Reversible Computing Technology is Essential for Sustainable Growth of the Digital Economy

Michael P. Frank, *Sandia National Laboratories* Thomas M. Conte, *Georgia Institute of Technology*

Abstract

The emergence of powerful new capabilities in areas such as artificial intelligence and machine learning (AI/ML) has created an apparently inexhaustible source of demand for ever-increasing "compute" (computational throughput), which will only increase further as AI-powered automation grows more efficient and ubiquitous, cost-effectively filling many economic needs and boosting the productivity of the entire economy. But this can only happen if computing technology itself becomes ever more energy-efficient, particularly given the environmental constraints limiting acceptable growth of aggregate deployed terrestrial power production capacity. However, low-level energy efficiency of conventional "non-reversible" digital technologies can only improve by about an order of magnitude or so before reaching practical physical limits. Improvement by further orders of magnitude will thus require digital bit energies to increasingly be *recovered and recycled* for use in multiple operations, which necessitates migrating to the unconventional computing paradigm called *reversible computing*.

In this position paper, we review the rationale for beginning intensive development of reversible computing technology, and the major challenges that will need to be addressed in its development, at all levels from devices through systems.

1. Introduction

The past few years have brought many impressive advances in the capabilities of compute-intensive systems, particularly those using artificial intelligence (AI) and machine learning (ML) techniques to effectively solve a wide variety of quite general types of problems with a level of competence that rivals, or in some cases even exceeds, human-level abilities. Example areas of fast-emerging proficiency include language arts (Brown *et al.*, 2020), visual arts (Ramesh *et al.*, 2022), music composition (Liu *et al.*, 2022), and visual scene interpretation (Alayrac *et al.*, 2022). In addition, AI-based approaches have greatly accelerated scientific discovery in areas ranging from molecular biology (Jumper *et al.*, 2021) to cosmology (Li *et al.*, 2021).

Further, the cost-efficiency of various AI/ML-based solutions is in many cases already competitive with that of human workers; for example, placing a maximally-sized query to OpenAI's largest GPT-3 natural language model costs[1] from $0.12–0.24 (v1 vs. v2), processes ~2–4 pages of written text input, and returns a high-quality written result. Compare this to a human knowledge worker earning $50K per year, who incurs a comparable cost in just 15–30 seconds of work (ignoring overhead, benefits, and downtime).

For this reason, it's easy to see that AI/ML-powered automation solutions will certainly rapidly expand in adoption and become ubiquitous throughout many areas of the economy, changing the nature of work as many routine, white-collar tasks become far less expensive per unit of output; we can imagine that a typical human desk worker's role will shift towards managing the operation of AI-powered linguistic or artistic tools, thereby amplifying per-worker productivity and overall GDP growth. Automated solutions will come to permeate increasingly many sectors of the economy, as well as enabling the development of large new markets, such as various segments of decentralized finance (DeFi), and entire new categories of economic activity centered around virtual reality "metaverses." We can foresee a burgeoning "digital economy," in which, over time, an increasing portion (and total amount) of civilization's economic activity takes place in the digital realm and is powered by computing technology.

As the digital economy grows more productive, demand for the continued expansion of its aggregate level of intensity (in terms of, *e.g.*, computational operations carried out, globally, per unit time) will only increase. But, at the same time, the need for responsible stewardship of our planetary environment and natural resources demands that we *not* allow the intensity of our energy production to increase nearly as much, or at nearly as high a rate, as our computing intensity.

Put simply, if we want the aggregate global number of computational operations per second to continue increasing at X% per year, but we can only tolerate deployed power production capacity increasing at Y% per year, with $X > Y$, then this can only happen sustainably if computational energy efficiency continues increasing at an adequate rate. As an example, if we would like for aggregate compute to continue to grow by 68%/year (an historically typical[2] rate), but deployed energy production capacity can only grow by 8%/year (which is the

[1] The pricing scale is at https://openai.com/api/pricing/; max query sizes at https://beta.openai.com/docs/engines/gpt-3.

