Energy-Efficient Circuit Design

Antonios Antoniadis* Neal Barcelo† Michael Nugent‡ Kirk Pruhs§
Michele Scquizzato¶

Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260, USA

ABSTRACT

We initiate the theoretical investigation of energy-efficient circuit design. We assume that the circuit design specifies the circuit layout as well as the supply voltages for the gates. To obtain maximum energy efficiency, the circuit design must balance the conflicting demands of minimizing the energy used per gate, and minimizing the number of gates in the circuit; If the energy supplied to the gates is small, then functional failures are likely, necessitating a circuit layout that is more fault-tolerant, and thus that has more gates. By leveraging previous work on fault-tolerant circuit design, we show general upper and lower bounds on the amount of energy required by a circuit to compute a given relation. We show that some circuits would be asymptotically more energy-efficient if heterogeneous supply voltages were allowed, and show that for some circuits the most energy-efficient supply voltages are homogeneous over all gates.

Categories and Subject Descriptors

F.0 [**Theory of Computation**]: General

General Terms

Theory

*Supported by a fellowship within the Postdoc-Programme of the German Academic Exchange Service (DAAD). E-mail: antonios@pitt.edu.

†This material is based upon work supported by the National Science Foundation Graduate Research Fellowship. E-mail: ncb30@cs.pitt.edu.

‡E-mail: mnugent@cs.pitt.edu.

§Supported in part by NSF grants CCF-1115575, CNS-1253218, and an IBM Faculty Award. E-mail: kirk@cs.pitt.edu.

¶Work supported, in part, by a fellowship of "Fondazione Ing. Aldo Gini", University of Padova, Italy. E-mail: scquizza@pitt.edu.

ITCS'14, January 12–14, 2014, Princeton, New Jersey, USA.

ACM 978-1-4503-2698-8/14/01 ...$15.00.
http://dx.doi.org/10.1145/2554797.2554826.

Keywords

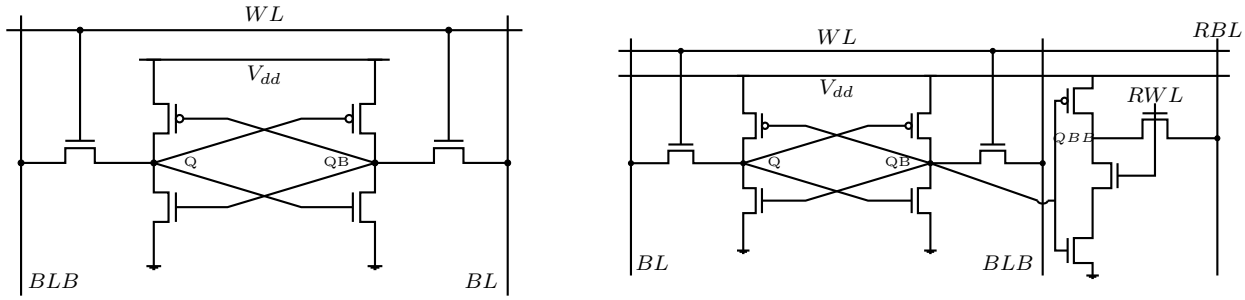
Energy Efficiency; Circuit Design; Near-Threshold Computing

1. INTRODUCTION

The number of transistors per unit volume on a chip continues to double about every two years. However, about a decade ago chip makers hit a thermal wall as the cost of cooling chips with these transistor densities became prohibitive. This has resulted in Moore's gap, namely that increased transistor density no longer directly translates into a similar increase in performance, and in energy becoming the first order design constraint in CMOS-based technologies.

One promising technique to attain more energy-efficient circuits is *Near-Threshold Computing* (NTC). The threshold voltage of a transistor is the minimum voltage at which the transistor starts to conduct current, around 0.2-0.3V for modern processors. Of course, even for identically-designed transistors, there can be variations in the actual threshold voltage due to manufacturing variations; And even for the same transistor, the actual threshold voltage will vary with environmental conditions. Further, actual supply voltages may differ from the designed voltage due to manufacturing and environmental variance. Thus if the designed supply voltage was exactly the ideal threshold voltage, some transistors would likely fail to conduct current as designed. For example, for a typical 65 nm SRAM circuit, halving the supply voltage from the nominal level to 0.5V typically increases the failure rate by about 5 orders of magnitude [6]. The traditional design approach to achieving fault tolerance is to set the supply voltage to be sufficiently high so that with sufficiently high probability no transistor fails. Near-Threshold Computing simply means that the supply voltages are designed to be closer to the threshold voltage. As the power used by a transistor is roughly proportional to the square of the supply voltage [2], Near-Threshold Computing can potentially offer orders of magnitude improvements in energy efficiency, provided another, more energy-efficient, solution for the fault-tolerance issue can be found.

One strategy to achieve fault-tolerance is to design fault-tolerant circuits, namely circuits that correctly compute the desired output if the number of failures is not significantly higher than the expected number of failures. The study of fault-tolerant circuits is not new. Starting with the seminal paper by von Neumann [14], several papers [4, 5, 10, 11, 12, 8, 7] have considered the question of how many faulty gates, each (independently) having a (small) fixed probability of failure, are required to mimic the computation of



(a) Standard 6-transistor design.

(b) A more fault-tolerant 10-transistor design from [3].

Figure 1: Two SRAM circuits with the same functionality.

an ideal circuit with some desired probability of correctness. In general, as the probability of gate failure increases one would expect that more gates will be required to achieve a fixed probability of failure for the circuit. As an example from [6], the circuit shown in Figure 1a is the traditional 6-transistor design for an SRAM cell, while the circuit shown in Figure 1b is a more fault-tolerant, and thus more suited for Near-Threshold Computing, 10-transistor design for an SRAM cell.

Our goal here is to initiate the study of the design of energy-efficient circuits. We assume that the design of the circuit specifies both the circuit layout as well as the supply voltages for the gates. To obtain maximum energy efficiency, the circuit design must balance the conflicting demands of minimizing the energy used per gate, and minimizing the number of gates in the circuit; If the energy supplied to the gates is small, then functional failures are likely, necessitating a circuit layout that is more fault-tolerant, and thus that has more gates. Thus the design should find a "sweet spot" for the supply voltages that balances the competing demands of small circuit size and low per-gate energy. In Section 2 we formalize this in the natural way.

The paper [6] gives an excellent survey on Near-Threshold Computing. As an example of a technology that has at least the spirit of Near-Threshold Computing, IBM's production POWER7 servers use a technique called *Guardband* to save energy by dynamically lowering operating voltage [1].

