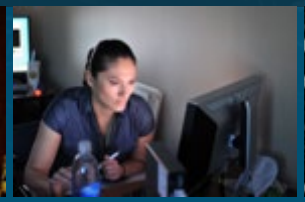
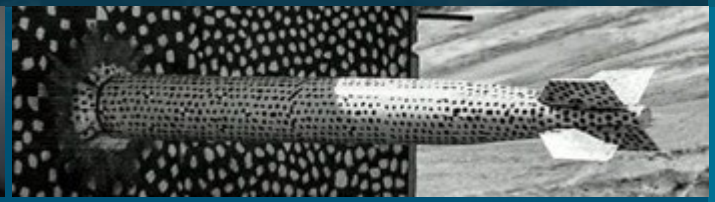
2022 Energy Consequences of Information Workshop
Santa Fe, New Mexico, USA

Sandia National Laboratories

The Reversible Computing Scaling Path: Challenges and Opportunities





Wednesday, February 22nd, 2022

Michael P. Frank, Center for Computing Research

with collaborators: Hannah Earley (Cambridge), Karpur Shukla (Brown)

Approved for public release, SAND2022-1769 C

U.S. DEPARTMENT OF ENERGY NNSA National Nuclear Security Administration

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) – now retired
    - Matt Wolak (now at Northrop-Grumman)
  - Brian Tierney (Rad Hard CMOS Technology)

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!

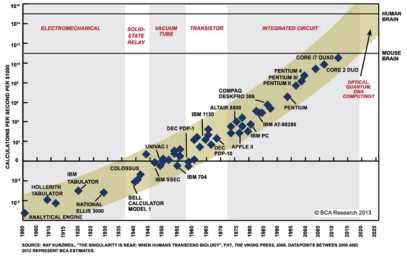
- Thanks are also due to the following colleagues & external collaborators:
  - Karpur Shukla (CMU/Brown U.)
    - w. Prof. Jingming "Jimmy" Xu
  - Hannah Earley (Cambridge U.)
  - Erik DeBenedictis
  - Joseph Friedman (UT Dallas)
    - X. Hu, B.W. Walker, F. Garcia-Sanchezy, P. Zhou, J.A.C. Incorviaz, A. Paler
  - Kevin Osborn (LPS/JQI)
    - Liuqi Yu
  - Steve Kaplan
  - Rudro Biswas (Purdue)
    - Dewan Woods & Rishabh Khare
  - 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

Key contributors to today's new material!

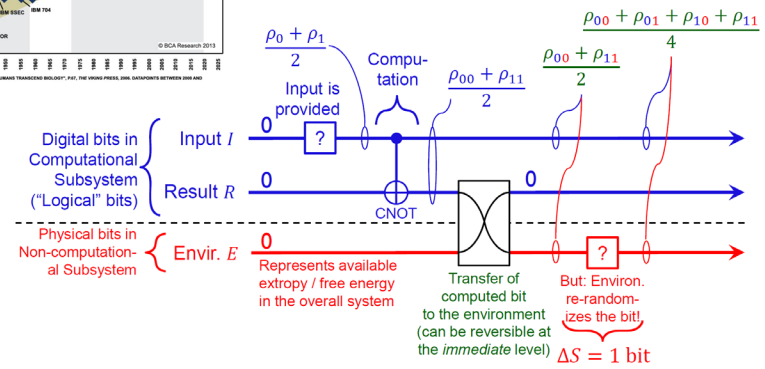
Talk Abstract/Outline

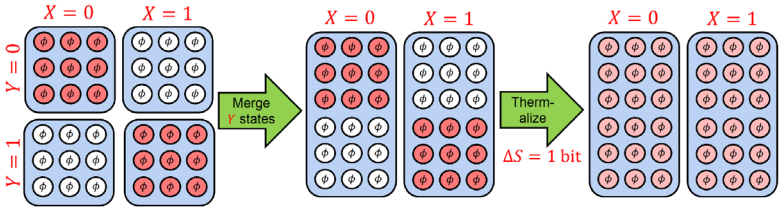
The Reversible Computing (RC) Scaling Path: Challenges and Opportunities

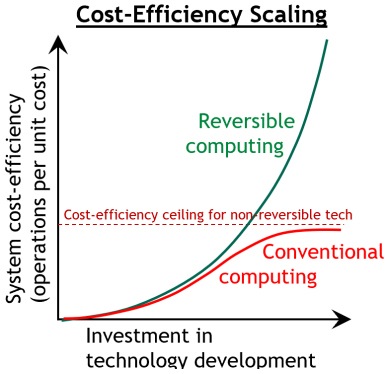
1. Avoiding limits on the energy efficiency of digital computing requires *reversible computing*
   - Exemplar design methodologies exist for both semiconducting & superconducting platforms
2. Critical to forecasting the opportunities in this area is to do a *careful* system scaling analysis
   - It's important to account for realistic overheads at various levels (device speed, device size, algorithmic…)
3. Various key scientific & engineering challenges still need to be addressed at this time:
   - Physics challenges:
     - Deriving fundamental limits of energy dissipation as a function of key physical parameters (Shukla @ Brown)
     - Clarifying the *asymptotic* scaling behavior of synchronous and asynchronous reversible machines (Earley @ Cambridge)
     - Searching for novel (*e.g.* quantum) mechanisms to suppress dissipation at finite scales
   - Engineering challenges:
     - Methods to systematically increase the effective quality factor of energy-recovering driving mechanisms for synchronous reversible machines
     - Further explore the potential engineering realizations of the asymptotic ballistic reversible paradigm
     - Extension of design automation tools and methodologies for reversible design
   - Workforce development challenges:
     - Educational tools & materials needed to train/retrain the engineering workforce to work in this unfamiliar technological paradigm
4. There are some potentially important synergies between RC and AI / machine learning:
   - Use of AI in design optimization/discovery, and workforce development, to speed up the introduction of RC.
   - Folding back reversible computing tech into improving the cost-efficiency of AI/ML training & inference.
5. Next steps & conclusion.

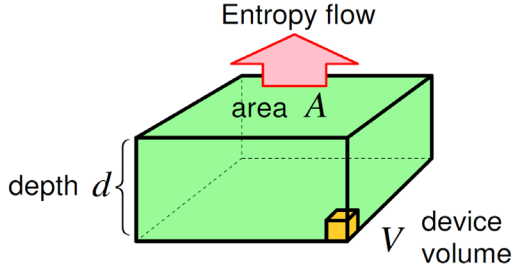


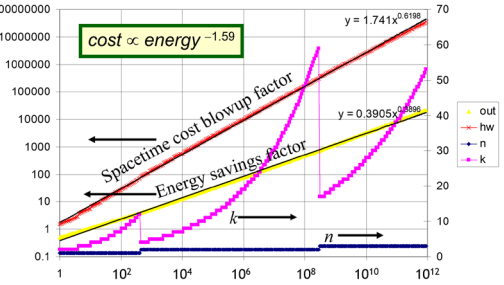
Sandia National Laboratories

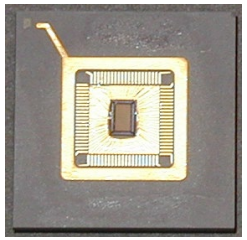




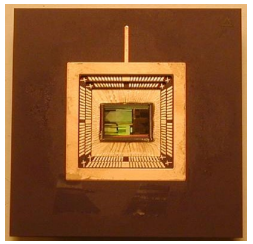




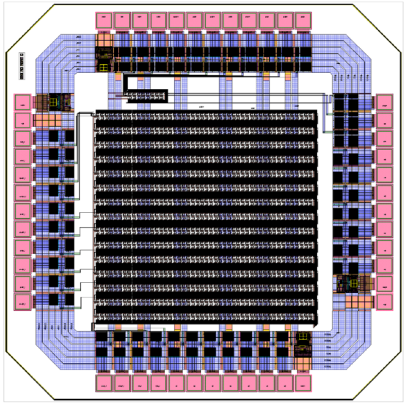




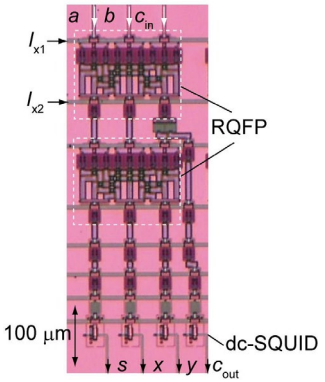
FlatTop (MIT '96)



Pendulum (MIT '99)



2LAL Shift Register (Sandia '20)



RQFP Full Adder (Yokohama '18)