[2] Using the aggregate maximum sustained performance of all systems on the TOP500 list (top500.org) as a proxy for the growth rate of total global compute intensity, the increase from 1.1 Tflop/s in June 1993 to 3.04 Eflop/s in Nov. 2021 corresponds to an average compute growth rate of ~68.4% per year over this period.

forecast for the near-term growth of renewables), then computational energy efficiency will need to grow at a rate of about 56%/year.[3] In the past, computational energy efficiency actually grew a bit faster than this (at ~60%/yr.) on average (Koomey *et al.*, 2010), but more recently, low-level efficiency growth slowed to about 30% per year (Koomey & Naffziger, 2016) and appears likely to slow further due to fast-approaching physical limits of conventional semiconductor-based logic.

Briefly, we can summarize the physical reasons for these limits, which are quite fundamental. The efficient operation of complex switching circuits requires suppressing switch conductance by a substantial factor (*e.g.,* 10^9) between "on" and "off" states. Then Boltzmann statistics requires that suppressing the probability of an event (*e.g.*, an electron occupying a channel state) by a factor of r requires an energy difference of at least $kT \ln r$;[4] for the example value, this is $\sim 20\ kT$. Due to various nonidealities, this then translates to closer to $\sim 100\ kT$ for a minimum-sized transistor, $\sim 1{,}000\ kT$ for a typical optimally sized transistor in high-performance logic, and $\sim 10{,}000\ kT$ after accounting for various other overheads (fanout, wiring, multiple transistors per logic gate). These are conservative estimates. Altogether, that's ~250 eV or 40 attojoules (aJ) at room temperature, but today, logic signal energies are already only ~700 aJ (IRDS, 2021), which gives us only at most 7 more years of improvement (if we assume a 50% per year rate) before we hit the physical limits.[5] After we hit this wall, (at least) one of the following must occur:

(1) The growth rate of aggregate compute intensity will have to slow, throttling back the growth rate of the digital economy, and of overall economic abundance;

(2) The fraction of deployed power production capacity devoted to computing will have to increase, progressively choking off other, more "analog" sectors of the economy from energy and growth, more and more so as computing's share of the total rises;

(3) The growth rate of *total* global deployed energy production capacity will have to accelerate, thereby making the attainment of environmental and resource sustainability far more elusive; or,

(4) Computing will be forced to shift to a new dominant paradigm, and more specifically, to one very different from today's digital technology.

It is our contention that (1)–(3) are all rather dramatically undesirable, and that therefore, (4) will be essential if we wish to maintain sustainable growth of the digital (and overall) economy, and of the aggregate global level of abundance.

Thus, we are in dire need of a new computing paradigm. Below we discuss why *reversible computing* is, indeed, the only new paradigm for *digital* computing that might potentially be long-term sustainable.

2. The Need for Reversible Computing

Above we discussed why, in practice, at least ~10,000 kT energy per logic signal is needed in field-effect transistor technology, *even at the very end of* the development roadmap for such devices. Moreover, this practical limit arises from fundamental thermodynamic and quantum-mechanical principles and cannot be trivially circumvented. Any technology that aims to replace today's CMOS (complementary metal-oxide-semiconductor) technology, for purposes of general digital computing, will need to solve the same problems, including how to build the equivalent of a digital logic gate, and how to interconnect gates together, while still accounting for the various nonidealities and overheads that may arise. Thus, any viable computing technology will still have a minimum digital signal energy, likely framed as a significant multiple of the thermal energy kT, that will be required for effective operation within complex computing hardware.

Further, we know from fundamental principles of statistical physics and information theory (Frank, 2018) that in *no physically possible technology* can the amount of available energy associated with a bit's worth of digital information, which is lost when the physical system encoding that bit is allowed to thermalize, ever be less than the *Landauer limit* of $kT \ln 2$ (or ~18 meV at room temperature). This follows directly from the modern understanding that *physical entropy* is just unknown *physical information*; this understanding arose out of the early work by Boltzmann, Planck, and other pioneers.

