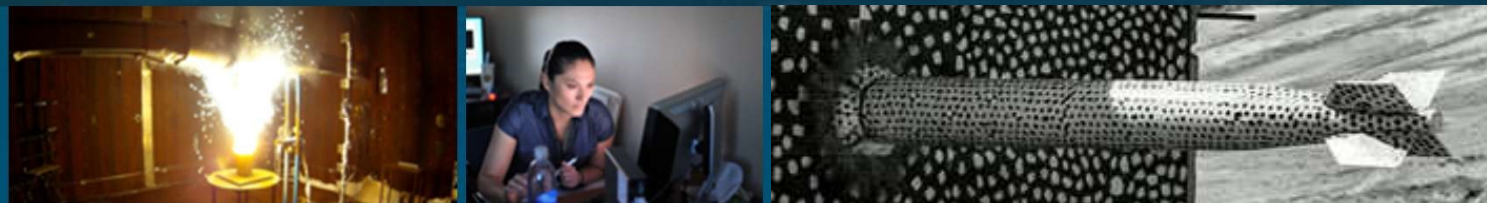


Semiconductor Research Corporation (SRC)
Decadal Plan Committee Meeting
Deep Dive Presentation



Current Status of Reversible Computing



Friday, May 14th, 2021

Michael P. Frank, Center for Computing Research

with collaborators: Robert Brocato, Rupert Lewis, Nancy Missert & Brian Tierney (Sandia), Kevin Osborn & Lingqi Yu (LPS), Erik DeBenedictis (Zettaflops, LLC), Karpur Shukla (Brown), Rudro Bismas & Dewan Woods (Purdue), Tom Conte & Anirudh Jain (Georgia Tech).

Approved for public release, SAND2021-5928 PE



Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

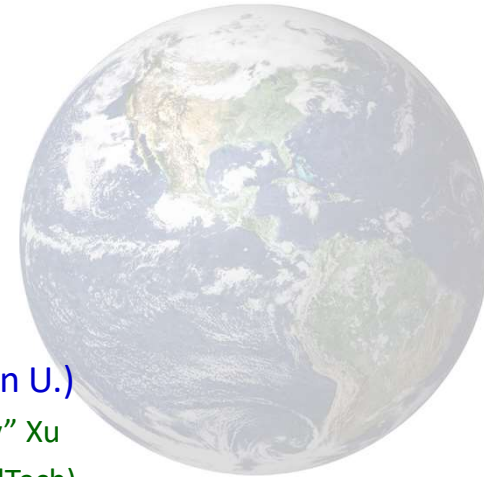
Contributors to the larger effort:

- Full group at Sandia:

- Michael Frank (Cognitive & Emerging Computing)
- Robert Brocato (RF MicroSystems)
- David Henry (MESA Hetero-Integration)
- Rupert Lewis (Quantum Phenomena)
- Nancy Missert (Nanoscale Sciences)
 - Matt Wolak (now at Northrop-Grumman)
- Brian Tierney (Rad Hard CMOS Technology)

- Thanks are also due to the following colleagues & external collaborators:

- Erik DeBenedictis
- Kevin Osborn (LPS/JQI)
 - Liuqi Yu
- Steve Kaplan
- Rudro Biswas (Purdue)
 - Dewan Woods
- Karpur Shukla (CMU/Brown U.)
 - w. Prof. Jingming “Jimmy” Xu
 - Also w. Victor Albert (CalTech)
- Tom Conte (Georgia Tech/CRNCH)
 - Anirudh Jain
- David Guéry-Odelin (Toulouse U.)
- FAMU-FSU College of Engineering:
 - Sastry Pamidi (ECE Chair)
 - Jerris Hooker (Instructor)
 - 2019-20 students:
 - Frank Allen, Oscar L. Corces, James Hardy, Fadi Matloob
 - 2020-21 students:
 - Marshal Nachreiner, Samuel Perlman, Donovan Sharp, Jesus Sosa



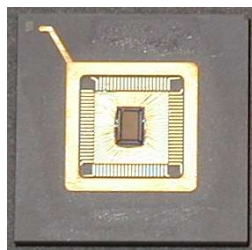
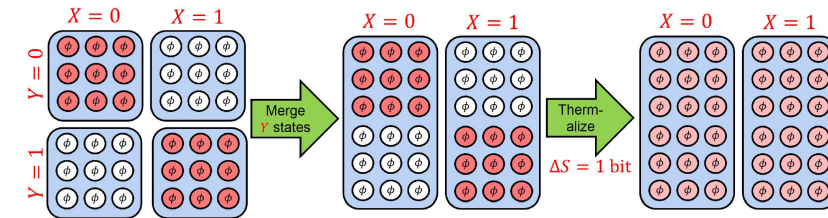
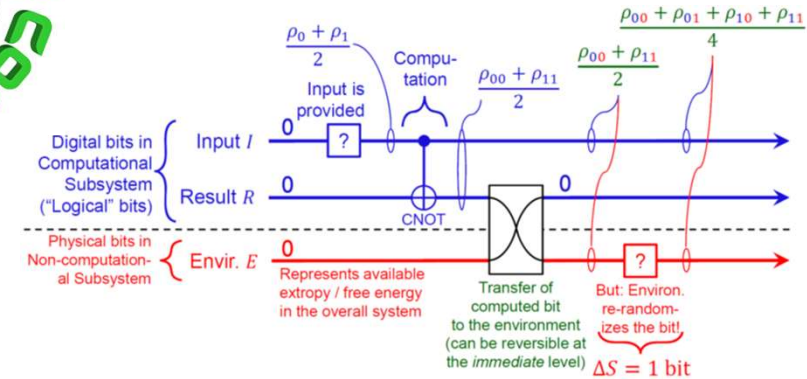
Thanks are due to Sandia’s LDRD program, DOE’s ASC program, and the DoD/ARO ACI (Advanced Computing Initiative) for their support of this line of research!

Talk Abstract/Outline

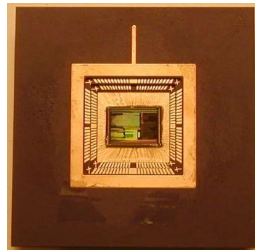


Current Status of (classical) Reversible Computing (RC)

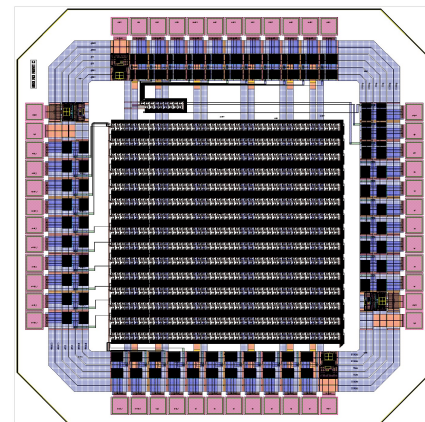
- Since the field's early developments in the 1970s-1980s, significant strides have been made at multiple levels:
 - Improved understanding of the fundamental physics of computing
 - Clear and rigorous formulations now exist for Landauer's Principle & the fundamental theorems of reversible computing
 - Analyses of the *asymptotic scaling* and the associated case for RC from an *economic and systems engineering* perspective
 - Bottom line: RC wins, in the long term, despite its overheads
 - Concrete demonstrations of how to implement RC exist for both adiabatic CMOS and superconducting platforms
- The field is now ready for a much more intensive level of development to begin.



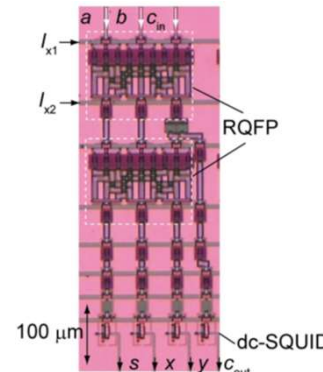
FlatTop (MIT '96)



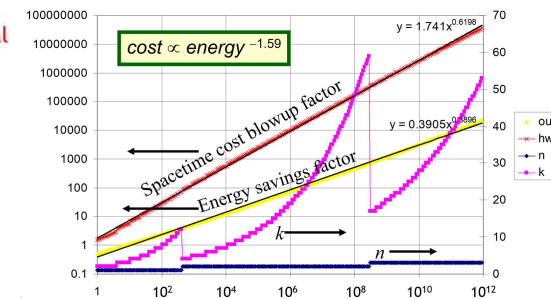
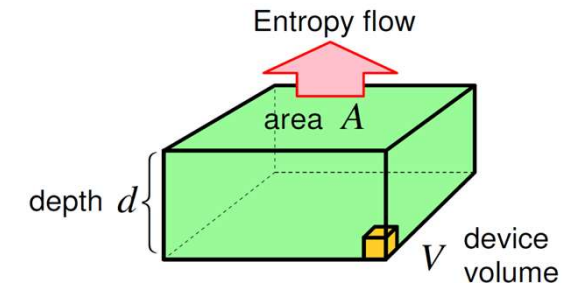
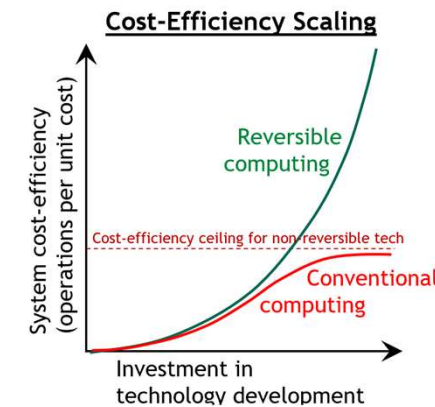
Pendulum (MIT '99)



2LAL Shift Register (Sandia '20)



RQFP Full Adder (Yokohama '18)



Semiconductor Roadmap is Ending...

Thermal noise on gate electrodes of minimum-width segments of FET gates leads to significant channel PES fluctuations if $E_g \lesssim 1-2 \text{ eV}$!

- This increases leakage, impairs practical device performance
 - Thus, roadmap has minimum gate energy asymptoting to $\sim 2 \text{ eV}$

Further, real logic circuits incur many *compounding* overhead factors *multiplying* this raw transistor-level limit:

- Transistor width $10-20\times$ minimum width for fastest logic.
- Parasitic (junction, etc.) transistor capacitances ($\sim 2\times$).
- Multiple (~ 2) transistors fed by each input to a given logic gate.
- Fan-out of each gate to a few (~ 3) downstream logic gates.
- Parasitic wire capacitance ($\sim 2\times$).

Due to all these overhead factors, the energy of each logic bit in real logic circuits is necessarily many times larger than the minimum-width gate energy!

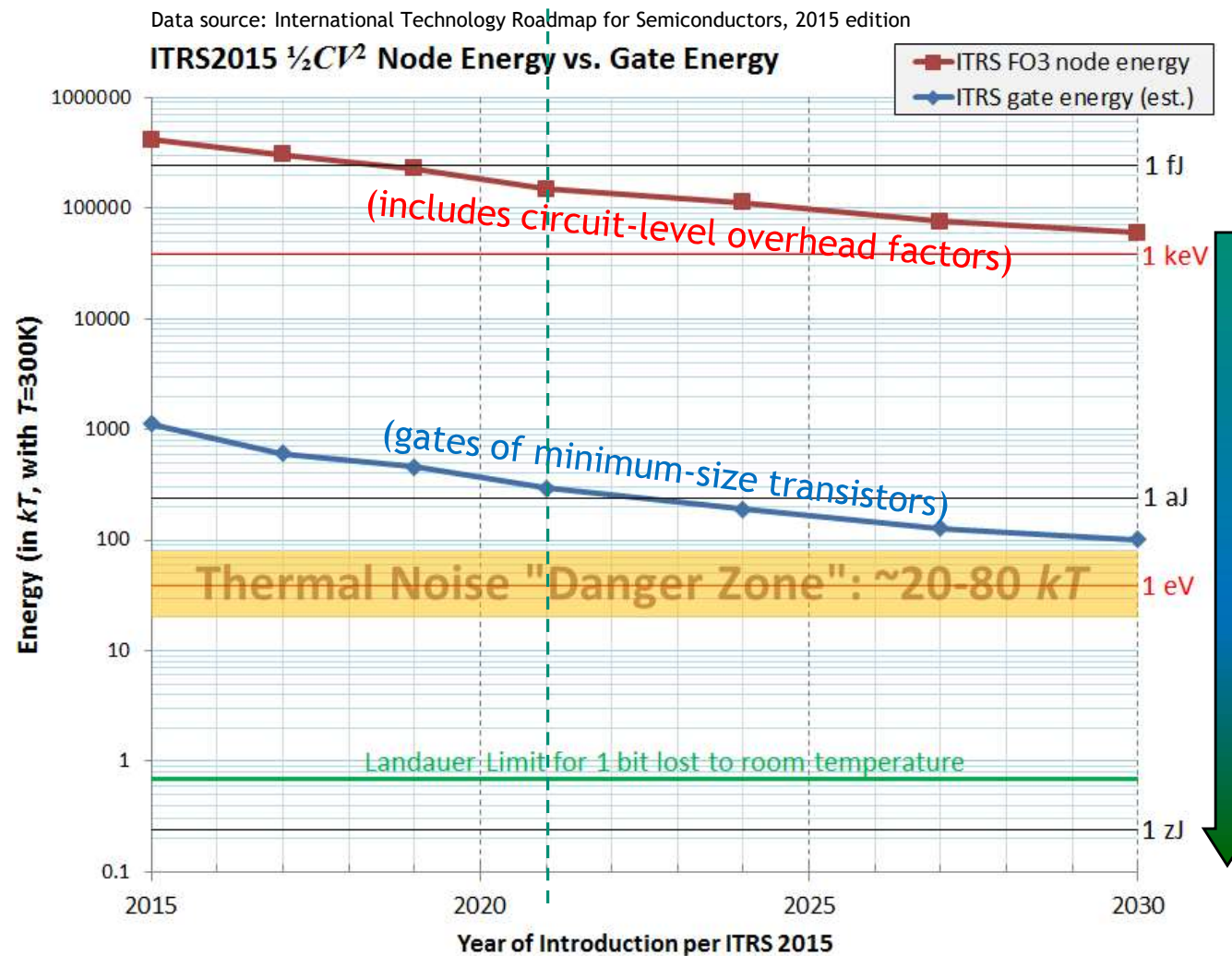
- $375-600\times$ (!) larger in ITRS'15.
 - \therefore Practical bit energy for irreversible CMOS logic asymptotes to $\sim 1 \text{ keV}$!

Practical, real-world logic circuit designs can't just magically cross this $\sim 500\times$ architectural gap!

- \therefore Thermodynamic limits imply much larger practical limits!
 - The end is near!

This is Now!

Only about a decade left...



Only reversible computing can take us from $\sim 1 \text{ keV}$ at the end of the CMOS roadmap, all the way down to $\ll kT$.



Section I. Fundamental Physics of Computing

Current Status of Reversible Computing

Fundamental Physics of Computing—Earliest Roots



This topic can be placed on a firm theoretical foundation using tools from the field of *non-equilibrium quantum thermodynamics* (NEQT), the theoretical formulation of which derives from the mathematical foundation first laid down by von Neumann (1927). →

- However, even before von Neumann, the roots of modern stat. mech., thermodynamics and quantum theory were *already inseparable*.
- What we know today as “Boltzmann’s constant” k was actually first derived by Planck, in the *very same* analysis that simultaneously first resolved the value of what we now call “Planck’s constant” h .
- Statistical mechanics could never possibly have become a complete, coherent foundation for thermodynamics without the concomitant discovery of quantum mechanics! *Quantization is crucial*.

Some key foundational principles of NEQT are the following:

- *Unitary time evolution* of all closed systems (including the whole universe \mathcal{U})
- **NOTE:** von Neumann entropy $S = -\text{Tr}(\rho \ln \rho)$ is conserved by unitary transforms.
- Environment \mathcal{E} of an open system \mathcal{S} is treated as *independent* and *thermal*.
- Entropy *increase* can be viewed as merely a natural consequence of our inability as modelers to track quantum correlations (incl. entanglement) with (or within) any complex thermal environment \mathcal{E} .

Perspective is summarized in the definition of *thermal operations* derived from the (1955) *Stinespring Dilation Theorem*:

$$\rho_{\text{in},\mathcal{S}} \mapsto \mathbb{E}[\rho_{\text{in},\mathcal{S}}] := \text{Tr}_{\mathcal{E}} \left[\hat{U}_{t,\mathcal{S}\mathcal{E}} (\rho_{\text{in},\mathcal{S}} \otimes \tau_{\mathcal{E}}) \hat{U}_{t,\mathcal{S}\mathcal{E}}^\dagger \right]$$

$$= \text{Tr}_{\mathcal{E}} \left[e^{-i\hat{H}t} (\rho_{\text{in},\mathcal{S}} \otimes \tau_{\mathcal{E}}) e^{i\hat{H}t} \right].$$

Trace out correlations w. environment → Thermal state of environment

Thermodynamik quantenmechanischer Gesamtheiten.

Von

J. v. Neumann, Berlin.

Vorgelegt von Max Born in der Sitzung vom 11. November 1927.

Einleitung.