1.1 Our Contribution

We give four main results, which we discuss in the following four paragraphs.

We show in Section 3 that the classic lower bound on circuit size from the fault-tolerant circuit literature can be extended to give a lower bound for energy. Gács and Gál [8], and independently Reischuk and Schmeltz [12], show that any Boolean function f with sensitivity s (roughly the number of input bits which affect the output) requires a circuit of size $\Omega(s \log s)$ to be reliably computed when the gates of the circuit fail independently with a fixed positive probability. We modify the techniques in [8] to prove an $\Omega\left(s \log\left(s(1-2\sqrt{\delta})/\delta\right)\right)$ lower bound on the energy required by any circuit that computes a relation with sensitivity s correctly with probability at least $(1-\delta)$.

We show in Section 4 that the classic upper bound on circuit size from the fault-tolerant circuit literature can be extended to give an upper bound on circuit energy consumption. Von Neumann [14] showed that given a Boolean function f and a circuit of size c which computes f, a circuit of size $O(c \log c)$ is sufficient for computing f correctly with high probability. Using techniques from [14] and from [10, 7], we show that a relation h that is computable by a circuit of size c can, with probability at least $(1-\delta)$, be computed by a circuit of faulty gates using $O(c \log(c/\delta))$ energy. In our construction, the supply voltages are homogeneous (that is, the same for all gates).

While it may not currently be practical, in principle the supply voltages need not be homogeneous over all gates, that is, different gates could be supplied with different voltages. This naturally leads to the question of whether allowing heterogeneous supply voltages might yield lower-energy circuits than is possible if the supply voltages are required to be homogeneous. In Section 5 we observe that there are relations, namely the parity function, for which allowing heterogeneous supply voltages will not allow one to achieve a circuit design that uses asymptotically less energy than is achievable with a circuit design with homogeneous supply voltages. Intuitively, the parity function has such high sensitivity that every gate in any reasonable circuit will be of equal importance, so nothing can be gained by heterogeneous supply voltages. Formally, the proof is essentially a corollary of our lower bounds from Section 3.

In contrast, in Section 6 we give a natural example where allowing heterogeneous supply voltages allows one to use asymptotically less energy than would be achievable using homogeneous supply voltages. In particular, we consider a natural super-majority relation, which outputs the majority of the input bits if this majority is sufficiently large, and the most natural circuit that computes this relation, a balanced tree of majority gates. We show that for homogeneous supply voltages the energy required by this circuit to compute this relation is $\Omega(nP(\delta))$, where n is the number of input bits, δ is the maximum probability of failure for the circuit that one is willing to tolerate, and $P(x)$ is the power required in order for one majority gate to fail with probability at most x. We then show that if supply voltages can be heterogeneous then this circuit can compute this relation using energy $O(n + 3^{1/\delta}P(\delta/10))$, which is asymptotically less than $nP(\delta)$ for functions P observed in current circuits. Intuitively, the gates closer to the output are more important, and one can obtain a greater increase in probability of correctness by investing more energy in these gates. In some sense this is our main result as the proof is not based on proofs in the fault-tolerant circuit literature.

2. PROBLEM DESCRIPTION AND NOTATION

We now formally define the problem. A *Boolean relation* h is a map from $\{0,1\}^n$ to $\{0,1\}$, where each input is mapped to 0, 1, or both 0 and 1. If x is mapped to both 0 and 1, this can be thought of as "don't care" (for example because the input x should not occur in a correctly functioning system). Note that a *Boolean function* is a Boolean relation where each input is uniquely mapped to either 0 or 1. For any input $x \in \{0,1\}^n$, denote by x^ℓ the input that has the same bits as x, except for the ℓ-th bit, which is flipped. A Boolean relation h is *sensitive* on the ℓ-th bit of x if neither $h(x)$ nor $h(x^\ell)$ is mapped to both 0 and 1, and $h(x) \neq h(x^\ell)$. The *sensitivity of h on x* is the number of bits of x that h is sensitive on. The *sensitivity of h* is the maximum over all x of the sensitivity of h on x.

A *gate* is a function $g : \{0,1\}^{n_g} \to \{0,1\}$, where n_g is the number of inputs (i.e., the *fan-in*) of the gate. We say that a gate *fails* when it produces an incorrect output, that is, when given an input x it produces an output other than $g(x)$. Every gate has some independent probability of failure ϵ. A gate that never fails is said to be *reliable*. A *Boolean circuit* C with n inputs is a directed acyclic graph in which every node is a gate. Among them there are n gates with fan-in zero, each of which outputs one of the n inputs of the circuit, i.e., the input gates are assumed to be reliable. One gate is designated as the output gate, which has out-degree zero. The *size* of a circuit is the number of gates it contains. Any Boolean relation can be represented by a Boolean circuit. Given a value $\delta \in (0, 1/2)$ (δ may not be constant), a circuit that computes a Boolean relation h is said to be $(1-\delta)$-*reliable* if for every input x on which $h(x)$ is not both 0 and 1 it outputs $h(x)$ with probability at least $1-\delta$. We say that a circuit is *reliable* if it is 1-reliable (for example, because all its gates are reliable).

Every gate g is supplied with a voltage v_g. We say that the supply voltages are *homogeneous* when every gate of the circuit is supplied with the same voltage, and *heterogeneous* otherwise. A voltage-to-failure function $\epsilon(v) : \mathbb{R}^+ \to (0, 1/2)$ maps a supply voltage to a probability of functional failure, that is, the probability that the gate fails when supplied with voltage v. For convenience, we refer to the probability of failure with ϵ when the supply voltage is implied by the context. Hence, every gate has a probability of failure of $\epsilon(v_g)$ which, to lighten notation, we denote as ϵ_g. There is also a voltage-to-energy function $E(v)$ mapping the supply voltage to the energy used by a gate with that supply voltage. The energy required by a circuit C is simply the aggregate energy used by the gates, $\sum_{g \in C} E(v_g)$ in our notation. For convenience, we define a failure-to-energy function $P(q) := E(\epsilon^{-1}(q))$, where ϵ^{-1} denotes the inverse of the function ϵ. Thus the energy of a circuit C can be rewritten as $\sum_{g \in C} P(\epsilon_g)$.

By observing a semi-log plot of voltage-to-failure for a current 65nm SRAM cell from [6] we can see that the relationship between voltage and the log of the failure is approximately linear. Hence, the error as a function of voltage is of the form of $\epsilon(v) = c^{-v}$, for some positive constant c. Thus using the fact that the energy is proportional to the square of the supply voltage, we conclude that the failure-to-energy function for a 65nm SRAM cell is approximately $P(\epsilon) = \Theta(\log^2(1/\epsilon))$.