As a result of this fundamental principle, the *only* long-term sustainable paradigm for digital computing is the *reversible computing* paradigm in which loss of digital bits, and their associated energy, is avoided. This idea was first shown to be theoretically coherent by Bennett (1973), and since then, research on the concept has proceeded, albeit at a modest pace. As a result, today we understand quite well how reversible computing can be implemented using various technologies, including the field-effect transistors used in CMOS (Frank *et al.*, 2020b), though much work remains to further develop reversible tech towards large-scale practicality.

[3] The calculation here is just $1.684/1.08 \cong 1.56$.

[4] Here, k is Boltzmann's constant of $\sim 1.4 \times 10^{-2}$ J/K, and T is the ambient temperature. At a typical operating temperature around 300 K, the thermal energy kT is 26 meV.

[5] The IRDS itself is actually more conservative than this, targeting only a ~2× reduction in signal energy (from 650 aJ to 330 aJ) from 2022 to 2034, which is only about a 5.8% efficiency boost per year—not even close to what's needed.

Note that here, we are *not* referring to reversible computing in the way that it's used in quantum computing, to help enable fast quantum algorithms for certain problems, but without concern for energy dissipation in more general computing workloads. Rather, here we are speaking of the use of reversible computing in a more *classical* (*i.e.*, non-quantum) mode, to make *all* digital computing far more energy efficient.

One widespread misconception about reversible computing is that it saves *only the last* $kT \ln 2$ of energy, and thus isn't useful when signal energies are much larger than this – but that isn't the case! When reversible computing principles are properly applied, one can save the vast majority of the *entire* signal energy, and approaching *all* of it, as the technology is further refined. How does this work? Simply put, it leverages a combination of two basic physical principles:

(1) Use of nearly *adiabatic* operation of individual digital elements, that is, transforming between distinct digital states at a speed that is slow compared to the rate at which the system naturally settles down to a stable state, but fast compared to the rate at which the element dissipates its energy to the thermal environment. In the case of CMOS, this means switching at a rate that is slow compared to the RC switching time constant of a given logic gate, but fast compared to the gate's leakage time constant. This is not difficult, as a typical leading-edge CMOS technology today has a switching time constant around 1 ps, and a leakage time constant of at least 1 ms (this difference corresponding to the $r = 10^9$ on/off conductance ratio mentioned earlier). At speeds in between these, the energy loss from switching can be reduced (Frank & Shukla 2021) by a factor of up to $\sim\sqrt{r}$, which in this example is $\sim 30{,}000\times$. Due to the increased energy efficiency, the relatively slow speed of individual device operations can be overcome by operating much larger numbers of devices in parallel, so long as per-device manufacturing cost continues to decrease.

(2) Use of nearly *ballistic* evolution of resonant energy-recovering driving systems. For a rough mental picture of what this would mean in the mechanical domain, think of a vibrating tuning fork, or a spinning flywheel. Resonant all-electrical circuits can be built as well (Frank *et al*., 2020). The efficiency of such periodically evolving systems can be characterized in terms of their *quality factor*, Q, which is roughly the number of cycles they can "coast" along, unpowered, before losing the lion's share ($1 - \mathrm{e}^{-1} = \sim 63\%$) of their energy; or equivalently, $\sim 1/Q$ is the fraction of energy lost per cycle. The attainable Q value also limits the system's energy efficiency.

Thus, fundamentally, highly energy-efficient computation in an adiabatic switching circuit, driven by a ballistic dynamical evolution encompassing it, "only" requires refining the technology so as to increase Q and $\sqrt{r}$ to commensurately high levels. This is by no means a trivial task in general, but it offers a *sustainable path* by which we can proceed forwards. That is, no fundamental *limits* are known on these quantities.

What is the connection between adiabatic switching and the reversible computing concepts mentioned earlier? It is simply that a fully adiabatic transformation of a physical system is necessarily also reversible – if you perform the same transformation in reverse, you'll get back to where you started. In particular, you can't merge together two possible physical states sitting at very different energies without irreversibly losing a substantial fraction of the energy of the higher-energy state. Thus, a non-reversible digital transformation *can't* be adiabatic, or even *approach* full adiabaticity. Again, the reasons for this go back to Landauer's principle, and the deep connection between entropy and information in physics.