1. Efficiency Limits of Conventional Digital Computing, and the Need for Reversible Computing

The Reversible Computing Scaling Path:
Challenges and Opportunities

Trend of Improving Cost-Efficiency of Computing

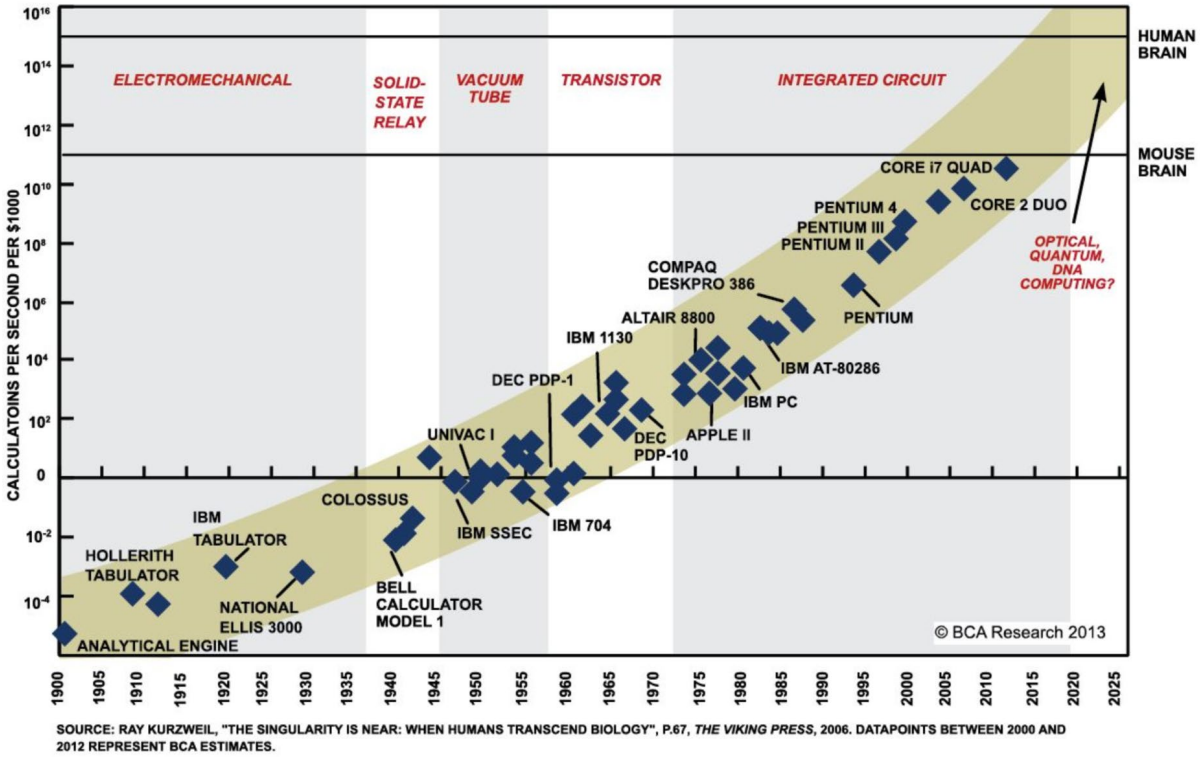
$$\eta_c = \frac{\#ops}{Cost} = \frac{1}{C_{op}}$$

Since at least 1950 (and really even longer), the *cost-efficiency* η_c of computing has improved exponentially…

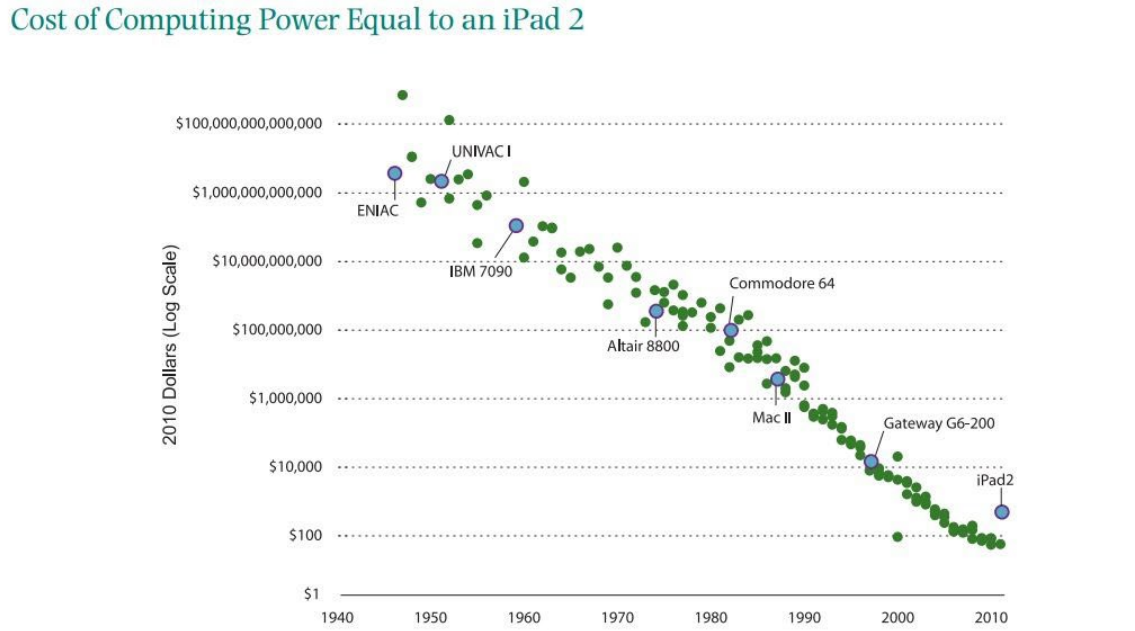
- Can generically define cost-efficiency in terms of *computational operations performed* (*e.g.*, FLOPs) *per dollar spent.*
  - Maximizing cost-efficiency equates to minimizing the cost to perform (some given number of) operations over the system's lifetime.
  - In general, this includes both costs to manufacture/deploy the system, and the lifetime cost of operating the system (including energy costs).

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

- In typical contexts today, the practical lifetime L of most computing systems is relatively fixed (a few years, say).
  - And also, for most applications, there is a maximum tolerable latency ℓ until the result of a given computational task must be obtained.
- So, generally we care about not *just* maximizing η_c, but also minimizing cost/op *for operations within some fixed timeframe,*
  - Which translates to increasing both *performance per unit (manufacturing) cost*, as well as (accounting for energy costs) *performance per unit power dissipation*.



SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.



Note: The iPad2 has computing power equal to 1600 million instructions per second (MIPS). Each data point represents the cost of 1600 MIPS of computing power based on the power and price of a specific computing device released that year.
Source: Moravec n.d..

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 eV!

- This increases leakage, impairs practical device performance
  - Thus, roadmap has minimum gate energy asymptoting to ~2 eV

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

- Transistor width 10-20× minimum width for fastest logic.
- Parasitic (junction, *etc.*) transistor capacitances (~2×).
- 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 (~2×).

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

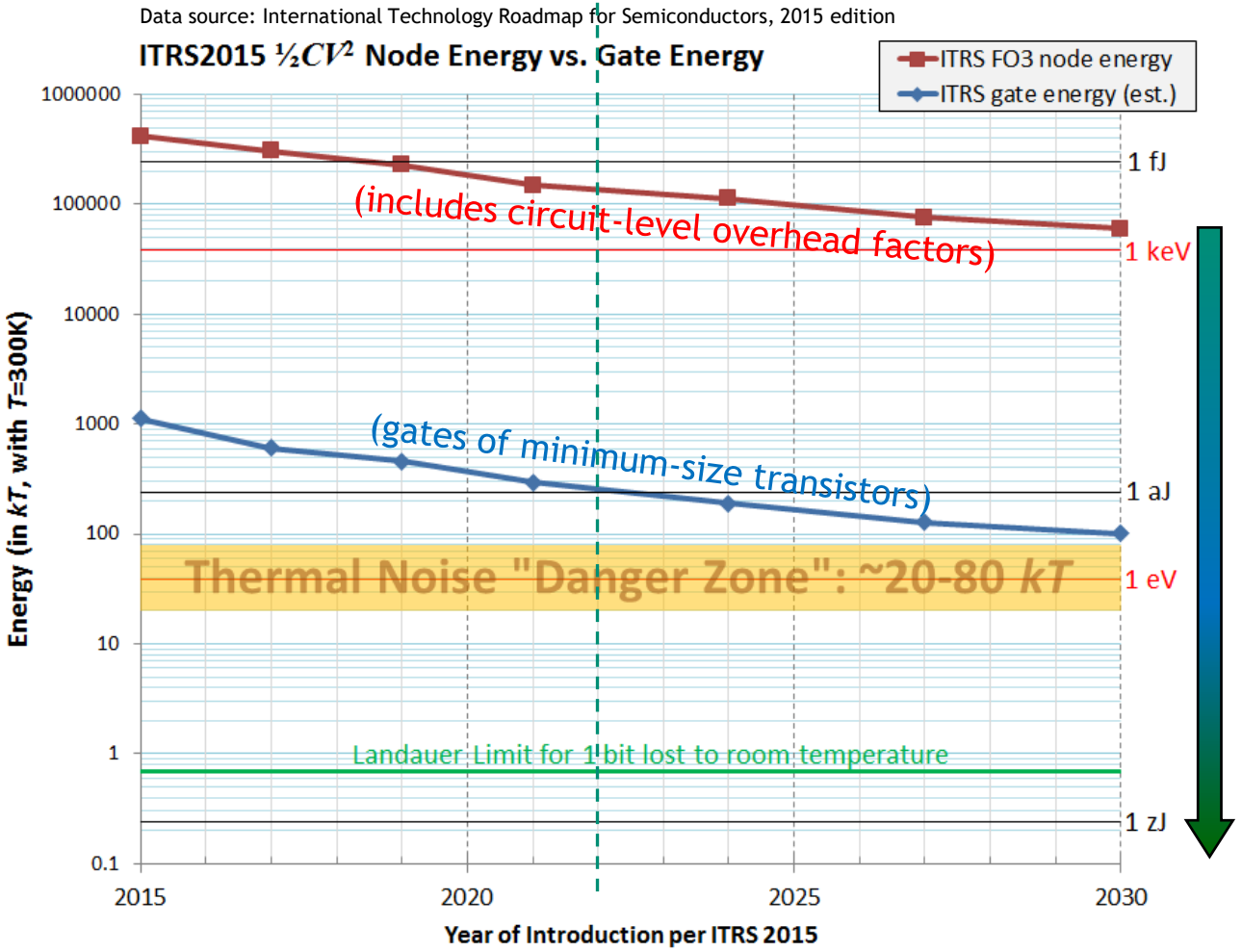
- 375-600× (!) larger in ITRS'15.
  - ∴ Practical bit energy for irreversible CMOS logic asymptotes to **~1 keV**!

Practical, real-world logic circuit designs can't just magically cross this ~500× architectural gap!

- ∴ Thermodynamic limits imply much *larger* practical limits!
  - The end is near!

This is Now!