3. A GENERAL ENERGY LOWER BOUND

Our main goal in this section is to prove Theorem 1, which roughly states that $\Omega(s \log s)$ energy is necessary to compute a relation with sensitivity s.

THEOREM 1. *Let $\delta < 1/4$, and let C be a circuit that $(1-\delta)$-reliably computes a relation h of sensitivity s. If each gate g of C fails independently with probability ϵ_g, and incurs an energy consumption of $P(\epsilon_g)$, with P being a proper failure-to-energy function, then C requires*

$$\Omega\left(s \log\left(s \frac{1-2\sqrt{\delta}}{\delta}\right)\right)$$

energy in order to $(1-\delta)$-reliably compute h.

The outline of this section is as follows. First, we define proper failure-to-energy functions (Definition 2), and discuss why proper functions are natural. Then, similarly to [8], we show how to translate our problem to an equivalent problem where the failures occur not only on gates, but on wires as well. This is formalized in Statement 3, which is implied by the proof of Lemma 3.1 in [4], and is also used in [8]. Lemma 6 then gives a lower bound on the energy necessary for $(1-\delta)$-reliable circuits within this new model with wire failures. The proof is based on the proof of Theorem 3.1 in [8], and uses a series of inequalities that relate the probability of an input being incorrectly transmitted to the probability of the circuit being incorrect. Using this, we can write the problem as a single-variable optimization problem and use standard techniques to give the desired lower bound. Finally, to prove Theorem 1 we show that given a $(1-\delta)$-reliable circuit C in our original model without wire failures, we can create a $(1-\delta)$-reliable circuit C' in the new model with wire failures, where the energy consumptions of C and C' differ only by a constant.

DEFINITION 2. *A failure-to-energy function P is called* proper *when it satisfies the four following properties:*

1. *P is nonincreasing,*
2. *$\lim_{\epsilon \to 0^+} P(\epsilon)/(\log 1/\epsilon) > 0$,*
3. *$\lim_{\epsilon \to 1/2^-} P(\epsilon) > 0$,*
4. *$P(\epsilon_1) + P(\epsilon_2) \geq 2P(\sqrt{\epsilon_1 \epsilon_2})$ for all $\epsilon_1, \epsilon_2 \in (0, 1/2)$.*

The first and third restrictions are natural, since they just require that the energy used decreases, but never becomes zero, as the probability of failure of a gate increases. The second property states that the energy must increase "quickly enough" as the probability of a gate's failure tends to 0, which is necessary in order to have any energy saving over gates that never (or almost never) fail. The last property provides a convexity constraint on the function P. We point out that failure-to-energy functions typically observed in real gates fall within this class of proper failure-to-energy functions [9, 6].

STATEMENT 3 ([4]). *Let g be a gate with fan-in n_g, in a circuit C where both gates and wires may fail. Furthermore, let $\epsilon \in (0, 1/2), \zeta_g \in [0, \epsilon/n_g]$ and let $g(t)$ be the output of gate g assuming that its input-wires receive input t, and both g and g's input-wires are reliable. Then there exists a unique value $\eta_g(y, \zeta_g) \in [0, 1]$ such that if*

- *the input wires of g fail independently with probability ζ_g, and*
- *gate g fails with probability $\eta_g(y, \zeta_g)$ when the gate receives input y,*

then the probability that g does not output $g(t)$ is equal to ϵ.

Note that in Statement 3, since we can now have failures on wires, the input y received by a gate g may be different than the input t received by the corresponding wires.

We need the following definition and technical lemma.

DEFINITION 4. *Given $x_{1,1}, x_{1,2}, \ldots, x_{1,n} \in \mathbb{R}$, we recursively define a sequence of numbers as follows. Let $m_j^u = \arg\max_i x_{j,i}$ and $m_j^l = \arg\min_i x_{j,i}$. Then, for all $i \neq \{m_j^u, m_j^l\}$, let $x_{(j+1),i} = x_{j,i}$, and let $x_{(j+1),m_j^u} = x_{(j+1),m_j^l} = \sqrt{x_{j,m_j^u} x_{j,m_j^l}}$.*

LEMMA 5. *Let $a_1, a_2, \ldots$ be a sequence of numbers such that $a_j = x_{j,m_j^u} - x_{j,m_j^l}$, with the terms x_{j,m_j^u} and x_{j,m_j^l} as defined above. Then,*

$$\lim_{j\to\infty} a_j = 0.$$

LEMMA 6. *Let f be a proper failure-to-energy function, and let C be a circuit that $(1-\delta)$-reliably computes a relation h of sensitivity s. If (i) each gate g of C fails independently with probability $\eta_g(y, \zeta_g)$ when receiving input y, (ii) g incurs an energy consumption of zero, and (iii) each wire i entering g fails independently with probability $\zeta_g \in (0, 1/4)$ and incurs an energy usage of $f(\zeta_g)$, then C requires*

$$\Omega\left(s \log\left(s\frac{1-2\sqrt{\delta}}{\delta}\right)\right)$$

energy in order to $(1-\delta)$-reliably compute h.

PROOF. We start by rephrasing our problem after borrowing a constraint on the number of wires and some notation from [8]. Specifically, let z be an input such that h has maximum sensitivity on z. Let $S \subset \{1, 2, \ldots, n\}$ be the set of indexes so that $\ell \in S$ if and only if h is sensitive to the ℓ-th bit on input z. Then $|S| = s$, where s is the sensitivity of h. For each $\ell \in S$ denote by B_ℓ the set of all wires originating from the ℓ-th input of the circuit. Let $m_\ell = |B_\ell|$. For any set $\beta \subset B_\ell$, let $H(\beta)$ be the event that the wires belonging to β fail and the other wires of B_ℓ are correct. Denote by β_ℓ the subset of B_ℓ where

$$\max_{\beta \subset B_\ell} \Pr[C(z^\ell) = h(z^\ell) \text{ s.t. } H(\beta)]$$

is obtained, where $C(z^\ell)$ is a random variable for the output of the circuit given input z^ℓ. Finally, let $H_\ell = H(B_\ell \setminus \beta_\ell)$. Note that since wires can now fail with different probabilities, we have that,

$$\Pr[H_\ell] = \prod_{i\in\beta_\ell} (1-\zeta_i) \prod_{i\notin\beta_\ell} \zeta_i \geq \prod_{i\in B_\ell} \zeta_i.$$

