Moore's law and the energy requirement of computing versus performance

L.B. Kish

Abstract: It has recently been recognised that speed, noise and energy dissipation are strongly interrelated entities. Following Moore's law of miniaturisation, at sizes below 40 nm, physics will impose fundamental and practical limits of performance by shrinking noise margin, increasing and quickening noise, and increasing power dissipation. It is important to locate the fundamental aspects of the problem, explore relevant practical problems and possible solutions, and investigate this situation, not only in microelectronics (CMOS etc.) but also in single-electron-transistor-based nanoelectronics and also in quantum informatics applications. The energy requirement of running classical and quantum logic gates is compared.

1 Thermal noise and the Rice equation

Thermal noise is an omnipresent small voltage fluctuation on resistors. It has been thought that thermal noise will never be an issue in digital electronics. This view has been re-evaluated and changed recently [1, 2]. On a parallel resistor–capacitor (RC) unit, the effective thermal noise voltage U_n is given as

$$U_n = \sqrt{kT/C} \tag{1}$$

and the bandwidth f_c of this noise is

$$f_c = \frac{1}{2\pi RC} \tag{2}$$

Gaussian noise processes, such as thermal noise, can cross large amplitude levels provided that sufficiently long time is available. In a logic circuitry, noise amplitudes reaching beyond the noise margin U_{th} cause false bit flips, which can result in bit errors. For a single, band-limited noise process, the mean frequency $\nu(U_{th})$ of bit errors can be obtained from the Rice formula [1, 2]:

$$\nu(U_{th}) = \frac{2}{\sqrt{3}} \exp\left(\frac{-U_{th}^2}{2U_n^2}\right) f_c \tag{3}$$

where U_n is the effective noise voltage and f_c is the bandwidth.

So, what happens during miniaturisation? In CMOS technology, the resistors are two-dimensional conductors, so the resistance stays (roughly) constant. Therefore, the supply voltage has to be decreased to keep the electrical field and dissipation at an acceptable level. The capacitances are decreasing. Together, these effects yield the following trends:

(i) a shrinking noise margin because U_{th} is only a fraction of the supply voltage

(ii) a growing noise because of (1)

(iii) the bandwidth is growing; see (2) – quickening of the noise.

All these phenomena, (i)–(iii), help toward radically increasing the frequency of bit errors via (3). To have a feeling for the nature of this problem, in Fig. 1 the bit error frequency versus the noise margin normalised to the noise voltage is shown for different bandwidths (clock frequency) and different numbers of transistors. The practical limit of usability is a certain U_{th}/U_n ratio when the bit error rate is around 1 error /year. Even a 10% decrease of the U_{th}/U_n ratio compared to its critical value yields an error rate increase to 10^5 when we consider all transistors in a modern PC (3×10^9–10^{10} transistors).

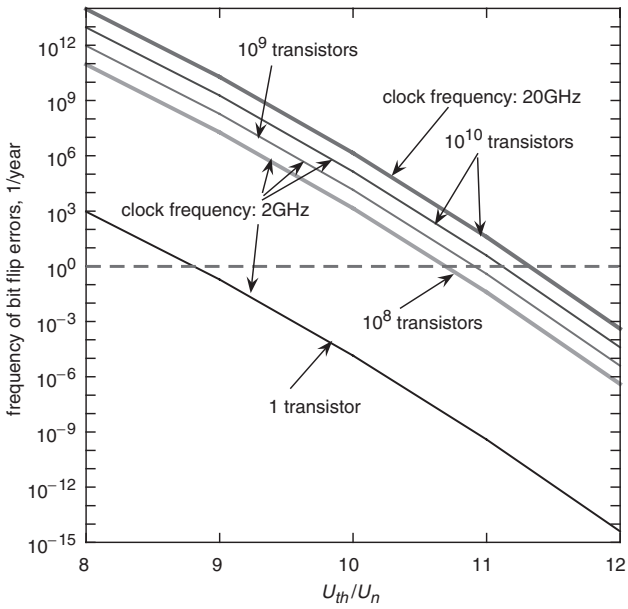

Fig. 1 *Bit errors against ratio of noise margin and effective noise voltage [1, 2]*

Using these results and a prediction of the evolution of capacitance and noise margin using present trends, recently a prediction of the end of Moore's law was published [1, 2], see Fig. 2.