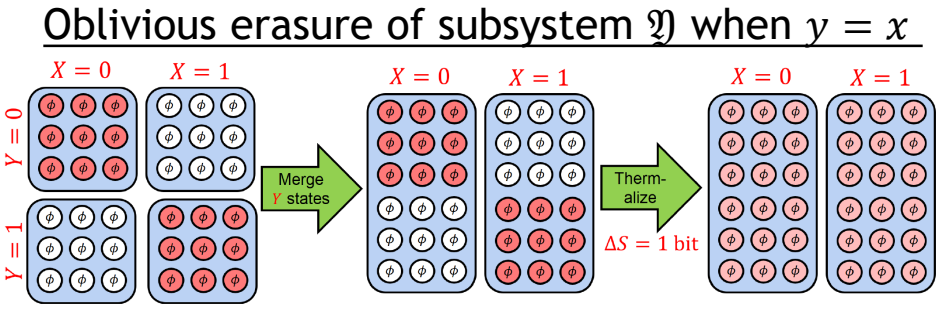
Only about a decade left…



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

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.*, isolated and unconditional) ***erasure*** (to a standard state) of a digital subsystem $\mathfrak{Y}$ that possesses ***marginal*** digital entropy $H(Y)$ (entropy after restriction of the joint $\mathfrak{X}\mathfrak{Y}$ distribution to $\mathfrak{Y}$) and was ***deterministically computed*** from another subsystem $\mathfrak{X}$ necessarily increases ***total physical entropy*** S by at least $H(Y)$. *(We can also generalize from this a little)*
  - **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 $\mathfrak{Y}$'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 $\mathfrak{Y}$ was computed ***deterministically*** from $\mathfrak{X}$, its conditional entropy $H(Y|X) = 0$, and therefore its marginal entropy is ***entirely*** accounted for by its mutual information with $\mathfrak{X}$, *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 $\mathfrak{E}$, 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 $I'(X;Y) = 0$.

Basic Reversible Computing Theory

(For full proofs, see 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 (*i.e.*, injective over its entire domain).

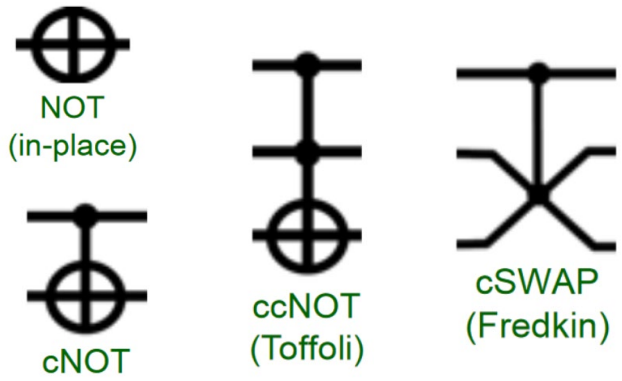
Fundamental theorem of generalized reversible computing:

- 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.

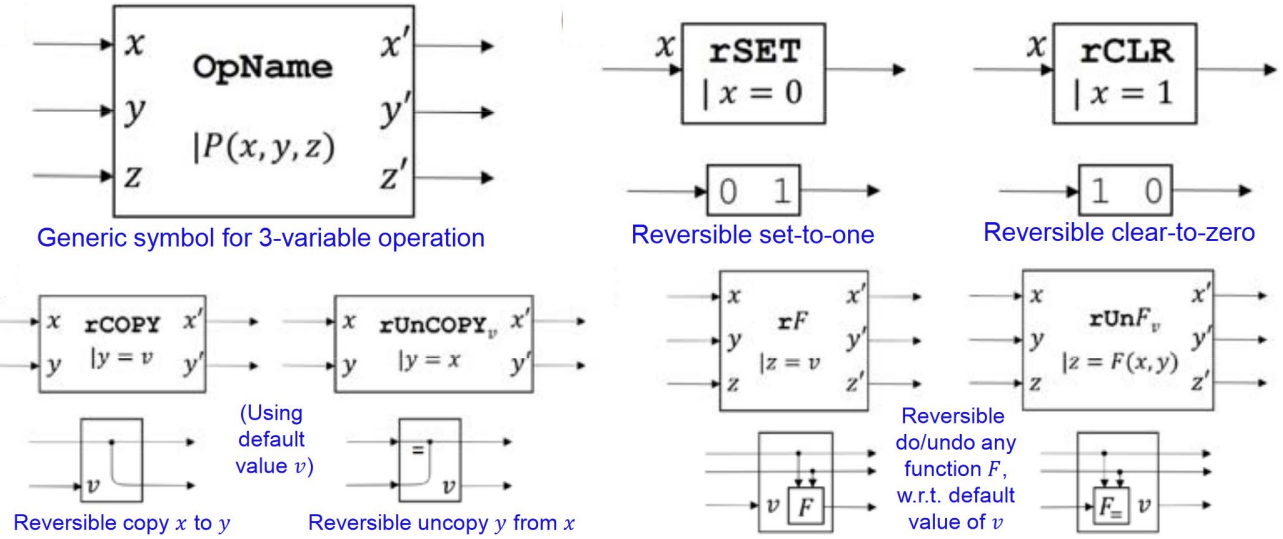
Bottom line: To avoid requiring Landauer costs, it is *sufficient to just have reversibility when some specified preconditions are satisfied.*

- This gives us a realistic (and more flexible!) basis for developing practical engineering implementations.
  - An example of this is provided by fully adiabatic CMOS.

Traditional *Unconditionally* Reversible "Gates" (Operations)

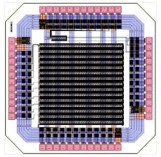


Some Generalized *Conditionally* Reversible Operations



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

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



Perfectly Adiabatic Reversible Computing in CMOS

To approach ideal reversible computing in CMOS…

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

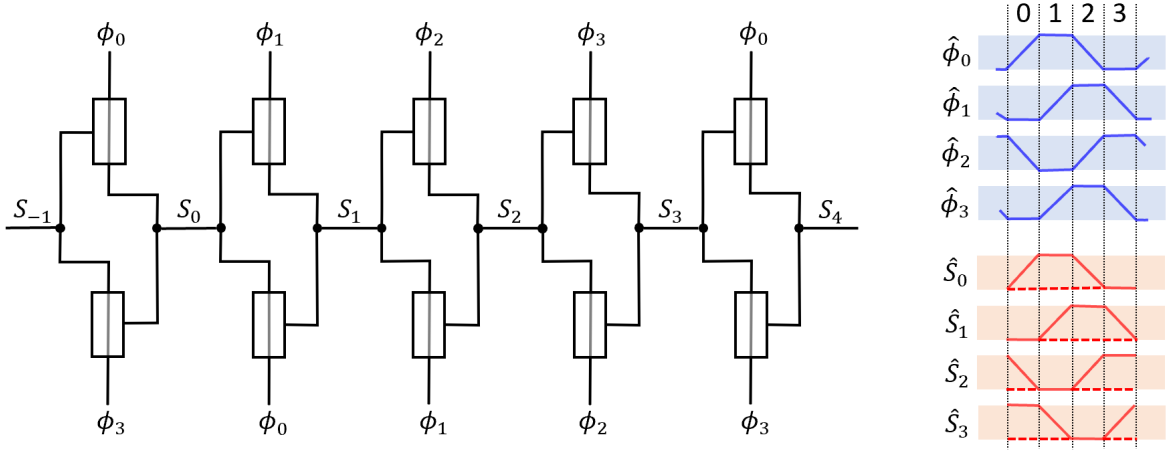
- Diodes in charging path, "sparking," "squelching,"
  - Eliminated by "**truly, fully adiabatic**" design. (*E.g.*, CRL, 2LAL).
    - Can suffice to get down to a few aJ (10s of eV) even *before* voltage optimization.
- Voltage level mismatches that dynamically arise on floating nodes before reconnection.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

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

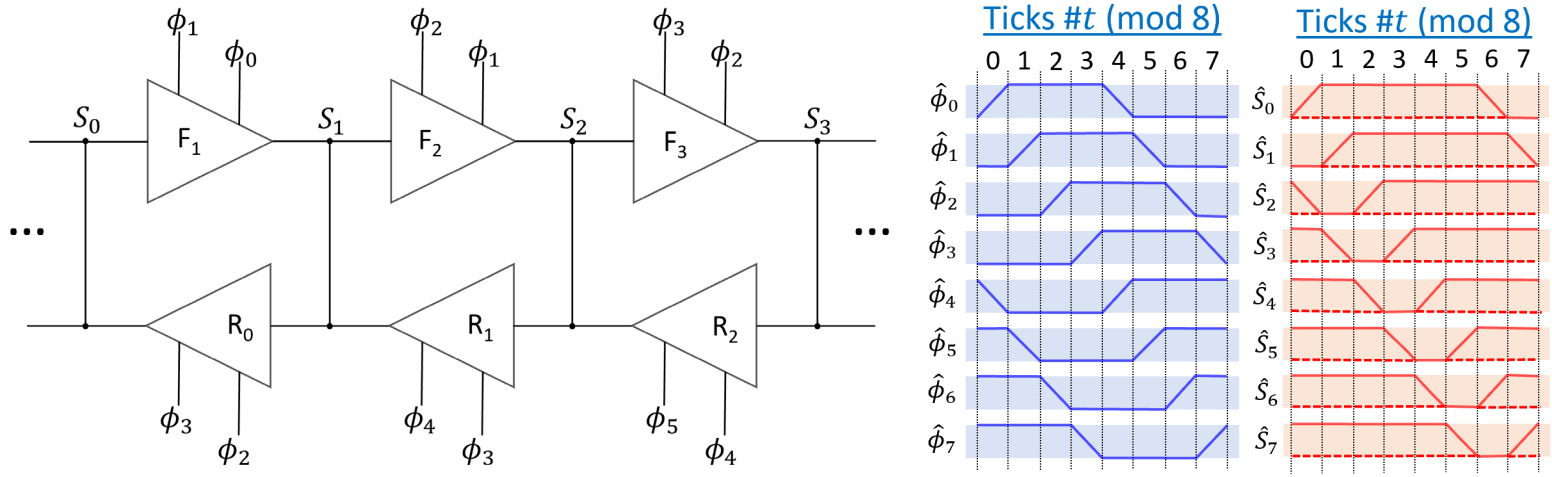
- Subthreshold channel currents.
  - Ultra-low-*T* (*e.g.* 4K) operation helps with this.
- Tunneling through gate oxide.
  - *E.g.*, use thicker gate oxides.

