Quid après les lois de Moore et de Koomey ?

Par Vincent MAZAURIC

Principal Scientist, Schneider Electric

Alexia AUFFÈVES

Directrice de recherche au CNRS,
directrice du laboratoire international MajuLab du CNRS
et cofondatrice de la Quantum Energy Initiative

Olivier EZRATTY

Consultant, auteur et cofondateur de la Quantum Energy Initiative

Et Sergio CILIBERTO

Directeur de recherche au CNRS, École normale supérieure de Lyon

Pendant près de cinquante ans, les lois de Moore et de Koomey ont caractérisé les progrès continus des performances computationnelles des microprocesseurs, et accompagné – voire fondé – l'exceptionnelle croissance de l'industrie du semi-conducteur. Ainsi, les ordinateurs sont devenus de plus en plus petits et de moins en moins coûteux, tout en étant de plus en plus rapides et puissants, alimentant ainsi un perpétuel « effet rebond » du secteur des technologies de l'information et de la communication (TIC) qui n'est toujours pas arrivé à « satiété » ! Néanmoins, les fabricants de microprocesseurs se heurtent, depuis quelques années, aux limites physiques des hypothèses qui avaient permis de conjecturer la loi de Koomey. Si bien que l'avenir de l'industrie des semiconducteurs, et plus généralement du secteur des TIC, doit désormais se construire au-delà de la loi de Moore.

Dans le même temps, la massification actuelle a conduit à identifier le secteur des TIC comme étant largement intensif en énergie électrique, et donc fortement émissif en CO_2, mais aussi extractif en matériaux critiques, alors qu'il était perçu comme « immatériel » il y a encore quelques années de cela. Pour envisager le rôle que peuvent jouer les technologies de l'information et de la communication en tant que réponse aux enjeux du développement durable, il faut donc relativiser le concept de performance computationnelle et revenir au lien entre information et énergie, qui a été énoncé, y compris dans le contexte digital, bien avant la loi de Moore. Les lois de Moore et de Koomey n'apparaissent alors que comme des « sentiers » conjoncturels menant à la maturité thermodynamique, qui est mesurée par une tendance vers la réversibilité. Afin que le « data deluge » ne se transforme pas en « mur de l'énergie », d'autres paradigmes devront être envisagés pour accompagner les futurs défis à relever par un secteur des technologies de l'information et de la communication engagé sur la voie de la soutenabilité.

Énergie et information

Le lien entre énergie et information a été pour la première fois évoqué dans le paradoxe du Démon de Maxwell (1867), sans pour autant avoir été élucidé. La résolution de ce paradoxe au cours du XX^{e} siècle a permis de clarifier le rôle thermodynamique des technologies de l'information et d'établir l'équivalence entre information manquante et entropie (Brillouin, 1956).

Du point de vue thermodynamique

Le rôle des technologies de l'information est de maintenir ou d'accroître la connaissance dont on dispose sur un système donné. Elles opèrent donc un traitement des mesures globalement imputé au « Démon de Maxwell ».

D'un point de vue thermodynamique, cette ambition contrarie l'évolution naturelle d'un système isolé vers une augmentation de son « information manquante », si

bien que, pour respecter le second principe de la thermodynamique, l'acquisition d'informations opérée par le biais de ce traitement est forcément moindre que la création d'entropie réalisée par ailleurs par la dégradation d'énergie noble en chaleur.

Le processeur est la machine thermodynamique qui réalise cette acquisition : pour accroître la connaissance d'un système donné ou pour, de manière équivalente, en faire baisser l'entropie, cette machine consomme de l'énergie noble (d'origine électrique) qu'elle dissipe dans un thermostat en conservant globalement les transferts d'énergie, selon le premier principe de la thermodynamique.

La machine duale du processeur étant le moteur de Szilard (voir la Figure 1 ci-contre), la classification thermodynamique suggère de voir un processeur comme une machine frigorifique, dont l'efficacité est définie par un coefficient de performance (CoP) pouvant varier de 0 à l'infini. Le CoP est utile pour calculer le service thermodynamique rendu par la numérisation.