IEE Proceedings online no. 20040434

doi:10.1049/ip-cds:20040434

Paper first received 9th June and in final revised form 29th October 2003

The author is with the Department of Electrical Engineering, Texas A&M University, College Station TX 77843-3128, USA

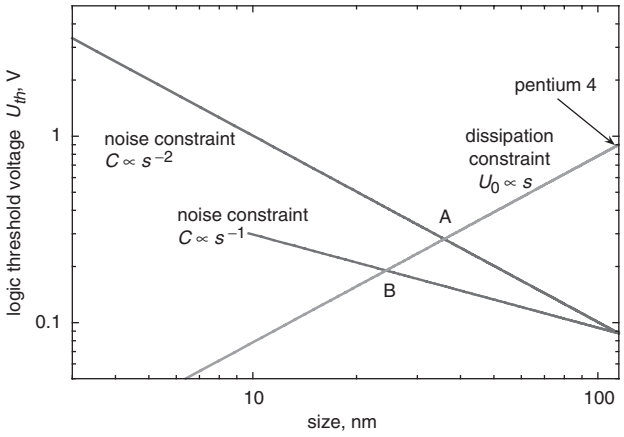

Fig. 2 *Prediction of the end of Moore's law [1, 2]*
The technology faces a difficult problem when the upper limit of noise margin set by the dissipation/field constraint and the lower limit of noise margin required by the noise/error constraint cross each other (between points A or B, depending on the evolution of gate oxide thickness)
A: 36 nm, 0.3 V; B: 25 nm, 0.2 V

The shrinking noise margin (logic threshold voltage) and the increasing noise margin required by the increasing and quickening noise pose conflicting requirements which would stop miniaturisation in 6–8 years if the trends of recent years continue.

In the rest of this paper, we outline the objectives that need further analysis and give some initiatives for extended research in the future.

2 Refinements of predictions for CMOS technology

It has been pointed out in [1, 2] that the prediction was based on strong approximations to keep generality and because of lack of information about certain device parameters. Since [1] was published new information from microprocessor makers about some previously unknown parameters have emerged and some new efforts have been made to reduce the supply voltage with a rate *less* than previously supposed. These efforts are, however, *controversial* because many large-scale users, for example in aviation electronics [3], would like to *increase* the rate of supply voltage reduction in order to improve the device failure rate, which has been steadily growing due to high electric fields in chips.

Therefore theoretical efforts have been made on refining the prediction for the bit error problem due to thermal noise. The refined model takes into the account the following aspects:

- further noise margin decrease due to the opening threshold voltages of the *P* and *N* type MOSFET transistors
- somewhat reduced noise due to parallel gate circuits
- the various supply voltage reduction strategies (predictions are controversial)
- the fact that the internal supply voltage and noise margin of microprocessors in 2002 was already less than supposed in [1].

Interestingly, the preliminary investigations indicate that the size range, where the problems begin, remains the same because the different corrections act in different directions and the effects compensate each other. Thus, the conclusions obtained in the context of Fig. 2 remain the same.

3 Single-electron transistor based microprocessors

The next question is the bit error situation of microprocessors based on single-electron transistors (SETs) [4]. SETs utilise tunnelling, which is not dissipative, but the processes coupled to it are dissipative. The electrons at the tunnel junction of a closed SET can be excited by thermal energy fluctuations to the energy level where tunnelling can occur and a single electron can cause a single bit error in a SET. Similarly, an open SET can be temporarily shut down by thermal fluctuations. So, the tentative expectation is that SETs will have similar types of bit error characteristics as CMOS. However, the picture is more complex [4]:

- Instead of a single capacitance, three different capacitances influence the bit errors, the two tunnel junction capacitances and the quantum dot capacitance (gate capacitance).
- The noise margin cannot be increased arbitrarily by increasing supply voltage. It has a practical maximum which is equal to the voltage difference between the totally closed and totally open transistor. Higher voltages cause multiple electron operation modes and large noise and dissipation.
- There are two different working ranges versus the quantum dot size. A larger size, Coulomb blockade controls the current transport, and smaller size (<10 nm) quantum confinement effects dominate.