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)

Examples of S2LAL Logic Gates

14-transistor AND gate, 16-transistor OR gate.

- Carefully designed to ensure that each internal node is always connected to either a constant or variable source.
  - The structures shown are minimal, given the design constraints.

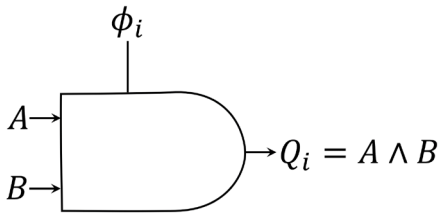
Inverting gates are done easily, by using signal pairs for complementary symbols:

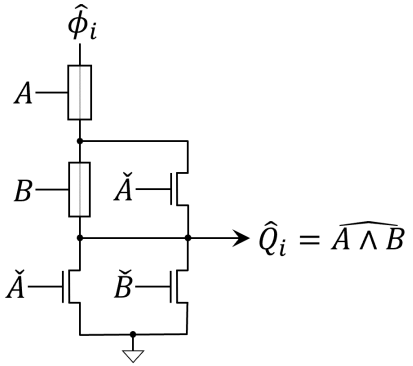
- $\mathrm{NOT}(A^1) = \mathrm{BUFFER}(A^0)$
- $\mathrm{NAND}(A^1, \mathrm{B}^1) = \mathrm{OR}(A^0, B^0)$
- $\mathrm{NOR}(A^1, \mathrm{B}^1) = \mathrm{AND}(A^0, B^0)$

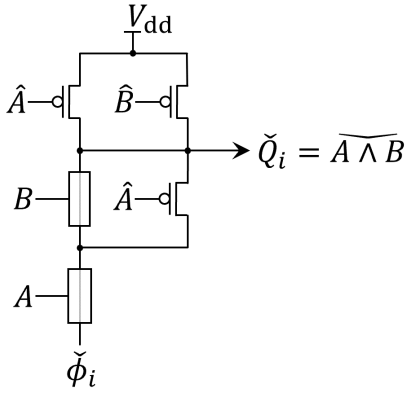
Also! Erik DeBenedictis invented an optimization to S2LAL that can compute the inverses as-needed, rather than always keeping both the 0,1 signal pairs around:

- See https://zettaflops.org/zf004/ .

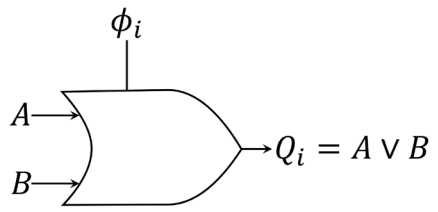
AND

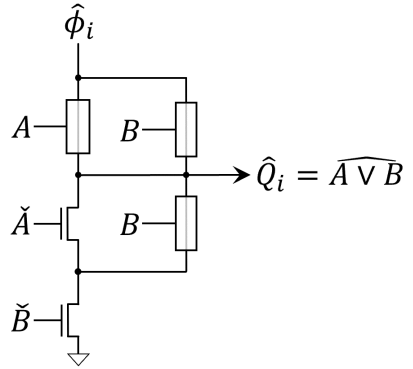


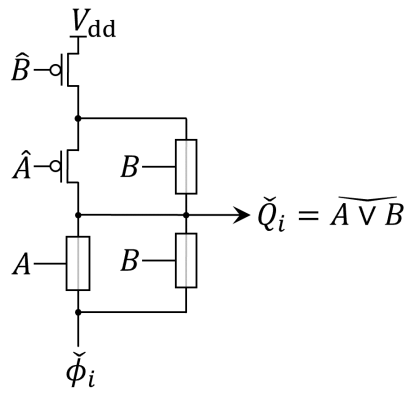




OR







Reversible Computing Technologies in Superconducting Platforms

The Reversible Computing Scaling Path: Challenges and Opportunities

Adiabatic Reversible Computing in Superconducting Circuits

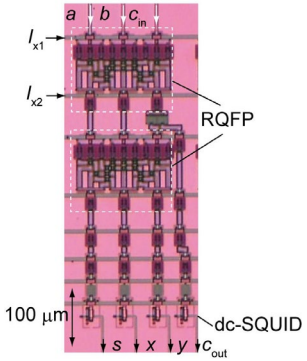
Work along this general line has roots that go all the way back to Likharev, 1977. (doi:10.1109/TMAG.1977.1059351)

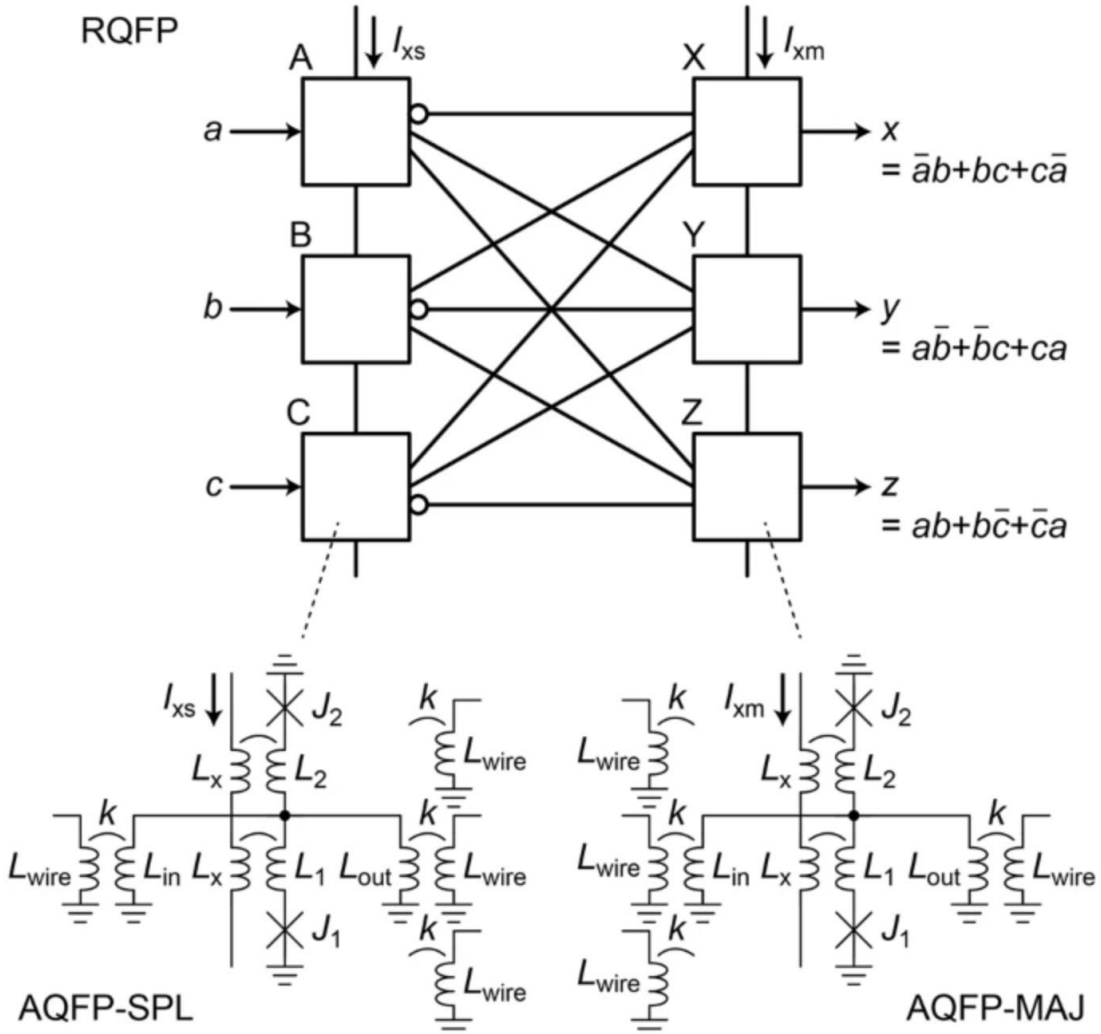
- Most active group recently is Prof. Yoshikawa's group at Yokohama National University in Japan.

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 noting that $T = 4\text{K}$), at speeds on the order of 10 MHz





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

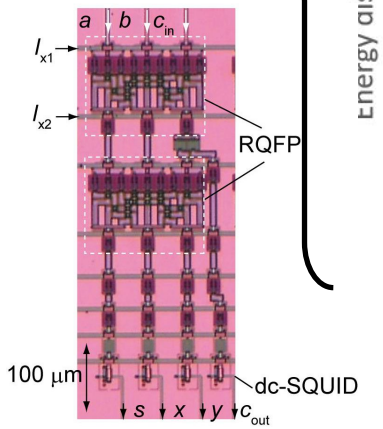
Reversible adiabatic superconductor logic:

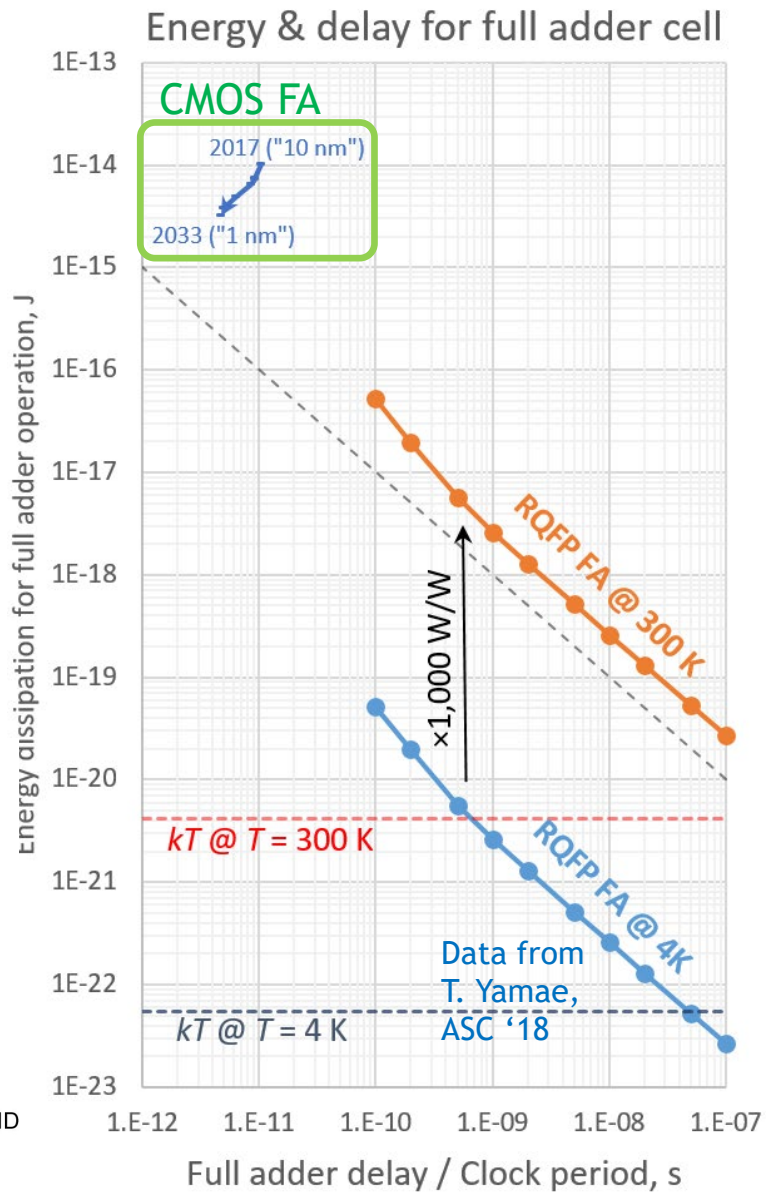
- State-of-the-art is the **RQFP** (Reversible Quantum Flux Parametron) technology from Yokohama National University in Japan.
  - Chips were fabricated, function validated.
- Circuit simulations predict DdP is >1,000× *lower* than even *end-of-roadmap* CMOS.
  - Dissipation extends *far below* the 300K 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×).

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?

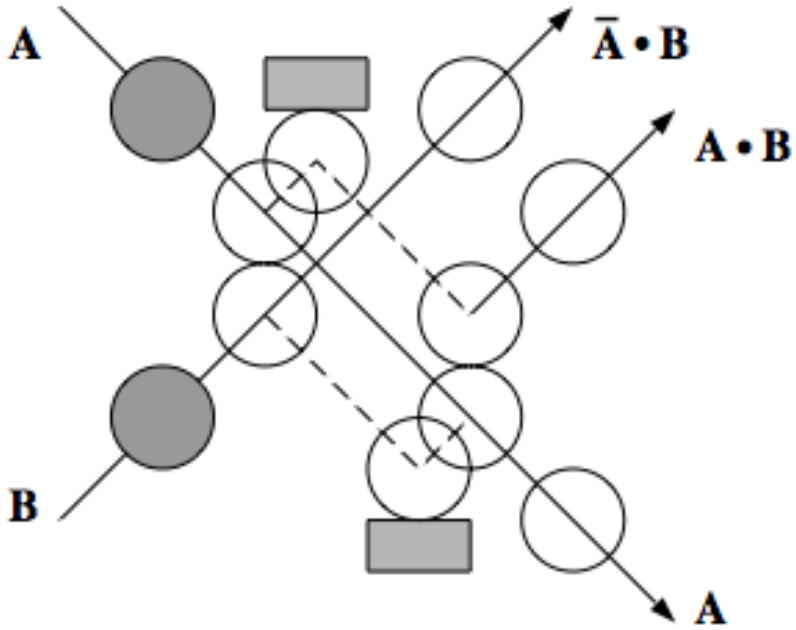
RQFP = Reversible Quantum Flux Parametron (Yokohama U.)

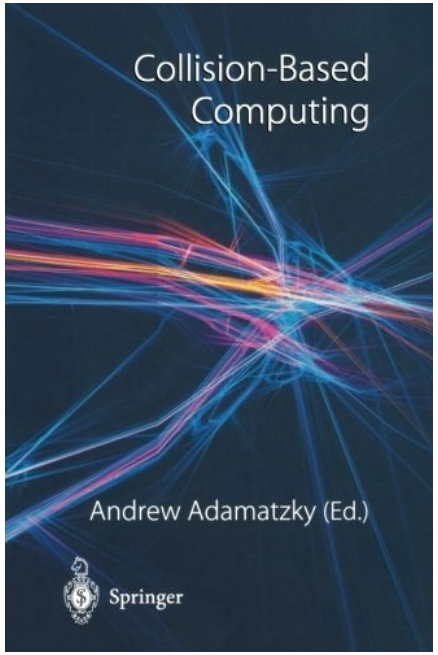




Ballistic Reversible Computing

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





Historical origin of this concept:

- Fredkin & Toffoli's *Billard 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 *re*synchronization 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

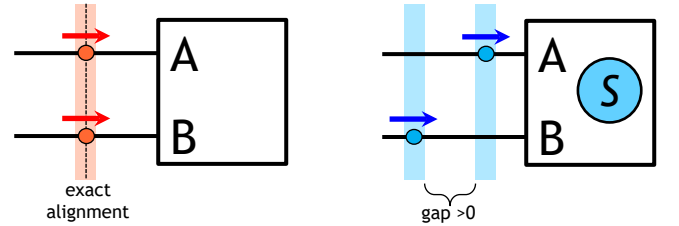
- Device's dynamical trajectory then becomes *independent* of the precise (absolute *and* relative) pulse arrival times
  - As a result, timing uncertainty per logic stage can now accumulate only *linearly*, not exponentially!
    - Only relatively occasional re-synchronization will be needed
- For devices 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!

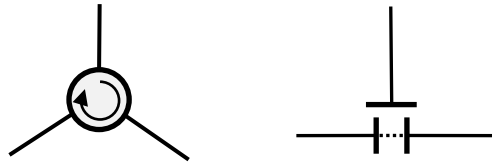
- Devices simply operate whenever data pulses arrive
- The operation energy is carried by the pulse itself
  - Most of the energy is preserved in outgoing pulses
    - Signal restoration can be carried out incrementally, or periodically

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

(BARCS effort)

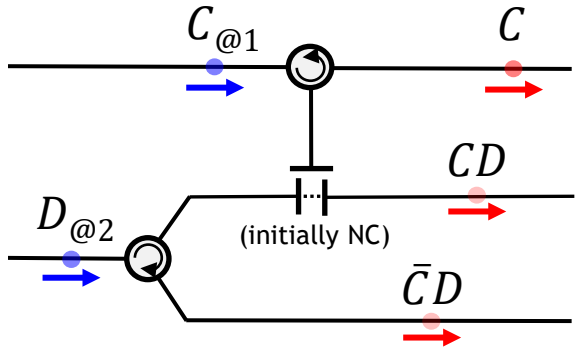


Synchronous Ballistic | Asynchronous Ballistic



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:

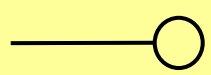
- (1) Bits encoded in fluxon polarity; (2) Bounded planar circuit conserving flux; (3) Physical symmetry.

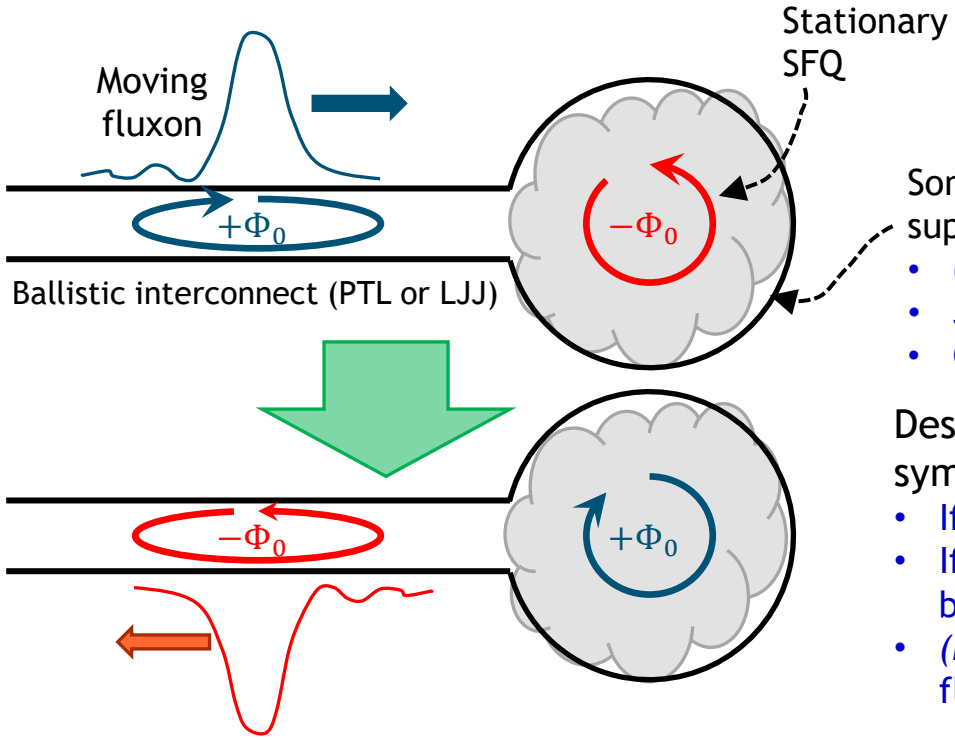
Determined through theoretical hand-analysis that the simplest such function is the ***1-Bit, 1-Port Reversible Memory Cell (RM):***

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

RM Transition Table

Input Syndrome		Output Syndrome
$+1(+1)$	$\rightarrow$	$(+1)+1$
$+1(-1)$	$\rightarrow$	$(+1)-1$
$-1(+1)$	$\rightarrow$	$(-1)+1$
$-1(-1)$	$\rightarrow$	$(-1)-1$

RM icon: 



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—First working (in simulation) implementation!

