Energy Limits of Message-Passing Error Control Decoders

Christian Schlegel
Dalhousie University
UMDCC
Halifax, Nova Scotia, Canada
Email: Christian.Schlegel@Dal.ca

Christopher Winstead
Department of Electrical and Computer Engineering
Utah State University
Logan, UT, United States of America
Email: Chris.Winstead@usu.edu

***Abstract*—Modern message-passing error control decoders are studied in regards to the processing energy required to extract digital information from a noisy received signal. It is shown that fundamental charge-based computational models, together with limits of error-free message passing along the processors' communications network, imply a lower limit to the energy efficiency achievable for such decoders in modern and future VLSI implementation technologies. The limiting energy of node processing is estimated for belief propagation decoding of LDPC codes, using estimates of nodes' internal processing activity. The limiting energy of message passing is estimated by using an energy-annotated density evolution procedure. For the class of decoders studied, the minimum energy is found to be on the order of 0.4 fJ per bit for node processing, and 9.86 aJ per bit for message passing.**

I. Introduction

As the miniaturization of very large scale integrated circuits continues to advance at an exponential pace, the power consumed by these circuits is becoming an ever more limiting problem, even as the computational resources — i.e. the density and number of switching devices — are becoming ever more available to accommodate even the most complex digital algorithms.

Integrated circuits (chips) with more than one billion transistors are now quite commonplace, and that growth is projected to continue at least for the next decade. The International Technology Roadmap for Semiconductors illustrates the exponential empirical law, known as *Moore's Law*, that appears to be underlying this progress.

The process is driven largely by the decreasing feature sizes of integrated circuits. Subsequent generations of fabrication technologies are known collectively as process nodes. Current leading-edge technology has arrived at the 22nm process node, and further miniaturization towards single digit nano-meter scales appears assured. Clearly there are immense challenges that are facing the industry in pushing this miniaturization forward, and we will not further concern ourselves with these, but accept this trend.

Instead, we will concentrate on the minimum power that is required to operate circuits of any size, in particular those at the most advanced nodes. In order to lead this discussion, we need to have a suitably general computational model of switching devices and the energy dissipated during an operation. The ubiquitous complementary metal-oxide semiconductor (CMOS) technology uses two complementary devices as shown in Fig. 1. This CMOS gate operates as a basic switch and thus is capable of processing binary information. The switch operates by charging and discharging a holding capacitance (usually the interconnect capacitance combined with the input capacitance of the next gate(s)).

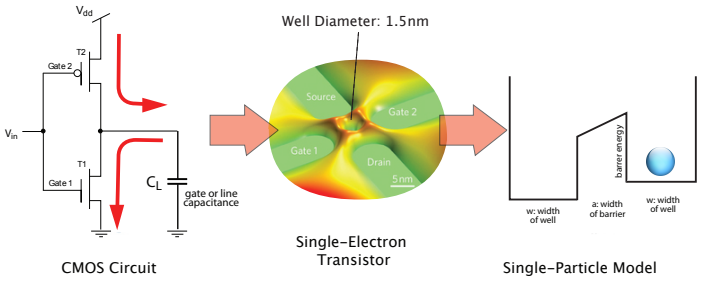

Fig. 1. Generic structure of a CMOS logic gate, and the evolution from MOS technologies toward single charge-based switching devices.

When the gate switches states it moves charge between the supply reservoir ($V_{\rm dd}$) and ground. At each such transition, an amount of energy equal to $C_I V_{\rm dd}^2$ is dissipated into heat. When considering the future evolution of digital technologies to the quantum scale, as illustrated in Fig. 1, the energy of transition may be described as the work done when moving a particle across a potential barrier. The evolution toward single charge switches is inevitable, particularly since single-electron transistors and other quantum-scale devices have been demonstrated in the laboratory (e.g. [1]).