Thus, we contend that, if the world wants to get very far beyond the practical limits of non-reversible CMOS technology, this is absolutely going to *require* diligently applying the principles of reversible computing and adiabatic/ballistic systems mentioned above. We further expect that systems based on these principles can begin to become competitive within a 5–10-year timeframe *if* there is aggressive investment to support their development between now and then.

2. Engineering Challenges

What is required for reversible computing to move forward and be ready to carry on the torch of ever-increasing computing efficiency once the progress of the roadmap for conventional non-reversible CMOS reaches its endpoint? The important challenges span multiple levels of the computing technology stack, from basic physics through systems.

(1) **Fundamental physics.** In anticipation of the long-term need for further improvement of minimum energy dissipation, energy-delay product, and other important figures of merit, the fundamental quantum thermodynamic limits of reversible computing need to be intensively studied (Frank & Shukla, 2021).

(2) **Fabrication process development.** In near-term technologies such as reversible CMOS, manufacturing processes need to be optimized with adiabatic operation in mind. In particular, the peak energy efficiency of reversible CMOS is influenced more strongly by the on-off ratio r than is conventional non-reversible CMOS. Thus, device selections with larger r values than is typical need to be developed (Frank *et al.*, 2020). Also, due to the improved energy efficiency, the number of layers of active logic can and should be

further increased, and cost per layer reduced, compared to what's justifiable for non-reversible technologies.

(3) **Device characterization and modeling.** For applications at ultra-low temperatures, *e.g.*, for control of qubits in quantum computer engineering (Frank & DeBenedictis 2018), the behavior of (preferably high-r) CMOS devices in adiabatic circuits at very low temperatures needs to be studied and characterized. Well-validated device models are needed that remain accurate at low temperatures.

(4) **Cell library development.** Well-optimized adiabatic CMOS logic cell libraries are needed, across a range of competitive fabrication processes, to support design development.

(5) **EDA tool support.** Electronic design automation (EDA) tool suites need to be extended to support the unique architectural needs and design constraints of the adiabatic/reversible approach throughout the entire design flow. This will require appropriate extensions to hardware description languages (HDLs), as well as support from tools for design synthesis and verification.

(6) **Reversible algorithms.** Much more study is needed on the design of well-optimized algorithms (both hardware algorithms for use in logic designs and IP blocks, and also software algorithms for reversible processing architectures) which embody well-analyzed tradeoffs between time, space, and energy resources in an adiabatic/reversible context (Demaine *et al.*, 2016).

(7) **IP blocks and architectural components.** Design components need to be developed to support a wide range of digital applications, including development of ASICs, configurable ICs such as FPGAs, and various processor types including MCUs, DSPs, CPUs, GPUs, and especially in the fast-growing market segment of specialized AI accelerator chips.

(8) **RF design, advanced packaging, and heterogeneous integration.** Design of high-quality resonant elements, and their integration and/or packaging together with adiabatic ICs, presents a formidable engineering challenge which will need to be tackled in order for room-temperature applications to be highly competitive in terms of system-level energy-efficiency, although some benefit might still be gained even at lower Q values via offloading thermal burdens from the IC, which can enable higher 3D packing densities and reduce parasitic loading on interconnects.

(9) **Globally asynchronous system design.** As the size of a digital system increases, it becomes more difficult to maintain globally synchronous operation, and the energy costs implied by synchronizing communication between different asynchronous domains could eventually limit the long-term scale-up of the benefits that can be attained through reversible operation (Earley, 2021). However, the issues here are not yet clear and require further, more detailed study.

(10) **Application and systems software.** As we pursue the reversible computing development path, and the degree of reversibility of hardware gradually increases, eventually we'll even require system and application software to become "reversibility-aware." Instruction set architectures and programming languages with explicit support for reversibility in the language will eventually be needed (Frank, 1999).

