministic, the relationship between its correctness probability (or error) and energy consumed does not arise. Rather, randomness is present in the extraneous pseudorandom or random source that is not part of the $\wedge$ operator itself, and furthermore, it is not connected to the energy consumption.

It quickly became clear to us that any useful extension has to satisfy a few constraints. Specifically, it must capture simultaneously and explicitly (i) the relationship between the probability parameter p, (ii) the manner in which p affects the output bits of the boolean operator, and (iii) the associated energy cost. Furthermore, a natural boolean logic should be a special case derived by appropriately restricting the probability parameter as $p \rightarrow 1$. Despite the significant amount of time that has elapsed since Boole first proposed his logic, we were surprised that we could not find a suitable generalization of these goals.

So, in 2008, we described a *probabilistic boolean logic* (PBL) [Chakrapani and Palem 2008] in which the operators have the probability parameter p tied to them inextricably, as shown in Figure 4(b). As illustrated there, operators such as AND ($\wedge$) and OR ($\vee$) are not deterministic and are associated instead with the probability of correctness parameter p. In our example, the $\wedge$ operator has an associated p value of 3/4. The results in the truth table have the correct outcomes occurring with a probability p, whereas the incorrect values occur with probability $(1-p)$. Thus, each probabilistic boolean operator, say $\wedge_p$ denoting a probabilistic AND, has an explicitly associated probability of correctness parameter p, which relates it simultaneously to its energy cost through the relationship discussed earlier in Figure 3. Therefore, each of our PBL operators—$\wedge_p$, $\vee_p$, and others—has an associated energy (thermodynamic) cost hand in hand with its correctness, represented by p.

Input x	y	z	Truth Value
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

(a)

Input x	y	z	Probabilities Truth Value = 1	Truth Value = 0
0	0	0	¼	¾
0	0	1	¼	¾
0	1	0	¼	¾
0	1	1	¾	¼
1	0	0	¼	¾
1	0	1	1	0
1	1	0	1	0
1	1	1	1	0

(b)

Fig. 4. (a) Conventional truth table for computing the carry of a full adder $(((x \wedge y) \vee (x \wedge z)) \vee (y \wedge z))$ built from reliable hardware with inputs x, y, and a carry-in z. (b) A probabilistic boolean truth table for $(((x \wedge_1 y) \vee_1 (x \wedge_1 z)) \vee_1 (y \wedge_{3/4} z))$ for the same circuit built from unreliable hardware [Chakrapani and Palem 2008].

4.1. From Gates and Logic to Applications

Once our PCMOS studies revealed the very nonlinear relationship between energy and error, the next obvious question was to see how best to use such hardware in mainstream computing. A natural and immediate direction was to consider applications that embodied probabilities naturally, such as decision systems and pattern recognition based on bayesian and random neural networks respectively, as well as cryptographic applications. Using a system-on-a-chip type architecture, we showed that significant energy and speed gains could be simultaneously achieved through PCMOS derived architectures [Chakrapani et al. 2007]. Beyond such applications, since we are increasingly consuming information through our human sensory pathways as we see and hear the results signals processed by computers as video and sound, we investigated the use of Principle 1 in the context of signal processing applications.

What is the meaning and effect of error in this case? To answer this question, we undertook a study which showed even to our surprise that PCMOS gates used in adders and multipliers to build signal processing primitives, such as filters, can yield significant—up to a factor of 5.6—energy savings when used to process video and image data [George et al. 2006]. So, while the results of computing can be erroneous, they can be perceptually acceptable, if not indistinguishable from those produced using correct hardware with greater energy cost. As far as we can tell, this is the first foray of using hardware to design a system that embodies error, and significantly, there is no intention or effort to correct it—the results are produced and consumed in their incorrect form.

This was achieved through our second principle as summarized in Figure 5: invest resources, energy in this case, proportional to the significance of information being computed [George et al. 2006; Palem et al. 2006]. Specifically, in hardware designed to implement arithmetic units, lower order bits received lower energy investments and were thus, more vulnerable to error. In contrast, bits of higher significance received larger energy investments and were, thus, much less error prone.

Inexact design guided by value of information

Principle 2: *The (energy) resource investment in various elemental components of an inexact circuit and the computing architecture built from it must be proportional to the value of the information they are meant to process and produce as outputs.*

- Jason George, Bo Marr, Bilge E. S. Akgul, Krishna V. Palem**, Probabilistic arithmetic and energy efficient embedded signal processing**, CASES 2006.
- Krishna V. Palem, Bilge E. S. Akgul, Jason George, Bo Marr**, Variable scaling for computing elements**, United States Patent 8316249 B2.

Introduced Biased Voltage Scaling (BiVOS) as a basis for designing inexact architectures for DSP.

Fig. 5. The second principle in the domain of inexact design and the paper and pending patent that established it.

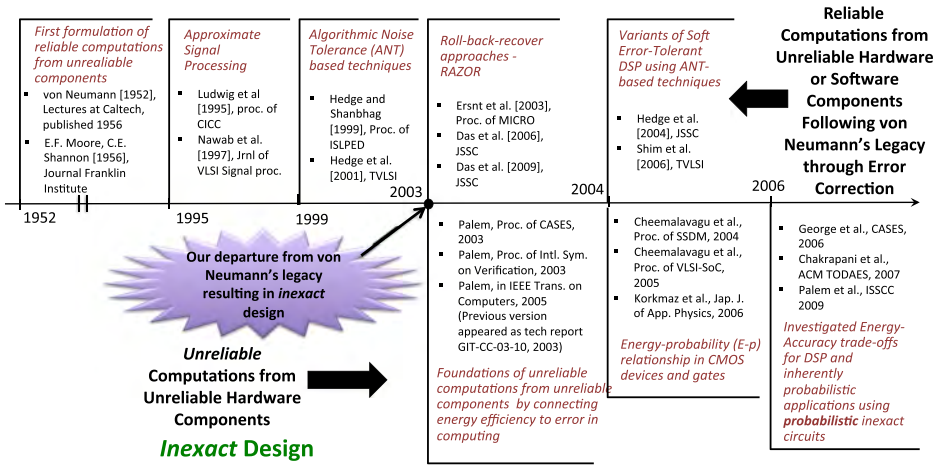

Fig. 6. Timeline of the papers and innovations that enabled the domain of inexact circuit design.

5. FROM PROBABILISTIC TO INEXACT CIRCUITS AND COMPUTING

Thanks to the tremendous success of Moore's law, transistors and VLSI circuits built upon it are reliable even today and allow us to build computing structures that are deterministic for all practical purposes. However, we were curious to think of ways of exploiting our two principles in the context of reliable VLSI technology that is readily available today. If we could do this, we could potentially achieve gains in energy and possibly in other physical attributes, such as area and speed by inducing error into otherwise reliable or error-free hardware. Specifically, since CMOS switches as they exist today are reliable, various innovative ways to *intentionally* make the components unreliable or "erroneous" through *deterministic* means, in exchange for significant savings in energy, and also, in speed and in area in some cases have emerged.