It follows from Inequalities (5) and (6) of [8] that

$$\frac{\delta}{1-2\sqrt{\delta}} \geq \sum_{\ell\in S} \prod_{i\in B_\ell} \zeta_i$$

and as in [8], using the inequality of arithmetic and geometric means, we have

$$\frac{\delta}{1-2\sqrt{\delta}} \geq s \left(\prod_{\ell\in S, i\in B_\ell} \zeta_i\right)^{1/s}.$$

Rewriting this to isolate the product term, we have

$$\prod_{\ell\in S, i\in B_\ell} \zeta_i \leq \left(\frac{\delta}{s(1-2\sqrt{\delta})}\right)^s.$$

Therefore, minimizing the energy consumption, is equivalent to the following optimization problem,

$$\begin{aligned} \text{minimize} \quad & \sum_{\ell\in S, i\in B_\ell} f(\zeta_i) \\ \text{subject to} \quad & \prod_{\ell\in S, i\in B_\ell} \zeta_i \leq \left(\frac{\delta}{s(1-2\sqrt{\delta})}\right)^s. \end{aligned}$$

Now, take some feasible solution ζ^* to the above optimization problem. Let ζ_1^* and ζ_2^* denote the minimum and maximum ζ_i^* respectively, and M denote the total number of wires, i.e., $M = \sum_{\ell\in S} m_\ell$. Note that since we assume that $f(p_1) + f(p_2) \geq 2f(\sqrt{p_1 p_2})$ for all $p_1, p_2 \in (0, 1/2)$, we can set $\zeta_1^* = \zeta_2^* = \sqrt{\zeta_1^* \zeta_2^*}$, without increasing the value of the objective, and further the constraint remains feasible. By Lemma 5 this process, if repeated, will converge to a solution where all ζ_i are equal. Therefore, we can rewrite the optimization problem as

$$\begin{aligned} \text{minimize} \quad & Mf(x) \\ \text{subject to} \quad & x^M \leq \left(\frac{\delta}{s(1-2\sqrt{\delta})}\right)^s. \end{aligned}$$

Isolating M in the constraint above, the problem is equivalent to that of minimizing

$$\left(\frac{s}{\log 1/x} \log\left(s\frac{1-2\sqrt{\delta}}{\delta}\right)\right) f(x).$$

Since the function satisfies properties 1, 2, and 3 of Definition 2, the above expression will be minimized either at some constant $x \in (0, 1/4)$, in which case $f(x)/\log(1/x) > 0$, or in the limit as x approaches 0, in which case

$$\lim_{x\to 0^+} f(x)/\log(1/x) > 0,$$

or in the limit as x approaches $1/4$, in which case

$$f(x)/\log(1/x) > 0.$$

The lemma follows. □

We are now ready to prove Theorem 1.

PROOF OF THEOREM 1. We start by constructing a new circuit C' for computing h, which is identical to C except that both wires and gates may fail, wires of C' incur some non-zero energy consumption (as a function of their probability of failure), and the gates in C' do not consume energy. First we argue that this can be done such that C' is $(1-\delta)$-reliable. Observe that if for each wire i entering gate g we set its probability of failure to $\zeta_g = \epsilon_g/n_g$, we can apply

Statement 3 and set the failure probability on gate g when receiving input y to $\eta_g(y, \zeta_g)$. The result is that when the input wires of gate g in C' receive input t, the probability that g does not output $g(t)$ is ϵ_g (the same as the probability of failure of g in the original circuit C). Thus by setting these failure probabilities for each gate and wire in C' we have that, for any input x, C and C' output $h(x)$ with the same probability, and so C' is $(1-\delta)$-reliable.

Now we set the energy consumption of the wires such that the energy of C' is at most the energy of C. First observe that if for each gate g we set the failure-to-energy function of the wires that are inputs to g to be $\tilde{P}_g(\zeta) = P(n_g \cdot \zeta)/n_g$, then since $\zeta_g = \epsilon_g/n_g$, the total energy of the wires entering g would be $n_g \tilde{P}_g(\zeta_g) = P(\epsilon_g)$ and the energy of C and C' would be equal. However, to apply Lemma 6, all wires must have the same failure-to-energy function. Therefore, let n_g^* be the maximum fan-in of any gate of C, i.e., $n_g^* = \max_{g \in C} n_g$. We set the failure-to-energy function of all wires in C' to be

$$\tilde{P}(\zeta) = \begin{cases} P(n_g^* \cdot \zeta)/n_g^* & \text{if } \zeta < \frac{1}{2n_g^*}, \\ \lim_{\epsilon \to 1/2^-} P(\epsilon)/n_g^* & \text{if } \zeta \geq \frac{1}{2n_g^*}. \end{cases}$$

First observe that $\tilde{P}_g(\zeta) \geq \tilde{P}(\zeta)$ for all $\zeta \in (0, 1/2)$ since P is nonincreasing so $P(n_g\zeta) \geq P(n_g^*\zeta)$. This implies that the energy of C' is at most the energy of C.

In order to apply Lemma 6, we need to verify that $\tilde{P}$ is a proper failure-to-energy function. The first property follows directly from the definition of $\tilde{P}$. For the second property, observe that

$$\begin{aligned} \lim_{\zeta \to 0^+} \frac{\tilde{P}(\zeta)}{\log\left(\frac{1}{\zeta}\right)} &= \frac{1}{n_g^*} \lim_{\zeta \to 0^+} \frac{P(n_g^*\zeta)}{\log\left(\frac{1}{n_g^*\zeta}\right)} \cdot \lim_{\zeta \to 0^+} \frac{\log\left(\frac{1}{n_g^*\zeta}\right)}{\log\left(\frac{1}{\zeta}\right)} \\ &= \lim_{\zeta \to 0^+} \frac{P(n_g^*\zeta)}{\log\left(\frac{1}{n_g^*\zeta}\right)} > 0. \end{aligned}$$

The third property follows from the fact that

$$\lim_{\zeta \to 1/2^-} \tilde{P}(\zeta) = \lim_{\epsilon \to 1/2^-} P(\epsilon)/n_g^* > 0,$$

where we exploited the definition of $\tilde{P}$ and the fact that, by hypothesis, P is a proper failure-to-energy function. For the fourth property, let $\zeta_1, \zeta_2 \in (0, 1/2)$, and, w.l.o.g., $\zeta_1 < \zeta_2$. There are four cases, depending on the relationship between ζ_1, ζ_2, and n_g^*. When $\zeta_1 < \zeta_2 < 1/2n_g^*$, by applying the definition of $\tilde{P}$ and since P by hypothesis is a proper failure-to-energy function, we have

$$\begin{aligned} \tilde{P}(\zeta_1) + \tilde{P}(\zeta_2) &= \frac{P(n_g^*\zeta_1)}{n_g^*} + \frac{P(n_g^*\zeta_2)}{n_g^*} \\ &\geq 2\frac{P\left(\sqrt{n_g^*\zeta_1 n_g^*\zeta_2}\right)}{n_g^*} \\ &= 2\tilde{P}\left(\sqrt{\zeta_1\zeta_2}\right). \end{aligned}$$