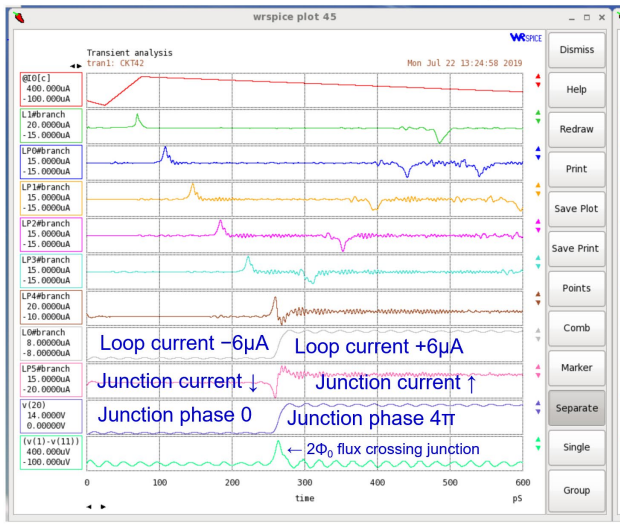
Erik DeBenedictis: "Try just strapping a JJ across that loop."
- This actually works!

"Entrance" JJ sized to = about 5 LJJ unit cells (~1/2 pulse width)
- I first tried it twice as large, & the fluxons annihilated instead…
  - "If a 15 μA JJ rotates by 2π, maybe ½ that will rotate by 4π" 🤔

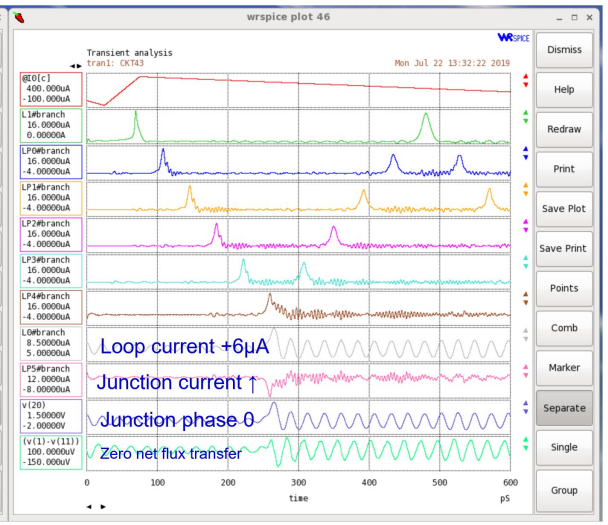
Loop inductor sized so ±1 SFQ will fit in the loop (but not ±2)
- JJ is sitting a bit below critical with ± 1

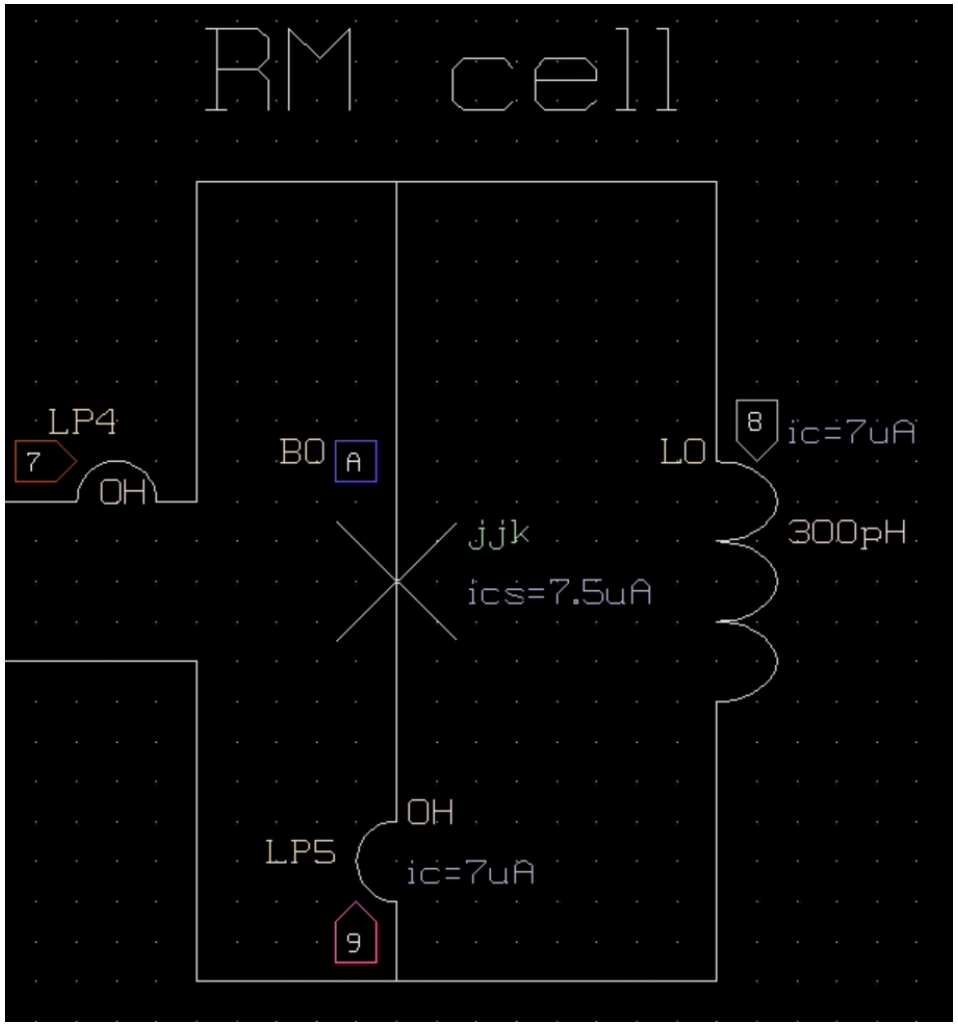
WRspice simulations with ±1 fluxon initially in the loop
- Uses `ic` parameter, & `uic` option to `.tran` command
  - Produces initial ringing due to overly-constricted initial flux
    - Can damp w. small shunt *G*

Polarity mismatch → Exchange



Polarity match → Reflect (=Exchange)

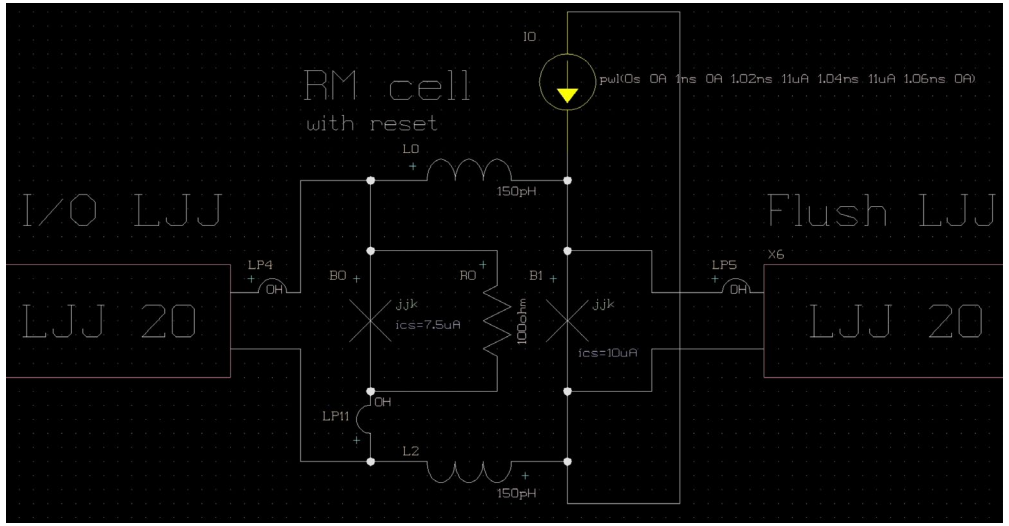


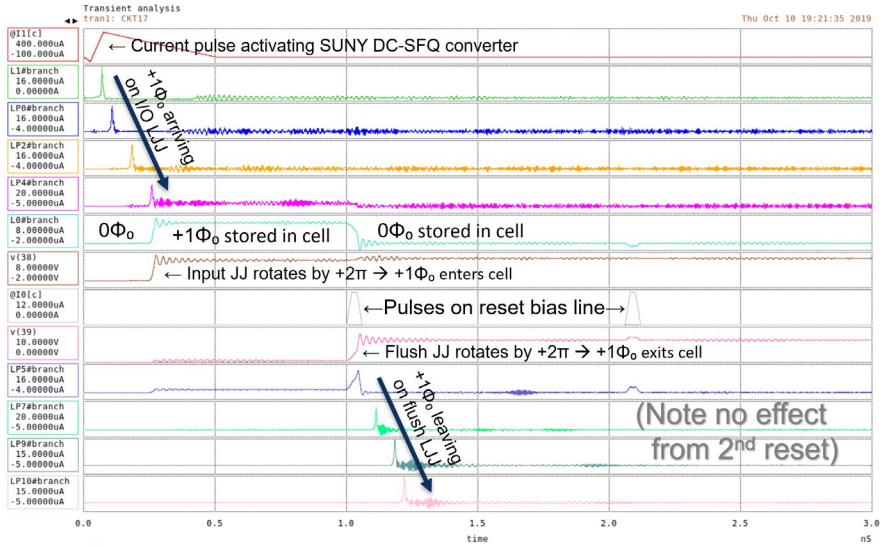


Resettable version of RM cell—Designed & Fabricated!