Preliminary studies [4] show that the requirement for small quantum dot sise becomes much harder to satisfy when not only the DC characteristics but also bit errors matter in a microprocessor with 10^8 or more SETs. To have SET based microprocessors, the characteristic quantum dot sise has to be 1 nm.

4 Ultimate limits of energy dissipation versus performance in classical and quantum computing

The ultimate and most fundamental questions are related to the power requirement of classical and quantum information processing. If CMOS and SET fail, it is a natural question if quantum computing and quantum information can help and how. Because, so far, the existing quantum computing architectures are not practical, the only question we may be able to address is regarding the ultimate limits of performance. Performance includes error rate (accuracy), speed (bandwidth) and power dissipation. It is very important to take temperature into account, and to compare the ultimate performance limits of classical and quantum computers at the same temperature [5].

Recent studies [5, 6] (see the Appendix, Section 7) show that quantum computers have problems at high accuracy due to Heisenberg's uncertainty principle. Classical computers perform much better as accuracy is concerned; however, quantum computers can balance this deficiency by a greater speed [5] (see in the Appendix). For a comparison of the energy requirements of a classical and a quantum gate when they run at the same clock frequency, see Figs. 3 and 4 [2, 5].

Recent studies [7] comparing general-purpose quantum and classical computers indicate that, even by using error correction, quantum computer will dissipate at least 100 times more energy at the same performance.

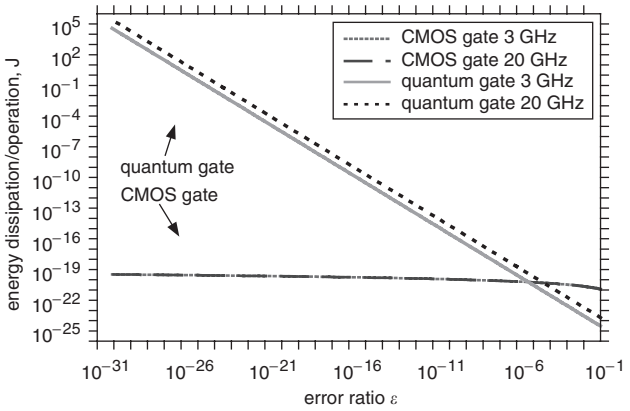

Fig. 3 *Minimal energy dissipation of single logical gates [2, 5, 6], classical (CMOS) and quantum, against the error ratio of the gate*
The quantum gate result is for zero temperature. See the derivation in the Appendix (Section 7.1)

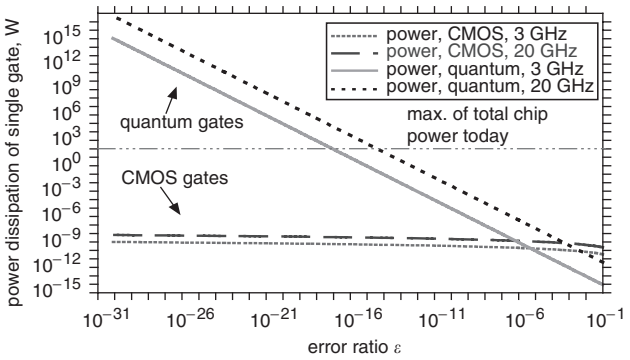

Fig. 4 *Minimal power dissipation of single logical gates [2, 5, 6], classical (CMOS) and quantum, against the error ratio of the gate*
The quantum gate result is for zero temperature. See the derivation in the Appendix (Section 7.1)