When $\zeta_1 < \zeta_2 = 1/2n_g^*$, by the previous case we have that

$$\lim_{\zeta_2 \to (1/(2n_g^*))^-} \left(\tilde{P}(\zeta_1) + \tilde{P}(\zeta_2) - 2\tilde{P}\left(\sqrt{\zeta_1\zeta_2}\right)\right) \geq 0,$$

and so in this case the property holds. When $\zeta_1 < 1/2n_g^* < \zeta_2$, we have that

$$\begin{aligned} \tilde{P}(\zeta_1) + \tilde{P}(\zeta_2) &= \tilde{P}(\zeta_1) + \tilde{P}\left(\frac{1}{2n_g^*}\right) \\ &\geq 2\tilde{P}\left(\sqrt{\frac{\zeta_1}{2n_g^*}}\right) \\ &\geq 2\tilde{P}\left(\sqrt{\zeta_1\zeta_2}\right), \end{aligned}$$

where the first equality holds by definition of $\tilde{P}$, the first inequality follows by the preceding case, and the second inequality holds since $\tilde{P}$ is nonincreasing ans since, in this case, $\sqrt{\zeta_1/2n_g^*} \leq \sqrt{\zeta_1\zeta_2}$. Finally, when $1/2n_g^* \leq \zeta_1 < \zeta_2$, $\sqrt{\zeta_1\zeta_2} > \zeta_1$ and thus, by definition, $\tilde{P}(\zeta_1) = \tilde{P}(\zeta_2) = \tilde{P}\left(\sqrt{\zeta_1\zeta_2}\right)$. We conclude that $\tilde{P}$ is a proper failure-to-energy function. The theorem then directly follows by applying Lemma 6 to C' and $\tilde{P}$. □

4. A GENERAL ENERGY UPPER BOUND

Our main goal in this section is to prove Theorem 7, which roughly states that $O(N \log N)$ energy is sufficient to simulate a circuit of size N.

THEOREM 7. *Given a reliable circuit C of size N and a non-trivial failure-to-energy function, it is possible to construct a circuit C' with homogeneous voltage supply that uses $O\left(N \log(N/\delta)\right)$ units of energy and that $(1-\delta)$-reliably computes the same function computed by C.*

To prove Theorem 7 we use an upper bound on the number of gates for fault-tolerant circuits originally stated by Pippenger [10] and later proved in full generality by Gács [7]. This upper bound is stated in Theorem 8. The energy upper bound follows by choosing the voltage supply that minimizes the product of the total number of gates in the circuit constructed in Theorem 8, and the energy used by each gate. More specifically, we want to set the gate failure probability ϵ so as to minimize $P(\epsilon)/(\log(1/\epsilon) - r_0)$ for some constant r_0. As long as P is *non-trivial*, i.e., if for any $p^* \in (0, 1/2)$ it holds that $P(p^*) < +\infty$, one can find an ϵ such that $P(\epsilon) = O(1)$. This setting of ϵ then implies the upper bound of $O\left(N \log(N/\delta)\right)$ on the energy used by this construction using homogeneous supply voltages.

THEOREM 8 ([7]). *There are constants $R_0, \epsilon_0, r_0 > 0$ such that for all $\epsilon < \epsilon_0$ and $\delta \geq 3\epsilon$, for every reliable circuit C of size N there is a circuit of size $R_0 \frac{N \log(N/\delta)}{\log(1/\epsilon^*) - r_0}$ that computes the same result as C with probability at least $1-\delta$ if gates fail independently with probability at most ϵ, where $\epsilon^* = \max\{\epsilon, \delta/N\}$.*

PROOF. The main idea of this construction is to replace each gate of the reliable circuit with a gadget in the fault-tolerant circuit. A constant k is chosen as the level of redundancy for the circuit, meaning that each gadget has k outputs and each input to a gate in the reliable circuit is replaced by k inputs to a gadget. Another constant $\theta \in (0, 1)$ is chosen such that, with high probability, θk wires exiting each gadget carry the same value as the corresponding gate in the reliable circuit. Each gadget contains k copies of the corresponding gate from the reliable circuit, as well as an additional circuit that ensures that at least a θ fraction of the wires exiting the gadget are correct. □

5. SUPPLY VOLTAGE HETEROGENEITY MAY NOT HELP

In this section we observe in Theorem 9 that there are relations, namely the parity function, where heterogeneous supply voltages do not allow for an asymptotic reduction in energy.

THEOREM 9. *Let $\delta < 1/4$. The energy used by any circuit to $(1-\delta)$-reliably compute the parity function is $\Omega(n\log(n/\delta))$, and this is achievable by a circuit with homogeneous supply voltages.*

PROOF. The parity function can be reliably computed by a perfect binary tree of $2n-1$ XOR reliable gates. Thus, by Theorem 7 there exists a $(1-\delta)$-reliable circuit for the parity function that uses homogeneous voltage supplies and that incurs $O(n\log(n/\delta))$ energy consumption. Since the sensitivity of the parity function is n, by Theorem 1 this is the best possible to within a constant factor. □

6. SUPPLY VOLTAGE HETEROGENEITY CAN HELP

The goal of this section is to prove Theorem 13, which roughly states that heterogeneous supply voltages allow the natural majority circuit to compute a super-majority with asymptotically less energy than is possible with homogeneous supply voltages.

We start by defining the circuit and the logarithmic supermajority relation (LSR). Lemma 17 shows that $\Omega(n \cdot P(\delta))$ energy is required to $(1-\delta)$-reliably compute the LSR with homogeneous voltage supplies. The intuition behind the proof is that the output gate of any $(1-\delta)$-reliable circuit cannot have a probability of failure greater than δ and, since the voltage supplies are homogeneous, neither can any other gate. Then in Lemma 18 we show that there is a heterogeneous setting of the supply voltages so that this circuit $(1-\delta)$-reliably computes LSR with energy $O\left(n + 3^{1/\delta}P(\delta/10)\right)$. Intuitively, we split the circuit into an "upper" part consisting of gates close to the output gate, and a "lower" part consisting of gates close to the input gates: Each gate in the lower part has a constant probability of failure and thus a small energy consumption which results in non-constant savings compared to the optimal homogeneous setting. With the help of technical Lemmas 14 and 15, we are able to show that although there exist gates in the lower part of the circuit that fail with a probability higher than δ, no such gate fails with a probability $o(1)$. This preserves enough information for the upper part of the circuit to still $(1-\delta)$-reliably compute LSR. In other words, in the upper part of the circuit we use a much smaller probability of failure for each gate in order to ensure that the circuit will "autocorrect" itself and output the correct result (see Lemma 16).