I. In meiner Arbeit „Wahrscheinlichkeitstheoretischer Aufbau der Quantenmechanik“¹⁾ wurde gezeigt, wie die quantenmechanische Statistik aus einigen einfachen und rein qualitativen physikalischen Grundannahmen²⁾, sowie dem folgenden formalen Prinzip: die physikalischen Größen α eines gegebenen Systems \mathcal{S} entsprechen eindeutig und umkehrbar den (hermiteisch-)symmetri-

9. Ueber das Gesetz der Energieverteilung im Normalspectrum; von Max Planck.

(In anderer Form mitgeteilt in der Deutschen Physikalischen Gesellschaft, Sitzung vom 19. October und vom 14. December 1900, Verhandlungen 2. p. 202 und p. 237. 1900.)

Einleitung.

Die neueren Spectralmessungen von O. Lummer und E. Pringsheim¹⁾ und noch auffälliger diejenigen von H. Rubens und F. Kurlbaum²⁾, welche zugleich ein früher von H. Beckmann³⁾ erhaltenes Resultat bestätigten, haben

Hieraus und aus (14) ergeben sich die Werte der Naturkonstanten:

$$(15) \quad h = 6,55 \cdot 10^{-27} \text{ erg} \cdot \text{sec},$$

$$(16) \quad k = 1,346 \cdot 10^{-16} \frac{\text{erg}}{\text{grad}}.$$

Das sind dieselben Zahlen, welche ich in meiner früheren Mitteilung angegeben habe.

Fundamental Physics of Computing—Earliest Roots, *cont.*

“Shannon’s” 1948 entropy formula $H = -\sum p \log p$ was historically rooted in Boltzmann’s 1872 “H-theorem” (*cf.* E^* quantity below)

- Its importance was well already established in statistical mechanics by the time of von Neumann’s (1920s) work on quantum thermodynamics.

However, Shannon did introduce key concepts such as mutual information, $I(X; Y) = H(X) + H(Y) - H(X, Y)$.

- The concept that information-bearing digital states can be identified with *sets* of (digitally interpreted) microstates also dates back to this era.

NOTE: Shannon *never once* addressed energy *dissipated*, only *invested*.

- There is nothing in Shannon’s (or von Neumann’s) work that contradicts RC.

Weitere Studien über das Wärmegleichgewicht unter Gas-
molekülen.

Von Ludwig Boltzmann in Graz.

(Mit 4 Holzschnitten.)

(Vorgelegt in der Sitzung am 10. October 1872.)

Die mechanische Wärmetheorie setzt voraus, dass sich die Moleküle der Gase keineswegs in Ruhe, sondern in der lebhaftesten Bewegung befinden. Wenn daher auch der Körper seinen Zustand gar nicht verändert, so wird doch jedes einzelne seiner Moleküle seinen Bewegungszustand beständig verändern, und

$$E^* = N \iint f^* \log f^* ds d\sigma.$$

The Bell System Technical Journal

Vol. XXVII

July, 1948

No. 3

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist¹ and Hartley²

Quantities of the form $H = -\sum p_i \log p_i$ (the constant K merely amounts to a choice of a unit of measure) play a central role in information theory as measures of information, choice and uncertainty. The form of H will be recognized as that of entropy as defined in certain formulations of statistical mechanics³ where p_i is the probability of a system being in cell i of its phase space. H is then, for example, the H in Boltzmann’s famous H theorem. We shall call $H = -\sum p_i \log p_i$ the entropy of the set of probabilities

* See, for example, R. C. Tolman, “Principles of Statistical Mechanics,” Oxford, Clarendon, 1938.

Communication in the Presence of Noise*

CLAUDE E. SHANNON†, MEMBER, IRE

Summary—A method is developed for representing any communication system geometrically. Messages and the corresponding signals are points in two “function spaces,” and the modulation process is a mapping of one space into the other. Using this representation, a number of results in communication theory are deduced concerning expansion and compression of bandwidth and the threshold effect. Formulas are found for the maximum rate of transmission of binary digits over a system when the signal is perturbed by various types of noise. Some of the properties of “ideal” systems which transmit at this maximum rate are discussed. The equivalent number of binary digits per second for certain information sources is calculated.

* Decimal classification: 621.38. Original manuscript received by the Institute, July 23, 1940. Presented, 1948 IRE National Convention, New York, N. Y., March 24, 1948; and IRE New York Section, New York, N. Y., November 12, 1947.

† Bell Telephone Laboratories, Murray Hill, N. J.

I. INTRODUCTION

A GENERAL COMMUNICATIONS system is shown schematically in Fig. 1. It consists essentially of five elements.

1. *An information source.* The source selects one message from a set of possible messages to be transmitted to the receiving terminal. The message may be of various types; for example, a sequence of letters or numbers, as in telegraphy or teletype, or a continuous function of time $f(t)$, as in radio or telephony.

2. *The transmitter.* This operates on the message in some way and produces a signal suitable for transmission to the receiving point over the channel. In teleph-

Landauer's Principle from Basic Statistical Physics & Information Theory

(For further details, see arXiv:1901.10327)



When stated *correctly*, proving Landauer's Principle is elementary...

- *I.e.*, it takes only a small handful of simple logical steps to prove;
- Depends *only* on basic facts of statistical physics and information theory.

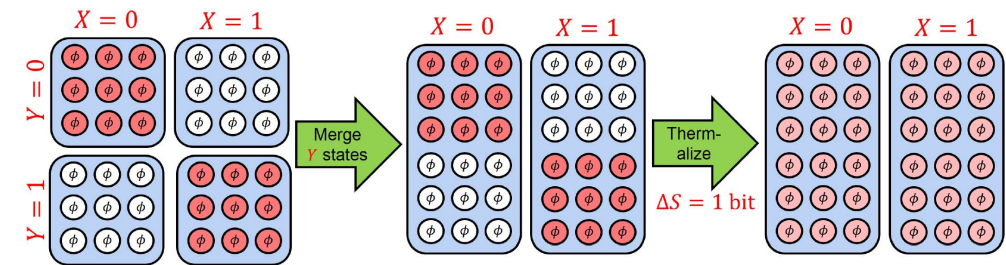
Here's a *correct* statement of Landauer's Principle:

- Within any computational process composed out of *local, digital* primitive transformations, the *oblivious* (*i.e.*, local and unconditional) *erasure* (to a standard state) of a digital subsystem \mathcal{Y} that possesses *marginal* digital entropy $H(Y)$ (entropy after restriction of the joint $\mathcal{X}\mathcal{Y}$ distribution to \mathcal{Y}) and was *deterministically computed* from another subsystem \mathcal{X} necessarily increases *total physical entropy* S by at least $H(Y)$.
 - **Corollary:** Free energy is reduced by $\Delta F = -H(Y) \cdot T$, and expulsion of entropy to environment results in heat $\Delta Q = H(Y) \cdot T$.
 - **Generalization:** Any local reduction of \mathcal{Y} 's marginal entropy by any amount $-\Delta H(Y)$ affects free energy and heat proportionately.

Here's a simple proof:

1. The Second Law of Thermodynamics ($\partial S / \partial t \geq 0$), together with the statistical definition of entropy, imply that microphysical dynamics *must* be *bijective* (this is reflected *e.g.* in the unitarity of quantum time-evolution).
2. Given that \mathcal{Y} was computed *deterministically* from \mathcal{X} , its conditional entropy $H(Y|X) = 0$, and therefore its marginal entropy is *entirely* accounted for by its mutual information with \mathcal{X} , *i.e.*, $H(Y) = I(X; Y)$.
3. Because microphysics is bijective, local transformations *cannot destroy* the information $I(X; Y)$ but can only *eject* it out to some other subsystem (if not part of the machine's stable, digital state, it's in the thermal state).
4. Thermal environments, by definition, *don't preserve* correlation information at all (as reflected by, *e.g.*, thermal operations *a la* Stinespring); therefore, the total universe entropy gets increased by $\Delta S = I(X; Y) = H(Y)$.
 - This can be seen through the trace operation over \mathcal{E} , or more simply by just observing that joint entropy $H(X, Y) = H(X) + H(Y) - I(X; Y)$ over two systems increases by $I(X; Y)$ if the original mutual information $I(X; Y)$ is replaced with a new value $I'(X; Y) = 0$.

Oblivious erasure of subsystem \mathcal{Y} when $y = x$



Basic Reversible Computing Theory

(For full proofs, see [arxiv.org:1806.10183](https://arxiv.org/1806.10183))

Fundamental theorem of traditional reversible computing:

- A deterministic computational operation is (unconditionally) non-entropy-ejecting if and only if it is *unconditionally* logically reversible (injective over its entire domain).

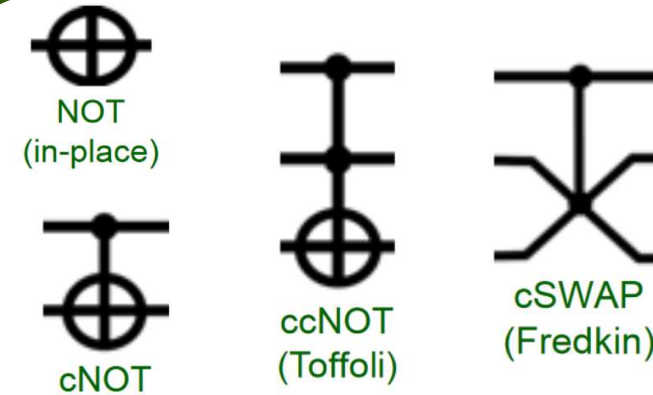
Fundamental theorem of generalized reversible computing:

- A *specific* (contextualized) deterministic computational process is (specifically) non-entropy-ejecting if and only if it is *specifically* logically reversible (injective over the set of *nonzero-probability* initial states).
- Also, for any deterministic computational operation, which is conditionally reversible under some assumed precondition, then the entropy required to be ejected by that operation approaches 0 as the probability that the precondition is satisfied approaches 1.

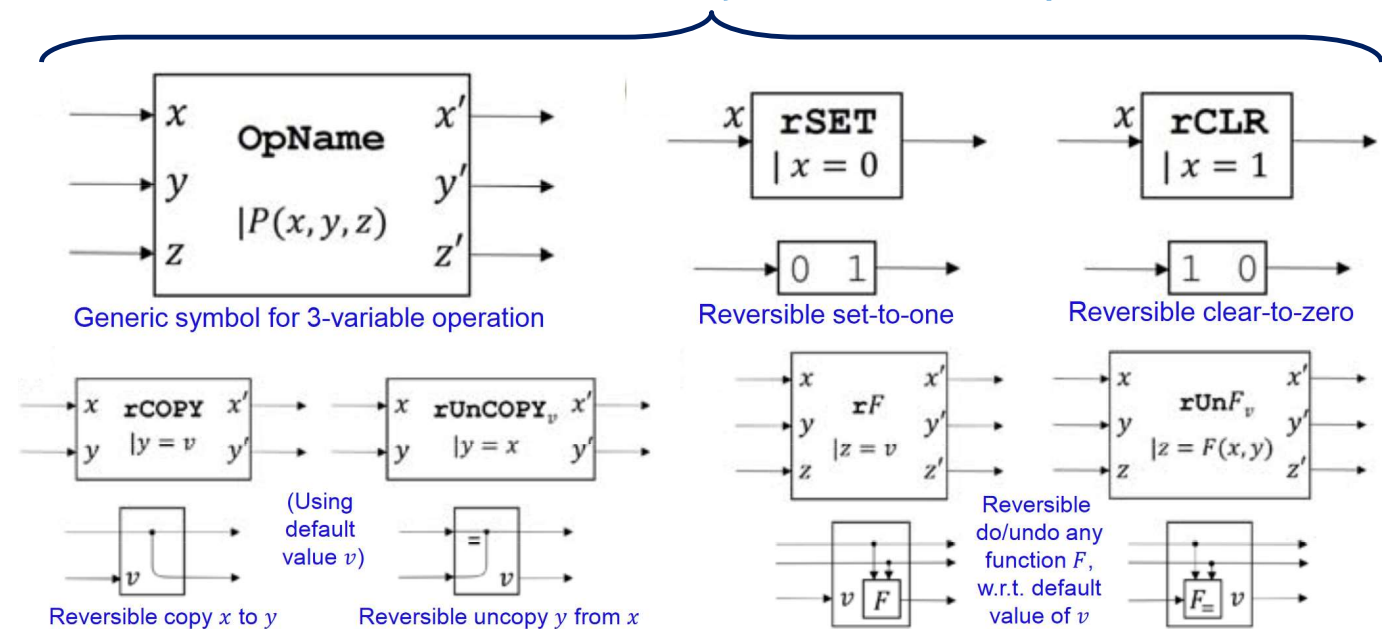
Bottom line: To avoid requiring Landauer costs, it is *sufficient to just have reversibility when some specified preconditions are satisfied*.

- Basis for practical engineering implementations.
- Exemplified by Adiabatic CMOS.

Traditional *Unconditionally* Reversible “Gates” (Operations)



Generalized *Conditionally* Reversible Operations



Latest Work! (with Karpur Shukla, Brown University)

This paper shows *rigorously* that the results summarized on the preceding slides are 100% consistent with the entirety of modern non-equilibrium quantum thermodynamics, incl. multiple theoretical treatments of Landauer's principle based on that framework that have been developed over the last few decades.



Article

Quantum Foundations of Classical Reversible Computing

Michael P. Frank ^{1,*} and Karpur Shukla ^{2,*}

- ¹ Center for Computing Research, Sandia National Laboratories, Albuquerque, NM 87185, USA
² Department of Electrical and Computer Engineering, Brown University, Providence, RI 02906, USA
 * Correspondence: mpfrank@sandia.gov (M.P.F.); karpur_shukla@brown.edu (K.S.);
 Tel.: +1-505-284-4103 (M.P.F.); +1-646-580-5277 (K.S.)
 † These authors contributed equally to this work.

Abstract: The reversible computation paradigm aims to provide a new foundation for general classical digital computing that is capable of circumventing the thermodynamic limits to the energy efficiency of the conventional, non-reversible digital paradigm. However, to date, the essential rationale for, and analysis of, classical reversible computing (RC) has not yet been expressed in terms that leverage the modern formal methods of non-equilibrium quantum thermodynamics (NEQT). In this paper, we begin developing an NEQT-based foundation for the physics of reversible computing. We use the framework of Gorini-Kossakowski-Sudarshan-Lindblad dynamics (a.k.a. *Lindbladians*) with multiple asymptotic states, incorporating recent results from resource theory, full counting statistics and stochastic thermodynamics. Important conclusions include that, as expected: (1) Landauer's Principle indeed sets a strict lower bound on entropy generation in traditional non-reversible architectures for deterministic computing machines when we account for the loss of correlations; and (2) implementations of the alternative *reversible* computation paradigm can potentially avoid such losses, and thereby circumvent the Landauer limit, potentially allowing the efficiency of future digital computing technologies to continue improving indefinitely. We also outline a research plan for identifying the fundamental minimum energy dissipation of reversible computing machines as a function of speed.

Keywords: non-equilibrium quantum thermodynamics; thermodynamics of computing; Landauer's principle; Landauer limit; reversible computing; resource theory of quantum thermodynamics; Gorini-Kossakowski-Sudarshan-Lindblad dynamics; Lindbladians; von Neumann entropy; Rényi entropy; open quantum systems



Citation: Frank, M.P.; Shukla, K. Quantum Foundations of Classical Reversible Computing. *Entropy* **2021**, *23*, 701. <https://doi.org/10.3390/e23060701>

Academic Editor: Neal G. Anderson

[doi:10.3390/e23060701](https://doi.org/10.3390/e23060701)



Quantum Foundations of Classical Reversible Computing

Volume 23 • Issue 6 | June 2021

mdpi.com/journal/entropy
ISSN 1099-4300

Can dissipation scale better than linearly with speed?



Some observations from Pidaparathi & Lent (2018) suggest Yes!

- Landau-Zener (1932) formula for quantum transitions in e.g. scattering processes with a missed level crossing...
 - Probability of exciting the high-energy state (which then decays dissipatively) scales down *exponentially* as a function of speed...

$$P_D = e^{-2\pi\Gamma}$$
 - This scaling is commonly seen in many quantum systems!
- Thus, dissipation-delay *product* may have *no lower bound* for quantum adiabatic transitions—*if* this kind of scaling can actually be realized in practice.
 - *I.e.*, in the context of a complete engineered system.
- **Question:** Will unmodeled details (e.g., in the driving system) fundamentally prevent this, or not?

J. Low Power Electron. Appl. 2018, 8(3), 30; <https://doi.org/10.3390/jlpea8030030>

Open Access Article

Exponentially Adiabatic Switching in Quantum-Dot Cellular Automata

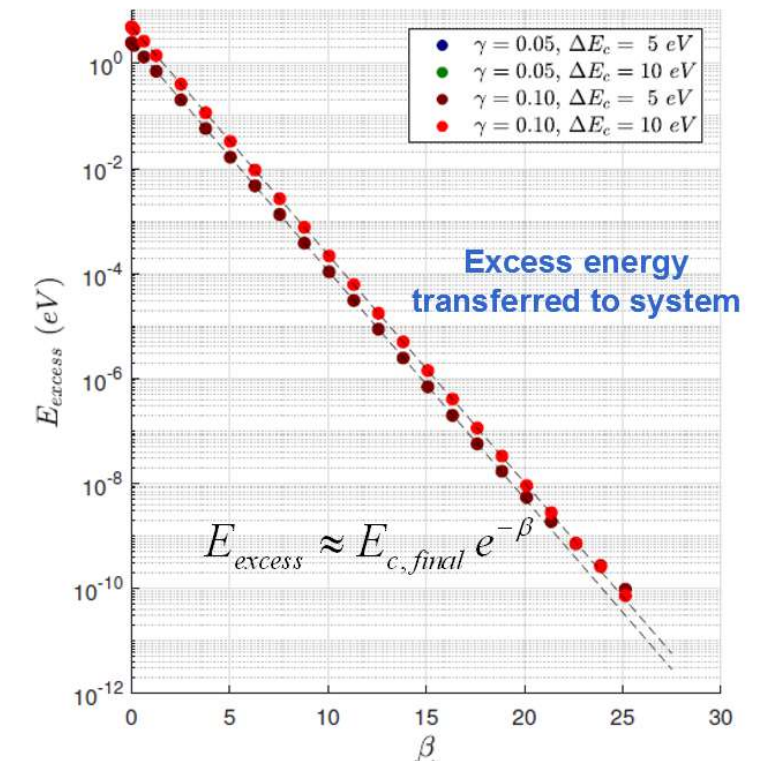
Subhash S. Pidaparathi and Craig S. Lent*

Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

* Author to whom correspondence should be addressed.