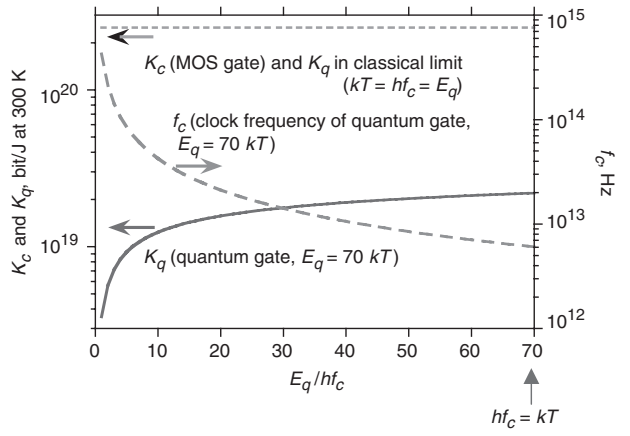

Fig. 5 *Bit/joule performance of classical and quantum logical gates [5]*
See derivation in the Appendix (Section 7.2)

However, it is important to note that the ultimate focus of this study, in the future, will not be the accuracy (error rate) of classical and quantum computers but the energy requirements of processing Shannon information (bit). This aspect induces many new questions including the problem of computer architectures which are not sensitive to noise. In the Appendix (Section 7.2), we show that the ultimate measure, Shannon's information channel capacity versus the energy dissipation, also gives better results in a classical than in a quantum gate (see Fig. 5).

5 Conclusion

We have all been enjoying the fast growth of speed and memory size of computers during the last few decades. Recently, the emerging fields of quantum computing and nanoelectronics have suggested that the future will be even more brilliant.

However, as soon as we confront physical laws and reality with expectations, the future seems to be less bright. The author of this paper has developed the opinion that 'nano' and 'quantum' will turn out to be dead ends as far as information technology is concerned. Most probably, the microelectronics in the 50–100 nm range will be proven to be the best technology, at least, for silicon. However, it was important to experience the nano and quantum scales in order to draw this conclusion and for the important fundamental scientific results and the limits found there.

6 References

1 Kish, L.B.: 'End of Moore's law: thermal (noise) death of integration in micro and nanoelectronics', *Phys. Lett.*, 2002, **305**, pp. 144–149
2 Kish, L.B.: 'Moore's law, performance and power dissipation', in 'Encyclopedia of nano science' (Marcel Dekker, in press)
3 Huang, B., Qin, J., Walters, J., and Bernstein, J.B.: 'Development of derating guidelines for semiconductor devices'. Aerospace Vehicle Systems Institute Report, 2003, unpublished
4 Kim, J., and Kish, L.B.: 'Can single electronic microprocessors ever work at room temperature?', *Proc. SPIE-Int. Soc. Opt. Eng.*, 2003, **5115**, pp. 174–182
5 Kish, L.B.: 'Moore's law is killed by classical physics; can quantum information save it?', *Proc. SPIE-Int. Soc. Opt. Eng.*, 2003, **5115**, pp. 167–173
6 Gea-Banacloche, J.: 'Minimum energy requirements for quantum computation', *Phys. Rev. Lett.*, 2002, **89**, p. 217901/1-4
7 Gea-Banacloche, J., and Kish, L.B.: 'Comparison of energy requirements for classical and quantum information processing', *Fluct. Noise Lett.*, 2003, **3**, pp. C3–C7
8 Landauer, R.: *IBM J. Res. Dev.*, 1961, **5**, p. 183
9 Allahverdyan, A.E., and Nieuwenhuizen, Th.M.: 'Testing the violation of the Clausius inequality in nanoscale electric circuits', *Phys. Rev. B, Condens. Matter Mater. Phys.*, 2002, **66**, p. 115309
10 Kiss, L.B.: 'To the problem of zero-point energy and thermal noise', *Solid State Commun.*, 1988, **67**, p. 749

7 Appendix

7.1 Accuracy versus power dissipation in CMOS and quantum computers

The lowest limit of the energy requirement, E_q, for a single operation of a quantum logic gate is given by Gea-Banacloche (see [2]) as

$$E_q \approx \frac{\hbar}{\varepsilon_q \tau_q} = \frac{\hbar f_{c,q}}{\varepsilon_q} \tag{4}$$

where ε_q and τ_q are the error probability and the time requirements of the quantum logic operation; $1/\varepsilon_q$ is related to the accuracy of the gate operation. Then, for the case of sequential operation of the gate, the maximum clock frequency is given as $f_{c,q} = 1/\tau_q$. Although the required energy is not the dissipated energy but the energy required to be put into the quantum gate, we can consider this energy as a practically dissipated energy. The situation is very similar in CMOS technology, where the electrical energy due to the gate capacitance charge is an ordered energy: it is most obvious that this energy is practically dissipated at each change of logic state.