État de l'art thermodynamique résultant du choix de la représentation digitale

Les technologies actuelles s'appuient toutes sur un codage binaire matérialisé par des bits[1] regroupés en registres, sur lesquels des opérations de calcul (respectivement de mémoire) séquencées par une horloge se succèdent sans ambiguïté, grâce à des portes d'électronique logique élémentaires, essentiellement NAND ou NOR[2], qui sont associées en circuits combinatoires (respectivement séquentiels) intégrés dans un processeur (Mange, 1995). Afin de disposer du processeur pour réaliser le calcul suivant, les machines classiques, inspirées de la machine de Turing, obéissent au principe de Landauer, qui fixe le coût énergétique minimal de l'effacement de l'information acquise (Gershenfeld, 1996 ; Landauer, 1961). Qualitativement, le principe de Landauer exprime l'idée qu'il n'est pas possible d'extraire du travail (selon la Figure 1 ci-contre) une fois (la certitude de) l'information binaire perdue, soit par nécessité de libérer le registre pour réaliser le calcul suivant, soit après l'exécution irréversible[3] d'une porte logique : cette opération de perte de mémoire de la valeur antérieure du registre ou de la valeur des entrées d'une porte logique conduit à une dégradation de l'énergie en chaleur, *a minima* correspondant à l'information acquise préalablement ; dans les faits, à l'énergie consacrée à l'opération d'effacement pour la technologie considérée.

Autrement dit, « savoir », pour ensuite « oublier » a un coût énergétique irréductible – en l'espèce $k_B T \ln 2$ (où $k_B = 1{,}38 \times 10^{-23}$J/K est la constante de Boltzmann, et où T°K correspond à la température absolue) – dissipé en chaleur à chaque réinitialisation d'un registre binaire ou à chaque exécution d'une porte logique élémentaire.

Le principe de Landauer a été validé expérimentalement par observation de la dissipation thermique associée à un cycle d'effacement, successivement, sur un dispositif optique piégeant l'information dans un double puits de potentiel (Berut *et al*., 2012 ; Bérut *et al*., 2015 ; Lutz et Ciliberto, 2015), puis sur un dispositif micromécanique (Dago *et al*., 2021). D'autres expériences ont confirmé ce résultat (Jun *et al*., 2014 ; Peterson *et al*., 2016), dont une réalisée sur une structure magnétique proche de celles utilisées pour les mémoires des ordinateurs (Hong *et al*., 2016).

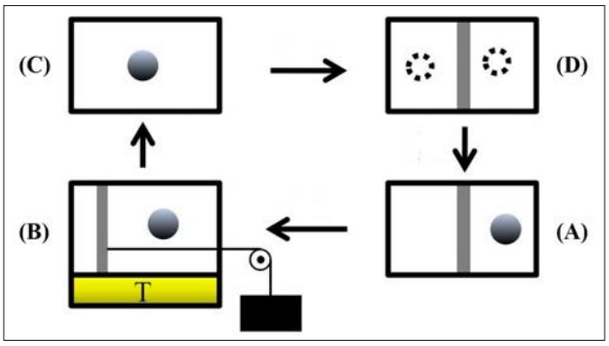

Figure 1 : Moteur de Szilard (Szilard, 1929 ; Szilard, 1964). La connaissance du compartiment occupé par le gaz (A) permet de récupérer un travail gravitaire grâce à l'agitation procurée par une source thermique T°K (B). Au terme du déplacement du piston, l'information sur le compartiment occupé par le gaz est perdue, puisque celui-ci se disperse dans l'ensemble du volume accessible. Mais elle a été convertie en travail, qui, en l'espèce, est au maximum égal à $k_B T \ln 2$ pour une évolution quasi statique, compte tenu du rapport double entre le volume accessible et celui initial (C). La réinitialisation du dispositif (positionnement du piston au centre (D) et assignation du gaz dans un compartiment (A)) permet de fermer le cycle de fonctionnement. La possibilité d'extraire du travail à partir d'informations est formalisée dans Sagawa et Ueda (2008) et dans Toyabe *et al*. (2010). Le cycle peut évidemment être décrit en sens inverse afin de proposer un « processeur mécanique » : un travail de compression (B) d'un gaz occupant initialement l'ensemble du volume accessible (C) permet d'accroître la connaissance sur sa position « moyenne » (A). Une fois le résultat copié, le cycle suivant n'est possible qu'à condition de « remonter » la source de travail mécanique d'au moins $k_B T \ln 2$, établissant ainsi la nécessité de disposer d'une source d'énergie noble pour acquérir à nouveau de l'information (k_B est la constante de Boltzmann).