One can ask the question what the limits are of such a charge-based computational model. This is precisely the exercise that Zhirnov et al. [2] conducted, using the Landauer limit for irreversible computing [3], which postulates that the minimum amount of energy released in the (irreversible) processing of one bit of information is bounded by

$$E \geq E_L = k_B T \ln 2 = 0.017 \text{ eV}, \tag{1}$$

where $T = 300$ K and $1 \text{ eV} = 1.602 \times 10^{-19}$ V. Applying Heisenberg's uncertainty principle Zhirnov computed minimal size and switching times for such an $E = E_L$ minimum-energy switch as

$$x_{\min} = \frac{h/2\pi}{\sqrt{2m_e E}} = 1.5\text{nm}; \quad t_{\min} = \frac{h/2\pi}{E} = 0.04\text{ps} \tag{2}$$

These figures in turn imply a maximum integration density $\nu_{\max}$ of such minimum size switches, and a maximum power density $P_{\max}$ of

$$\nu_{\max} = 4.7 \times 10^{13} \text{ devices/cm}^2; \quad P_{\max} = 3.7 \times 10^6 \text{ W/cm}^2 \tag{3}$$

The Landauer limit applies to a single charge storage model shown in Fig. 1, where a charged particle is in one of two wells, separated by an energy barrier. The minimum energy needed to move the charge is precisely E_L. Meindl also showed that the Landauer limit can be obtained when considering only the operation of an "ideal" MOSFET device operating in its subthreshold region [4]. Hence the Landauer model is quite applicable to the CMOS structure from Fig. 1.

Since the Landauer limit (1) is applicable to irreversible operations, one might speculate that it can be circumvented by using reversible or adiabatic computational circuits. Meindl and Davis [4] disposed of this possibility by showing that the Landauer limit can also be obtained by solving the Shannon capacity of an interconnect wire in the presence of thermal noise, and hence can be interpreted as a limit on signalling between gates and modules, one that cannot be circumvented by technology choices.

In the sequel we will study the two main aspects of energy consumption that affect an error control decoder, in particular a decoder for low-density parity check (LDPC) codes. LDPC codes are among a class of very powerful error control codes which can achieve the theoretical limits on communications, known as the Shannon limit. They have a low-complexity iterative decoding algorithm, which not only allows the construction of decoders for very large codes, but also has made these codes the de-facto standard in many communications applications. With error control coding now applied to very high speed applications, such as the 802.3 10Gbit/s modems, or optical communications systems at even larger rates, their power consumption has become a targeted concern that appears to be limiting future applications.

Fig. 2 shows the processing structure of an LDPC code. The two processes that need to be executed are local computation, summarized in the figure, and communication to connected nodes. These will be analyzed from a power consumption viewpoint.

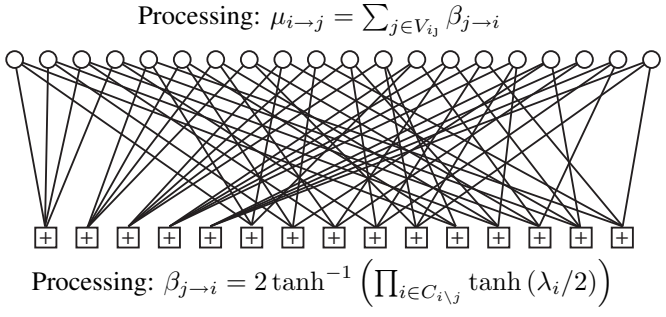

Fig. 2. Network of an LDPC code. Their typical size is 1000–10,000 nodes.

II. Node Processing

Without going into the full implementation details of the local processing operations at the nodes, it can be shown that the computational complexity of each node is $\mathcal{O}(W)$, where W is the number of binary digits used in the number representation of a digital decoder. W is typically between 6–12 bits, and has a subtle impact on the performance of the code – see [8]. More specifically, activity simulations showed that there were on the average 2.7 digital transitions per bit and message line at the variable nodes (top nodes in Fig. 2), and 3.3 digital transitions per outgoing message in the check nodes (bottom nodes).

We approach the computational power limit in the following way: for each transition that occurs during decoding, a minimal amount of energy proportional to E_L is dissipated. The problem now is that if the barrier is set at the minimum energy E_L, there is a significant over-barrier probability $P = e^{-E_{\text{barrier}}/kT}$, which makes such a cell very unreliable. In fact, the over-barrier probability reaches 50% at E_L, making the cell quite useless for computation. To keep our nodes operating at acceptable levels of reliability, the barrier energy needs to exceed E_L. We somewhat arbitrarily assume that a factor $K = 10$, which leads to an over-barrier probability of 10^{-4}, can both be realized in the future, and is acceptable in the algorithm.