Received: 15 August 2018 / Revised: 5 September 2018 / Accepted: 5 September 2018 / Published: 7 September 2018

(This article belongs to the Special Issue Quantum-Dot Cellular Automata (QCA) and Low Power Application)



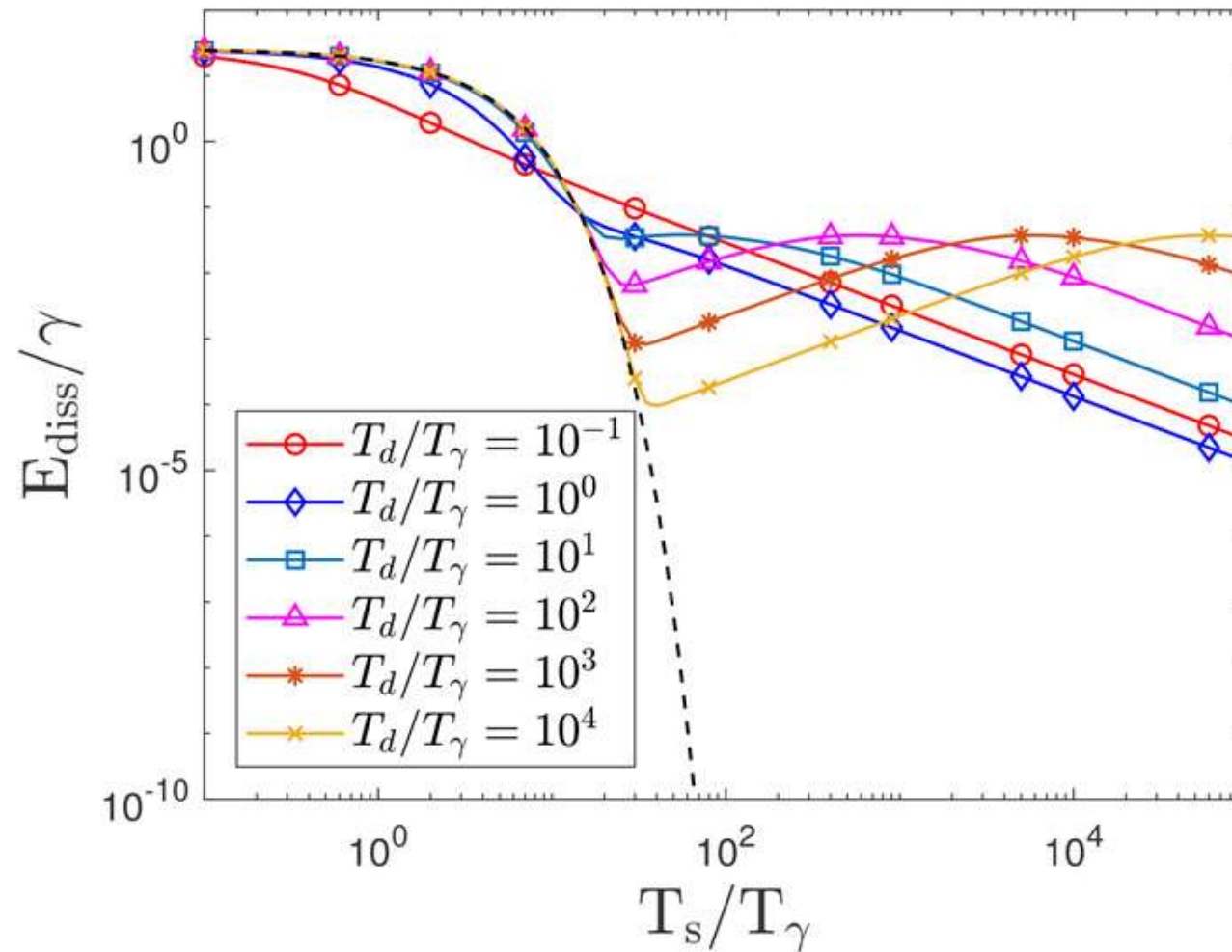
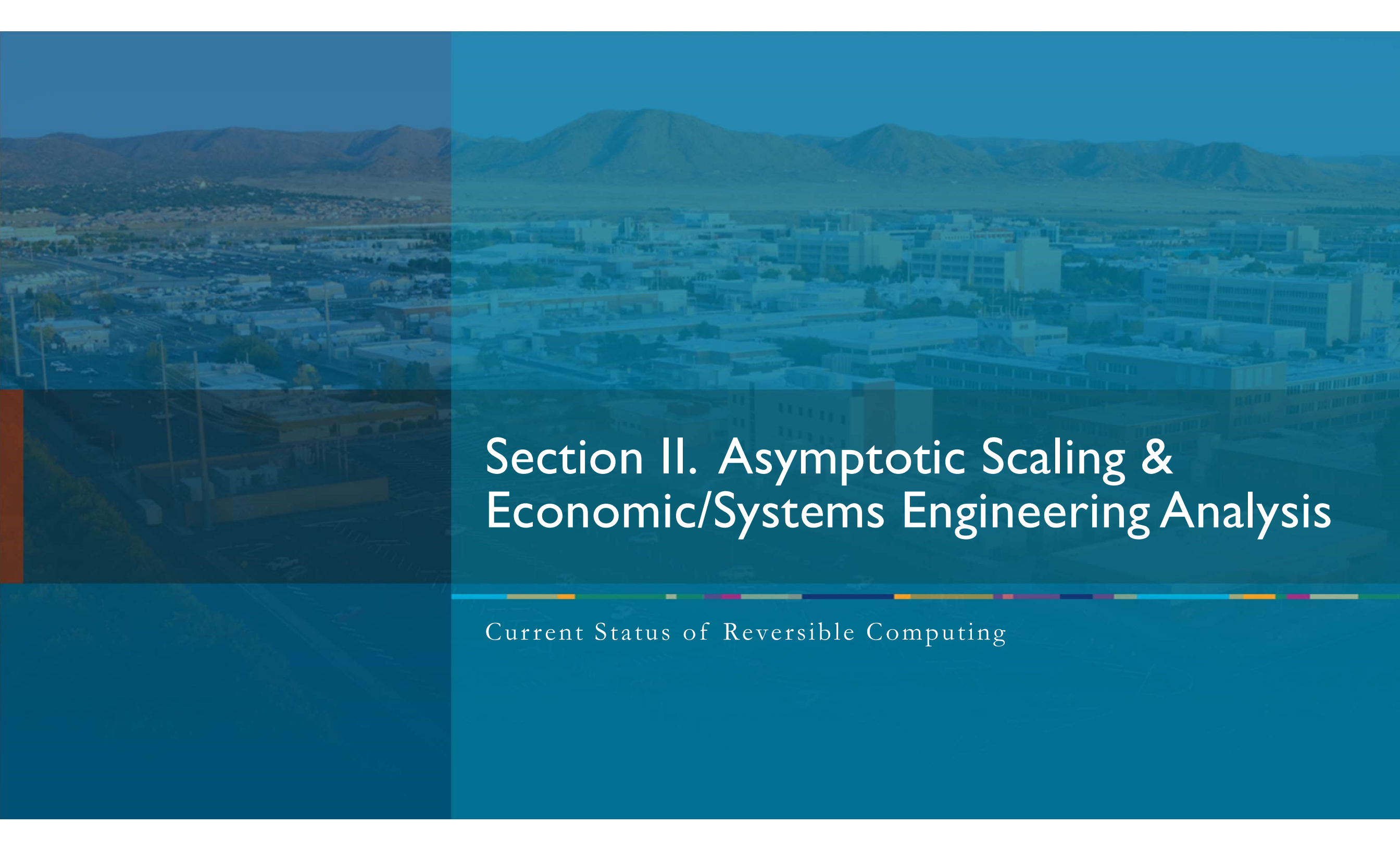


FIG. 10. Dissipated energy of an open system as a function of switching speed for different dissipation time constants. The dashed line is the excess energy of an isolated system. Here, the environmental temperature $k_B T / \gamma = 0.5$.



Section II. Asymptotic Scaling & Economic/Systems Engineering Analysis

Current Status of Reversible Computing

Motivation from Economics / Systems Engineering



In general, *efficiency* η of any process can be defined as the amount P of some valued *product produced* by the process, divided by the amount C of *cost consumed* (in terms of resources, or dollars) by the process.

$$\eta = \frac{P}{C}$$

- For a computing system,
 - P can be amount of useful *information processing performed* (e.g., number of operations) by the system over its operating lifetime, and
 - C can be expressed the sum of manufacturing (& deployment) costs, plus operating costs over the system lifetime.
- We can also annualize the costs, in terms of, e.g. time-amortized manufacturing cost.
 - More sophisticated variations that account for net present value of future returns, depreciation curves, *etc.*, not considered here.
- Operating costs largely amount to *energy-proportioned costs*: $C_{\text{oper}} = c_{\text{en}} \cdot E_{\text{oper}}$
 - c_{en} = operating cost per unit of energy dissipated; E_{oper} = total energy dissipated during a given period of operation.

$$C = C_{\text{tot}} = C_{\text{mfg}} + C_{\text{oper}} \quad (\text{may be time-amortized})$$

We can thus reduce the efficiency formula $\eta = P/C_{\text{tot}}$ for computing to the form at right:

- E_{op} = Energy dissipated due to *one* primitive device operation (or by one primitive device in time t_d).
- $c_{\text{dev},t}$ = Amortized manufacturing cost per primitive device per unit time t .

$$\begin{aligned} \eta &= \frac{1}{c_{\text{en}} \cdot E_{\text{op}} + c_{\text{dev},t} \cdot t_d} \\ &= \frac{1}{E_{\text{op}} t_d \left(\frac{c_{\text{en}}}{t_d} + \frac{c_{\text{dev},t}}{E_{\text{op}}} \right)} \end{aligned}$$

Some observations from this equation.:

- There are *diminishing* efficiency returns from decreasing *either* E_{op} or the $c_{\text{dev},t} \cdot t_d$ term in isolation
 - \therefore Continuing to push non-reversible technologies will ultimately reach a dead end!
- Note that if *both* E_{op} and $c_{\text{dev},t}$ were decreased by $N\times$, overall efficiency would be increased by $N\times$. (All else being equal.)
- Decreasing $E_{\text{op}} \cdot t_d$ (dissipation-delay product, DdP) is *often* (but not always!) a win.
 - E.g., in scenarios where total lifetime cost of operation starts out very heavily energy-dominated, total cost can be reduced by lowering E_{op} , *even* in cases where $E_{\text{op}} t_d$ stays the same, or even increases somewhat!
- However, at any given per-device cost, decreasing $E_{\text{op}}(t_d)$ (dissipation as a function of delay) for any given delay value t_d is *always a win*.
 - Thus, this will be our focus in future work.

Why Reversible Computing Wins Despite Its Overheads!

$$\eta = \frac{P}{C}$$



Bumper-sticker slogan: “*Running Faster by Running Slower!*” (Wait, what?) More precisely:

- Reversible technology is so energy-efficient that we can overcome its overheads (including longer transition times!) by using much greater parallelism to increase overall performance within system power constraints.
- This is borne out by a detailed economic/systems-engineering analysis.

Bottom line: The computational *performance per unit budgetary cost* on parallelizable computing workloads can become as large as desired, given only that *both terms* in this expression for total *cost per operation* C_{op} can be made sufficiently small:

$$C_{op} = c_E \cdot E_{diss,op} + c_M (s_{elem} \cdot t_{delay}).$$

where:

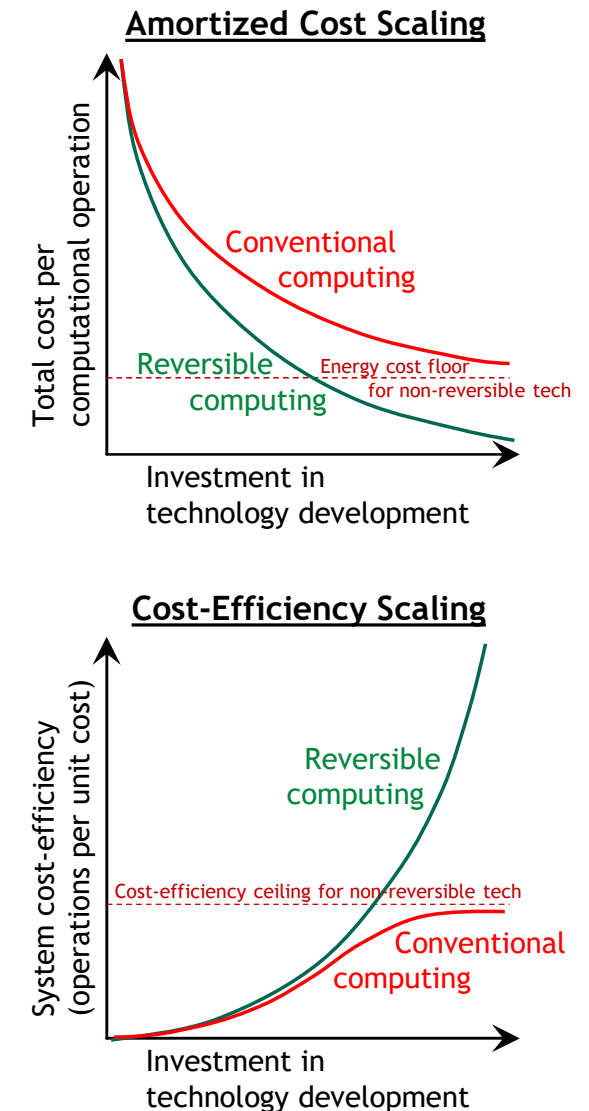
- c_E is the operating cost C_{oper} attributable to supplying power/cooling, divided by energy delivered.
- $E_{diss,op}$ is the system energy dissipation, divided by number of operations performed.
- c_M is the total cost C_{mfg} for system manufacturing & installation, *divided by* the number n_{elem} and physical size s_{elem} (in appropriate units) of individual computing elements, & the system’s total useful lifetime t_{life} .
- t_{delay} is the average time delay between instances of re-use of each individual computing element.

Two key observations:

- The cost per operation of *all* conventional computing *approaches a hard floor* due to Landauer.
 - Assuming *only* that the economic cost of operation *per Joule delivered* cannot become arbitrarily small.
- But, there is no clear barrier to making the manufacturing cost coefficient c_M *ever smaller* as manufacturing processes are refined (and/or the deployed lifetime of the system increases).