We have been using the phrase *inexact design* to broadly refer to approaches that let us garner resource gains by trading accuracy using our two principles (see Figures 1 and 5) at the hardware level, and hence compute unreliably, both from hardware which is error-prone or error-free. By extension, we have been referring to the resulting designs as *inexact circuits*. In Figure 6, we have summarized the

chronology of work related to the use of unreliable hardware *reliably* going back to 1952 and have highlighted our departure in 2003 that resulted in the branch of inexact (or unreliable yet useful) computing. Traditionally, as shown in the Figure 6, realizing or synthesizing reliable systems from unreliable switches entailed boosting the probability of correctness by compensatory error correction mechanisms. In contrast, following our work from 2003, there is a distinct split from the legacy of von Neumann's lectures, resulting in a branch of design and engineering which explicitly seeks to build *unreliable* yet useful computing hardware and systems from unreliable components, without compensating through error correction.

At this point, we will digress briefly from unreliable hardware to dwell on two innovative ideas (Figure 6) towards realizing *reliable* or error-free low-cost DSP systems. The first idea of *approximate signal processing* [Ludwig et al. 1995] was inspired by the approaches to designing resource-constrained (typically execution time) systems in the areas of artificial intelligence and real-time computing. The second idea of *algorithmic noise tolerance* [Hegde and Shanbhag 1999] allows circuits to be unreliable in the first instance but uses algorithmic error-control schemes based on the statistics of the system and the input-output behavior to correct the resulting errors. A third approach that appeared around the same time as our work in 2003 is the RAZOR effort from University of Michigan [Ernst et al. 2003]. In this approach, a roll-back-recover scheme is used to achieve deterministic computing by correcting errors. Specifically, a circuit is operated below the safe voltage for a given frequency of operation, and delay latches are utilized to detect circuit errors for subsequent correction. Sometimes, we have found references in literature that indicate a close affinity between these three ideas and our work. We note that these approaches are fundamentally different from our work described so far since they eventually realize reliable computations. The distinction ought to be very clear since we do not aspire or propose to develop reliable systems with the associated encumbrance and costs that go with it. Therefore, as shown in Figure 6, these prior efforts do not follow our split from von Neumann's legacy but significantly in contrast, follow that legacy.

Returning to inexact design and the use of our two principles, the field has been growing since our initial foray in 2003 and has been increasingly applied to current reliable CMOS technologies. Building on these ideas, inexact design has now been applied at the level of physical design, logic design, and at the architectural layers of abstraction. We refer the reader to Figures 7, 8, and 9 where we have listed some of the salient landmark papers that embody Principle 1 and/or 2 as an annotated bibliography. Some of the other works which build on these initial papers are summarized in Figures 10 and 11. We will now give a quick overview of the main ideas that have been developed. In doing so, we will, by necessity, use technical terms that are specific to the communities that contributed to the individual layers of abstraction.

5.1. Inexact Techniques at the Physical Layer

Most of the initial approaches to inducing erroneous behavior in return for resource savings from correctly functioning hardware were variations of *voltage overscaling*, which involved computing elements being operated at a frequency higher than the permitted level required to achieve conventional or correct execution. This enabled a trade-off between error that is deliberately introduced by *overclocking* the circuit and the energy it consumed. In Chakrapani et al. [2008], we provided a mathematically rigorous foundation for this technique by characterizing this trade-off between the energy consumed and the error and hence the quality of the solution for adders. We showed analytically that by biasing the energy investments as determined by Principle 2, our approach could be more efficient by an exponential factor in its energy consumption, over conventional uniform voltage scaling, wherein all the bits are given the same

1. Lakshmi N. Chakrapani, Bilge E. S. Akgul, Suresh Cheemalavagu, Pinar Korkmaz, Krishna V. Palem, Balasubramanian Seshasayee, **Ultra-efficient (embedded) SOC architectures based on probabilistic CMOS (PCMOS) technology**, DATE 2006. *(Principle 1).*
Advocated the idea of using inexact co-processors in an SoC architecture that compute the error-tolerant or probabilistic components of the application.
2. N Banerjee et al. **Process variation tolerant low power DCT architecture**, DATE 2007. *(Principle 1,2).*
Presents a novel DCT architecture that allows aggressive voltage scaling and trades resulting timing error and variations in process parameters for significant power savings.
3. G Karakonstantis et al., **Design methodology to trade off power, output quality and error resiliency: application to color interpolation filtering**. ICCAD 2007. *(Principle 1,2).*
Presents a novel color interpolation filter architecture in which the less important computations are susceptible to vdd-scaling and variations in exchange for significant power savings.
4. F. J. Kurdahi et al., **Error-aware design**, Euromicro Digital System Design 2007. *(Principle 1).*
Demonstrates the accuracy-energy trade-offs through voltage-overscaling in a wide variety of applications such as H.264 video decoder, JPEG 2000 image encoder & 3G wireless receiver.
5. A. K. Djahromi et al.**, Cross layer error exploitation for aggressive voltage scaling,** ISQED 2007. *(Principle 1).*
Demonstrates the accuracy-energy trade-offs through voltage-overscaling in the memories in the context of a 3GPP WCDMA modem.
6. Lakshmi N. B. Chakrapani and Krishna V. Palem, **A probabilistic boolean logic and its meaning,** Rice University, Department of Computer Science Technical Report, No. TR08-05, June 2008. *(Principle 1)*
Also appeared as - LNB Chakrapani, Krishna Palem. **A probabilistic Boolean logic for energy efficient circuit and system design**, ASPDAC 2010. *(Principle 1)*
Proposes a novel Probabilistic Boolean Logic that provides a theoretical basis for probabilistic circuit designs and derives the corresponding properties and principles that govern the applicability to design automation.
7. Lakshmi N. B. Chakrapani et al., **Highly energy and performance efficient embedded computing through approximately correct arithmetic: A mathematical foundation and preliminary experimental validation**, CASES 2008. *(Principle 1,2).*
Demonstrates the usage of Biased Voltage Scaling and the underlying mathematical foundations in the context of arithmetic adders.
8. K Palem et al., **Sustaining moore's law in embedded computing through probabilistic and approximate design: retrospects and prospects**, CASES 2009. *(Principle 1,2).*
Presents an overview of the domain of probabilistic and inexact computing and extends the application of BiVOS to larger arithmetic blocks.
9. Mohapatra et al. **Significance driven computation: a voltage-scalable, variation-aware, quality-tuning motion estimator**, ISLPED 2009. *(Principle 1,2).*
Presents a novel motion estimator architecture in which the less important computations are susceptible to vdd-scaling and variations in exchange for significant power savings.
10. Ik Joon Chang et al., **A voltage-scalable & process variation resilient hybrid SRAM architecture for MPEG-4 video processors**, DAC 2009. *(Principle 1,2).*
Presents a hybrid SRAM design with the less important data being stored in 6-T cells that are more error-prone due to vdd scaling but consumer lesser power.
11. M. A. Makhzan et al.**, A low power JPEG2000 encoder with iterative and fault tolerant error concealment,** IEEE Trans. VLSI Systems, 2009. *(Principle 1,2).*
Presents a hybrid SRAM design with the less important data being stored in 6-T cells that are more error-prone due to vdd scaling but consumer lesser power.
12. M. Lau et al., **Energy-aware probabilistic multipliers: design and analysis**, CASES 2009. *(Principle 1).*
Demonstrates power savings in JPEG2000 application by utilizing aggressive voltage overscaling at the encoder and leveraging the in-built error resiliency of the decoder.
13. Vaibhav Gupta et al., **VEDA: Variation-aware energy-efficient discrete wavelet transform architecture**, ICCD 2010. *(Principle 1, 2).*
Presents a Discrete Wavelet transform architecture in which the less important computations are susceptible to vdd-scaling and variations in exchange for significant power savings.