For CMOS transistors, the mean frequency of bit errors [6] is given as

$$\nu = \frac{2}{\sqrt{3}} \exp\left(\frac{-U_{th}^2}{2U_n^2}\right) f_c \tag{5}$$

where U_{th} (logic threshold voltage) is the noise margin between the logic low (0) and high (1) levels, U_n is the effective thermal noise voltage on the CMOS transistors' resultant gate capacitance, and f_c is the RC time constant of the gate capacitance and its driving resistance, which is the maximal clock frequency. Then, if we drive the system at frequency f_c, the error probability is

$$\varepsilon_c = \frac{\nu}{f_c} = \frac{2}{\sqrt{3}} \exp\left(\frac{-U_{th}^2}{2U_n^2}\right) \tag{6}$$

Because U_{th}^2 and U_n^2 are related to the static and thermodynamic energies in the capacitor, respectively, this result can be generalised for any kind of available digital technology where, without exception, the logic threshold energy is dissipated during the change of logic state. Taking into the account that the logic voltage change is the voltage change on the gate capacitance, the lower limit of energy dissipated during the change of logic state satisfies the relation

$$E_c \geq \frac{1}{2} C U_{th}^2 \equiv E_{c,\min} \tag{7}$$

where C is the resultant capacitance at the gate electrode. Note that the actual value is

$$\frac{1}{2} C\left[(U_0 + U_{th})^2 - U_0^2\right] = \frac{1}{2} C\left(U_{th}^2 + 2U_0 U_{th}\right) \geq \frac{1}{2} C U_{th}^2$$

where U_0 is the low (0) logic voltage level, so (7) indeed gives the lower limit (the $U_0 = 0$ case). The mean thermal noise energy in the capacitor is

$$\frac{1}{2} C U_n^2 = \frac{kT}{2}$$

Using (2), for room temperature, the minimum U_{th}/U_n ratio for error-free operation (<1 false-bit-flip/year due to thermal noise) was given 12 [1], which corresponds to

$$\frac{E_c}{kT} \geq 72 \approx 70 \tag{8}$$

Thus, the error rate can be given as

$$\varepsilon_c = \frac{2}{\sqrt{3}} \exp\left(\frac{-U_{th}^2}{2U_n^2}\right) = \frac{2}{\sqrt{3}} \exp\left(\frac{-CU_{th}^2}{2CU_n^2}\right) = \frac{2}{\sqrt{3}} \exp\left(\frac{-E_{diss,min}}{kT}\right) \tag{9}$$

which is a generalised result for digital gates, independent of technology. We can then express the minimum dissipated energy per bit-flip as follows:

$$E_{c,\min} = -kT \ln\left(\frac{\sqrt{3}}{2} \varepsilon_c\right) \tag{10}$$

If, on the average, N transistors are changing their logical state in the processor during one clock period, the total dissipated power is

$$P_{total,min} = -N f_c kT \ln\left(\frac{\sqrt{3}}{2} \varepsilon_c\right) \tag{11}$$

In Fig. 3, the energy dissipation of one logical gate, for a single classical or quantum operation, versus the error ratio of the gate, is shown. The quantum case is given by the results of Gea-Banacloche [6]. At today's clock frequencies and those expected in the near future, the quantum gate dissipates more energy than the classical gate for error rates $<10^{-6}$. At the required error rate (10^{-25}) of today's classical gates, the quantum gate would need ~ 100 J for a single operation!

In Fig. 4, the power dissipation of single logical gates, classical (CMOS) and quantum, driven with a given clock frequency, versus the error ratio of the gate, is shown. It is supposed that the classical gate changes its state in each clock frequency period. Already at a modest error rate of 10^{-16}, a single quantum gate would require more power than today's microprocessors. At the error rate of a classical logical gate ($\approx 10^{25}$), the single quantum gate would dissipate over 104 megawatts.