En confirmant expérimentalement le lien entre théorie de l'information et thermodynamique, le coefficient de performance (CoP) d'une technologie de l'information donnée s'évalue, après « fermeture » du cycle de fonctionnement du processeur, par le rapport entre la valeur énergétique de l'information élémentaire donnée par la limite de Landauer et l'énergie qui lui a été consacrée par la polarisation des circuits d'électronique logique.

Rapportée à un transistor élémentaire[4], l'énergie de commutation en logique CMOS (Complementary

[1] Pour Binary digIT.

[2] Respectivement « NON ET » et « NON OU ».

[3] Une opération logique est dite irréversible, quand la connaissance de la sortie ne permet pas de « remonter » à l'entrée de manière univoque : par exemple, la sortie a NAND b = 1 peut être le résultat de trois entrées différentes : 1 NAND 0, 0 NAND 1 et 0 NAND 0.

[4] Pour fixer les idées, il faut, en logique CMOS, quatre transistors pour réaliser une porte NAND ou NOR, et donc huit transistors pour une mémoire à bascule monostable.

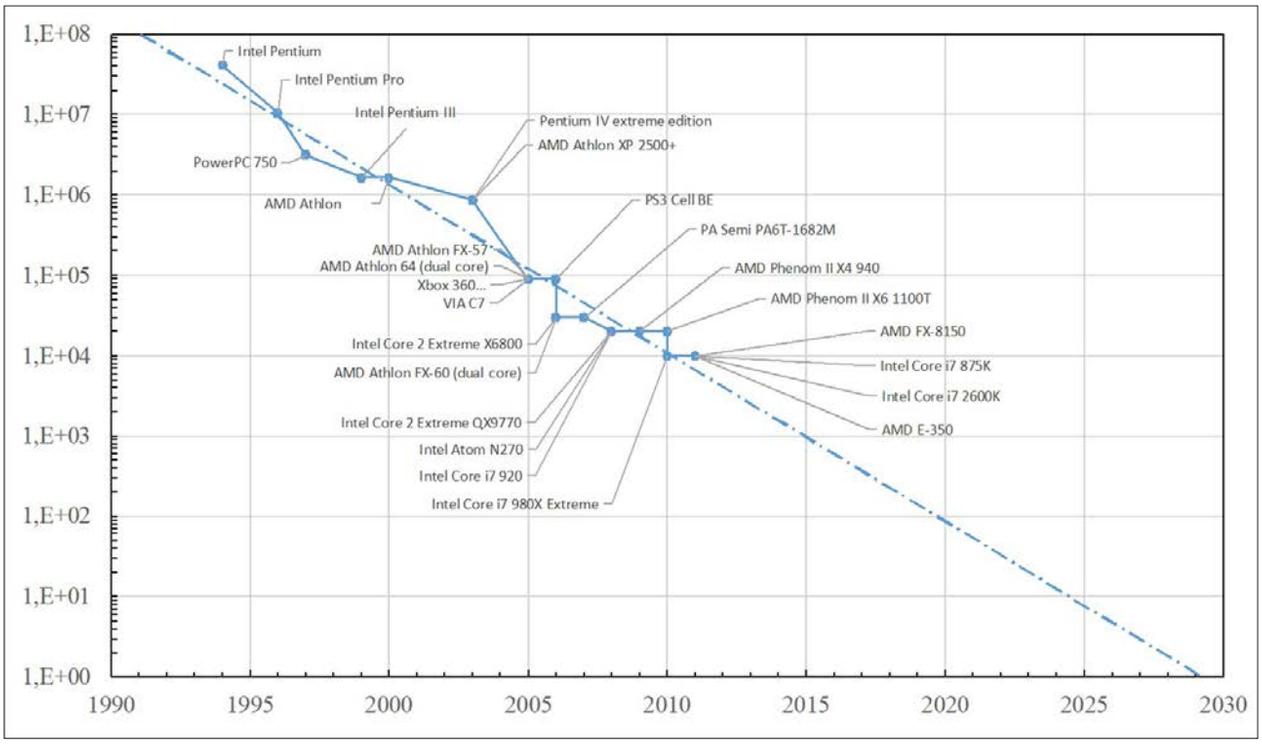

Figure 2 : Énergies de commutation des transistors à effet de champ en technologie CMOS (Complementary Metal Oxyde Semi-conductor) rapportées à la limite de Landauer calculée à la température ambiante (300°K) pour différentes générations de microprocesseurs : l'énergie nécessaire à une commutation est divisée par deux chaque 1,45 année avec un coefficient de détermination R^2=0,95, selon la régression linéaire représentée par le trait mixte – Source : Données extraites de International Technology Roadmap for Semiconductors (ITRS), http://www.itrs2.net

- Si les lois d'échelle de Dennard avaient été respectées, la réduction par deux des énergies de commutation aurait été constatée sur 1,33 année.
- Si les progrès réalisés entre 1994 et 2011 devaient se poursuivre tendanciellement, la limite de Landauer serait atteinte en 2029.

D'autres régressions portant sur des périmètres fonctionnels et des périodes d'analyse différents sont données dans Delahaye (2017) et Zhirnov *et al*. (2014). Elles aboutissent aux mêmes conclusions tendancielles.

Metal Oxyde Semi-conductor) est actuellement de 4 à 5 ordres de grandeur au-dessus de la limite de Landauer (voir la Figure 2 ci-dessus). Avec l'électronique de spin (Joshi, 2016), l'énergie de commutation gagnerait de 2 à 3 ordres de grandeur supplémentaires. Si la performance absolue reste faible, un tel gain permet de gagner un temps précieux face à l'urgence de soutenabilité.