With these assumptions, the node processing per information bit is lower bounded by

$$E_b > KE_L IW(2.7d_v + 3.3(d_c - 1)) \tag{4}$$

where I is the number of iterations in the code network, usually on the order of 5–20. Clearly, varying W, I, and the code parameters d_v, d_c can move this number by an order of magnitude or so, and the limit is to be seen as that for an average code. With $K = 20, I = 10, d_v = 3, d_c = 6, W = 8$, we obtain $E_b > 0.1$fJ.

Expressed in kT, the processing requirements are on the order of 25,000 kT per bit. In the next section we will argue that transporting the messages between the local nodes will incur an energy effort that is comparable to the energy expended in the computations themselves, thus arriving at a lower energy limit for a charge-based error control decoder.

III. Network Communication

In this section we address the second power-intensive portion of a message-passing error control decoder, which is the communication of the node messages along branches of the code network (see Fig. 2), by evaluating the minimum energy cost associated with transporting messages. The decoder is parameterized by the traditional degree distribution, by symbol and check node update algorithms, and by the *message representation* used for transmitting messages between check and symbol nodes. The message representation is defined as the mapping from possible messages to a set of corresponding physical signals. A similar approach has been taken previously to minimize activity in digital LDPC decoders (e.g. by Gaudet and Crowley [5], [6]).

The density evolution method is modified to examine the effects of internal signalling near the kT noise limit in electrical interconnects. By considering the physical signals passed between nodes, we obtain technology-independent conclusions about the minimum energy associated with error-free decoding. Low-energy signals are subject to upsets from kT noise, but a decoder is able to self-correct many such upsets.

In order to evaluate signalling energy limits in message

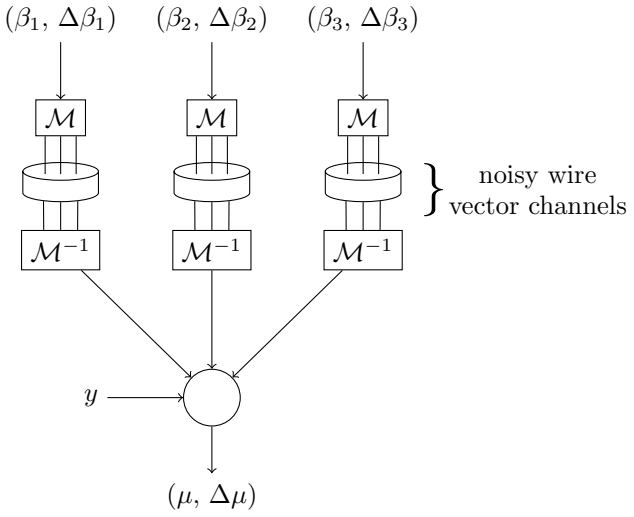

Fig. 3. Illustration of the density estimation procedure for a symbol node with $d_v = 3$. The incoming messages $(\beta_i,\, \Delta\beta_i)$ are generated from the joint distribution. The β messages are used to compute outgoing message μ. Then the $\beta + \Delta\beta$ messages are used to compute $\mu + \Delta\mu$. The energy is a function of $\Delta\beta$ and $\Delta\mu$.

passing, we consider that each signal wire is affected by Gaussian white thermal noise with energy kT Joules per sample. Physical signals are assumed to be voltages, specified in units of $(kT)^{1/2}$. The normalized energy of physical signals is estimated by assuming unit interconnect capacitance, so that $E_{\text{signal}} = \alpha V_{\text{dd}}^2$, where α is the *activity* defined as the wire's frequency of transitions. The unknown capacitance C_I is treated as a technology-dependent scale constant. The precise value of C_I has no influence on the energy calculations, since V_{dd} can be varied proportional to $C_I^{-1/2}$ (we could equivalently say that V_{dd} has units $(kT/C_I)^{1/2}$).