As an example, let us now estimate the classical power limit for today's microprocessors (2003). The number of transistors is $\approx 1.5 \times 10^8$. Suppose that, at maximum load, all transistors are effectively changing their logic state at each clock period. The clock frequency is 3 GHz and let the allowed total bit error frequency in the system of the $\approx 1.5 \times 10^8$ transistors be 1/year [1]. Then

$$\varepsilon_c = 1/(3 \times 10^9 \times 3600 \times 24 \times 365 \times 1.5 \times 10^8) = 7.05 \times 10^{-26} \tag{12}$$

and from (11) we obtain

$$P_{total,\min} = -1.5 \times 10^8 \times 3 \times 10^9 \times 4 \times 10^{-21} \ln\left(\frac{\sqrt{3}}{2} \times 7.05 \times 10^{-26}\right) \approx 0.105\,\text{W} \tag{13}$$

This calculation indicates that the energy efficiency of today's microprocessors with ≈ 100 W power dissipation, at the given error probability, is $\sim 0.1\%$. We will see in the following Section that this low error rate, which is necessary for the error-free running of the programs and processor routines, is not necessary for the information content of data. In computers which would be able to utilise the Shannon information, the required power is even less.

7.2 Energy requirement of Shannon-information transfer in single classical (CMOS) and quantum gates

In this Section, we evaluate the energy requirement of Shannon information transfer. This measure, which cannot be improved by error correcting algorithms, is the ultimate one, the real characteristic of performance versus power dissipation. How to see this? Shannon's information channel capacity depends on two factors: the bandwidth and the signal to noise ratio (see below). A greater error probability can be compensated by a greater bandwidth to keep good performance. For example, the poor accuracy of quantum computers can, in principle, be compensated by a sufficiently higher speed of operation. Although the practical realisation of such systems is not obvious, it is still interesting to explore the ultimate limits of performance for Shannon information. If the signal to noise ratio (SNR), which is the ratio of the signal power, P_s, to the noise power, P_N, and the frequency bandwidth B is known, then Shannon's information channel capacity can be calculated by the Shannon formula. For the quantum gate, the SNR is $1/\varepsilon_q$, so using the best case given by Gea-Banacloche (see [2]), we obtain

$$C_q = B \log_2\left(1 + \frac{P_s}{P_N}\right) = \frac{1}{\tau_q} \log_2\left(1 + \frac{1}{\varepsilon_q}\right) = \frac{1}{\tau_q} \log_2\left(1 + \frac{E_q \tau_q}{\hbar}\right) = f_c \log_2\left(1 + \frac{E_q}{f_c \hbar}\right) \tag{14}$$

where the maximum clock frequency is $f_c = 1/\tau_q$. Let us introduce the normalised information channel capacity K (bit/s/W or bit/J), which is the information channel capacity divided by the power dissipation. In the case of quantum gate, it is as follows:

$$K_q = \frac{C_q}{P_q} = \frac{C_q}{f_c E_q} = \frac{f_c}{f_c E_q} \log_2\left(1 + \frac{E_q}{\hbar f_c}\right) = \frac{1}{E_q} \log_2\left(1 + \frac{E_q}{\hbar f_c}\right) \quad (15)$$

where P_q is the power dissipation of the quantum gate when driven by the maximum clock frequency. At nonzero temperature, the following natural limitations occur for the quantum limit:

$$\left.\begin{array}{l} \hbar f_c \gg kT \text{ (thermal (classical) decoherence constraint)}, \\ \hbar f_c \leq E_q \text{ (quantum uncertainty error constraint; compare with equation (1) [2]} \\ E_q \gg kT \text{ (thermodynamical error constraint for quantum gate)} \end{array}\right\} \quad (16)$$

The last condition is required to avoid flipping the logic state of the quantum gate by thermal activation and requires the same kind of thermal noise considerations as (3) and (5). As a practically motivated example, in Fig. 3, the maximum information transfer rate is at 1 W power dissipation (bit/s/W) when the clock frequency is varied and E_q is set so that flipping the quantum gate's state by thermal noise can be neglected ($E_q = 70\,\mathrm{kT}$). The broken line shows how the clock frequency should be decreased to improve performance. The Figure contains the whole range of meaningful working range as expressed by (16). The right hand end of the X-axis corresponds to the classical thermodynamical (thermal decoherence) decoherence limit, and the left hand end to the case limited by quantum uncertainty, $\varepsilon_q = 1$ (see (4).