DEFINITION 10. *The* Logarithmic Supermajority Relation (LSR) *is the following Boolean relation:*

$$LSR(x) = \begin{cases} 0 & \text{if the number of } 0\text{'s in } x \text{ is at least } n - \frac{1}{2}\log_3 n, \\ 1 & \text{if the number of } 1\text{'s in } x \text{ is at least } n - \frac{1}{2}\log_3 n, \text{ and} \\ 0 \text{ and } 1 & \text{otherwise,} \end{cases}$$

where x is the input and $|x| = n$.

This relation outputs 1 when the input contains at least $n - (1/2)\log_3 n$ ones, 0 when the input contains at least $n - (1/2)\log_3 n$ zeros, and otherwise we "don't care".

DEFINITION 11. *A* majority tree *is a Boolean circuit where the gates form a perfect ternary tree in which the leaves represent the inputs, and each internal gate, called* majority gate*, outputs the majority of its three children.*

DEFINITION 12. *A failure-to-energy function P is called* easy-going *when the following hold:*

- $\lim_{x\to 0} P(x) = +\infty$
- *There exists a constant $c > 0$ such that $\frac{P(x/10)}{P(x)} \le c$ for all $x \in (0, 1/2)$.*

Note that this class of failure-to-energy functions contains many natural functions. For example $P(x) = 1/x^\alpha$ and $P(x) = (\log 1/x)^\alpha$ are both easy-going.

THEOREM 13. *Let P be an easy-going failure-to-energy function, and let $c \in (0,1)$ be a constant. Furthermore, let $E_1(\delta)$ be the optimal energy consumption of the $(1-\delta)$-reliable majority tree on n leaves where all the gates must have the same failure probability, and $E_2(\delta)$ be the optimal energy consumption of the same $(1-\delta)$-reliable majority tree when each gate can have an arbitrary failure probability. Then, for $\delta' = \frac{1}{(1-c)\log_3 n}$, there holds*

$$\frac{E_1(\delta')}{E_2(\delta')} = \omega(1).$$

Let p_i be the probability that a gate of height i in a majority tree outputs 1. Notice that

$$p_{i+1} = p_i^3(1-\epsilon) + 3p_i^2(1-p_i)(1-\epsilon) + 3p_i(1-p_i)^2\epsilon + (1-p_i)^3\epsilon = (3p_i^2 - 2p_i^3)(1-2\epsilon) + \epsilon.$$

Also, let $R(p_i) := p_{i+1}$, and let $\ell^*(\epsilon)$ be the largest real number such that $R(\ell^*(\epsilon)) = \ell^*(\epsilon)$. Note that $\ell^*(\epsilon)$ only exists when $\epsilon < 1/6$. Therefore for the following we assume that $\epsilon < 1/6$.

LEMMA 14. *It holds that $\ell^*(\epsilon) = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{1-6\epsilon}{1-2\epsilon}}$. Furthermore, if $1/2 \le p_i \le \ell^*(\epsilon)$ then $p_i \le p_{i+1} \le \ell^*(\epsilon)$, and if $p_i \ge \ell^*(\epsilon)$ then $p_i \ge p_{i+1} \ge \ell^*(\epsilon)$, for all $i \in \{1, 2, \ldots, \log_3 n\}$.*

Note that the above technical lemma implies that $\ell^*(\epsilon)$ is a *stable* fixed point. The next two lemmas will be useful for setting the failure probabilities and analyzing the upper part of the tree.

LEMMA 15. *Let G be a majority gate with input gates g_1, g_2 and g_3 which output 1 with probability $q_1 > 1/2, q_2 > 1/2$, and $q_3 > 1/2$, respectively. Furthermore let q_G be the probability that G outputs 1 (for the given probabilities of the inputs to output 1). If we alter g_1, g_2, and g_3 to have probabilities $q_1' > q_1, q_2' > q_2$, and $q_3' > q_3$ of outputting 1, then for the new probability q_G' of G outputting 1 it holds that $q_G' \ge q_G$.*

LEMMA 16. *Consider a majority tree T of height $\lfloor 1/\delta \rfloor$ (for δ small enough) where each input of T is 1 with probability at least 0.79, and suppose that each gate of T has a failure probability of $\delta/10$. Then T outputs 1 with probability at least $1-\delta$.*

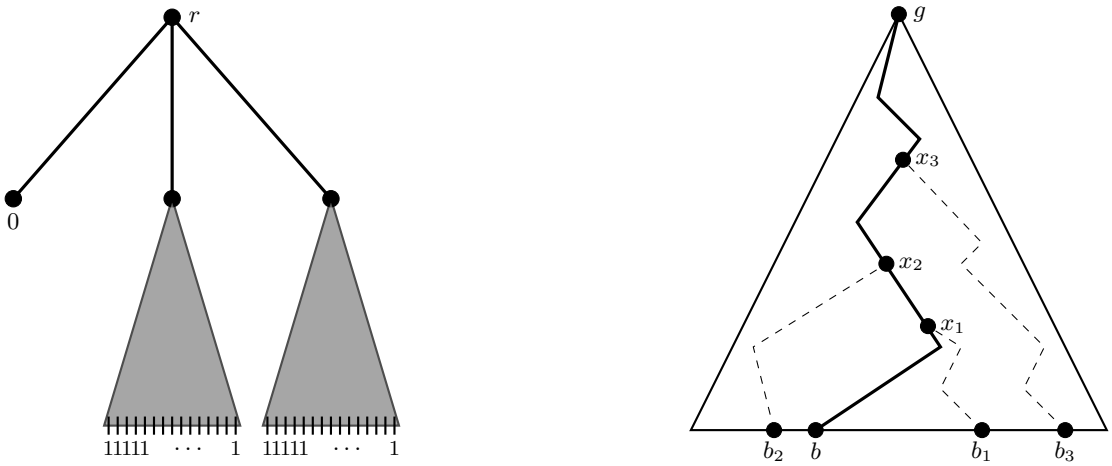


Figure 2: (a) $\Pr[r \text{ outputs } 1] \geq 1 - p$. (b) The path from b to g. The input gates b_i receive input 0.

With the help of the above lemmas, we are now ready to bound E_1 and E_2.