Based on this model, each signal wire is represented as an additive white Gaussian noise channel with zero mean and unit variance. For example, suppose a binary value $x \in \{0,\, 1\}$ is transmitted as part of the message from a symbol node to a check node. The value is represented as a voltage $v_x \in \{0,\, V_{\text{dd}}\}$, and transmitted across the unit-variance AWGN channel, which adds a noise sample n. At the channel's output, the received signal $v_y = v_x + n$ is resolved to a binary value $y \in \{0,\, 1\}$ by applying a threshold at $V_{\text{dd}}/2$.

In order to account for the energy per message in a fully-parallel LDPC decoder, it is necessary to track the joint distributions of messages and their transitions, as shown in Fig. 3. The detailed procedure was described by Gaudet [5], and is only briefly summarized here. The symbols $\beta\,(t)$ and $\Delta\beta\,(t+1)$ refer to messages passed from a check to symbol nodes, and the changes in those messages, respectively, during iteration t. Similarly, the symbols $\mu\,(t)$ and $\Delta\mu\,(t+1)$ refer to messages passed from symbol to check nodes and their changes during iteration t. Note that the dependence on t is dropped when there is no ambiguity.

To estimate the energy required for message passing, density evolution is performed as usual while tracking the joint distributions for $(\beta,\, \Delta\beta)$ and $(\mu,\, \Delta\mu)$. The symbol node estimation is performed as follows. Random samples are generated jointly for both β and $\Delta\beta$. The symbol node update equations are used to obtain μ and $\Delta\mu$. To compute the sample's transition energy, both μ and $\mu + \Delta\mu$ are mapped to their corresponding message representations. If a wire's signal value at iteration t is $v_x\,(t)$, then the wire's normalized transition energy during iteration t is equal to $(v_x\,(t+1) - v_x\,(t))^2$. The sample's total energy is the sum over all wire transition energies.

The same procedure is used to estimate the energy of sample check-to-symbol messages. The samples are accumulated to compute the mean energy per symbol message and check message, $\mathcal{E}_\mu$ and $\mathcal{E}_\beta$, respectively. Finally the average energy per message is

$$\mathcal{E}_m = \frac{d_v}{d_v + d_c}\mathcal{E}_\mu + \frac{d_c}{d_v + d_c}\mathcal{E}_\beta.$$

IV. Results

Simulations were performed across 30 iterations using the modified density evolution procedure described in Sec. III. The channel noise parameter σ was varied following the procedure described by Richardson [7]. Regular $(3, 6)$ code ensembles were considered. The message representation is a W-bit word with sign-magnitude encoding and uniform quantization in the interval $[-L_{\max},\, +L_{\max}]$. The maximum LLR magnitude for each simulation was calculated as $(d_v + 1) * 4.0/N_0$, where $N_0 = 2\sigma^2$ is the power spectral density of channel noise.

Fig. 4 shows the message error rate (MER) as a function of both σ and V_{dd} for the case $W = 8$. Fig. 5 shows the MER plotted against the corresponding $\mathcal{E}_m$ for $W = 8$. Figs. 6 and 7 show the MER vs V_{dd} and $\mathcal{E}_m$, respectively, for the case $W = 10$. The average $\mathcal{E}_m$ was calculated from the mean transition activity averaged over 30 iterations. The MER was calculated as the average rate of messages with erroneous sign during the final six iterations (time steps 26 to 30). Fig. 5 shows that the MER is a function of V_{dd} and decreases below our measurement threshold (i.e. MER $< 10^{-6}$) when $V_{\text{dd}} > 10$, which we consider to be successful convergence.

Counting the number of messages that need to be exchanged during the course of a decoding cycle normalized per bit, we obtain a bound for the minimum energy per symbol to drive the interconnect network as

$$E_n > 2Id_v\mathcal{E}_m \tag{5}$$

Using the thresholds evaluated in Figs. 5 and 7 of approximately 40 kT, we obtain $E_n > 2400kT$ for $I = 10$ and $d_v = 3$.