En d'autres termes, les circuits microélectroniques les plus efficients sur un plan énergétique actuellement envisagés consommeraient mille fois plus d'énergie que la limite de Landauer établie pour des machines de Turing irréversibles. Il existe donc d'autres sources de dissipation de l'énergie, qui sont toujours largement dominantes et sur lesquelles portent les efforts de recherche en microélectronique (Ernst, 2022). La réduction de cette dissipation permettrait d'atteindre une « maturité thermodynamique » qui signifierait la fin des lois de Moore et de Koomey.

Les lois de Moore et de Koomey

Depuis l'avènement de la technologie CMOS et des microprocesseurs à la fin des années 1960, les progrès réalisés par les industries de la microélectronique au niveau des procédés lithographiques des substrats semi-conducteurs ont permis d'augmenter l'intégration des transistors selon la loi de Moore (1965), qui prévoyait un doublement de la densité des transistors dans les microprocesseurs tous les deux ans (Moore, 1965). Dans les faits, la loi de Moore a été effectivement vérifiée avec un doublement tous les deux ans (plus précisément au rythme de 1,96 année) entre 1976 et 2001.

Les caractéristiques des transistors CMOS (énergie de polarisation, tension de grille, fréquence de commutation) permises par la réduction de leur taille s'expriment au travers des lois d'échelle de Dennard (Dennard *et al.*, 1974), qui sont à l'origine de la loi de Koomey (Koomey *et al*., 2011). Selon cette loi, pour une surface donnée de semi-conducteur, la performance des microprocesseurs exprimée par le nombre de calculs par unité d'énergie croît encore plus rapidement que la loi de Moore. Dans la Figure 2 ci-dessus, la régression linéaire (en trait mixte) constitue une analyse compatible avec l'énoncé de la loi de Koomey. Alors que l'estimation initiale d'un doublement de la performance des microprocesseurs tous les 18 mois se soit avérée sous-évaluée sur la période 1945-2000 – le nombre des calculs par Joule dépensé ayant doublé à un rythme d'environ 1,57 an (Koomey, 2010), en tenant

compte des effets de la miniaturisation des logiques à transistors des années 1950 sur la performance –, on constate depuis un essoufflement de la loi de Koomey, dont le rythme de progression désormais plus lent que celui de la loi de Moore selon les dernières estimations, exprime la perte de pertinence des lois de Dennard à des échelles inférieures aux lithographies nanométriques, et ce en raison :