Fig. 7. An annotated bibliography highlighting key papers on inexact circuit design that embody Principles 1 and/or 2 (Figures 1 and 5).

level of importance independent of their significance—the computational speed or performance is the same in both cases.

Using Principle 2, voltage-overscaled circuits were used as the basis for realizing cost-accuracy trade-offs for a variety of applications, such as discrete cosine transform [Banerjee et al. 2007], traditional datapath adders [Chakrapani et al. 2008; Palem et al. 2009a], motion estimation [Mohapatra et al. 2009], image processing [Kim et al.

14. M. Lau et al., **A general mathematical model of probabilistic ripple-carry adders,** DATE 2010. (*Principle 1).*
Presents an energy assignment scheme for PCMOS-based Ripple carry adders.
15. Vinay K. Chippa et al., **Scalable effort hardware design: exploiting algorithmic resilience for energy efficiency**. DAC 2010. *(Principle 1, 2).*
Presents a cross-layer approach that finds control knobs at the algorithm, architecture and the circuit levels and identifies the possible cross-layer optimizations.
16. G Karakonstantis et al.**, HERQULES: system level cross-layer design exploration for efficient energy-quality trade-offs**. ISLPED 2010. *(Principle 1, 2).*
Another paper that presents a cross-layer approach that finds control knobs at the algorithm, architecture and the circuit levels and identifies the possible cross-layer optimizations.
17. A B. Kahng et al., **Recovery-driven design: a power minimization methodology for error-tolerant processor modules**, DAC 2010. *(Principle 1, 2).*
Proposes a design approach that optimizes a target processor module for target timing error rate where errors are caused by voltage overscaling.
18. A B. Kahng et al., **Designing a processor from the ground up to allow voltage/reliability tradeoffs**, HPCA 2010. *(Principle 1, 2).*
Proposes "soft architectures" that fail gracefully over an extended range of voltage overscaling enabled by power-aware slack redistribution.
19. S Narayanan et al., **Scalable stochastic processors**, DATE 2010. *(Principle 1, 2).*
Proposes architectures for processor modules that produce stochastically correct outputs taking into consideration the performance and power constraints.
20. Leem et al., **ERSA: Error resilient system architecture for probabilistic applications**, DATE 2010. *(Principle 1, 2).*
Demonstrates the ERSA hardware prototype that uses asymmetric reliability in a many-core architecture to run the core of the probabilistic applications from the RMS application suite.
21. Z Kedem et al., **Optimizing energy to minimize errors in dataflow graphs using approximate adders**, CASES 2010. *(Principle 1, 2).*
Proposes an optimization scheme for energy investment through biased voltage scaling in dataflow graphs of adders.
22. A Lingamneni et al., **Energy parsimonious circuit design through probabilistic pruning**, DATE 2011. *(Principle 1, 2).*
Proposes a novel technique of probabilistic pruning that "prunes" or deletes circuit components that are less significant and have less switching probability during runtime and achieves significant savings with zero hardware implementation overheads.
23. A Lingamneni et al., **Parsimonious circuits for error-tolerant applications through probabilistic logic minimization**, PATMOS 2011. *(Principle 1, 2).*
Proposes an inexact logic minimization algorithm that uses the notion of bitflips to replace less significant components in a circuit to a less resource consuming logic function.
24. V. K. Chippa et al., **Dynamic effort scaling: managing the quality-efficiency tradeoff,** DAC 2011. *(Principle 1, 2).*
Presents a cross-layer approach that finds dynamic control knobs through sensors at the algorithm, architecture and the circuit levels and identifies the possible cross-layer optimizations.
25. R Venkatesan et al, **MACACO: Modeling and analysis of circuits for approximate computing**, ICCAD 2011. *(Principle 1, 2).*
Proposes ways to systematically analyze the inexact techniques for various error metrics.
26. Ik Joon Chang et al., **A Priority-Based 6T/8T hybrid SRAM architecture for aggressive voltage scaling in video applications**, IEEE Trans. Circuits Systems for Video Technology 2011. *(Principle 1, 2).*
Presents a hybrid SRAM design with the less important data being stored in 6-T cells that are more error-prone due to vdd scaling but consumer lesser power.
27. Kedem et al., **An approach to energy-error tradeoffs in approximate ripple carry adders**, ISLPED 2011. *(Principle 1, 2).*
Proposes an optimization scheme for energy investment through biased voltage scaling in ripple carry adders.

Fig. 8. An annotated bibliography highlighting key papers on inexact circuit design that embody Principles 1 and/or 2 (Figures 1 and 5) (continued).

2009], SRAM memories [Kurdahi et al. 2010], and support vector machines [Chippa et al. 2010].

Other recent efforts involved applying voltage overscaling at the coarser granularity of processor modules, referred to as *stochastic computation* [Narayanan et al. 2010; Sartori et al. 2011]. In this approach, the following techniques were considered: *(a)* facilitating extended voltage scaling through cell upsizing on critical and frequently exercised paths while downsizing the others, and *(b)* reshaping the slack distribution

28. John Sartori et al., **Stochastic computing: embracing errors in architecture and design of processors and applications**. CASES 2011. *(Principle 1, 2).*
Presents an overview of the research on stochastic computing and its relevance in the design of processors.
29. G Karakonstantis et al., **Significance driven computation on next-generation unreliable platforms**, DAC 2011. *(Principle 1, 2).*
Explores a co-design methodology between the software and hardware to dynamically switch between exact and inexact operation mode based on the significant of the task as determined by the OS.
30. A Lingamneni et al., **Algorithmic methodologies for ultra-efficient inexact architectures for sustaining technology scaling.**, ACM Computing Frontiers 2012. *(Principle 1, 2).*
Proposes a symbiotic co-design algorithmic framework between the architecture-layer and physical-layer techniques that can realize highly energy-parsimonious inexact circuits with zero implementation overheads.
31. Heinig et al., **Classification-based improvement of application robustness and quality of service in probabilistic computer systems**, ARCS 2012. *(Principle 1, 2).*
Proposes an approach for mapping error-tolerant portions of an H.264 video decoder on the probabilistic components of a simulated ARM-based architecture.
32. Joseph Sloan et al., **On software design for stochastic processors**, DAC 2012. *(Principle 1, 2).*
Presents three approaches- analysis of application characteristics, detecting and correcting faults and transforming applications to error tolerant forms– for designing robust software for stochastic processors.
33. G Karakonstantis et al., **On the exploitation of the inherent error resilience of wireless systems under unreliable silicon**, DAC 2012. *(Principle 1, 2).*
Investigates the impact of inexactness from parametric variations and voltage overscaling on performance of wireless communication systems.

Fig. 9. An annotated bibliography highlighting key papers on inexact circuit design that embody Principles 1 and/or 2 (Figures 1 and 5) (continued).

of a circuit to create a gradual failure under voltage overscaling, as opposed to having an abrupt failure point, thus enabling a larger design space to be explored.