For a CMOS (classical) gate the error probability ε, when it is small, is

$$\varepsilon_c = \frac{\nu(E_c)}{f_c} = \tau_c \nu(E_c) \quad (17)$$

Because the signal to noise ratio (SNR) is equal to the error probability, the information channel capacity can be given as

$$C_c = B \log_2\left(1 + \frac{P_s}{P_N}\right) = f_c \log_2\left(1 + \frac{f_c}{\nu}\right) = f_c \log_2\left[1 + \frac{\sqrt{3}}{2}\exp\left(\frac{U_{th}^2}{2U_n^2}\right)\right] = f_c \log_2\left[1 + \frac{\sqrt{3}}{2}\exp\left(\frac{E_c}{kT}\right)\right] \quad (18)$$

and for $E_c \gg kT$

$$C_c \approx f_c \frac{E_c}{kT} \quad (19)$$

Thus, in the case of a classical gate, the normalised information channel capacity is

$$K_c = \frac{C_c}{P_c} = \frac{C_c}{f_c E_c} = \frac{f_c}{f_c E_c} \log_2\left[1 + \frac{\sqrt{3}}{2}\exp\left(\frac{E_c}{kT}\right)\right] = \frac{1}{E_c} \log_2\left[1 + \frac{\sqrt{3}}{2}\exp\left(\frac{E_c}{kT}\right)\right] \approx \frac{1}{kT} \quad (20)$$

Equation (20) contains an exact analytical result. It is of interest to note that the dissipation is the same as Landauer's conjecture [8] about the energy dissipation during information erasure at reversible computing, which is about kT per bit. However, it is important to emphasise that our result is not for reversible computing, so it is not directly relevant for Landauer's case. The total energy in the CMOS capacitor is dissipated during discharge which is an irreversible process.

Equation (20) indicates that the classical gate performs better by more than an order of magnitude than the thermal noise free quantum gate ($E_q = 70\,kT$); see Fig. 3. It is important to note that in the limit of

$$kT = E_q = \hbar f_c \quad (21)$$

the quantum gate would have the same performance as the classical, because then $K_q = 1/kT$. Altough this case is excluded by (16), it can be approached. From another angle of view, the conditions described by (21) set the limit when the quantum system becomes classical. This finding and the fact that then (21) yields the same results as (20), confirms that both the classical and quantum theories are on the right track. However, in this limit the error probability ε_q approaches 1 (see Equation (4)), and thus the quantum gate is useless for practical applications. To improve the accuracy to an acceptable level, we have to move toward the strongly quantum limit, but then the value of K_q decreases. A simple estimation shows that for <0.1 error probability caused independently by both the thermal excitation and the quantum measurement, we would need $E_q \geq 2.4\,kT$ (3) and $E_q \geq 10\hbar f_c$ ((4)). As, according to (16), $\hbar f_c \gg kT$, the condition $E_q \geq 10\hbar f_c \gg 10\,kT$ satisfies both relations and then the quantum error would be dominant (0.1). Similar considerations would lead to $E_q \geq 70\,kT$ for an error probability requirement of 1.5% dominated by the quantum error again.

Finally, let us estimate how much power today's microprocessors would need in the most ideal case of a classical CMOS gate for processing of Shannon information. The same conditions as in (13), yield

$$P_{total,min} \approx 1.5 \times 10^8 \times 3 \times 10^9 \times 4 \times 10^{-21} \approx 0.002\,\mathrm{W} \quad (22)$$

which is 50 times less than the minimal power required by today's error rate; see (13). So, compared to the energy requirement of the processing of Shannon information, the energy efficiency of today's microprocessors is $\sim 0.002\%$.

The general conclusion of these derivations is that, in general purpose applications, where we have to have access to each memory element in the computer, classical computers perform many orders of magnitude better when high data accuracy is required. This is not really surprising. The phenomena of zero-point thermal noise [9, 10] show that 'quantum' can produce noise even when 'classical' is certainly silent, which is at zero temperature.