V. Conclusions

We have used fundamental physical and information theoretic limits to bound the minimal energy required to operate a (large) message passing error control decoder. Based on average switching activity within a simulated decoder design, the limiting energy per bit is on the order of 30,000kT, which equals 0.1 femto Joules per bit. This bound may even be an overestimate because the decoder may be able to tolerate some noise-induced errors if the barrier parameter K is lowered. Additionally, the average switching activity tends to diminish across iterations when the decoder converges. Adiabatic circuit techniques could conceivably be employed to reduce the irreversible information loss. In that case, the

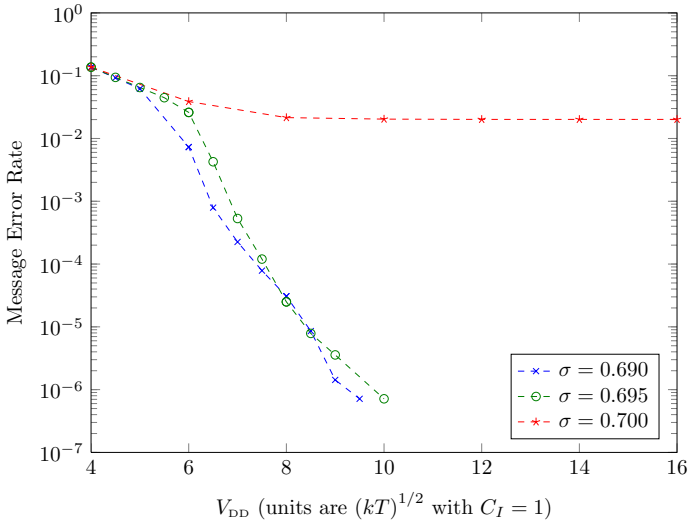

Fig. 4. MER for a regular $(3, 6)$ code ensemble with 8-bit linear-quantized message representation, varying the normalized supply voltage (V_{dd}) and the channel noise standard deviation (σ). $\sigma = 0.7$ exceeds the ensemble's decoding threshold. Under ideal message passing, the threshold is $\sigma^* = 0.88$ [7].

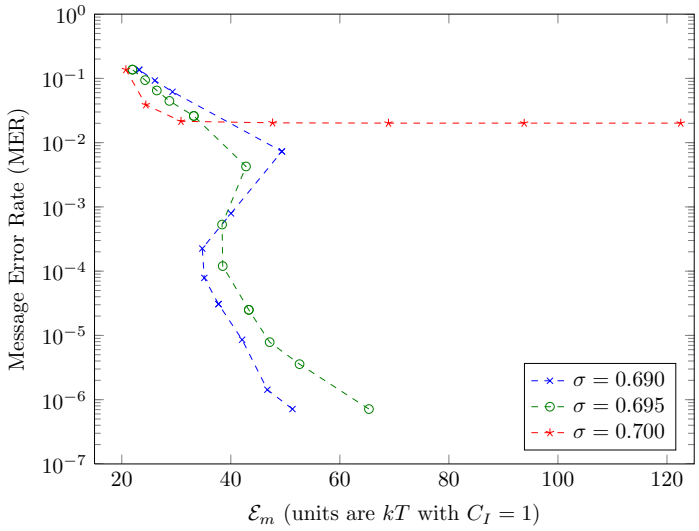

Fig. 5. MER for a regular $(3, 6)$ code ensemble with 8-bit linear-quantized message representation, plotted against the average energy per message.

energy in the communications network would be limiting and density evolution would provide an inescapable lower bound on the energy required for message passing, approximately $2400kT$ or 10 atto Joules per bit. The results in this analysis represent only a single class of codes and decoding algorithms. The methods presented here can conceivably be extended and applied to other codes and algorithms as a subject for future research.

Acknowledgment

C. Winstead's work was supported by the US National Science Foundation under award ECCS-0954747, and by the US Fulbright program. C. Schlegel's work was supported by the Canadian National Science and Engineering Research Council, under an Industrial Research Chair award.