A dynamically reconfigurable SRAM array, called the *accuracy-aware* SRAM, was proposed for mobile multimedia applications using spatial voltage scaling. Following the ideas originally published as *Biased Voltage Scaling* (or BIVOS) from 2006 [George et al. 2006] (Principle 2), in this SRAM, the bit cells storing lower-order bits of an image pixel are operated at a lower voltage while the higher-order bits are operated at a higher or recommended voltage [Cho et al. 2009]. This SRAM architecture was shown to save 45% in terms of the power consumed, with marginal and tolerable image quality degradation. In order to facilitate more aggressive voltage overscaling, a hybrid SRAM was also proposed [Chang et al. 2009]. This included a combination of 8T bit cells for storing the more significant bits; these are more robust at overscaled supply voltages. Conventional 6T bit cells are used for storing the less significant bits. This hybrid SRAM was shown to achieve 32% savings in power over an SRAM which designed exclusively using 6T cells.

5.2. Inexact Techniques at the Architecture and Logic Layers

Canonically, voltage overscaling provides a fine-grained approach to enable energy-accuracy trade-offs, but it has associated overheads. This arises from the fact that in all of its variants, level-shifters, metastability resolution circuits, and the routing of multiple voltage lines, are all essential features. As a result, the practical utility of voltage overscaling is typically limited to few of well-characterized voltage levels. Consequently, even though voltages can be scaled at very fine granularities in principle, this opportunity is limited by the associated overheads. To circumvent these overheads, more recent inexact design approaches have considered higher levels of abstraction, such as the architecture as well as the logic layers.

Our group has pursued two novel techniques at the architectural and logic layers called *probabilistic pruning* [Lingamneni et al. 2011a, 2012b] and *probabilistic logic minimization* [Lingamneni et al. 2011b, 2013]. While the former is used to systematically *prune* or delete components and their associated wires along the paths of the

34. K. Palem et al., **Realizing ultra-low energy application specific SoC architectures through novel probabilistic CMOS (PCMOS) technology,** SSDM 2005.
35. B E S Akgul et al., **Probabilistic CMOS technology: A survey and future directions**, VLSI-SoC 2006.
36. P Korkmaz et al., **Advocating noise as an agent for ultra low-energy computing: Probabilistic CMOS devices and their characteristics**, Jap. Journal of Applied Physics 2006.
37. P Korkmaz et al., **Ultra-low energy computing with noise: energy-performance-probability Trade-offs**, ISVLSI 2006.
38. S Cheemalavagu et al., **A probabilistic CMOS switch and its realization by exploiting noise,** VLSI-Soc 2005.
39. Korkmaz et al., **Analysis of probability and energy of nanometre CMOS circuits in the presence of noise**, Electronics Letters, 2007.
40. Chakrapani et al. **Probabilistic system-on-a-chip architectures**, ACM TODAES 2007.
41. Lakshmi N. B. Chakrapani et al., **Probabilistic design: A survey of probabilistic CMOS technology and future directions for terascale IC design**, Research Trends in VLSI and Systems on Chip, Springer 2008.
42. M. A. Makhzan et al., **Architectural and algorithm level fault tolerant techniques for low power high yield multimedia devices**. ICSAMOS 2008.
43. A.K. Djahromi et al., **Cross-layer co-exploration of exploiting error resilience for video over wireless applications**, ESTImedia 2008.
44. B. Marr et al.**, Increased energy efficiency and reliability of ultra-low power arithmetic** , MWSCAS 2008.
45. N Banerjee et al., **Design methodology for low power and parametric robustness through output-quality modulation: application to color-interpolation filtering**. IEEE Trans. on CAD of Integrated Circuits and Systems 2009.
46. G Karakonstantis et al., **System level DSP synthesis using voltage overscaling, unequal error protection & adaptive quality tuning,** Signal Processing Systems 2009.
47. B. Marr et al.**, Error immune logic for low-power probabilistic computing**, Journal of VLSI Design 2010.
48. G Karakonstantis et al., **Process-variation resilient and voltage-scalable DCT architecture for robust low-Power Computing**, IEEE Trans. VLSI Systems 2010.
49. John Sartori et al., **Overscaling-friendly timing speculation architectures**, ACM Great Lakes Symposium on VLSI 2010.
50. Kahng et al., **Slack redistribution for graceful degradation under voltage overscaling**, ASPDAC 2010.
51. Nicolas Ze et al., **Optimal power/performance pipelining for error resilient processors**, ICCD 2010.
52. M. Lau et al., **Error rate prediction for probabilistic circuits with more general structures**, SASIMI 2010.
53. A. Bhanu et al., **A more precise model of noise based CMOS errors**, DELTA 2010.
54. J. George et al., **Fixed-point arithmetic on a budget: Comparing probabilistic and reduced-precision addition,** MWSCAS 2010.
55. Kulkarni et al., **Trading accuracy for power with an underdesigned multiplier architecture**, VLSI Design 2011.
56. D Shin et al., **A new circuit simplification method for error tolerant applications**. DATE 2011.
57. C. Dhoot et al., **Fault tolerant design for low power hierarchical search motion estimation algorithms**, VLSI-SoC 2011.
58. A. Singh et al., **A novel and fast method for characterizing noise based PCMOS circuits,** ASQED 2011.
59. A. Gupta et al., **Low power probabilistic floating point multiplier design,** ISVLSI 2011.
60. C. Dhoot et al, **Low power motion estimation with probabilistic computing**, ISVLSI 2011.
61. A. Singh et al., **Modeling multi-output filtering effects in PCMOS**, VLSIDAT 2011.
62. M.Cho et al., **Reconfigurable SRAM architecture with spatial voltage scaling for low power mobile multimedia applications**, IEEE Trans. On VLSI 2011.
63. V Gupta et al.**, Impact: imprecise adders for low-power approximate computing**, ISLPED 2011.
64. D Mohapatra et al., **Design of voltage-scalable meta-functions for approximate computing**. DATE 2011.
65. G Karakonstantis et al., **Voltage over-scaling: A cross-layer design perspective for energy efficient systems**, ECCTD 2011.
66. A B. Kahng et al., **Recovery-driven design: Exploiting error resilience in design of energy-efficient processors**, IEEE Trans. on CAD for Integrated Circuits and Systems, 2011.

Fig. 10. A bibliography of other important papers on inexact circuit design that embody Principles 1 and/or 2 and build on the papers from Figures 7, 8, and 9.

circuit that have a lower significance or a lower probability of being active during circuit operation or both, the latter transforms logic functions to lower-cost variants by manipulating the corresponding boolean functions. Through these approaches, we have been able to demonstrate cumulative savings of a multiplicative factor of 7.5 through the energy-delay-area product (EDAP) metric by trading the *relative error magnitude percentage* (as defined in [Lingamneni et al. 2011a]), in the context of

67. P.K. Krause et al., **Adaptive voltage over-scaling for resilient applications**, DATE 2011.
68. John Sartori et al., **Stochastic computing**, Foundations and Trends in Electronic Design Automation, 2011.
69. Sartori et al., **Architecting processors to allow voltage/reliability tradeoffs**, CASES 2011.
70. J Huang et al., **A methodology for energy-quality tradeoff using imprecise hardware,** DAC 2012.
71. A.B. Kahng et al., **Accuracy-configurable adder for approximate arithmetic designs**, DAC 2012.
72. G Karakonstantis et al., **Logic and memory design based on unequal error protection for voltage-scalable, robust and adaptive DSP systems**, Signal Processing Systems 2012.