\therefore Nothing prevents system-level cost efficiency of reversible machines from becoming *arbitrarily* larger than conventional ones, *even* if we have to scale t_{delay} and/or s_{elem} up as we scale $E_{diss,op}$ down!

$$C_{tot} = C_{mfg} + C_{oper}$$



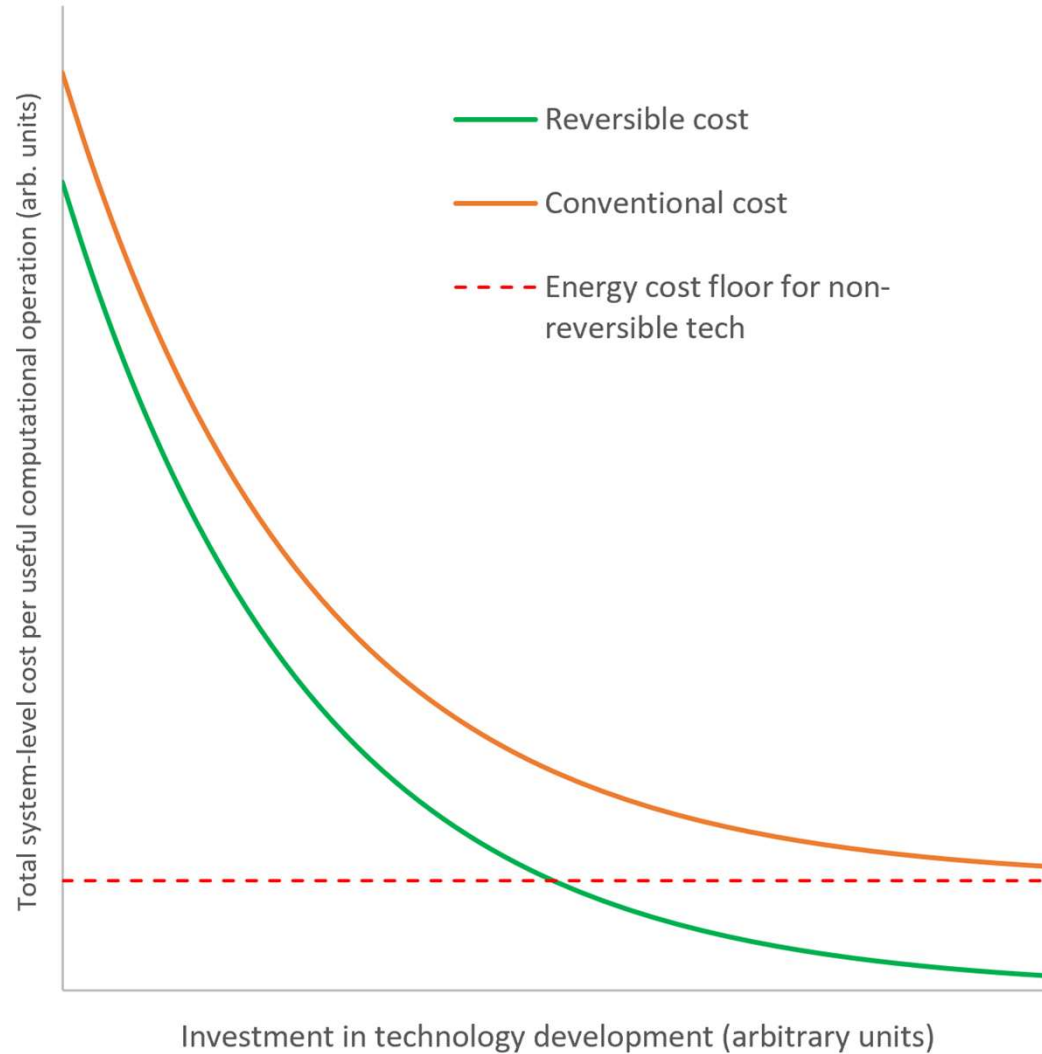
Economic Analysis at a Glance



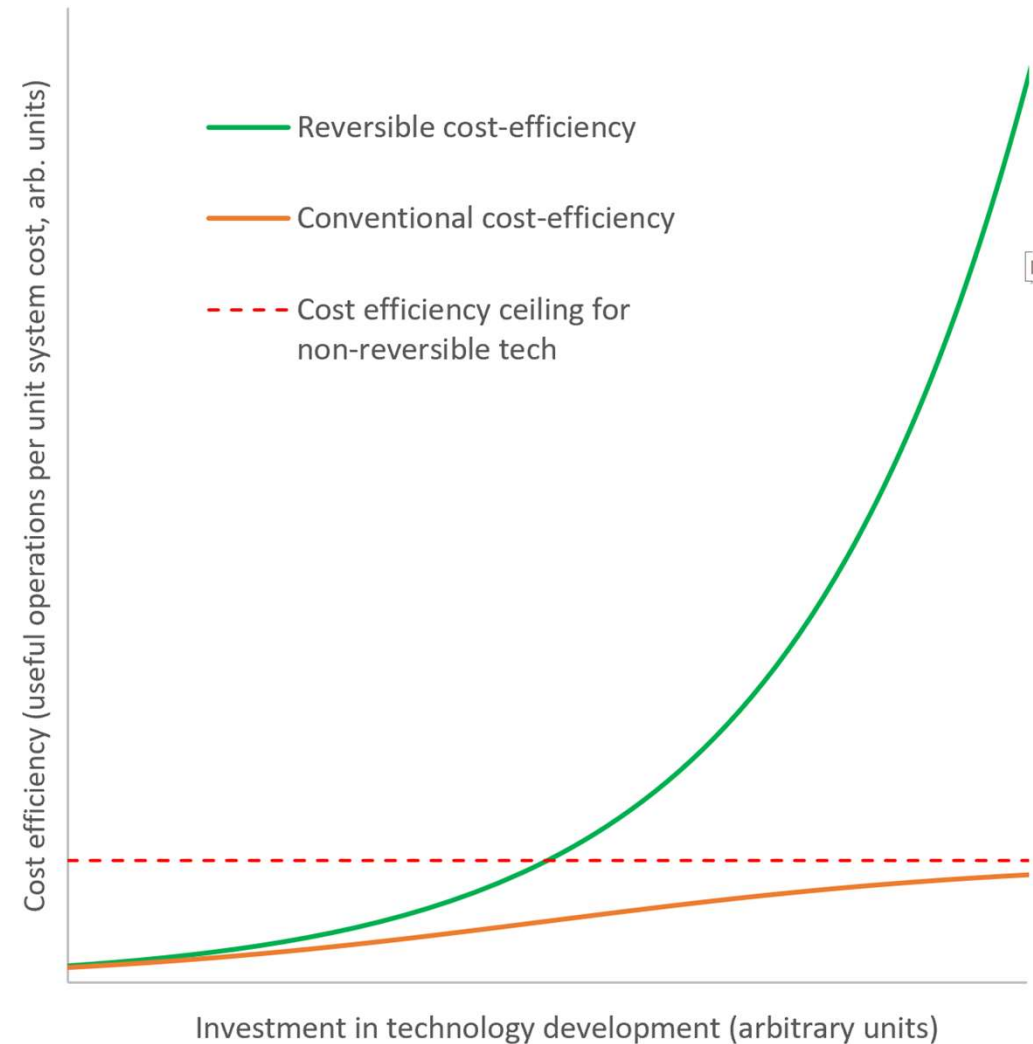
Same charts generated in Excel, using exponential decline in above-floor costs with investment.

- However, *any* rates of approach to 0 above-floor cost still lead to indefinitely-large long-term efficiency advantages for RC.

Amortized Cost Scaling



Cost-Efficiency Scaling



What is dissipation-delay efficiency, and why is it important?

Typically, the *total cost* $\$_{\text{tot}} = \$_E + \$_M$ to perform a computation is minimized when energy-related costs $\$_E$ and manufacturing-related costs $\$_M$ are roughly on the same order.

- Because, there are *diminishing returns* from individually reducing *either one* of these two cost components far below the other one.
 - And, doing so actually makes the total *larger*, if the other cost component gets *increased* as a result.

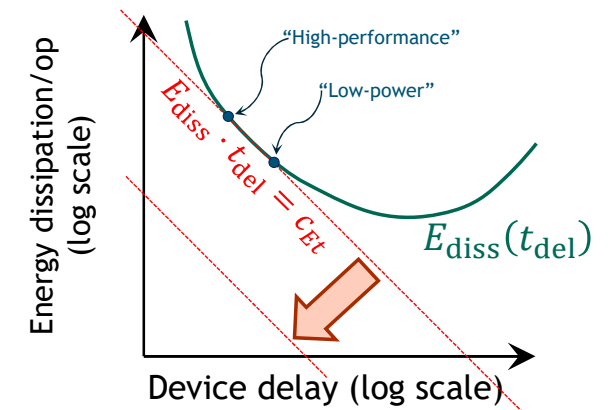
Can express total cost in terms of device parameters: $\$_{\text{tot}} = k_E E_{\text{diss}} + k_M t_{\text{del}}$

For *any* technology that permits tradeoffs between energy efficiency and serial performance, there will be *some* region of the energy-delay curve where the tangent line (on a log-log chart) has slope -1 .

- In this region, the *energy-delay product* is roughly constant.
 - This is even true for voltage scaling in standard irreversible CMOS.
 - But, fully adiabatic techniques can extend this scaling region over a much wider range.
- Different operating points in this linear scaling region will be suitable for applications with different cost *coefficients* k_E, k_M that apply to energy vs. manufacturing cost.
 - *E.g.*, in spacecraft, the effective cost of energy vs. hardware is much greater than in grid-tied applications.

NOTE: If you can move to a new technology whose energy-delay frontier (curve) touches a min. energy-delay product line that is $N \times$ lower than before,

- Then it follows that *total cost* for some applications is reduced by at least $\sqrt{N} \times$!



Dissipation-delay product:

$$C_{Et} = E_{\text{diss}} \cdot t_{\text{del}}$$

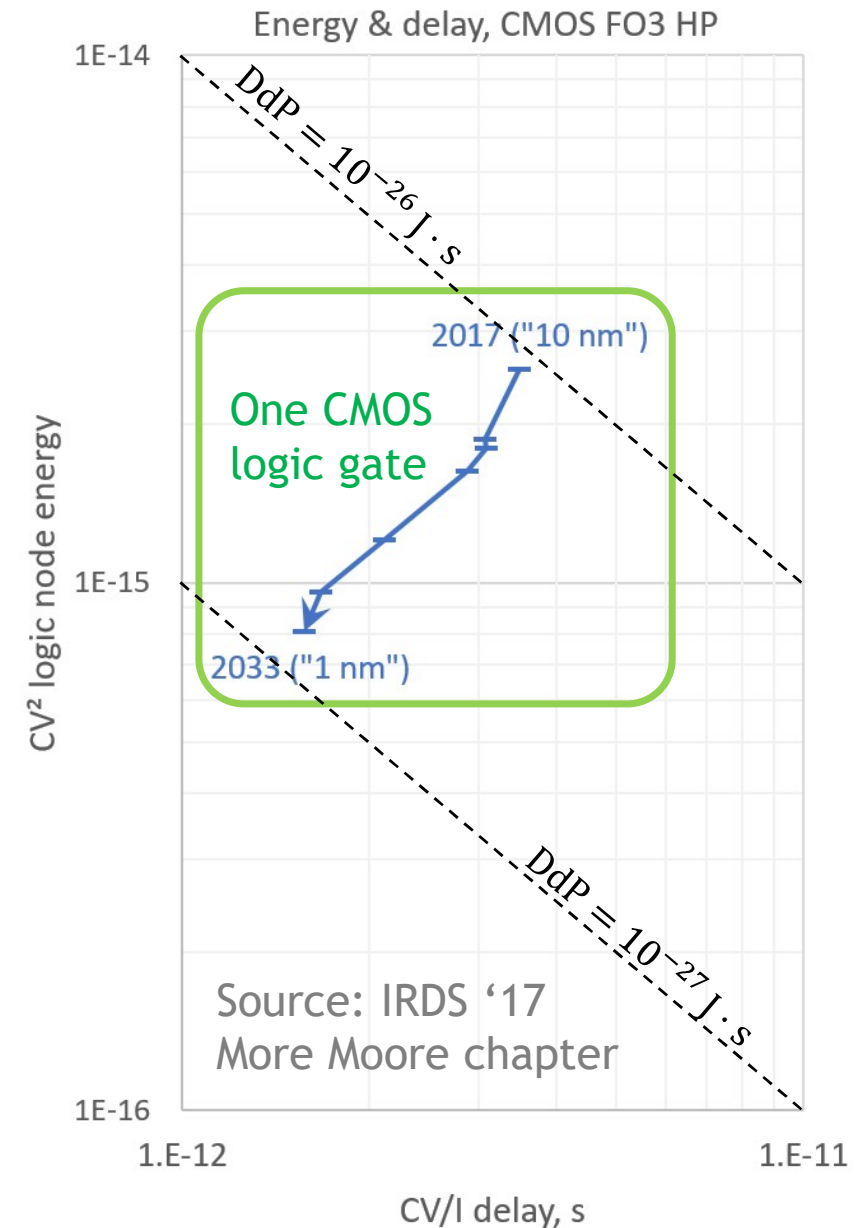
Dissipation-delay efficiency:

$$\eta_{Et} = \frac{1}{C_{Et}}$$

Existing Dissipation-Delay Products (DdP) —Non-reversible Semiconductor Circuits

Conventional (non-reversible) CMOS Technology:

- Recent roadmaps (e.g., IRDS '17) show Dissipation-delay Product (DdP) decreasing by only $< \sim 10\times$ from now to the end of the roadmap (~ 2033).
 - Note the typical dissipation (per logic bit) at end-of-roadmap is projected to be $\sim 0.8 \text{ fJ} = 800 \text{ aJ} = \sim 5,000 \text{ eV}$.
- Optimistically, let's suppose that ways might be found to lower dissipation by an additional $10\times$ beyond even that point.
 - That still puts us at $80 \text{ aJ} = \sim 500 \text{ eV}$ per bit.
- We need at least $\sim 1 \text{ eV} \approx 40 kT$ electrostatic energy at a minimum-sized transistor gate to maintain reasonably low leakage despite thermal noise,
 - And, typical *structural* overhead factors *compounding* this within fast random logic circuits are roughly $500\times$,
 - so, $\sim 500 \text{ eV}$ is *indeed* probably about the practical limit.
 - At least, this is a reasonable order-of-magnitude estimate.



Performance per-area scaling with machine thickness



Frank & Knight 1997, doi:[10.1088/0957-4484/9/3/005](https://doi.org/10.1088/0957-4484/9/3/005)

Assumptions of this simple analysis include:

- Classic adiabatic ($E_{\text{diss,op}} \propto 1/t$) scaling
- Fixed operating temperature
- Constant volume and mass per device
- Bounded entropy flux density F_S
- No algorithmic overheads for reversibility

Later, we will discuss the impact of considering the algorithmic overheads of reversibility.

- Spoiler: Reversible computing still wins!

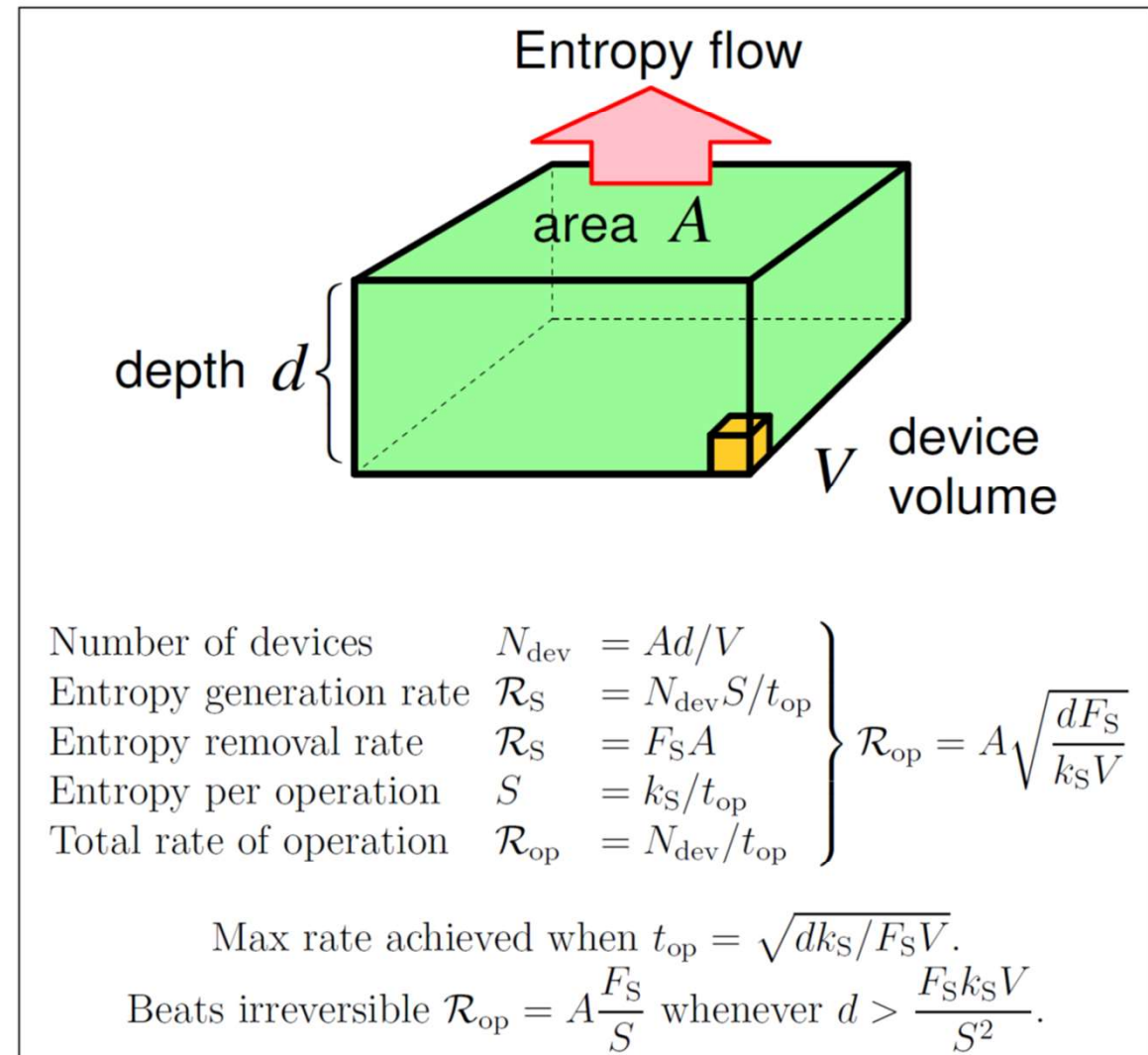


Figure 6.1: Speed limit for reversible machines of minimum-surface area $\Theta(A)$ and thickness $d \lesssim A^{1/2}$. The maximum rate of computation scales as $\Theta(A\sqrt{d})$.

Accounting for Nonidealities

Earlier analyses assumed that leakage can be engineered to be as small as necessary for it not to be limiting (which may be an OK assumption for *some* technologies) and negligible algorithmic overheads (which may be an OK assumption for *some* problems).