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

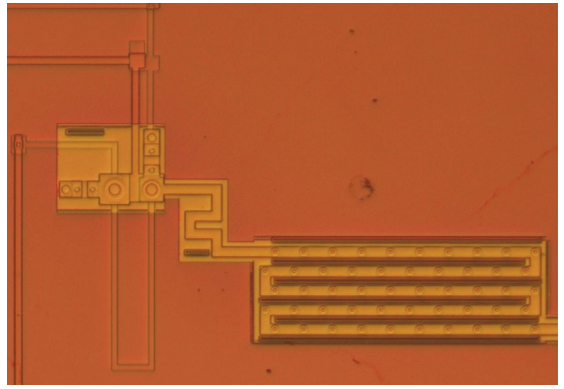
- To flush either polarity → Do both (±) resets in succession

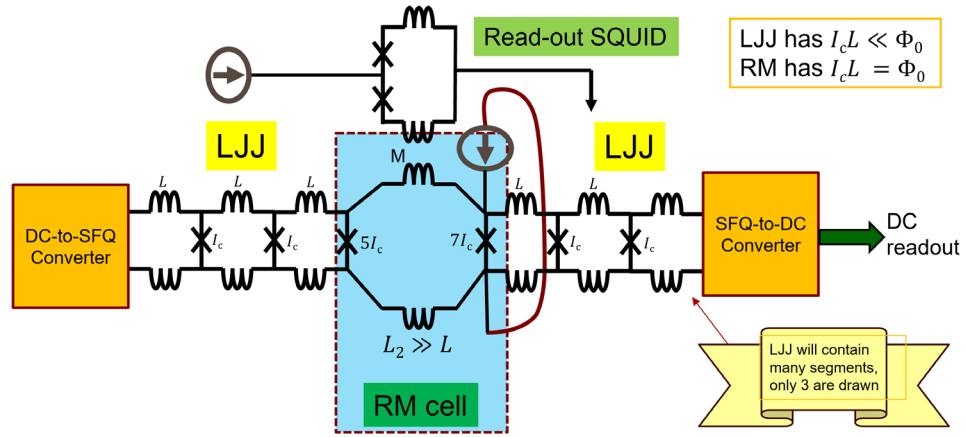


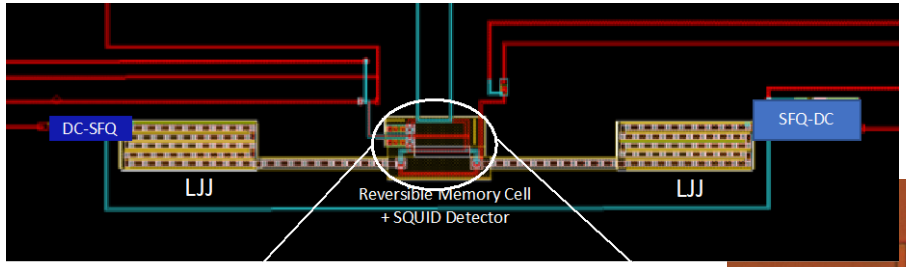


Fabrication at SeeQC with support from ACI

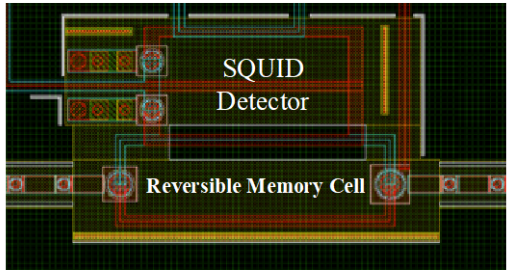
DC-SFQ & LJJ

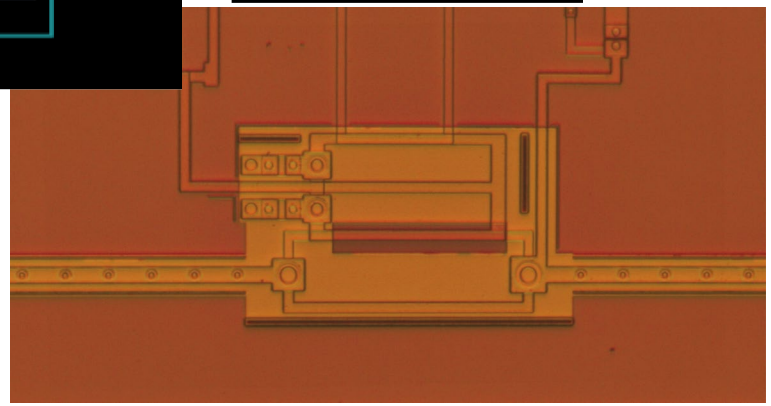






RM Cell & SQUID





2. Scaling Analysis of Reversible vs. Non-reversible Machines

The Reversible Computing Scaling Path:
Challenges and Opportunities

Why Reversible Computing Wins Despite Its Overheads!

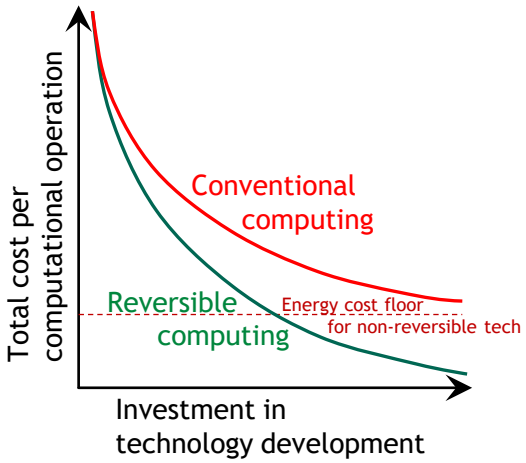
$$\eta_c = \frac{\#ops}{Cost}$$

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 *aggregate* performance within system power constraints.
  - This is borne out by a detailed economic/systems-engineering analysis.

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

Amortized Cost Scaling



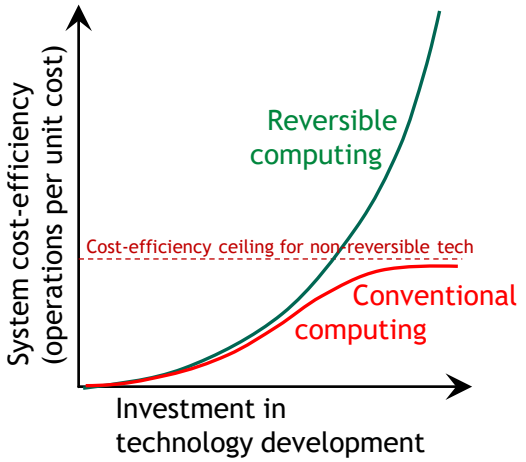
Bottom line: The computational *performance (ops./sec.) per unit budgetary cost (e.g. $/yr.)* on parallelizable computing workloads can become as large as desired, given only that *both terms* in this expression for total *cost per operation* C_{op} can be made sufficiently small:

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

where:

- c_E is the operating cost C_{oper} attributable to supplying power/cooling, divided by energy delivered.
- $E_{\text{diss,op}}$ is the system energy dissipation, divided by number of operations performed.
- c_M is the total cost C_{mfg} for system manufacturing & deployment, *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.

Cost-Efficiency Scaling



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 our continuing to make the manufacturing cost coefficient c_M *ever smaller* as manufacturing processes are refined (and/or the deployed lifetime of the system increases).

∴ *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 $E_{\text{diss,op}}$ down!

Minimum Energy Scaling for Adiabatic CMOS

From M. Frank & K. Shukla, doi:10.3390/e23060701

Appendix A. Minimum-Energy Scaling for Classical Adiabatic Technologies

In this appendix, we briefly present the derivation for the scaling of minimum energy dissipation for reversible technologies such as RA-CMOS (Section 2.3.1) that obey classic adiabatic scaling and that can be characterized in terms of relaxation and equilibration timescales.[44]

First, we assume (as is the case for "perfectly adiabatic" technologies such as [48]) that the total energy dissipation per clock cycle E_{diss} in a reversible circuit can be expressed as a sum of *switching losses* and *leakage losses*,

$$E_{\text{diss}} = E_{\text{sw}} + E_{\text{lk}}, \tag{A1}$$

and further, that switching and leakage losses depend on the signal energy E_{sig} and transition time t_{tr} approximately as follows:

$$E_{\text{sw}} \simeq E_{\text{sig}} \cdot c_{\text{sw}} \cdot \frac{\tau_{\text{r}}}{t_{\text{tr}}}, \tag{A2}$$

$$E_{\text{lk}} \simeq E_{\text{sig}} \cdot c_{\text{lk}} \cdot \frac{t_{\text{tr}}}{\tau_{\text{e}}}, \tag{A3}$$

where $\tau_{\text{r}}, \tau_{\text{e}}$ are the relaxation and equilibration timescales, respectively, and $c_{\text{sw}}, c_{\text{lk}}$ are small dimensionless constants characteristic of a particular reversible circuit in a specific family of technologies, such as [48]. In practice, although these specific formulas are only approximate, they approach exactness in the regime $\tau_{\text{r}} \ll t_{\text{tr}} \ll \tau_{\text{e}}$.