Fig. 11. A bibliography of other important papers on inexact circuit design that embody Principles 1 and/or 2 and build on the papers from Figures 7, 8, and 9 (continued).

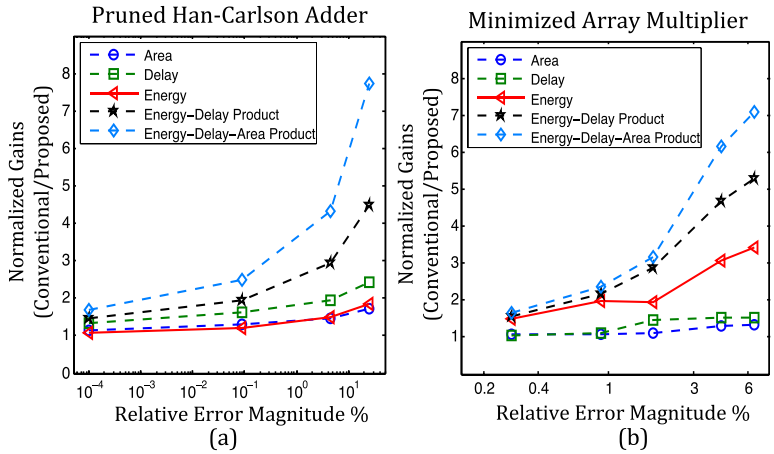

Fig. 12. (a) Results of the probabilistic pruning technique on 64-bit Han-Carlson Adder; (b) Results of probabilistic logic minimization technique on 16-bit array multiplier [Lingamneni et al. 2013].

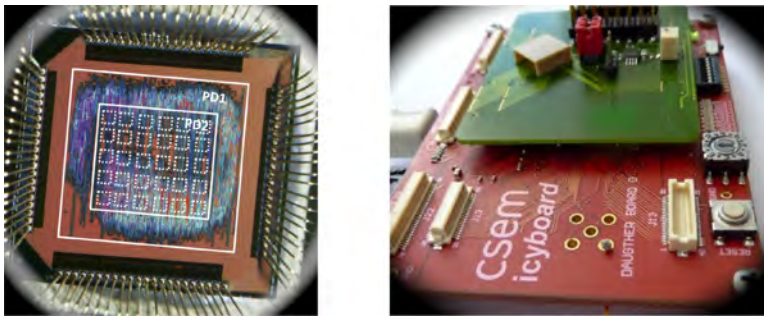

Fig. 13. (a) A die photograph of the fabricated prototype chip. (b) The prototype chip integrated into the *icy*board test platform [Lingamneni et al. 2012b].

datapath elements, such as adders and multipliers, the results for two examples are shown in Figure 12.

We also applied our pruning technique to a variety of standard 64-bit adder designs, and a prototype chip has been fabricated as a result, using TSMC 180 nm (low power) technology. A photograph of the 86-pin fabricated chip implementing our pruned adders along with its testing framework is shown in Figure 13(a). Our testing

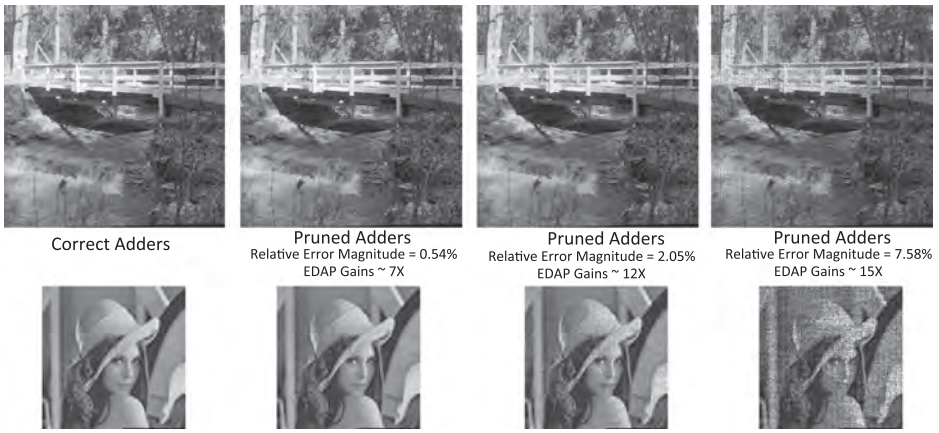

Fig. 14. Results for two sets of images after an FFT operation by correct adders and inexact pruned adders from Lingamneni et al. [2012b].

framework is based on the *icy*board [Gyger et al. 2009] platform shown in Figure 13(b) [Lingamneni et al. 2012b].

To understand the impact of error as we trade it for energy and speed gains, we also used the adders to construct a fast fourier transform (FFT) and have applied it to sample images. The corresponding output images for varying amounts of error are shown in Figure 14 [Lingamneni et al. 2012b]. As evident from these figures, the output of the inexact circuit design guided by Principle 2 [George et al. 2006] results in output images whose visual quality degrades gradually as we increase the admissible relative error magnitude to 7.58%. We note in passing that even at these high error magnitudes, visual data can remain discernible. The corresponding EDAP gains of the building blocks range between a multiplicative factor of 2 up to 8.

5.3. Cross-Layer Inexact Design Techniques

More recently, progress has been made in methods for realizing inexact circuits using methods that are not localized to single layers but rather span multiple layers of abstraction. Some of these approaches introduce inexactness at the architectural and physical layers by combining voltage overscaling and precision reduction [Chippa et al. 2010; Karakonstantis et al. 2010]. We have reported a *cross-layer co-design framework* (CCF) [Lingamneni et al. 2012b] by combining probabilistic pruning, essentially an architecture and logic-layer technique, with *confined voltage scaling*, a novel technique applied at the physical layer. Our approach allows us to vary parameters across more layers than those previously attempted, and yielded with higher EDAP gains when compared to using pruning in isolation in the context of 64-bit adders, as shown in Figure 15.

5.4. Application- and System-Level Inexact Design Techniques

A powerful extension of our Probabilistic SoC (PSOC) architecture [Chakrapani et al. 2007], named Error Resilient System Architecture (ERSA), was proposed recently [Leem et al. 2010]. This architecture has one reliable processor core with a large number of unreliable counterparts. It essentially uses a combination of three specific ideas: asymmetric reliability in a many-core architecture, error-resilient algorithms, and software optimizations. ERSA is shown to achieve resilience to higher-order bit-errors and maintains sufficient accuracy even at very high error rates with minimum impact on execution time.

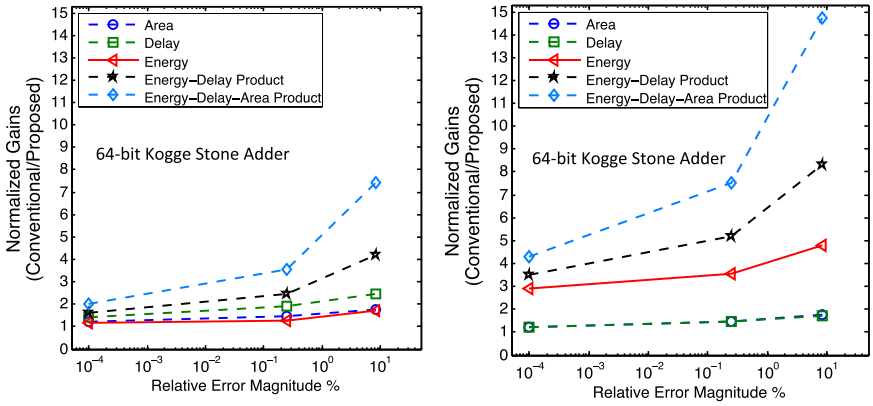

Fig. 15. A comparison of gains enabled by the architecture and logic-layer approach (probabilistic pruning) and the cross-layer co-design framework (CCF) approach for Kogge-Stone adders [Lingamneni et al. 2012b].