- But, can we still show an advantage even when making more pessimistic/realistic assumptions?
 - Answer is yes!

Even for worst-case problems, we can always at least still use the “Frank ‘02” algorithm (Bennett ‘89 variant).

- And, even better general “reversiblization” algorithms may yet be discovered in the future.

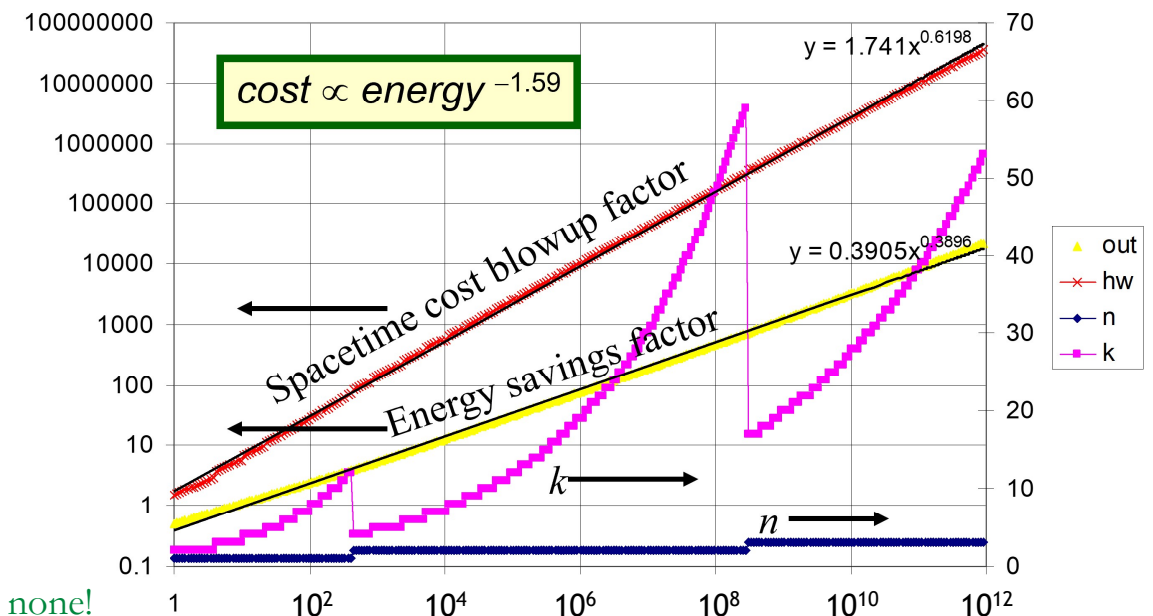
Then, as the technology is improved, and leakage is reduced, we can adjust the parameters of the algorithm to minimize the total cost

- Including both energy and spacetime/mfg. associated costs.

We find that we can reduce total lifetime *system* cost by any factor of N if we just reduce leakage by $\sim N^{2.56}$ and time-amortized per-device manufacturing cost by $\sim N^{1.59}$.

- Example: To achieve an $N = 1,000 \times$ overall efficiency boost, reduce leakage by $47.8M \times$ and mfg. cost/device by $59,000 \times$.
 - Ambitious but doable!! This gives us a way forward, where otherwise there is none!

Worst-Case Energy/Cost Tradeoff (Optimized Bennett-89 Variant)





Section III. Reversible Computing Technologies in Semiconducting Platforms

Current Status of Reversible Computing

Adiabatic Circuits in CMOS: A Brief History



A selection of some early papers:

Fredkin and Toffoli, 1978

(DOI:10.1007/978-1-4471-0129-1_2)

- Unfinished circuit concept based on idealized capacitors and inductors
 - How to control switches to do logic was left unspecified
 - Large design overhead—Roughly one inductor per gate

Seitz *et al.*, 1985

(CaltechCSTR:1985.5177-tr-85)

- Realistic MOSFET switches; more compact integration (off-chip L)
- Not yet known to be general-purpose; required careful tuning

Koller and Athas, 1992

(DOI:10.1109/PHYCMP.1992.615554)

- Not yet fully-reversible technique; limited efficiency
- Combinational only; conjectured reversible *sequential* logic impossible

Hall, 1992; Merkle, 1992

(DOIs:10.1109/PHYCMP.1992.615549;
10.1109/PHYCMP.1992.615546)

- General-purpose reversible methods, but for combinational logic only

Younis & Knight, 1993

(<http://dl.acm.org/citation.cfm?id=163468>)

- First fully-reversible, fully-adiabatic *sequential* circuit technique (CRL)

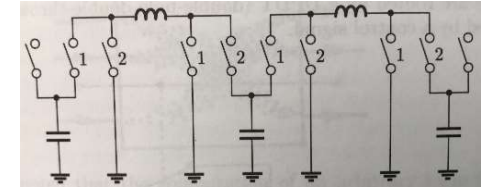


Figure reproduced with permission

Adiabatic Circuits in CMOS: History, cont.



Younis & Knight, 1994

- Simplified 3-level adiabatic CMOS design family (SCRL)
 - However, the original version of SCRL contained a small non-adiabaticity bug which I discovered in 1997
 - This problem is easily fixed, however

Subsequent work at MIT, 1995-99

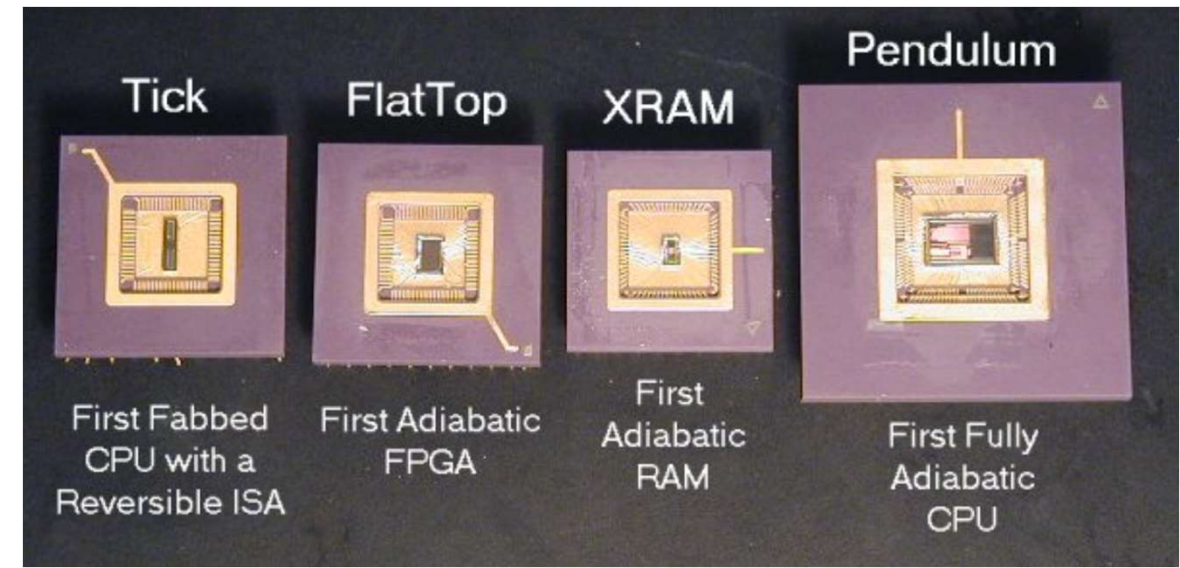
- Myself and fellow students
- Various chips designed using SCRL →
- Reversible processor architectures

Substantial literature throughout the late 90s / early 2000s...

- Too many different papers / groups to list them all here!
 - Most of the proposed schemes were not truly/fully adiabatic, though

Researchers recently active in adiabatic circuits include:

- A couple I know in the US:
 - Greg Snider (Notre Dame)
 - Himanshu Thapliyal (U. Kentucky)
- Also some groups in Europe, India, China, Japan...
- My group at Sandia (new work reported on slide #18)

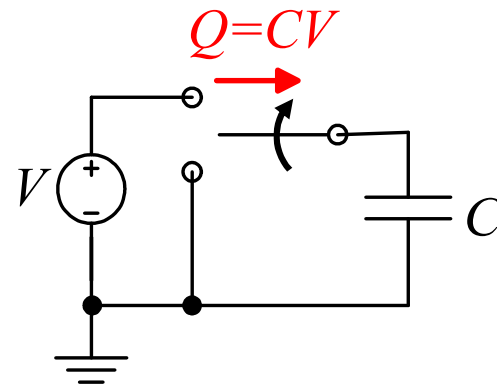


Conventional vs. Adiabatic Charging

For charging a capacitive load C through a voltage swing V

Conventional charging:

- Constant *voltage* source

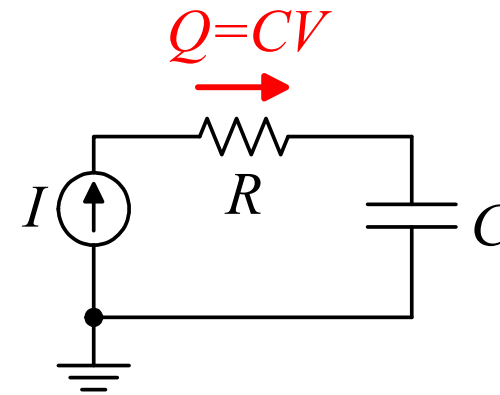


- Energy dissipated:

$$E_{\text{diss}}^{\text{conv}} = \frac{1}{2} CV^2$$

Ideal *adiabatic* charging:

- Constant *current* source



- Energy dissipated:

$$E_{\text{diss}}^{\text{adia}} = I^2 R t = \frac{Q^2 R}{t} = CV^2 \frac{RC}{t}$$

Note: Adiabatic charging beats the energy efficiency of conventional by advantage factor:

$$A = \frac{E_{\text{diss}}^{\text{conv}}}{E_{\text{diss}}^{\text{adia}}} = \frac{1}{2} \frac{t}{RC}$$

Adiabatic Charging via MOSFETs

A simple voltage ramp can *approximate* an ideal constant-current source.

- Note that the load gets charged up *conditionally*, if the MOSFET is turned on (gate voltage $V_g \gtrsim V + V_t$) during ramp.
- V_t is the transistor's threshold, typically $< 1/2$ volt

Can discharge the load later using a similar ramp.

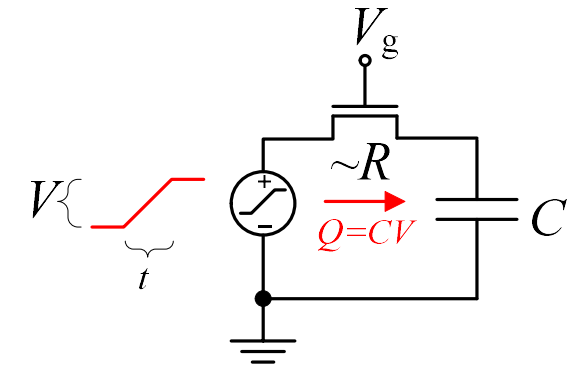
- Either through the same path, or a different path.

$$t \gg RC \Rightarrow E_{\text{diss}} \rightarrow CV^2 \frac{RC}{t}$$

$$t \ll RC \Rightarrow E_{\text{diss}} \rightarrow \frac{1}{2} CV^2$$

The (ideal) operation of this circuit approaches *physical reversibility* ($E_{\text{diss}} \rightarrow 0$) in the limit $t \rightarrow \infty$, but *only* if a certain *precondition* on the initial state is met (namely, $V_g \gtrsim V_{\text{max}} + V_t$)

- How does the possible physical reversibility of this circuit relate to its *computational* function, and to some *appropriate* concept of logical reversibility?
 - Traditional (Landauer/Fredkin/Toffoli) reversible computing theory does **not** adequately address this question, so, we need a more powerful theory!
 - The theory of **Generalized Reversible Computing** (GRC) meets this need.



Exact formula for linear ramps:

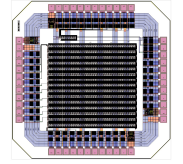
$$E_{\text{diss}} = s[1 + s(e^{-1/s} - 1)]CV^2$$

given *speed fraction* $s = RC/t$.

See [arxiv:1806.10183](https://arxiv.org/abs/1806.10183) for the full GRC model.

Perfectly Adiabatic Reversible Computing in CMOS

2LAL test chip
taped out at
Sandia, Aug. '20



To approach ideal reversible computing in CMOS...

We must aggressively eliminate *all* sources of non-adiabatic dissipation, including:

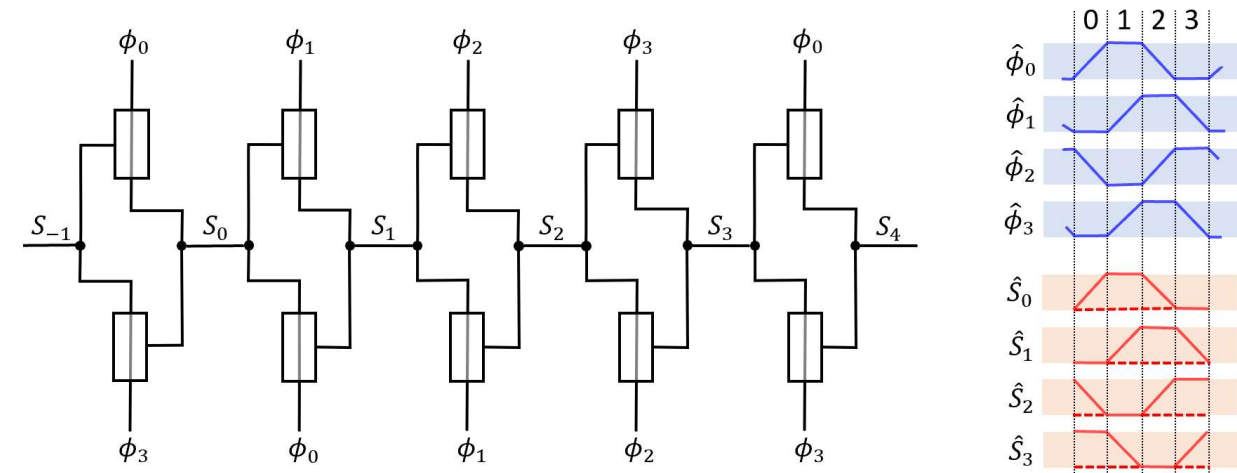
- Diodes in charging path, “sparking,” “squelching,”
 - Eliminated by “**truly, fully adiabatic**” design. (E.g., CRL, 2LAL).
 - Suffices to get to a few aJ (10s of eV) even *before* voltage optimization.
 - Voltage level mismatches that dynamically arise on floating nodes before reconnection.
- Eliminated by static, “**perfectly adiabatic**” design. (E.g., S2LAL).

We must also aggressively minimize standby power dissipation from leakage, including:

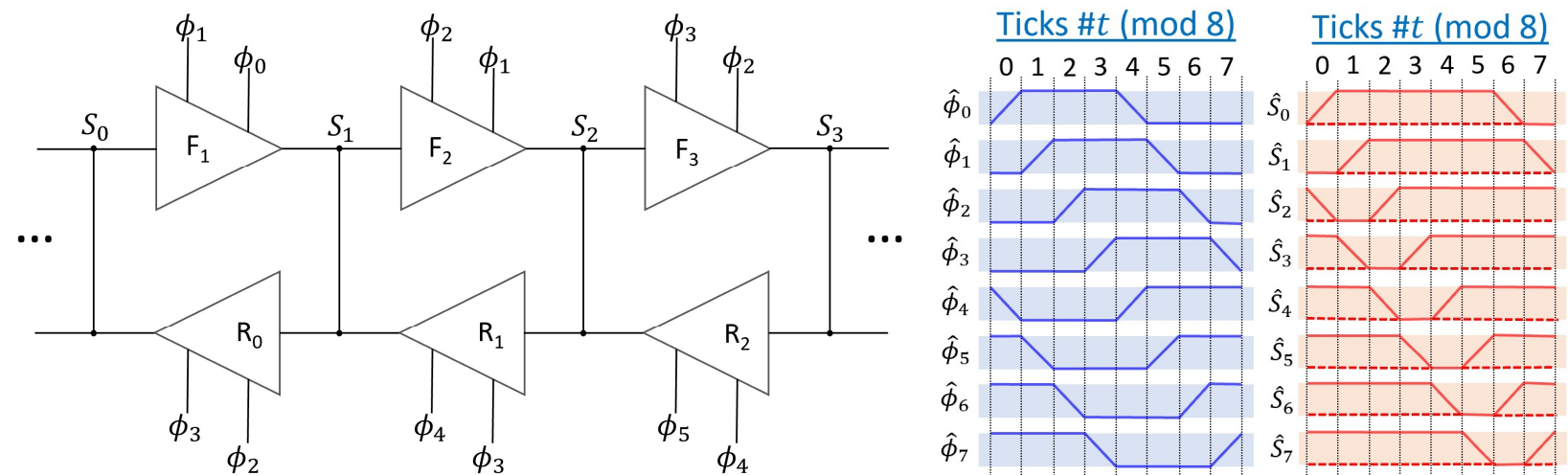
- Subthreshold channel currents
 - Low- T operation helps with this
- Tunneling through gate oxide
 - E.g., use thicker gate oxides

Note: (Conditional) logical reversibility *follows from* perfect adiabaticity.