- de la limite quantique de la théorie semi-classique décrivant les phénomènes de transport des charges dans les semi-conducteurs à l'origine des équations de Dennard ;
- et de la complexité croissante des liaisons métalliques entre les composants, conduisant à évoluer vers des structures tridimensionnelles et, par conséquent, contrariant l'analyse d'échelle bidimensionnelle proposée par Dennard. Par ailleurs, l'allongement de ces liaisons augmente les temps de propagation des signaux et, de fait, bride la fréquence d'exécution d'un calcul.

Globalement, alors que les pertes dynamiques liées aux commutations des transistors décroissent avec la finesse de la gravure, les pertes statiques causées par les courants de fuite liés à la polarisation des transistors CMOS nanométriques et aux effets capacitifs entre les liaisons métalliques s'accroissent à mesure que l'intégration augmente.

Finalement, la miniaturisation des logiques CMOS n'apparaît plus comme le facteur réconciliant la performance computationnelle avec une efficacité énergétique qui progresserait vers la limite de Landauer, réputée être la limite thermodynamique des machines classiques. Si, à court et moyen terme, l'hybridation avec des technologies issues de l'électronique de spin constitue un levier d'efficacité énergétique, d'autres paradigmes de calcul sont également à envisager dans un souci de soutenabilité des technologies de l'information.

Vers d'autres paradigmes de calcul

Deux causes d'irréversibilité ont été mises en évidence :

- l'irréversibilité d'origine physique induite par la commutation de transistors en un temps fini qui conditionne en partie la performance computationnelle escomptée[5] ;
- l'irréversibilité de nature computationnelle correspondant à l'énergie dissipée en chaleur à chaque réinitialisation d'un registre binaire ou à chaque exécution d'une porte logique élémentaire non bijective, selon le principe de Landauer. Même si cette cause d'irréversibilité reste aujourd'hui marginale (voir la Figure 2 de la page précédente), l'existence d'un « talon » irréductible de dissipation limiterait fondamentalement le recours à la numérisation pour accompagner la soutenabilité des infrastructures, de plus en plus dispersées, des commodités (Mazauric et Ciliberto, 2022).

En adoptant des polarisations par des rampes de tension plutôt que des échelons (Dickinson et Denker, 1995), l'irréversibilité d'origine physique – qui se manifeste, en l'espèce, par la dissipation dans les résistances de charge – peut être arbitrairement réduite en ralentissant suffisamment les processus de charge et de décharge des transistors. Pour que l'exécution du programme s'effectue dans un temps comparable, on doit adopter une parallélisation qui sera d'autant plus importante que les processeurs auront été ralentis (Konopik *et al*., 2023).

Intrinsèque au codage binaire, puis à l'utilisation de portes logiques irréversibles, l'irréversibilité de nature computationnelle pourrait être réduite par l'adoption du calcul réversible (paradigme de Bennett). Sous réserve d'utiliser des portes logiques réversibles et de conserver en mémoire tous les résultats intermédiaires du calcul, l'exécution du programme serait alors « rembobinée » après copie du résultat final afin de récupérer *a posteriori* l'énergie dépensée à chaque opération logique, à l'image de ce que permettrait une machine de Turing réversible (Bennett, 1973 ; Bennett, 1988).

Dans les deux cas, on aboutit à un report de la performance computationnelle sur le « hardware » (surcroît de mémoire, parallélisation massive de processeurs « multi-cores ») si bien que le compromis technologique entre énergie et matière doit être arbitré par une analyse de cycle de vie (ACV) complète, c'est-à-dire incluant les aspects logiciel et matériel, qui reste, à ce jour, encore balbutiante (Orlov *et al*., 2019).