Recently, we proposed a classification-based approach for improvement of application robustness and quality of service in computer systems that embody error [Heinig et al. 2012]. In our approach, static analysis ensures that errors only affect operations that are error resilient and thus can be tolerated while avoiding propagation of such errors to critical operations. To evaluate, we analyzed the robustness and quality-of-service of an H.264 video decoder. Using our classification results, we mapped unreliable arithmetic operations onto erroneous components injected into a simulated ARM-based architecture, while the remaining operations used deterministically reliable components.

6. FROM INEXACT CIRCUITS TO DESIGN AUTOMATION

While holding great promise, the proliferation of inexact design to influence the broader milieu of computing is limited by the fact that all of the approaches previously reported rely on ad hoc hand designs. There is a great need for algorithmically well-characterized automated design methodologies. Also, existing design approaches were limited to particular layers of abstraction, such as physical, architectural, and algorithmic or, more broadly, software. However, in designing computing systems traditionally, it is well known that significant gains can be achieved by optimizing across the layers of abstraction through co-design. Notably, automatic algorithmic methods have been used widely across the layers of abstraction to achieve gains to great effect [De Micheli 1994]. We took a first step in addressing these two needs through our cross-layer co-design framework (CCF) for automatically designing inexact datapath elements [Lingamneni et al. 2012b]. We believe that more work remains to be done in melding the twin ideas of Electronic Design Automation (EDA) and co-design into the milieu of inexactness.

In addition, given an energy budget or error constraints, we have also proposed optimization frameworks to design and select optimal inexact circuits. The first approach [Kedem et al. 2010] is based on a method to optimally distribute a given energy budget globally among adders in a dataflow graph (with results derived for a graph representing a finite impulse response (FIR) filter and an FFT in our case), while minimizing expected errors. Our approach included new formal mathematical models and algorithms for quantitatively characterizing the relative importance of the adders (Principle 2) in a circuit. The resulting optimized energy distribution yielded a factor of 2.05 lower average error in a 16-point FFT, over the error achieved by the best possible

approach using locally optimized adders to realize the same circuit. For these and other extensions of this line of work, the readers are referred to Kedem et al. [2011].

7. SOME INTERESTING DIRECTIONS FOR FUTURE WORK

Since this is an emerging area, we will identify several directions for possible future research.

— *Resilient applications and neurobiologically guided design*. Intuitively, inexactness is only useful in domains where error is tolerable. Clearly, in audio and video—rapidly growing domains for computing—the concept of error is not as absolute since it can be associated with quality. In such domains, trading accuracy or error and quality for resource savings can therefore occur naturally. Furthermore, our own neurobiological systems play a significant information processing role in these contexts and can potentially compensate for inexactness in a variety of ways. Thus, designing inexact computing systems that take advantage of the processing done by our neurobiological pathways when the information being computed interacts with our senses—we coined the word *sensoptimzation* to capture this notion—to mitigate the impact of error is a novel and unchartered direction in digital design. This approach can benefit and can synergistically shed light on the major agenda being pursued to map the human brain.

— *Algorithmic and optimization questions*. First, the algorithmic techniques used in our co-design framework, CCF, are heuristic. Given their simplicity, this framework has yielded surprisingly good results. We believe that a rigorous characterization of pruning functions, analyzing them using methods rooted in the average case analysis of algorithms [Karp 1977], and most significantly, removing the need for simulation during each of the pruning steps would be interesting and valuable. We also believe that conventional algorithms for (computer) arithmetic and associated designs for signal processing will have to be reexamined when inexactness is permitted and will likely result in novel algorithms. Notably, widely used algorithms and implementations for arithmetic such as addition, multiplication, and the FFT obvious candidates.

— *Hardware-software co-design and synthesis for inexact applications*. There is a need for developing a co-design framework for efficiently mapping parts of an application algorithm into corresponding inexact hardware by determining its resilience. Here, the problem of characterizing and managing inexact control-flow beyond the datapath component poses challenges. Efficient software compilers and design automation tools for realizing such a mapping, both statically and at runtime, would be of significant interest. Novel hardware support in the resulting platforms would also be crucial. Finally, the entire area of logic synthesis based on probabilistic boolean logic is nascent and there is a significant need for innovative work to be done here.

— *Mixed-signal and inexact SoC design*. While most of the current research in inexact design has focused on digital systems and their memories, analog components which are quite often inexact have not been studied in the same spirit. A mixed-signal SoC integrating digital logic for computing and memories combined with analog elements would allow for even greater opportunities to exploiting the value of inexactness. Similarly, domain-specific coprocessors building on the work of Kaul et al. [2008] are also well-suited candidates for incorporating inexactness. We also believe that there is significant potential for research in extending the use of inexact computing principles to general-purpose computing by adapting the Instruction Set Architecture (ISA) as well as the microarchitectures [Hennessy and

Patterson 2003] of the modern-day microprocessors; this is largely unchartered territory with significant potential for garnering efficiencies. Here, identifying general purpose workloads as benchmarks that are resilient and can admit error is also an important direction to investigate.

— *Verification and test of inexact systems*. As inexact systems allow error, the notion of correctness is redefined in a fundamental way. Therefore, conventional verification and test algorithms and their evaluation metrics have to be redesigned and extended to allow for inexactness.

8. REMARKS

When we were invited to submit an article on the genesis of inexact design, we considered three possible approaches. The first was to present a survey of the field. An alternative was to discuss our own technical work in some depth. Upon reflection, we concluded that both of these alternatives could be easily done by compiling and perusing existing publications. Therefore, we decided to write this article in a style that does not serve either of these purposes. However, what seemed to be missing, and therefore of some value as a contribution to the scholarship of this emerging field, was to connect the emerging domain of inexact computing to its rich legacy spanning the last six decades or so. This seemed a worthwhile endeavor, especially given the powerful historical ideas that served as a basis. Here, we were guided by Longfellow's inspirational comments from his poem "A Psalm of Life" where he eloquently characterizes the legacy of great ideas and the influence of people behind them, in that their lives "... remind us, We can make our lives sublime. And, departing, leave behind us. Footprints on the sands of time." Guided by this compelling thought and given the fast and furious pace at which the CMOS world has been progressing, we felt that pausing and reflecting on the history behind topics in VLSI design and its automation wherein probability and inexactness at the hardware are playing a central role would be a worthwhile endeavor. This determined the subject of our paper and we hope to have done some justice to this goal. In passing, we note that in our usage, the word "unreliable" has the connotation of imprecision or inexactness, in the same spirit as von Neumann [von Neumann 1956]. Therefore, unreliability in our sense is not tied to its utility; an unreliable computation in this sense can still be useful.