Shift Register Structure and Timing in 2LAL



Shift Register Structure and Timing in S2LAL



(arxiv:2009.00448)

An SRC-funded study done at the University of Florida (2004)



Simulation results *appeared* to show that 2LAL in TSMC 180nm could get to as low as 1 ev (!) dissipation/FET/ clock cycle.

We now believe (thanks to a current NSCI-funded study at Sandia) that this specific result was most likely unrealistic, because the BSIM3 models we had in '04 (we think) probably significantly underestimated the actual gate leakage resulting from tunneling.

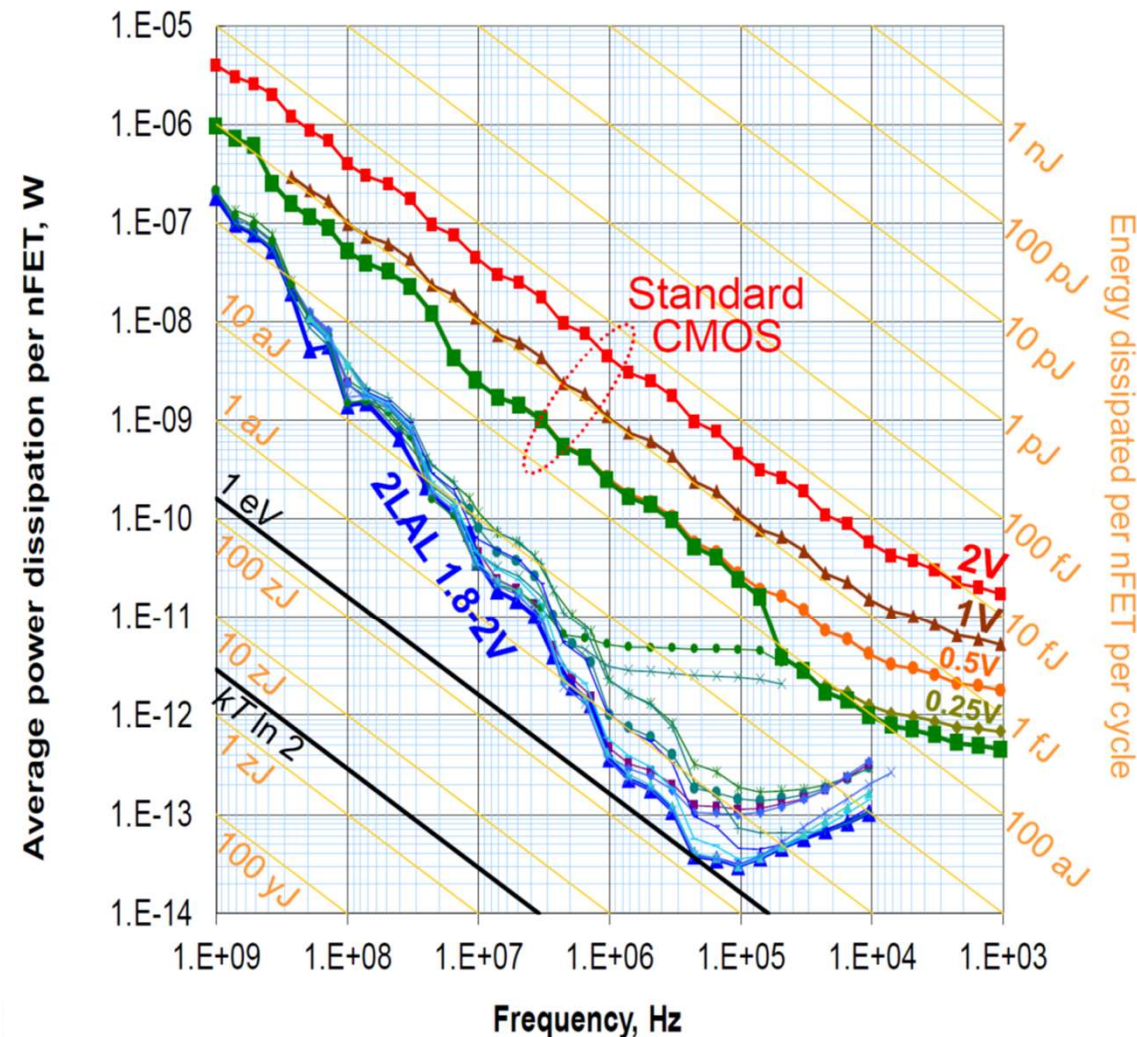
- We think that specific BSIM3 model did not capture gate leakage at all.

However, we do still believe that, in a real process that was well optimized for low leakage, we would be able to achieve similarly impressive results to this.

Simulation Results (Cadence/Spectre) **UF** UNIVERSITY of FLORIDA

Power vs. freq., TSMC 0.18, Std. CMOS vs. 2LAL

2LAL = Two-level adiabatic logic (invented at UF, '00)



- Graph shows per-FET power dissipation vs. frequency
 - in an 8-stage shift register.
- At moderate freqs. (1 MHz),
 - Reversible uses $< 1/100^{\text{th}}$ the power of irreversible!
- At ultra-low power levels (1 pW/transistor)
 - Reversible is $100 \times$ faster than irreversible!
- Minimum energy dissipation per nFET is **< 1 electron volt!**
 - $500 \times$ lower dissipation than best irreversible CMOS!
 - $500 \times$ higher computational energy efficiency!
- Energy transferred per nFET per cycle is still on the order of 1-10 fJ (10-100 keV)
 - So, energy recovery efficiency is at least 99.99%!
 - Quality factor $Q > 10,000!$
 - Note this does not include any of the parasitic losses associated with power supply and clock distribution yet, though

Latest Results from the *Adiabatic Circuits Feasibility Study*

Simulation Efforts at Sandia, funded via NSCI (2017-2021)



Created schematic-level fully-adiabatic designs for Sandia's in-house processes, including:

- Older, 350 nm process (**blue** curve)
 - FET widths = 800 nm
- Newer, 180 nm process (**orange, green** curves)
 - FET widths = 480 nm

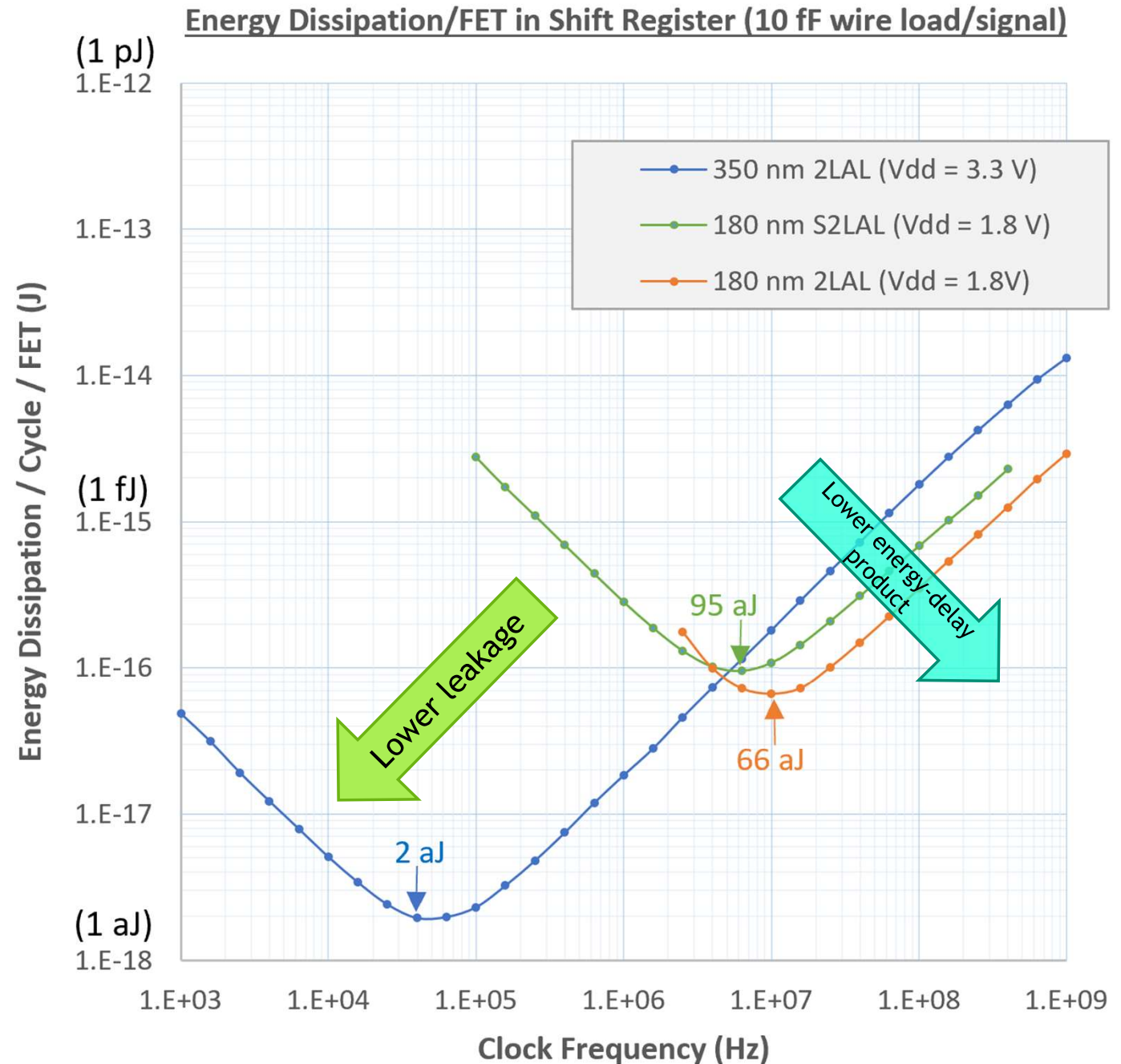
Plotted energy dissipation per-transistor in shift registers at 50% activity factor (alternating 0/1)

- 2LAL (**blue, orange** curves)
- S2LAL (**green** curve)

In all of these Cadence/Spectre simulations,

- We assumed a 10 fF parasitic wiring load capacitance on each interconnect node.
- Logic supply (V_{dd}) voltages were taken at the processes' nominal values.
 - 3.3V for the 350nm process; 1.8V in the 180nm process.

We expect these results could be significantly improved by exploring the parameter space over possible values of V_{dd} and V_{sb} (substrate bias).



See Frank *et al.* “Exploring the Ultimate Limits of Adiabatic CMOS”, 38th IEEE Int’l Conf. on Computer Design (ICCD’20), [10.1109/ICCD50377.2020.00018](https://doi.org/10.1109/ICCD50377.2020.00018)



Resonator design effort, in progress...

Goal of this effort:

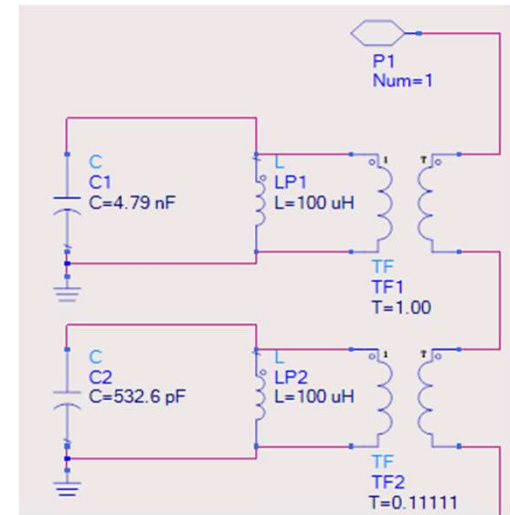
- Design & validate a high-efficiency resonant oscillator (for low-to-medium RF frequencies) that approximates a trapezoidal output voltage waveform.

Innovative design concept:

- Transformer-coupled** assemblage of LC tank circuits with resonant frequencies corresponding to odd multiples of the fundamental frequency, excited in the right relative amplitudes to approximate the target wave shape

Some detailed requirement specifications:

- Initial target operating point: 230 kHz, 1.8V (optimal point for minimum dissipation in the UF study) **(MET.)**
 - However, our circuit technique should be adaptable over a wide range of frequencies and voltages.
- Tops and bottoms of trapezoidal wave should be within $\leq 5\%$ of flatness throughout $\frac{1}{4}$ clock period. **(MET.)**
- The 10-90% rise/fall time should be between 75 & 100% of its nominal value (80% of $\frac{1}{4}$ clock period) **(MET.)**
- Efficiency goals:
 - Quality factor of resonator during unpowered ring-down should be $\geq 1,000$. **(MET. Simulated value: $\sim 3,000$.)**
 - Total energy dissipation per cycle during steady-state powered operation should be $\leq 1\%$ of magnetically-stored energy in the resonator, when the oscillator is running in isolation. (Still needs validation.)
 - Total energy dissipation per cycle during steady-state powered operation should be $\leq 10\%$ of the capacitively-stored energy on an appropriately-sized model (RC) load, when the oscillator is coupled to the load. (Needs validation.)

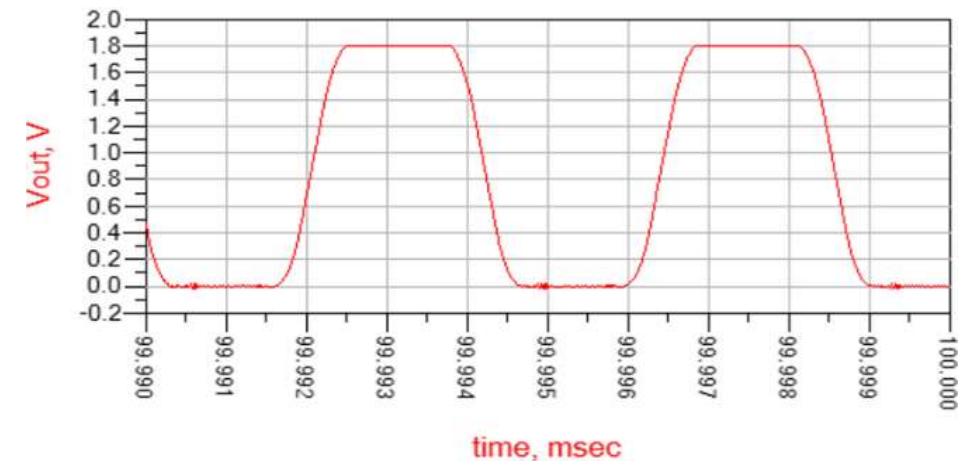


A number of significant design challenges that have been encountered so far:

- How to tune the relative amplitudes of the component resonant modes **(Solved.)**
- How to prevent phase drift and transfer of energy between modes **(Solved.)**
- Identifying/tailoring components to have precise-enough L , C values
- Designing a driver circuit that meets efficiency goals during steady-state operation
- Packaging & integration for a complete system including a resonator & a 2LAL die.

A provisional patent application has been filed on our resonator design.

- We invite industry firms to partner with us under NDA/CRADA.





Section III. Reversible Computing Technologies in Superconducting Platforms

Current Status of Reversible Computing

Adiabatic Reversible Computing in Superconducting Circuits



Work along this general line has roots that go all the way back to Likharev, 1977.

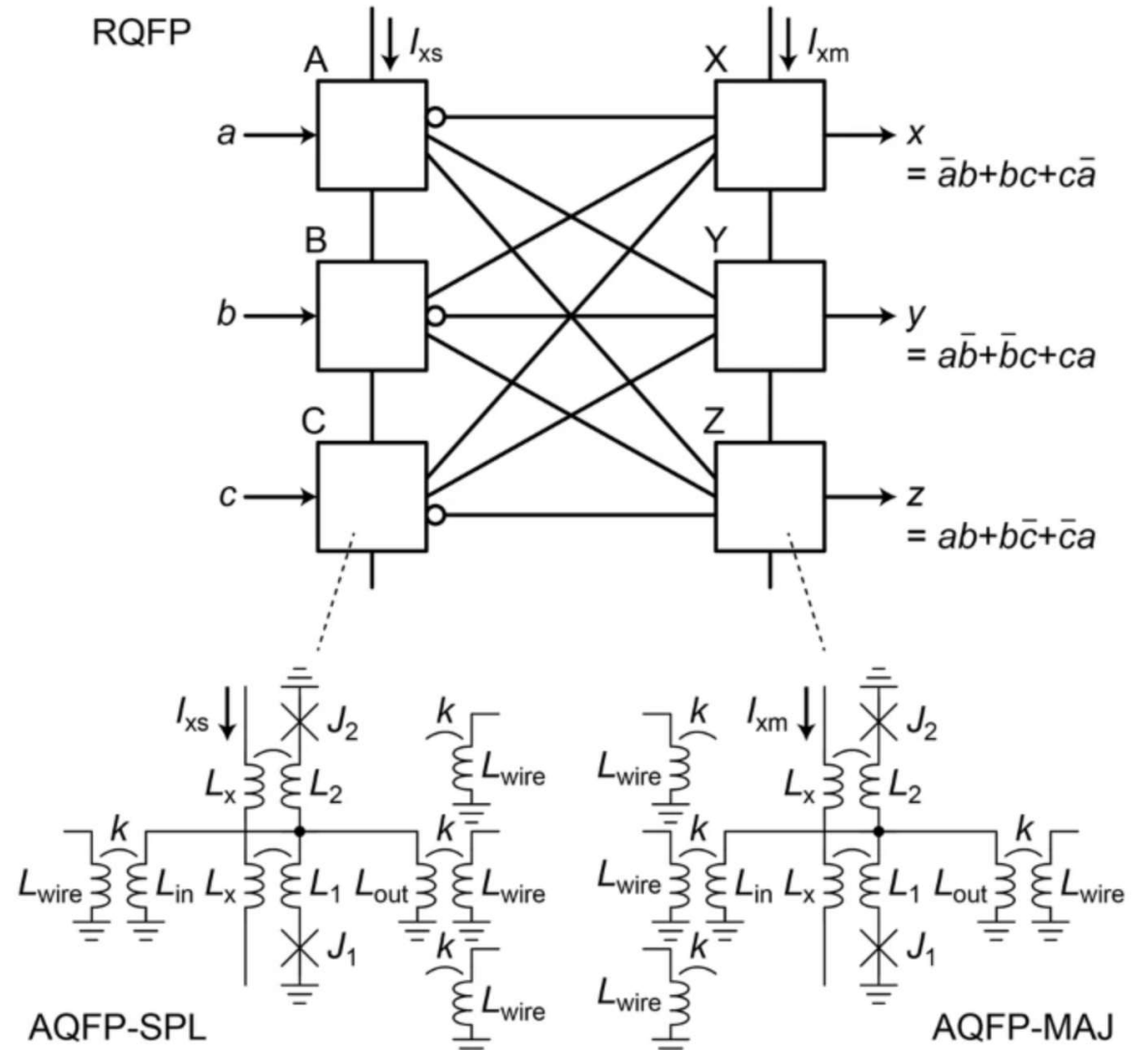
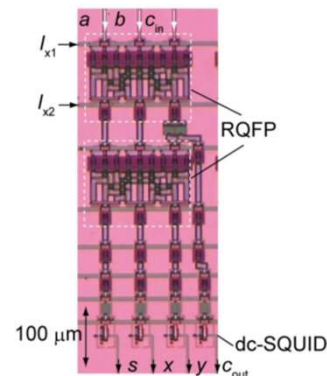
Most active group at present is Prof. Yoshikawa's group at Yokohama National University in Japan.