LEMMA 17. *It holds that* $E_1(\delta) = \Omega\left(n \cdot P(\delta)\right)$.

PROOF. In order to lower bound E_1, we will consider the case where each input bit is 1. Note that the root cannot have a probability of failure greater than δ, since then even with all its inputs being correct it would not give the right output with the desired probability, and by Lemma 15 this probability can only decrease as the probability that the input gates are correct decreases. Because all gates must have the same probability, we have that each of the $O(n)$ gates has an energy consumption of at least $P(\delta)$, and the lemma follows. □

LEMMA 18. *It holds that* $E_2(\delta) = O\left(n + 3^{1/\delta} P(\delta/10)\right)$.

PROOF. Assume without loss of generality that the input contains at least $n - (1/2)\log_3 n$ 1's, so that the desired output is 1. We assign a failure probability of $\delta/10$ to each gate located at a height of at least $\log_3 n - 1/\delta$, and a failure probability of 0.12 to each gate at a height strictly less than $\log_3 n - 1/\delta$. By Lemma 16, it suffices to show that each gate at height $\lfloor \log_3 n - 1/\delta \rfloor$ outputs 1 independently with probability at least 0.79.

Let $p = 0.36$. Consider a majority tree where each gate has a failure probability of 0.12. Then, by Lemma 14, and since for this tree $p_0 = 0.88 > \ell^*(0.12)$, we have that the root of the tree outputs 1 with probability at least $\ell^*(0.12) > 0.8$. Thus, a reliable majority gate whose inputs are one 0 and the outputs of two arbitrarily sized majority trees whose inputs are all 1's outputs 0 with probability at most $1 - 0.8^2 = p$. See Figure 2a.

Consider a majority tree of height h rooted at gate g, and fix an input to this tree that contains exactly d zeros as input, with $0 < d < h$. Let b be any of the input gates of the tree that was assigned a zero for this input. We first show that the probability that the path from b to gate g contains only 0's after each gate has computed is at most p^{h-d}. Let b_i for $i = 1, 2, \ldots, d-1$ be the other input gates that were assigned 0's. We may assume that the path from each b_i to g intersects the path from b to g, at a distinct gate x_i. Furthermore we may assume that each such x_i outputs a 0. See Figure 2b for an example. The probability of such a path from b to g to contain only 0's is equal to the probability that the $h - d - 1$ non-x_i gates on the path from b to g output a 0. Note that these non-x_i gates either receive a 0 and two inputs from majority subtrees whose inputs are all 1, or three inputs from majority subtrees whose inputs are all 1. Therefore, by the above observation about p and Lemma 15, the probability of such a path of all 0's is at most p^{h-d}.

Let T be any full (but not necessarily complete) majority tree of some height h_T. For any $h_A \geq h_T$, we can "complete" a copy of tree T by adding extra gates in order to obtain a perfect majority tree A of height h_A. We associate each gate in T with the corresponding gate in A. We claim that if the input at the leaves of both T and A consists of only 1's, then each gate of T is at least as likely to output 1 as the corresponding gate in A. We prove this claim by induction over the heights of gates in T. The base case, i.e., if a gate in T is a leaf, is straightforward. Assume now that each gate of T up to some height h', has a higher probability of outputting 1 than its corresponding gate in A, and consider a non-leaf gate g' of T at height $h' + 1$. Since T is a full tree, and g' is not a leaf, g' must have three children. The inductive step now directly follows from Lemma 15.

Next, consider any subtree B of our original majority tree that is rooted at a gate g of height $\lfloor \log_3 n - 1/\delta \rfloor$. Since we assumed that the input to the original tree contains at most $(1/2)\log_3 n$ 0's, clearly this holds for B as well. See Figure 3 for an example of a tree B.

Now we want to lower bound the probability that g outputs a 1 when there is no path of all 0's from a leaf to g. We note that if there is no path of all 0's from a leaf to g then there exists a full subtree T' of B that is also rooted at g and whose inputs can be assumed to be all 1's. The subtree T' can be constructed by truncating each leaf-to-root path in B at the first node that outputs 1. The existence of T' follows from the fact that there is no path of all 0's from a leaf in B to g, but the structure T' depends on the random events occurring at each gate in $B \setminus T'$ and the leaves of T'.

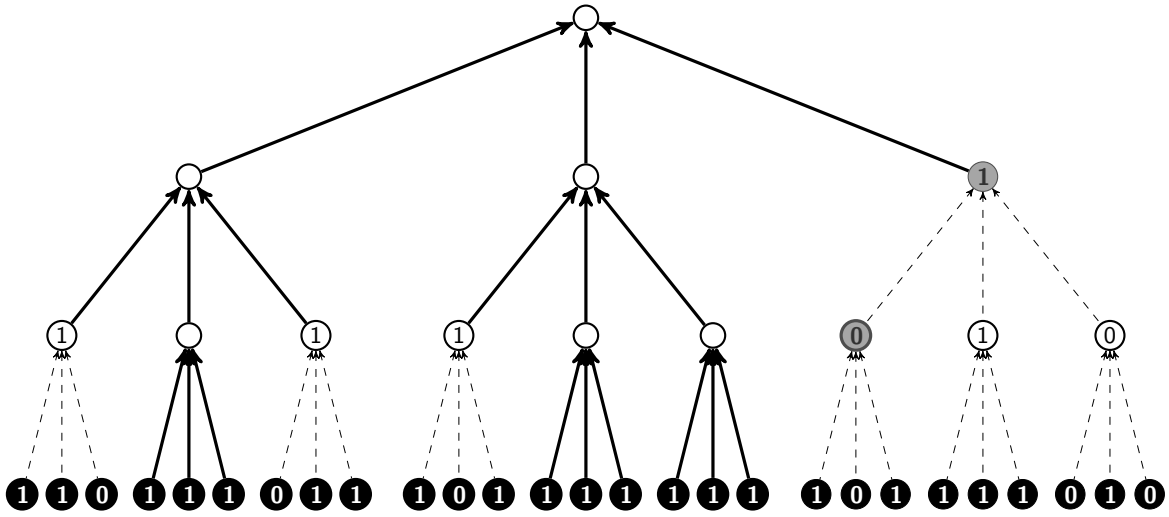


Figure 3: A subtree B. The solid edges denote the full ternary subtree $T \in \Gamma$. Note that T has 1's as inputs on its leafs. The dashed edges denote the edges in $B \setminus T$. The gray nodes denote gates that failed.

Conditioning on those random events, we have that B and T' output a 1 with the same probability.