The physical implications of CMOSdevices and their deterministic or reliable switching has been analyzed comprehensively and in depth by Meindl et al. [2010]. The reader is referred to this paper to gain a deeper understanding of the issues underlying our discussion within the context of reliable CMOSswitches. Finally, probabilistic methods have been a topic of great import and utility in the software and algorithmic domains. Starting with Monte Carlo simulations [Ulam et al. 1947] through the concepts of randomized algorithms [Rabin 1963; Solovay and Strassen 1977] to average case analysis [Karp 1977], the fruitful use of probabilistic methods and inexactness has deep and powerful roots. In contrast to these classical concepts, the ideas what we have discussed represents a novel foray of probabilistic computing concepts and inexactness based on the two principles summarized earlier into the hardware domain. We are happy to note that so far this has resulted in over seventy publications that we are aware of.

REFERENCES

Akgul, B. E. S., Chakrapani, L. N., Korkmaz, P., and Palem, K. V. 2006. Probabilistic CMOS technology: A survey and future directions. In *Proceedings of the IFIP International Conference on VLSI*. 1–6.

Banerjee, N., Karakonstantis, G., and Roy, K. 2007. Process variation tolerant low power DCT architecture. In *Proceedings of the Design, Automation and Test in Europe Conference*. 630–635.

Boltzmann, L. and Brush, S. 1995. *Lectures on Gas Theory, English translation by S.G. Brush.* Dover Publications, Mineola, NY.

Boole, G. 1847. The mathematical analysis of logic: Being an essay towards a calculus of deductive reasoning. http://www.gutenberg.org/ebooks/36884.

Chakrapani, L. N., Korkmaz, P., Akgul, B. E. S., and Palem, K. V. 2007. Probabilistic system-on-a-chip architectures. *ACM Trans. Design Autom. Electron. Syst. 12,* 3, 1–28.

Chakrapani, L. N., Muntimadugu, K. K., Lingamneni, A., George, J., and Palem, K. V. 2008. Highly energy and performance efficient embedded computing through approximately correct arithmetic: A mathematical foundation and preliminary experimental validation. In *Proceedings of the IEEE/ACM International Conference on Compilers, Architectures and Synthesis for Embedded Systems*. 187–196.

Chakrapani, L. N. B. and Palem, K. V. 2008. A probabilistic boolean logic and its meaning. Tech. rep. TR08-05, Rice University, Department of Computer Science, Houston, TX.

Chang, I. J., Mohapatra, D., and Roy, K. 2009. A voltage-scalable and process variation resilient hybrid sram architecture for mpeg-4 videoprocessors. In *Proceedings of the Design Automation Conference*. 670–675.

Cheemalavagu, S., Korkmaz, P., and Palem, K. V. 2004. Ultra low-energy computing via probabilistic algorithms and devices: CMOS device primitives akrishna v palemnd the energy-probability relationship. In *Proceedings of the International Conference on Solid State Devices and Materials*. 402–403.

Cheemalavagu, S., Korkmaz, P., Palem, K. V., Akgul, B. E. S., and Chakrapani, L. N. 2005. A probabilistic CMOS switch and its realization by exploiting noise. In *Proceedings of the IFIP International Conference on VLSI-SoC*. 452–457.

Chippa, V. K., Mohapatra, D., Raghunathan, A., Roy, K., and Chakradhar, S. T. 2010. Scalable effort hardware design: Exploiting algorithmic resilience for energy efficiency. In *Proceedings of the Design Automation Conference*. 555–560.

Cho, M., Schlessman, J., Wolf, W., and Mukhopadhyay, S. 2009. Accuracy- aware sram: A reconfigurable low power sram architecture for mobile multimedia applications. In *Proceedings of the Asia and South Pacific Design Automation Conference*. 823–828.

De Micheli, G. 1994. *Synthesis and Optimization of Digital Circuits*. McGraw-Hill, New York, NY.

Ernst, D., Kim, N. S., Das, S., Pant, S., Rao, R., Pham, T., Ziesler, C., Blaauw, D., Austin, T., Flautner, K., and Mudge, T. 2003. Razor: A low-power pipeline based on circuit-level timing speculation. In *Proceedings of the IEEE/ACM International Symposium on Microarchitecture*. 7–18.

Faggin, F. and Hoff, M. E. 1972. Standard parts and custom design merge in a four-chip processor kit. *Electron. Mag*.

Gell-Mann, M. 1997. Trying to make a reliable computer out of unreliable parts. http://www.webofstories.com/play/10585?o=MS.

George, J., Marr, B., Akgul, B. E. S., and Palem, K. V. 2006. Probabilistic arithmetic and energy efficient embedded signal processing. In *Proceedings of the IEEE/ACM International Conference on Compilers, Architectures and Synthesis for Embedded Systems*. 158–168.

Gibbs, J. 1902. *Elementary Principles in Statistical Mechanics*. Scribner, New York, NY.

Gyger, S., Corbaz, A., and Beuchat, P.-A. 2009. Hardware development kit for systems based on an icyflex processor. CSEM Scientific and Tech. rep.

Hegde, R. and Shanbhag, N. R. 1999. Energy-efficient signal processing via algorithmic noise-tolerance. In *Proceedings of the International Symposium on Low Power Electronics and Design*. 30–35.

Heinig, A., Mooney, V. J., Schmoll, F., Marwedel, P., Palem, K. V., and Engel, M. 2012. Classification-based improvement of application robustness and quality of service in probabilistic computer systems. In *Proceedings of the Architecture of Computing Systems Conference*. 1–12.

Hennessy, J. L. and Patterson, D. A. 2003. *Computer Architecture: A Quantitative Approach*. Morgan Kaufmann Publishers Inc., Burlington, MA.

Jonietz, E. 2008. Probabilistic chips. *Technol. Rev.* MIT, Cambridge, MA. http://www.technologyreview.com/energy/20246/.

Karakonstantis, G., Panagopoulos, G., and Roy, K. 2010. Herqules: System level cross-layer design exploration for efficient energy-quality trade-offs. In *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*. 117–122.

Karp, R. 1977. Probabilistic analysis of partitioning algorithms for the traveling-salesman problem in the plane. *Math. Oper. Res. 2,* 3, 209–224.

Kaul, H., Anders, M., Mathew, S., Hsu, S., Agarwal, A., Krishnamurthy, R., and Borkar, S. 2008. A 320 mv 56 μw 411 gops/watt ultra-low voltage motion estimation accelerator in 65 nm cmos. *IEEE J. Solid-State Circuits*.

Kedem, Z. M., Mooney, V. J., Muntimadugu, K. K., Palem, K. V., Devarasetty, A., and Parasuramuni, P. D. 2010. Optimizing energy to minimize errors in dataflow graphs using approximate adders. In *Proceedings of the International Conference on Compilers, Architectures and Synthesis for Embedded Systems*. 177–186.

Kedem, Z. M., Mooney, V. J., Muntimadugu, K. K., and Palem, K. V. 2011. An approach to energy-error tradeoffs in approximate ripple carry adders. In *Proceedings of the International Symposium on Low Power Electronics and Design*. 211–216.

Kim, S. H., Mukohopadhyay, S., and Wolf, W. 2009. Experimental analysis of sequence dependence on energy saving for error tolerant image processing. In *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*.

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

Korkmaz, P. 2007. Probabilistic cmos (PCMOS) in the nanoelectronics regime. Ph.D. thesis, Georgia Institute of Technology, Atlante, GA.