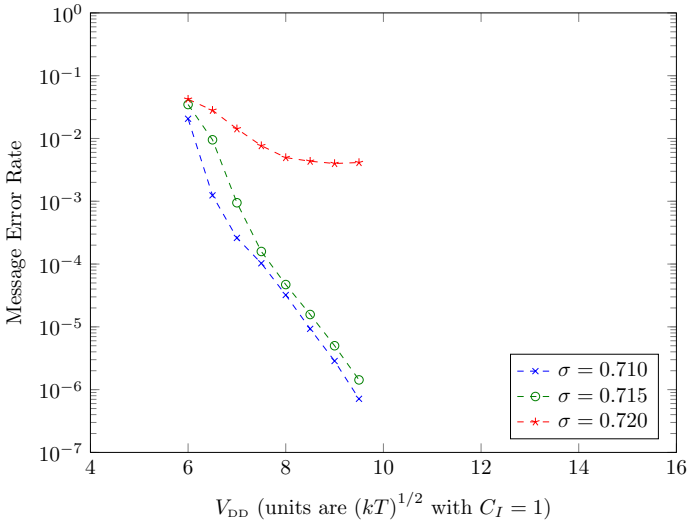

Fig. 6. MER for a regular $(3, 6)$ code ensemble with 10-bit linear-quantized message representation, while varying the normalized supply voltage (V_{dd}) and the channel noise standard deviation (σ).

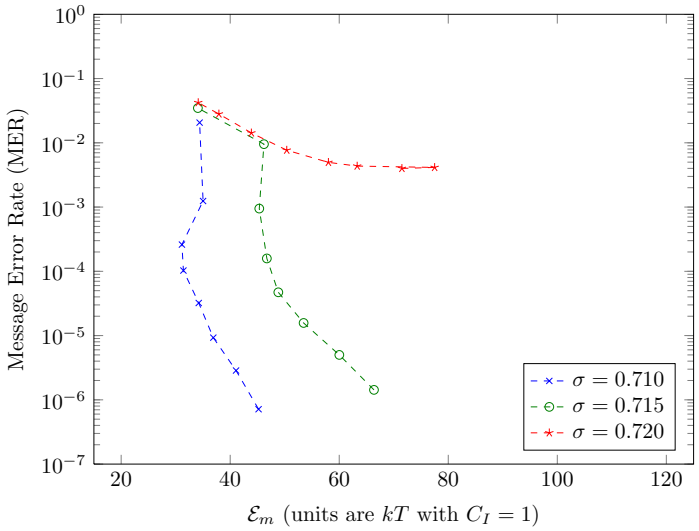

Fig. 7. MER for a regular $(3, 6)$ code ensemble with 10-bit linear-quantized message representation, plotted against the average energy per message.

References

[1] G. Cheng et al., "Sketched oxide single-electron transistor," *Nature Nanotechnology 6*, 343–347 (2011) doi:10.1038/nnano.2011.56.

[2] V. Zhirnov et al., "Limits to binary logic switch scaling – A gedanken model," *Proceedings of the IEEE*, Vol. 91, No. 11, November 2003.

[3] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM J. Research and Development*, vol. 5, pp. 181–191, 1961.

[4] J. D. Meindl, J. A. Davis, "The fundamental limit on binary switching energy for terascale integration (TSI)," *Solid-State Circuits, IEEE Journal of* , vol. 35, no. 10, pp.1515,1516, October 2000.

[5] V. C. Gaudet, C. Schlegel, R. Dodd, "LDPC Decoder Message Formatting Based on Activity Factor Minimization Using Differential Density Evolution," *IEEE Inform. Theory Wrkshp (ITW)*, pp. 571,576, Sep. 2007.

[6] B. Crowley, V. C. Gaudet, "Switching Activity Minimization in Iterative LDPC Decoders," *Journal of Signal Processing Systems*, vol. 68, no. 1, pp. 63–73, July 2012.

[7] T. J. Richardson, R. L. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *Information Theory, IEEE Transactions on*, vol. 47, no. 2, pp.599,618, February 2001.

[8] S. Zhang and C. Schlegel, "Controlling the Error Floor in LDPC Decoding," *IEEE Trans. Commun.*, Vol. 61, No. 9, pp. 3566–3575, 2013.