Logic style called *Reversible Quantum Flux Parametron* (RQFP).

Shown at right is a 3-output *reversible majority gate*.

Full adder circuits have also been built and tested.

Simulations indicate that RQFP circuits can dissipate $< kT \ln 2$ even at $T = 4\text{K}$, at speeds on the order of 10 MHz



Existing Dissipation-Delay Products (DdP)— Adiabatic Reversible Superconducting Circuits

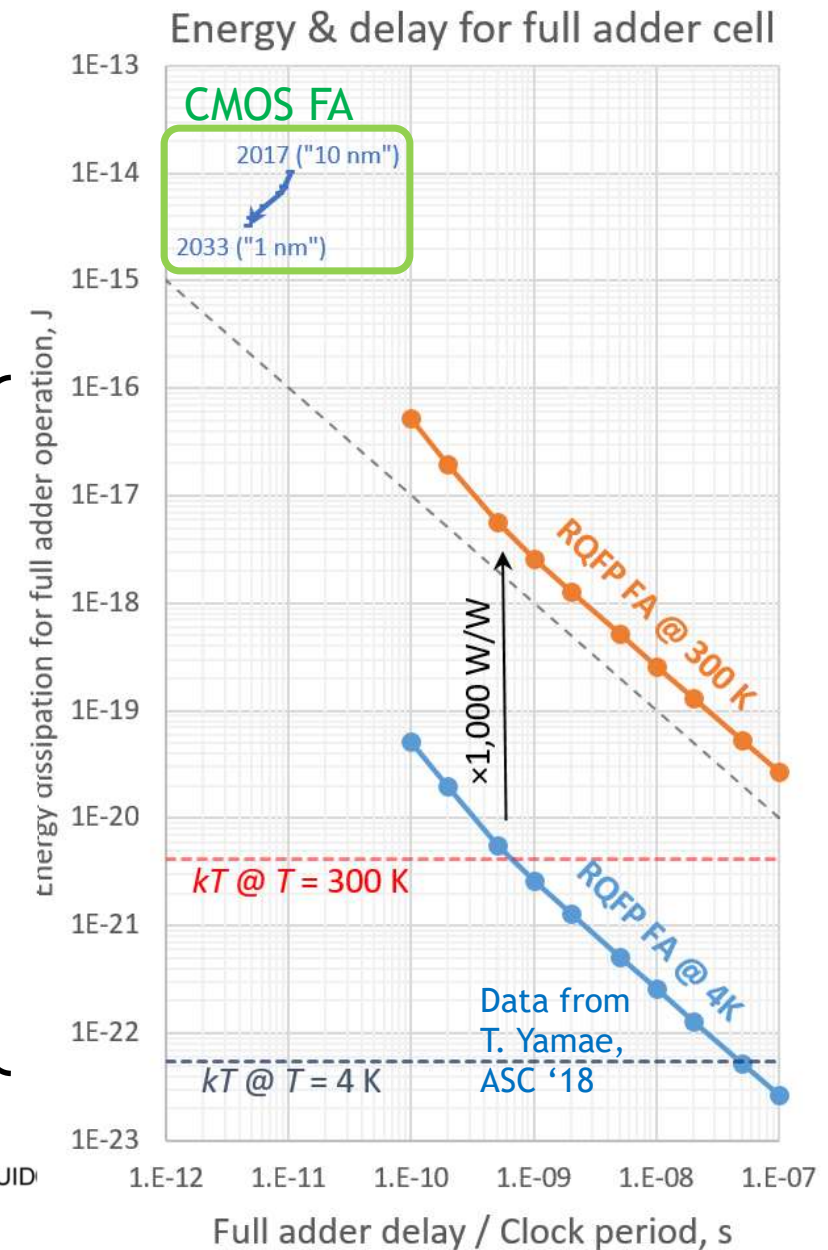
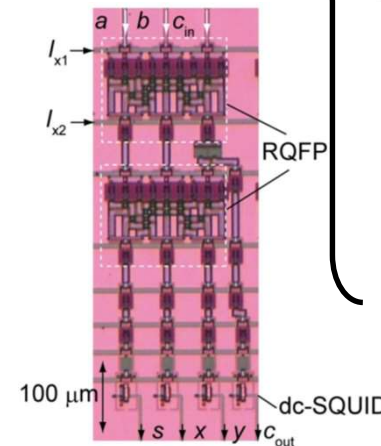
Reversible adiabatic superconductor logic:

- State-of-the-art is the **RQFP** (Reversible Quantum Flux Parametron) technology from Yokohama National University in Japan.
- Chips were fabricated, function validated.
- Circuit simulations predict DdP is $>1,000\times$ lower than even *end-of-roadmap* CMOS.
- Dissipation extends *far below* the 300K Landauer limit (and even below the Landauer limit at 4K).
- DdP is *still* better than CMOS even after adjusting by a conservative factor for large-scale cooling overhead (1,000 \times).

Question: Could some *other* reversible technology do even better than this?

- We have a project at Sandia exploring one possible superconductor-based approach for this (more later)...
- But, what are the *fundamental* (technology-independent) limits, if any?

RQFP =
Reversible
Quantum Flux
Parametron
(Yokohama U.)



Ballistic Reversible Computing

Can we envision reversible computing as a *deterministic* elastic interaction process?

Historical origin of this concept:

- Fredkin & Toffoli's *Billiard Ball Model* of computation ("Conservative Logic," IJTP 1982).
 - Based on elastic collisions between moving objects.
 - Spawned a subfield of "collision-based computing."
 - Using localized pulses/solitons in various media.

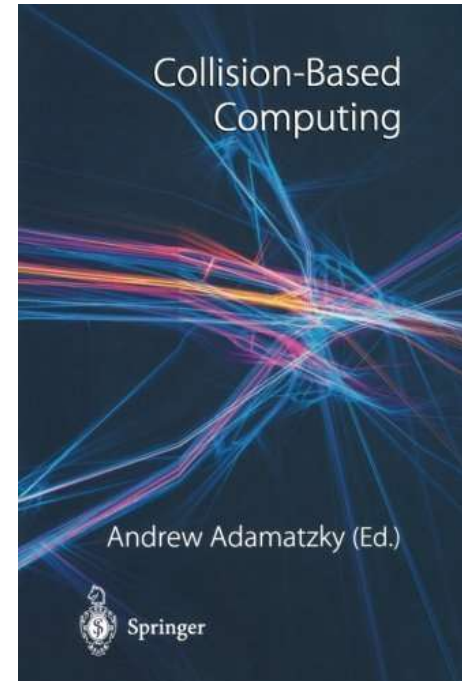
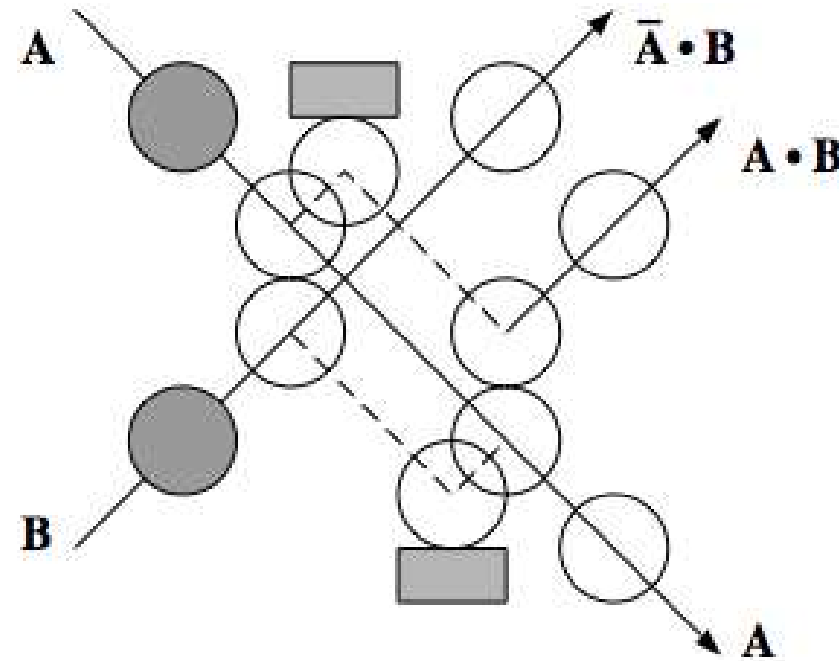
No power-clock driving signals needed!

- Devices operate when data signals arrive.
- The operation energy is carried by the signal itself.
 - Most of the signal energy is preserved in outgoing signals.

However, all (or almost all) of the existing design concepts for ballistic computing invoke implicitly *synchronized* arrivals of ballistically-propagating signals...

- Making that approach work in reality presents some serious difficulties, however:
 - Unrealistic in practice to assume precise alignment of signal arrival times.
 - Thermal fluctuations & quantum uncertainty, at minimum, are always present.
 - Any relative timing uncertainty leads to chaotic dynamics when signals interact.
 - Exponentially-increasing uncertainties in the dynamical trajectory.
 - Deliberate *resynchronization* of signals whose timing relationship has become uncertain incurs an inevitable energy cost.

Can we come up with a *new* ballistic model of reversible computing that avoids these problems?



Ballistic Asynchronous Reversible Computing (BARC)



Problem: Conservative (dissipationless) dynamical systems generally tend to exhibit chaotic behavior...

- This results from direct nonlinear *interactions* between multiple continuous dynamical degrees of freedom (DOFs), which amplify uncertainties, exponentially compounding them over time...
- *E.g.*, positions/velocities of ballistically-propagating “balls”
 - Or more generally, any localized, cohesive, momentum-bearing entity: Particles, pulses, quasiparticles, solitons...

Core insight: In principle, we can greatly reduce or eliminate this tendency towards dynamical chaos...

- We can do this simply by *avoiding* any direct interaction between continuous DOFs of different ballistically-propagating entities

Require localized pulses to arrive *asynchronously*—and furthermore, at clearly distinct, *non-overlapping* times

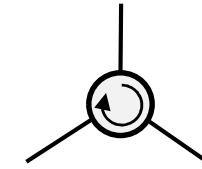
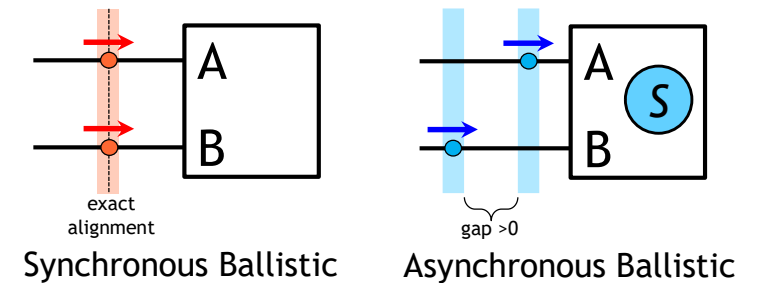
- Device’s dynamical trajectory then becomes *independent* of the precise (absolute *and* relative) pulse arrival times
 - As a result, timing uncertainty per logic stage can now accumulate only *linearly*, not exponentially!
 - Only relatively occasional re-synchronization will be needed
- For devices to still be capable of doing logic, they must now maintain an internal discrete (digitally-precise) state variable—a stable (or at least metastable) stationary state, *e.g.*, a ground state of a well

No power-clock signals, unlike in adiabatic designs!

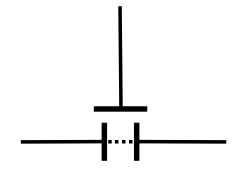
- Devices simply operate whenever data pulses arrive
- The operation energy is carried by the pulse itself
 - Most of the energy is preserved in outgoing pulses
 - Signal restoration can be carried out incrementally, or periodically

Goal of current effort at Sandia: Demonstrate BARC principles in an implementation based on fluxon dynamics in Superconducting Electronics (SCE)

(BARCS  effort)

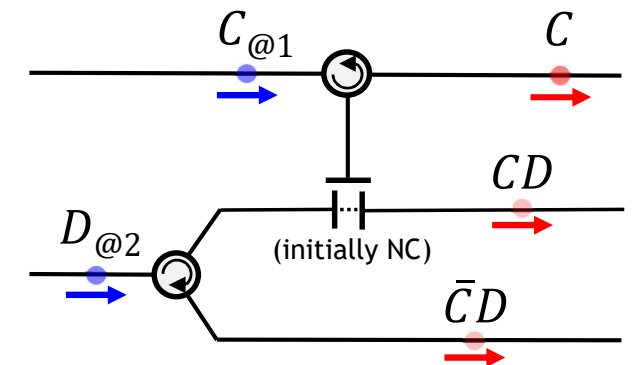


Rotary
(Circulator)



Toggled
Barrier

Example BARC device functions



Example logic construction

Simplest Fluxon-Based (bipolarized) BARC Function

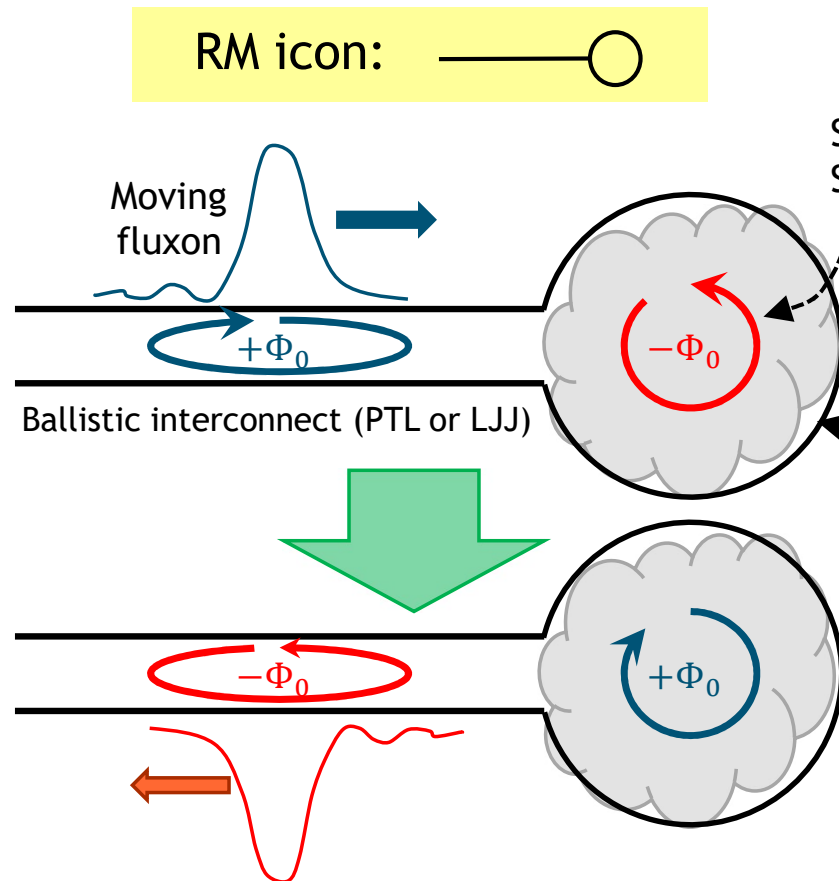


One of our early tasks: Characterize the simplest nontrivial BARC device functionalities, given a few simple design constraints applying to an SCE-based implementation, such as:

- (1) Bits encoded in fluxon polarity; (2) Bounded planar circuit conserving flux; (3) Physical symmetry.