Korkmaz, P., Akgul, B. E. S., Palem, K. V., and Chakrapani, L. N. 2006. Advocating noise as an agent for ultra low-energy computing: Probabilistic CMOS devices and their characteristics. *Japan. J. App. Physics 45,* 4B, 3307–3316.

Kurdahi, F. J., Eltawil, A., Yi, K., Cheng, S., and Khajeh, A. 2010. Low-power multimedia system design by aggressive voltage scaling. *IEEE Trans. VLSI Syst. 18,* 5, 852–856.

Landauer, R. 1961. Irreversibility and heat generation in the computing process. *IBM J. Res. Develop. 3*. 183–191.

Leem, L., Cho, H., Bau, J., Jacobson, Q., and Mitra, S. 2010. Ersa: Error resilient system architecture for probabilistic applications. In *Proceedings of the Design, Automation and Test in Europe Conference*. 1560–1565.

Lingamneni, A., Enz, C. C., Nagel, J.-L., Palem, K. V., and Piguet, C. 2011a. Energy parsimonious circuit design through probabilistic pruning. In *Proceedings of the Design, Automation and Test in Europe Conference*. 764–769.

Lingamneni, A., Enz, C. C., Palem, K. V., and Piguet, C. 2011b. Parsimonious circuit design for error-tolerant applications through probabilistic logic minimization. In *Proceedings of the International Workshop on Power And Timing Modeling, Optimization and Simulation*. 204–213.

Lingamneni, A., Muntimadugu, K. K., Enz, C. C., Karp, R. M., Palem, K. V., and Piguet, C. 2012. Algorithmic methodologies for ultra-efficient inexact architectures for sustaining technology scaling. In *Proceedings of the ACM International Conference on Computing Frontiers*.

Lingamneni, A., Enz, C. C., Palem, K. V., and Piguet, C. 2013. Synthesizing parsimonious inexact circuits through probabilistic design techniques. *ACM Trans. Embed. Comput. Syst. (spl. issue on Probabilistic Embedded Computing)*.

Ludwig, J., Nawab, S., and Chandrakasan, A. 1995. Low power filtering using approximate processing for DSP applications. In *Proceedings of the IEEE Custom Integrated Circuits Conference*. 185–188.

Mauchly, J. and Eckert, J. 1947. Electrical numerical integrator and calculator. U.S. Patent 3120606, filed June 26, 1947 and issued Feb 4, 1964.

Maxwell, J. C. 1871. *Theory of Heat*, Greenword Press, Westport, CT.

McCulloch, W. and Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophysics 7*, 115–133.

Meindl, J., Naeemi, A., Bakir, M., and Murali, R. 2010. Nanoelectronics in retrospect, prospect and principle. In *Proceedings of the IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*. 31–35.

Minsky, M. 1967. *Computation: Finite and Infinite Machines.* Prentice-Hall, Inc., Englewood Cliffs, NJ.

Mohapatra, D., Karakonstantis, G., and Roy, K. 2009. Significance driven computation: A voltage-scalable, variation-aware, quality-tuning motion estimator. In *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*. 195–200.

Moore, E. and Shannon, C. 1956a. Reliable circuits using less reliable relays I. *J. Franklin Instit. 262*, 191–208.

Moore, E. and Shannon, C. 1956b. Reliable circuits using less reliable relays II. *J. Franklin Instit. 262*, 281–297.

Moore, G. E. 1965. Cramming more components onto integrated circuits. *Electron. Mag. 38,* 8.

Mudge, T. 2001. Power: A first-class architectural design constraint. *Computer 34,* 4, 52–57.

Narayanan, S., Sartori, J., Kumar, R., and Jones, D. L. 2010. Scalable stochastic processors. In *Proceedings of the Design, Automation and Test in Europe Conference*. 335–338.

Nikolic, K., Sadek, A., and Forshaw, M. 2001. Architectures for reliable computing with unreliable nanodevices. In *Proceedings of the IEEE Conference on Nanotechnology*. 254–259.

Palem, K. V. 2003a. Energy aware algorithm design via probabilistic computing: From algorithms and models to Moore's law and novel (semiconductor) devices. In *Proceedings of the International Conference on Compilers, Architectures and Synthesis for Embedded Systems*. 113–116.

Palem, K. V. 2003b. Proof as experiment: Probabilistic algorithms from a thermodynamic perspective. In *Proceedings of the International Symposium on Verification (Theory and Practice)*.

Palem, K. V. 2005. Energy aware computing through probabilistic switching: A study of limits. *IEEE Trans. Comput. 54,* 9, 1123–1137 (abridged form appeared as K.V. Palem, Energy aware computing through randomized switching, Tech. rep. GIT–CC–03–16, Georgia Inst. of Technology.

Palem, K. V., Cheemalavagu, S., Korkmaz, P., and Akgul, B. E. 2005. Probabilistic and introverted switching to conserve energy in a digital system. U.S. Patent 20050240787, filed April 27, 2005 and issued October 27, 2005.

Palem, K. V., Akgul, B. E. S., and George, J. 2006. Variable scaling for computing elements. *Invention Disclosure*.

Palem, K. V., Chakrapani, L. N., Kedem, Z. M., Lingamneni, A., and Muntimadugu, K. K. 2009a. Sustaining moore's law in embedded computing through probabilistic and approximate design: Retrospects and prospects. In *Proceedings of the International Conference on Compilers, Architectures and Synthesis for Embedded Systems*. 1–10.

Palem, K. V., Korkmaz, P., Yeo, K.-S., and Kong, Z.-H. 2009b. Probabilistic cmos (pcmos) logic for nanoscale circuit design. In *Proceedings of the International Solid State Circuits Conference: Advanced Solid-State Circuits Forum*.

Palem, K. V. and Lingamneni, A. 2012. What to do about the end of moore's law, probably! In *Proceedings of the 49th Design Automation Conference*.

Rabin, M. O. 1963. Probabilistic automata. *Inform. Control 6*, 230–245.

Randall, V. A. 2006. A lost interview with ENIAC co-inventor *J. Presper Eckert*. http://www.computerworld.com. April 2011.

Sartori, J., Sloan, J., and Kumar, R. 2011. Stochastic computing: Embracing errors in architecture and design of processors and applications. In *Proceedings of the International Conference on Compilers, Architectures and Synthesis for Embedded Systems*, 135–144.

Shannon, C. 1937. A symbolic analysis of relay and switching circuits. M.S. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Solovay, R. and Strassen, V. 1977. A fast monte-carlo test for primality. *SIAM J. Comput.*

Szilard, L. 1929. Reduction in entropy of a thermodynamic system caused by the interference of intelligent beings. *Z. Physik 53*, 840–856.

Turing, A. M. 1936. On computable numbers, with an application to the entscheidungsproblem. In *Proceedings of the London Mathematical Society 42*, 230–265.

Ulam, S., Richtmyer, R. D., and von Neumann, J. 1947. Statistical methods in neutron diffusion. Los Alamos Scientific Laboratory report LAMS551.

von Neumann, J. 1956. Probabilistic logics and the synthesis of reliable organisms from unreliable components. In *Automata Studies*, C. E. Shannon and J. McCarthy Eds., Priceton Univ. Press, Princeton, N.J.

Zuse, K. 1993. *Der Computer. Mein Lebenswerk* 3rd Ed. Number 978-3-540-56292-4. Springer-Verlag.

Received May 2012; revised January 2013; accepted January 2013