Une même approche systémique mérite d'être adoptée pour les ordinateurs quantiques, avec un nouveau paradigme de calcul promettant une accélération polynomiale ou exponentielle de certains traitements, notamment pour les applications de simulation de la matière, des optimisations diverses ou pour l'apprentissage supervisé. En effet, si la réversibilité est intrinsèque au calcul quantique, elle est obérée par le bruit qui affecte les qubits[6] à l'échelle quantique et impose une forte redondance associée à des codes de correction d'erreurs. Cette dernière amène à un équilibre délicat à réaliser à grande échelle entre la dimension quantique des qubits et celle, classique, de leur contrôle à l'aide de systèmes de génération de micro-ondes, de tensions, de lumière laser et de systèmes informatiques classiques qui sont naturellement irréversibles et énergivores (Auffèves, 2021). L'enjeu est de concevoir des architectures matérielles et logicielles équilibrées, du quantique au classique, ayant le potentiel d'apporter un avantage énergétique par rapport au calcul classique (Auffèves *et al.*, 2022). C'est ce qui motive la « Quantum Energy Initiative » (https://quantum-energy-initiative.org/). Lancée à partir de la France, celle-ci ambitionne de fédérer à l'échelle internationale les savoir-faire

[5] Cette formulation peut sembler contradictoire avec la loi de Koomey. En fait, la miniaturisation permet une décroissance de l'énergie par commutation plus rapide que la croissance de la fréquence, selon les lois de Dennard. Si, toutes choses égales par ailleurs, la fréquence avait augmenté, les pertes se seraient elles aussi accrues, au moins proportionnellement.

[6] Abréviation donnée au bits quantiques.

dans toutes ces disciplines, allant de la recherche fondamentale aux entreprises du secteur, pour analyser, évaluer et optimiser la performance énergétique des ordinateurs quantiques d'aujourd'hui et de demain. Dans un monde aux ressources finies, il est désormais indispensable d'intégrer la question énergétique dès la conception de ces innovations disruptives.

Bibliographie

AUFFÈVES A. (2021), « Optimiser la consommation énergétique des calculateurs : un défi interdisciplinaire », *Reflets de la physique* 69, pp. 16-20.

AUFFÈVES A., EZRATTY O. & WHITNEY R. (2022), « Vers des technologies quantiques responsables », *Revue de l'électricité et de l'électronique* 5, pp. 109-113.

BENNETT C. H. (1973), "Logical Reversibility of Computation", *IBM Journal of Research and Development* 17(6), pp. 525-532.

BENNETT C. H. (1988), "Notes on the history of reversible computation", *IBM Journal of Research and Development* 32(1), pp. 16-23.

BERUT A., ARAKELYAN A., PETROSYAN A., CILIBERTO S., DILLENSCHNEIDER R. & LUTZ E. (2012), "Experimental verification of Landauer's principle linking information and thermodynamics", *Nature* 483, pp. 187-192.

BÉRUT A., PETROSYAN A. & CILIBERTO S. (2015), "Information and thermodynamics: Experimental verification of Landauer's erasure priciple", *Journal of Statistical Mechanics: Theory and Experiment*, P06015.

BRILLOUIN L. (1956), *Science and information theory*, New York (USA), Academic Press.

DAGO S., PEREDA J., BARROS N., CILIBERTO S. & BELLON L. (2021), "Information and Thermodynamics: Fast and Precise Approach to Landauer's Bound in an Underdamped Micromechanical Oscillator", *Physical Review Letters* 126(17), 170601.

DELAHAYE J.-P. (2017), « Vers un calcul sans coût énergétique », *Pour la science* 471, pp. 78-83.

DENNARD R. H., GAENSSLEN F. H., YU H. N., RIDEOUT V. L., BASSOUS E. & LEBLANC A. R. (1974), "Design of ion-implanted MOSFET's with very small physical dimensions", *IEEE Journal of Solid-State Circuits* 9(5), pp. 256-268.

DICKINSON A. G. & DENKER J. S. (1995), "Adiabatic dynamic logic", *IEEE Journal of Solid-State Circuits* 30(3), pp. 311-315.

ERNST T. (2022), « Vers une électronique soutenable dans un monde digital : enjeux et perspectives », *Revue de l'électricité et de l'électronique* 5, pp. 89-95.