Determined through theoretical hand-analysis that the simplest such function is the ***1-Bit, 1-Port Reversible Memory Cell (RM)***:

- Due to its simplicity, this was then the preferred target for our subsequent detailed circuit design efforts...



Some planar, unbiased, reactive SCE circuit w. a continuous superconducting boundary

- Only contains L's, M's, C's, and *unshunted* JJs
- Junctions should mostly be *subcritical* (avoids R_N)
- Conserves total flux, approximately nondissipative

Desired circuit behavior (NOTE: conserves flux, respects T symmetry & logical reversibility):

- If polarities are opposite, they are swapped (shown)
- If polarities are identical, input fluxon reflects back out with no change in polarity (not shown)
- (*Deterministic*) *elastic 'scattering'* type interaction: Input fluxon kinetic energy is (nearly) preserved in output fluxon

RM Transition Table

Input Syndrome	Output Syndrome
+1(+1)	→ (+1)+1
+1(-1)	→ (+1)-1
-1(+1)	→ (-1)+1
-1(-1)	→ (-1)-1

RM—First working (in simulation) implementation!



Erik DeBenedictis: “Try just strapping a JJ across that loop.”

- This actually works!

“Entrance” JJ sized to = about 5 LJJ unit cells ($\sim 1/2$ pulse width)

- I first tried it twice as large, & the fluxons annihilated instead...
 - “If a $15 \mu\text{A}$ JJ rotates by 2π , maybe $1/2$ that will rotate by 4π ” 🤔

Loop inductor sized so ± 1 SFQ will fit in the loop (but not ± 2)

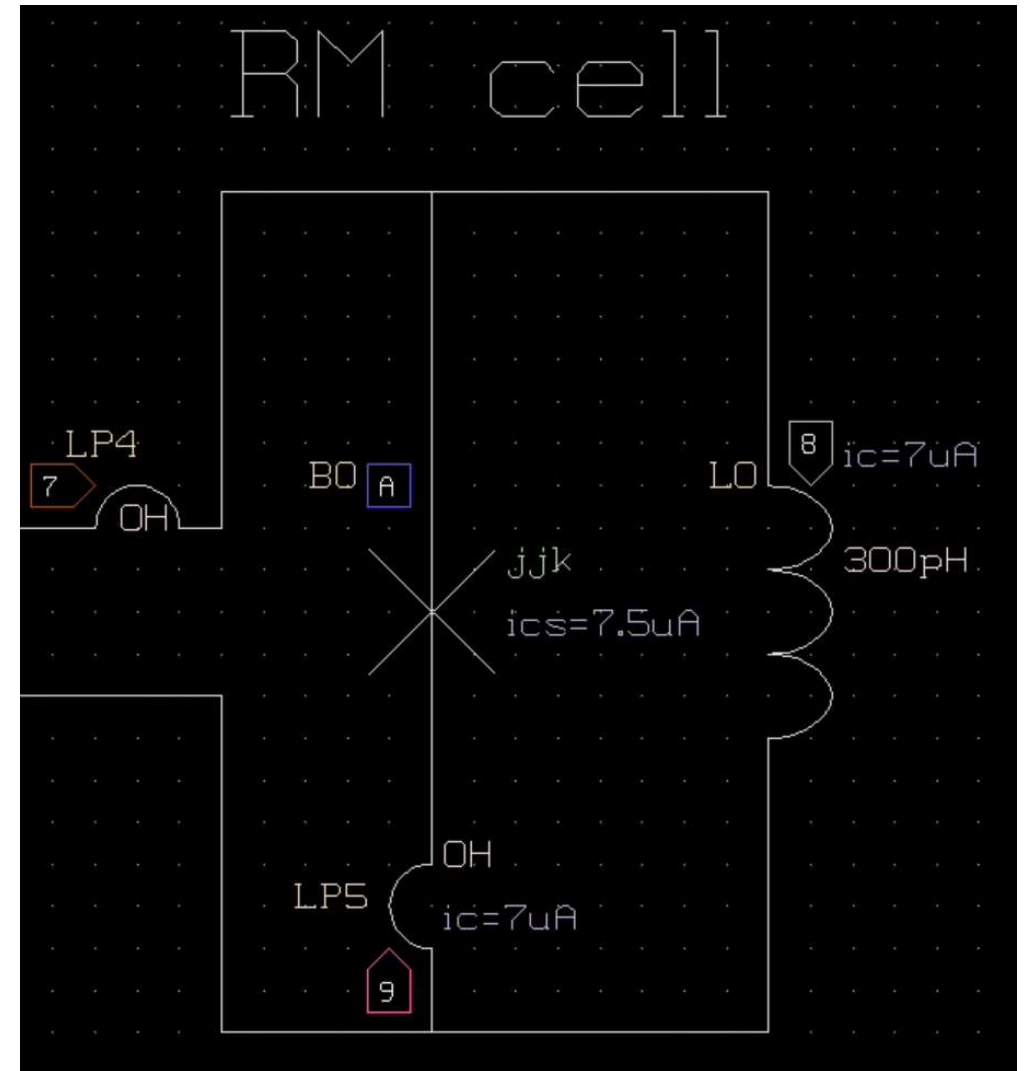
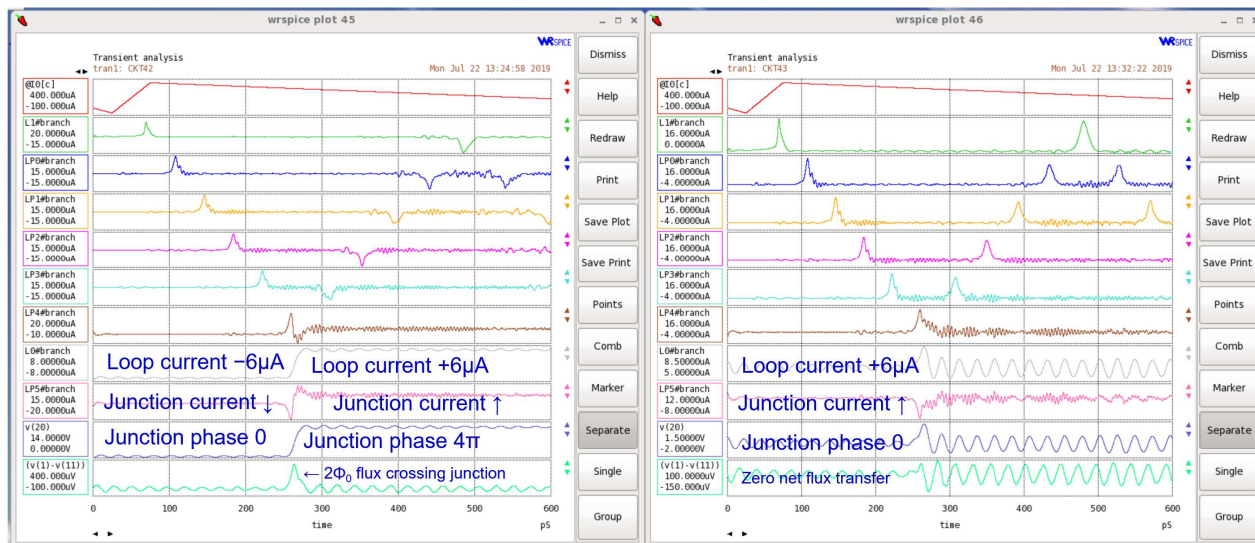
- JJ is sitting a bit below critical with ± 1

WRspice simulations with ± 1 fluxon initially in the loop

- Uses `ic` parameter, & `uic` option to `.tran` command
 - Produces initial ringing due to overly-constricted initial flux
 - Can damp w. small shunt G

Polarity mismatch \rightarrow Exchange

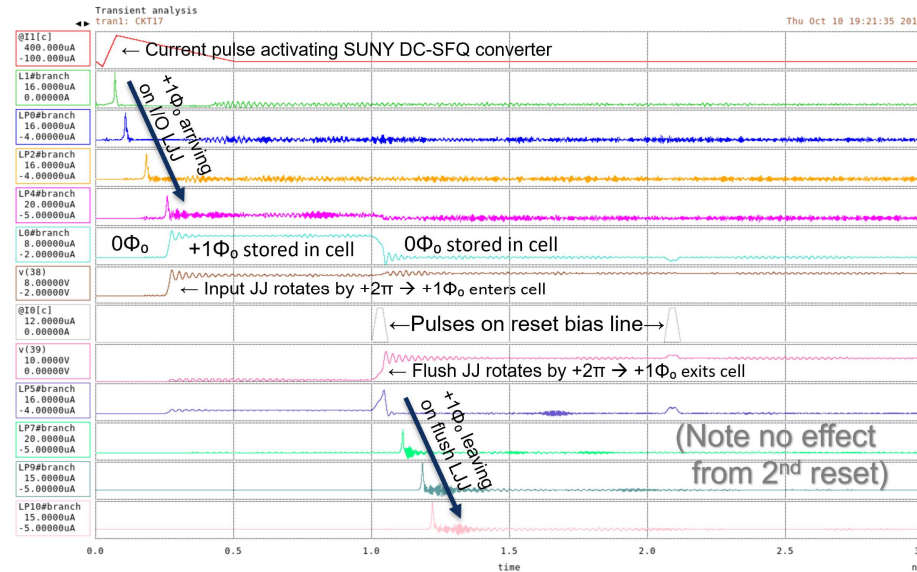
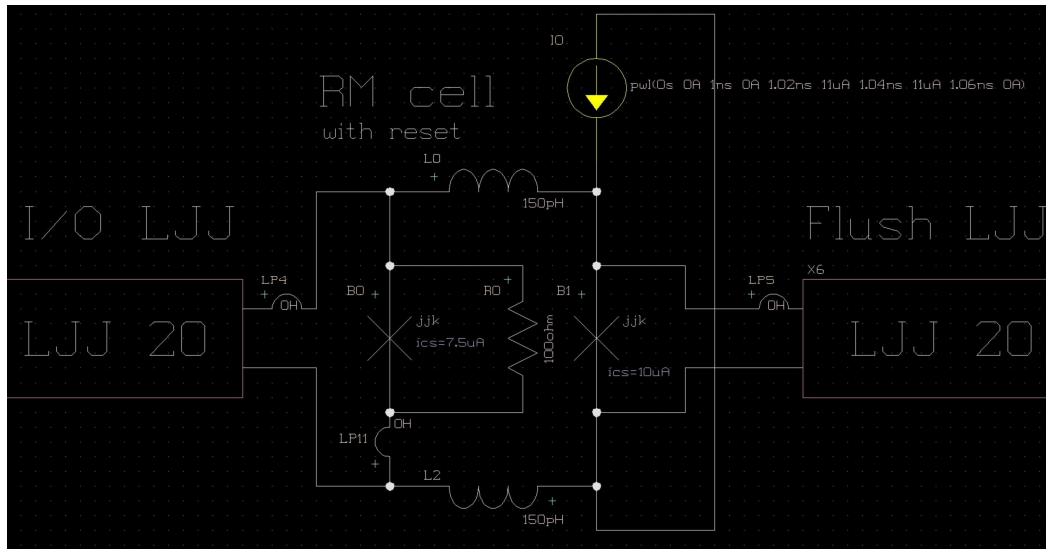
Polarity match \rightarrow Reflect (=Exchange)



Resettable version of RM cell—Designed & Fabricated!

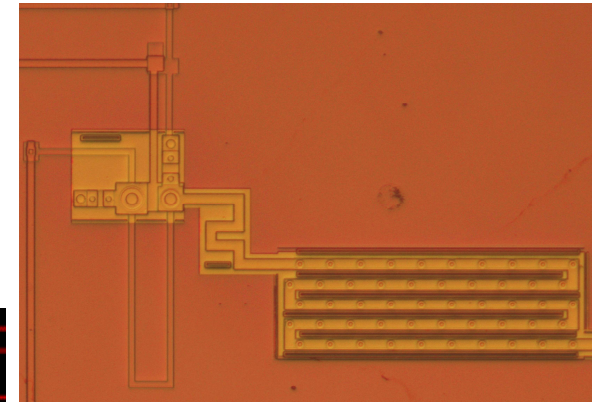
Apply current pulse of appropriate sign to flush the stored flux (the pulse here flushes out positive flux)

- To flush either polarity → Do both (\pm) resets in succession

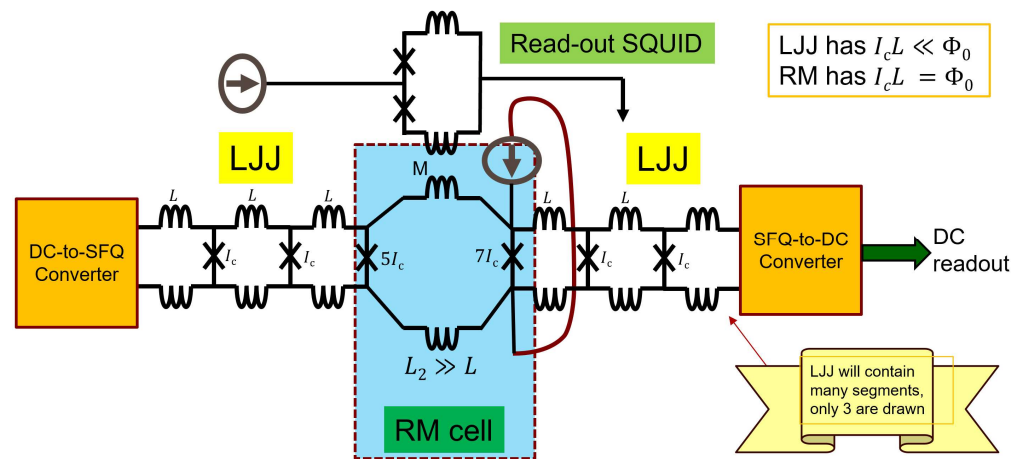
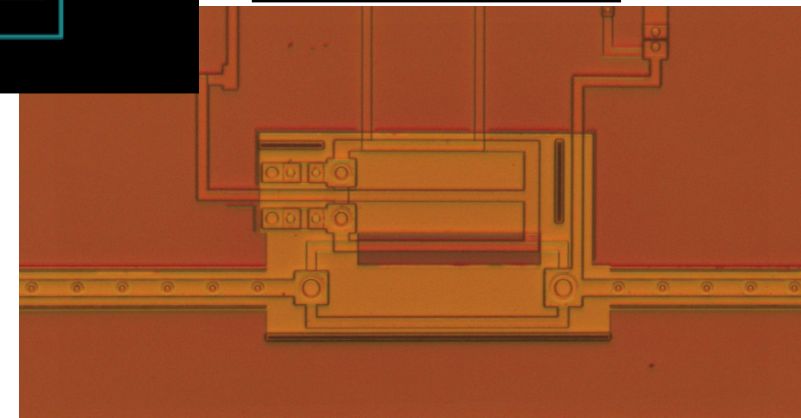
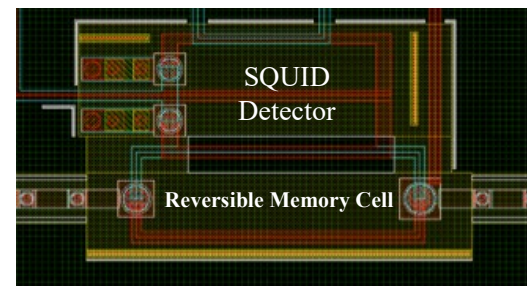
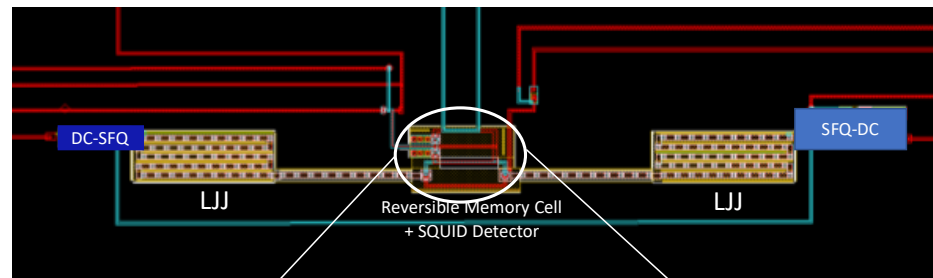


Fabrication at SeeQC with support from ACI

DC-SFQ & LJJ



RM Cell & SQUID





Section IV. Conclusion

Current Status of Reversible Computing

Conclusion

The theoretical underpinnings behind Landauer's Principle and reversible computing rest solidly on the most unshakable, bedrock, foundational principles of physics (as well as its very most cutting-edge, modern formulations).

- ∴ No other method for general digital computing **except** for (various forms of) reversible computing can possibly circumvent the thermodynamic limits of the conventional (non-reversible) paradigm, within the laws of physics.

Simple analyses from economics/systems engineering/asymptotic complexity imply that reversible computing can also yield ongoing improvements in **system-level** cost-efficiency, *despite* its various overheads.

- No limit** to the long-term cost-efficiency advantages that can be provided through the use of RC techniques along this new scaling trajectory is yet known.

Clear, compelling, energy-efficient engineering implementations of the principles of reversible computing have already been developed and demonstrated for both semiconducting and superconducting technology platforms.

- Nothing** prevents the field from *much more quickly* continuing to make progress towards commercial applications—except for funding!

I would strongly encourage SRC (and industry more broadly) to soon begin investing heavily in this area, for the benefit of humanity.