3. Conclusion

Due to the significant environmental impacts of increased energy production, achieving long-term sustainable growth of the terrestrial economy will soon require the trillion-dollar computing hardware industry to shift its focus from traditional, non-reversible digital computing towards the unconventional *reversible* computing paradigm. Far from breaking any laws of physics, reversible computing simply leverages our existing well-developed understanding of adiabatic and ballistic physical processes and applies it carefully in the design of novel mechanisms and structures for digital computing. There is nothing preventing this technology from being rapidly brought from concept to practical reality in time to keep the growth of the digital economy from stalling; it will only take aggressive investment.

In essence, the difference between aggressively pursuing reversible computing technology, versus not doing so, is the difference between enabling our local civilization's level of abundance to continue increasing to arbitrarily high levels, versus consigning it to eventual technological stagnation and catastrophic resource overshoot, risking civilizational collapse. Thus, in our view, there can be no greater or clearer motivation to immediately begin a major new push to overhaul the very foundations of computing and rejuvenate the flourishing of the digital economy with a fresh new vision.

Acknowledgement

This work is supported in part by the Advanced Simulation and Computing (ASC) program at the U.S. Department of Energy's National Nuclear Security Administration (NNSA). Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for NNSA under contract DE-NA0003525. This document describes objective technical results and analysis. Any subjective views or opinions that might be expressed in this document do not necessarily

represent the views of the U.S. Department of Energy or the United States Government. Approved for public release, SAND2022-7826 O.

References

Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: a Visual Language Model for Few-Shot Learning. *arXiv preprint arXiv:2204.14198*.

Bennett, C. H. (1973). Logical reversibility of computation. *IBM journal of Research and Development*, *17*(6), 525-532.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Demaine, E. D., Lynch, J., Mirano, G. J., & Tyagi, N. (2016, January). Energy-efficient algorithms. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science* (pp. 321-332).

Earley, A. H. (2021). *On the performance and programming of reversible molecular computers* (Doctoral dissertation, University of Cambridge).

Frank, M. P. (1999). *Reversibility for efficient computing* (Doctoral dissertation, Massachusetts Institute of Technology).

Frank, M. P. (2018, September). Physical foundations of Landauer's principle. In *International Conference on Reversible Computation* (pp. 3-33). Springer, Cham.

DeBenedictis, E. P., & Frank, M. P. (2018). The national quantum initiative will also benefit classical computers [rebooting computing]. Computer, 51(12), 69-73.

Frank, M. P., & Shukla, K. (2021). Quantum Foundations of Classical Reversible Computing. *Entropy*, *23*(6), 701.

Frank, M. P., Brocato, R. W., Conte, T. M., Hsia, A. H., Jain, A., Missert, N. A., ... & Tierney, B. D. (2020, October). Special Session: Exploring the Ultimate Limits of Adiabatic Circuits. In *2020 IEEE 38th International Conference on Computer Design (ICCD)* (pp. 21-24). IEEE.

Frank, M. P., Brocato, R. W., Tierney, B. D., Missert, N. A., & Hsia, A. H. (2020b, December). Reversible computing with fast, fully static, fully adiabatic CMOS. In *2020 International Conference on Rebooting Computing (ICRC)* (pp. 1-8). IEEE.

IRDS, the IEEE International Roadmap for Devices and Systems (2021). https://irds.ieee.org/editions/2021.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589.

Koomey, J., Berard, S., Sanchez, M., & Wong, H. (2010). Implications of historical trends in the electrical efficiency of computing. *IEEE Annals of the History of Computing*, *33*(3), 46-54.

Koomey, J., & Naffziger, S. (2016). Energy efficiency of computing: what's next. *Electronic Design*, *28*.

Li, Y., Ni, Y., Croft, R. A., Di Matteo, T., Bird, S., & Feng, Y. (2021). AI-assisted superresolution cosmological simulations. *Proceedings of the National Academy of Sciences*, *118*(19).

Liu, J., Dong, Y., Cheng, Z., Zhang, X., Li, X., Yu, F., & Sun, M. (2022). Symphony generation with permutation invariant language model. *arXiv preprint arXiv:2205.05448*.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*.