GERSHENFELD N. (1996), "Signal entropy and the thermodynamics of computation", *IBM Systems Journal* 35(3&4), pp. 577-586.

HONG J., LAMBSON B., DHUEY S. & BOKOR J. (2016), "Experimental test of Landauer's principle in single-bit operations on nanomagnetic memory bits", *Science Advances* 2(3), e1501492.

JOSHI V. K. (2016), "Spintronics: a contemporary review of emerging electronics devices", *Engineering Science and Technology, an International Journal* 19, pp. 1503-1513.

JUN Y., GAVRILOV M. & BECHHOEFER J. (2014), "High-Precision Test of Landauer's Principle in a Feedback Trap", *Physical Review Letters* 113(19), 190601.

KONOPIK M., KORTEN T., LUTZ E. & LINKE H. (2023), "Fundamental energy cost of finite-time parallelizable computing", *Nature Communications* 14(1), 447.

KOOMEY J. G., BERARD S., SANCHEZ M. & WONG H. (2011), "Implications of Historical Trends in the Electrical Efficiency of Computing", *IEEE Annals of the History of Computing* 33(3), pp. 46-54.

KOOMEY J. G. (2010), "Outperforming Moore's Law", *IEEE Spectrum* 47(3), p. 68.

LANDAUER R. (1961), "Irreversibility and heat generation in the computing process", *IBM Journal of Research and Development* 5(3), pp. 261-269.

LUTZ E. & CILIBERTO S. (2015), "Information: From Maxwell's demon to Landauer's eraser", *Physics Today* 68(9), pp. 30-35.

MANGE D. (1995), *Analyse et synthèse des systèmes logiques*, Lausanne (Suisse), Presses Polytechniques et Universitaires Romandes.

MAZAURIC V. & CILIBERTO S. (2022), « Proposition thermodynamique vers un monde plus électrique : II – Des contraintes de l'exploitation aux externalités », *Revue de l'électricité et de l'électronique* 5, pp. 102-108.

MOORE G. E. (1965), "Cramming More Components Onto Integrated Circuits", *Electronics Magazine* 38(8).

ORLOV A. O., HÄNNINEN I. K., CAMPOS-AGUILLON C. O., CELIS-CORDOVA R., MCCONNELL M. S., SZAKMANY G. P., THORPE C. C., APPLETON B. T., BOECHLER G. P., LENT C. S. & SNIDER G. L. (2019), "Experimental tests of the Landauer Principle in electron circuits, and quasi-adiabatic computing systems", in "Energy limits in computation: A review of Landauer's Principle", *Theory and Experiments* (ed. LENT C. S., ORLOV A. O., POROD W. & SNIDER G. L.), Cham (Switzerland), Springer, pp. 177-230.

PETERSON J. P. S., SARTHOUR R. S., SOUZA A. M., OLIVEIRA I. S., GOOLD J., MODI K., SOARES-PINTO D. O. & CÉLERI L. C. (2016), "Experimental demonstration of information to energy conversion in a quantum system at the Landauer limit", *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 472(2188), 20150813.

SAGAWA T. & UEDA M. (2008), "Second Law of Thermodynamics with Discrete Quantum Feedback Control", *Physical Review Letters* 100(8), 080403.

SZILARD L. (1929), "Über die entropyverminderung in einem thermodynamischen system bei eingriffen intelligenter wesen", *Zeitschrift für Physik* 53, pp. 840-856.

SZILARD L. (1964), "On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings", *Behavioral Science* 9(4), pp. 301-310.

TOYABE S., SAGAWA T., UEDA M., MUNEYUKI E. & SANO M. (2010), "Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality", *Nature Physics* 6(12), pp. 988-992.

ZHIRNOV V., CAVIN R. & GAMMAITONI L. (2014), "Minimum Energy of Computing, Fundamental Considerations", in *ICT – Energy – Concepts Towards Zero – Power Information and Communication Technology* (eds. FAGAS G., GAMMAITONI L., PAUL D. & BERINI G. A.), Rijeka (Croatia), IntechOpen, pp. 139-178.