Let Γ be the set of all full ternary trees of height at most $\lfloor \log_3 n - 1/\delta \rfloor$, and for $T \in \Gamma$ let X_T be the event that the truncated tree described above is T. We have that:

$$\Pr[B \text{ outputs } 1] \geq$$
$$\Pr[B \text{ outputs } 1 | \nexists \text{ path of all 0's}] \Pr[\nexists \text{ path of all 0's}] \geq$$
$$\Pr[\nexists \text{ path of all 0's}] \Pr[A \text{ outputs } 1].$$

The second inequality follows because

$$\Pr[B \text{ outputs } 1 | \nexists \text{ path of all 0's}] =$$
$$\sum_{T \in \Gamma} \Pr[X_T] \Pr[T \text{ outputs 1 when given only 1's as input}] \geq$$
$$\sum_{T \in \Gamma} \Pr[X_T] \Pr[A \text{ outputs } 1] =$$
$$\Pr[A \text{ outputs } 1].$$

It follows by the union bound over all possible leaf-to-root paths of all 0's that g, and therefore every gate of height $\lfloor \log_3 n - 1/\delta \rfloor$, outputs 1 independently with probability at least $\ell^*(\epsilon) \cdot (1 - \frac{1}{2}(\log_3 n) p^{\frac{1}{2}\log_3 n - \frac{1}{\delta}})$. For n large enough this is at least 0.79. By Lemma 16, the upper part of the majority tree outputs 1 with probability at least $1 - \delta$. The total energy of the circuit is at most $P(0.12)n + 3^{1/\delta} P(\delta/10)$. □

Proof of Theorem 13. By Lemmas 17 and 18, we have that for $\delta' = \frac{1}{(n-c)\log_3 n}$,

$$\begin{aligned}
\frac{E_1(\delta')}{E_2(\delta')} &= \Omega\left(\frac{n \cdot P(\delta')}{n + 3^{1/\delta'} P(\delta'/10)}\right) \\
&= \Omega\left(\frac{n \cdot P\left(\frac{1}{(1-c)\log n}\right)}{n + n^{1-c} P\left(\frac{1}{(1-c)\log n}\right)}\right) \\
&= \omega(1),
\end{aligned}$$

where the second equality follows by the second property of easy-going functions, and the third equality by the first property of easy-going functions when taking n large enough. □

We note that there are more trivial examples where heterogeneous supply voltages help. For example, consider a circuit that is a balanced binary tree of gates that each output the first bit, and the relation that outputs the first input bit. As most gates in this circuit are irrelevant to computing the desired relation, one can get an asymptotic energy saving by setting the supply voltages of the irrelevant gates to zero. Our example is more natural as one cannot simply power-off most of the gates. Although one might argue that our example is still not fully satisfactory as a more energy-efficient way to compute the super-majority relation is to use the majority circuit from [13] to compute the majority of the first $\log n$ bits with the supply voltages on each gate set so that the probability that any gate fails is at most δ. So a natural question is, "For every relation, is there is an asymptotically energy-optimal circuit for computing this relation that uses homogeneous supply voltages?"

7. CONCLUSIONS

This paper initiated the theoretical study of energy-efficient circuits. At this point, it seems that there are many interesting research lines in this area. Perhaps the most natural direction is to look at circuit design, namely determining the most energy-efficient circuit to compute a particular relation. But this design question is probably too difficult as a special case of this problem is the problem of designing a circuit with the least number of gates to compute a particular relation, which is a known hard problem. Perhaps questions related to circuit analysis are more amenable. A natural circuit analysis question would be: given a (perhaps special type of) relation, and (perhaps a special type of) circuit, how does one set the supply voltage, or voltages, so that the circuit $(1 - \delta)$-reliably computes the relation using approximately minimal energy. Another natural direction is to try to better understand which relations and circuits can benefit from heterogeneous supply voltages. An additional natural direction is to study the effect of switching delays of gates, which increase as the supply voltage decreases.

8. ACKNOWLEDGEMENT

We thank Rami Melhem for insightful discussions about Near-Threshold Computing.

9. REFERENCES

[1] G. Anthes. Inexact design: beyond fault-tolerance. *Communications of the ACM*, 56(4):18–20, 2013.

[2] J. Butts and G. Sohi. A static power model for architects. In *Proceedings of the 33rd annual ACM/IEEE International Symposium on Microarchitecture (MICRO)*, pages 191–201, 2000.

[3] B. Calhoun and A. Chandrakasan. A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation. *IEEE Journal of Solid-State Circuits*, 42(3):680–688, 2007.

[4] R. L. Dobrushin and S. I. Ortyukov. Lower bound for the redundancy of self-correcting arrangements of unreliable functional elements. *Problems of Information Transmission*, 13:59–65, 1977.

[5] R. L. Dobrushin and S. I. Ortyukov. Upper bound for the redundancy of self-correcting arrangements of unreliable functional elements. *Problems of Information Transmission*, 13:203–218, 1977.

[6] R. G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. N. Mudge. Near-threshold computing: Reclaiming Moore's law through energy efficient integrated circuits. *Proceedings of the IEEE*, 98(2):253–266, 2010.

[7] P. Gács. *Algorithms in Informatics*, volume 2, chapter Reliable Computation. ELTE Eötvös Kiadó, Budapest, 2005. Electronic version also in English: `http://www.cs.bu.edu/faculty/gacs/papers/iv-eng.pdf`.

[8] P. Gács and A. Gál. Lower bounds for the complexity of reliable boolean circuits with noisy gates. *IEEE Transactions on Information Theory*, 40(2):579–583, 1994.

[9] W. Ibrahim and V. Beiu. Reliability of NAND-2 CMOS gates from threshold voltage variations. In *Proceedings of the International Conference on Innovations in Information Technology (IIT)*, pages 310–314, 2009.

[10] N. Pippenger. On networks of noisy gates. In *Proceedings of the 26th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 30–38, 1985.

[11] N. Pippenger, G. D. Stamoulis, and J. N. Tsitsiklis. On a lower bound for the redundancy of reliable networks with noisy gates. *IEEE Transactions on Information Theory*, 37(3):639–643, 1991.

[12] R. Reischuk and B. Schmeltz. Reliable computation with noisy circuits and decision trees–A general $n \log n$ lower bound. In *Proceedings of the 32nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 602–611, 1991.

[13] L. G. Valiant. Short monotone formulae for the majority function. *Journal of Algorithms*, 5(3):363–366, 1984.

[14] J. von Neumann. Probabilistic logics and the synthesis of reliable organisms from unreliable components. In C. E. Shannon and J. McCarthy, editors, *Automata Studies*, pages 329–378. Princeton University Press, 1956.