Then, now treating (A2), (A3) as exact, we can write:

$$E_{\text{diss}} = E_{\text{sig}} \left(c_{\text{sw}} \tau_{\text{r}} \cdot \frac{1}{t_{\text{tr}}} + \frac{c_{\text{lk}}}{\tau_{\text{e}}} \cdot t_{\text{tr}} \right). \tag{A4}$$

We can collect the constants, absorbing them into adjusted timescales $\tau'_{\text{r}} = c_{\text{sw}} \tau_{\text{r}}$ and $\tau'_{\text{e}} = \tau_{\text{e}} / c_{\text{lk}}$, so

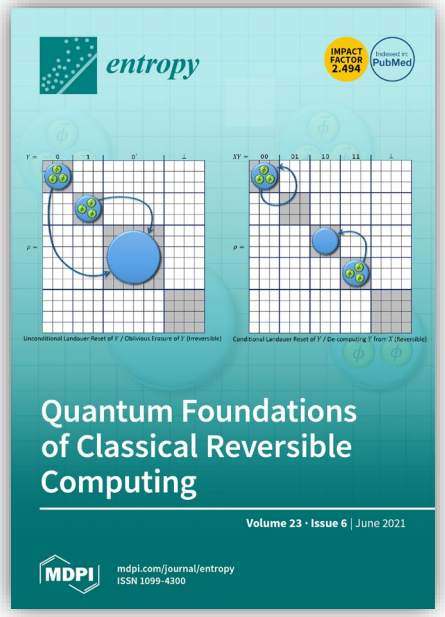
$$E_{\text{diss}} = E_{\text{sig}} \left(\tau'_{\text{r}} \cdot \frac{1}{t_{\text{tr}}} + \frac{1}{\tau'_{\text{e}}} \cdot t_{\text{tr}} \right). \tag{A5}$$

Setting the derivative of (A5) with respect to t_{tr} equal to zero, we find that E_{diss} is minimized when

$$\tau'_{\text{r}} \frac{1}{t_{\text{tr}}^2} = \frac{1}{\tau'_{\text{e}}}, \tag{A6}$$

Upshot for CMOS: As each device's leakage conductance I_{off} is decreased, the equilibration timescale τ_{e} increases, and the technology's minimum energy (given perfectly adiabatic, reversible designs) scales down with square-root proportionality.

$$E_{\text{diss,min}} \propto \frac{1}{\sqrt{\tau_{\text{e}}}} \propto \sqrt{I_{\text{off}}}$$



or in other words, when

$$t_{\text{tr}} = \sqrt{\tau'_{\text{r}} \tau'_{\text{e}}}, \tag{A7}$$

at which point E_{sw} and E_{lk} are equal. The minimum energy dissipation per cycle is then

$$E_{\text{diss}} = 2E_{\text{sig}} \sqrt{\frac{\tau'_{\text{r}}}{\tau'_{\text{e}}}}. \tag{A8}$$

Thus, for any given reversible circuit design in a family of technologies with given values of the constants $c_{\text{sw}}, c_{\text{lk}}$, in order for E_{diss} to approach 0 as the technology develops, we must have that the ratio of equilibration/relaxation timescales $\tau_{\text{e}}/\tau_{\text{r}} \to \infty$, and, if the relaxation timescale τ_{r} is fixed, this implies that also the (minimum-energy) value of the transition time $t_{\text{tr}} \to \infty$. These requirements were mentioned in Section 2.3.1.

More specifically, in order to increase the peak energy efficiency of a reversible circuit by a factor of $N\times$, in a given family of technologies obeying classic adiabatic scaling, this requires that the timescale ratio $\tau_{\text{e}}/\tau_{\text{r}}$ must be increased by $N^2\times$, and (assuming τ_{r} is fixed) the transition time t_{tr} for minimum energy will increase by $N\times$.

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 CMOS processes, including:

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

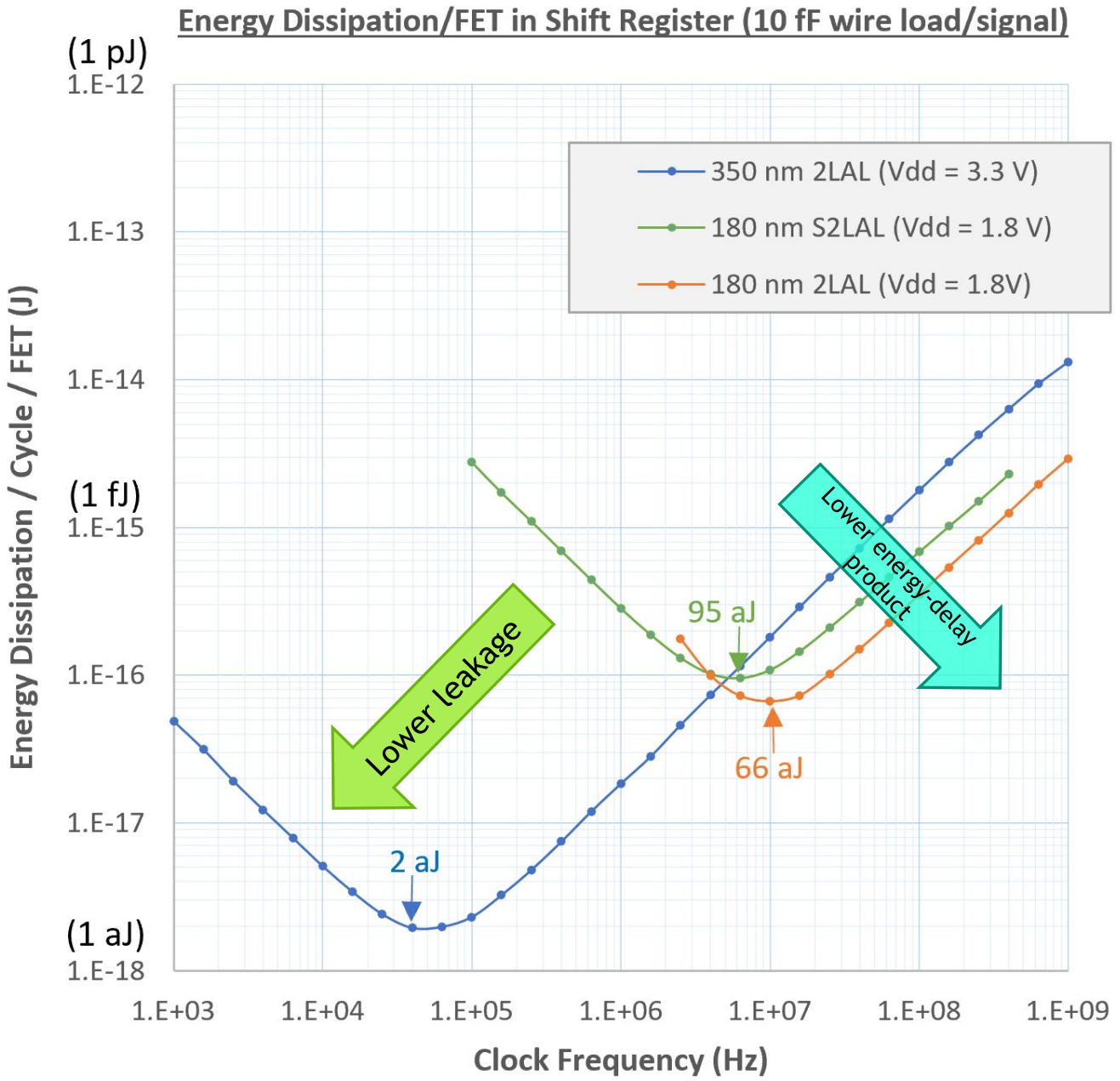
Plotted energy dissipation per-transistor in shift registers at 50% activity factor (alternating 0/1)

- 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).



Performance Per-Area Scaling with Machine Thickness

Frank & Knight 1997, doi:10.1088/0957-4484/9/3/005

Assumptions of this simple analysis include:

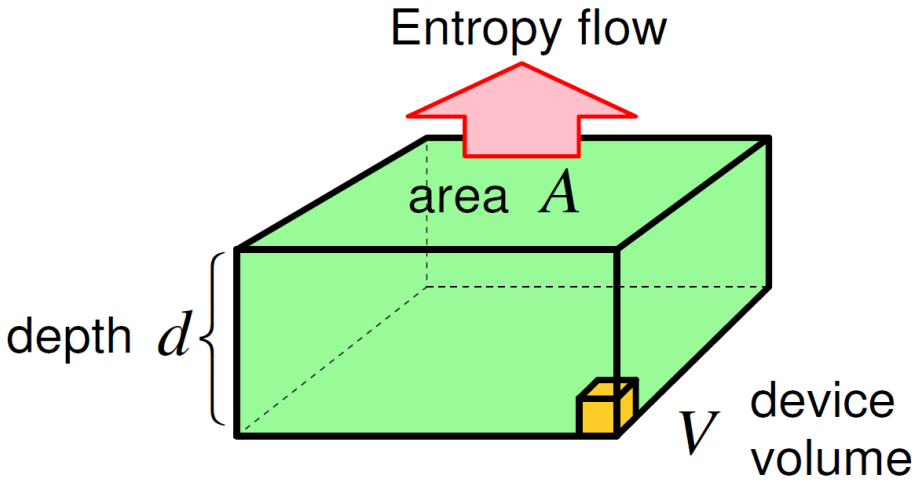
- Classic adiabatic ($E_{\text{diss,op}} \propto 1/t$) scaling.
- Fixed operating temperature.
- Constant volume and mass per device.
- Bounded entropy flux density F_S.
- No algorithmic overheads for reversibility.

Upshot: Sustained performance of reversible machines asymptotically scales as $A\sqrt{d}$, which is $\sqrt{d}\times$ better than scaling of irreversible machines.

- Here, A is the area of the machine's minimal bounding surface, and d is the *depth* or thickness of the machine (along its thinnest dimension).

More detailed analyses also account for the impact of considering the algorithmic overheads of reversibility.

- Spoiler: Reversible computing still wins!



Number of devices $N_{\text{dev}} = Ad/V$
Entropy generation rate $\mathcal{R}_S = N_{\text{dev}} S/t_{\text{op}}$
Entropy removal rate $\mathcal{R}_S = F_S A$
Entropy per operation $S = k_S/t_{\text{op}}$
Total rate of operation $\mathcal{R}_{\text{op}} = N_{\text{dev}}/t_{\text{op}}$

$$\mathcal{R}_{\text{op}} = A\sqrt{\frac{dF_S}{k_S V}}$$

Max rate achieved when $t_{\text{op}} = \sqrt{dk_S/F_S V}$.

Beats irreversible $\mathcal{R}_{\text{op}} = A\frac{F_S}{S}$ whenever $d > \frac{F_S k_S V}{S^2}$.

Figure 6.1: Speed limit for reversible machines of minimum-surface area $\Theta(A)$ and thickness $d \precsim A^{1/2}$. 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).

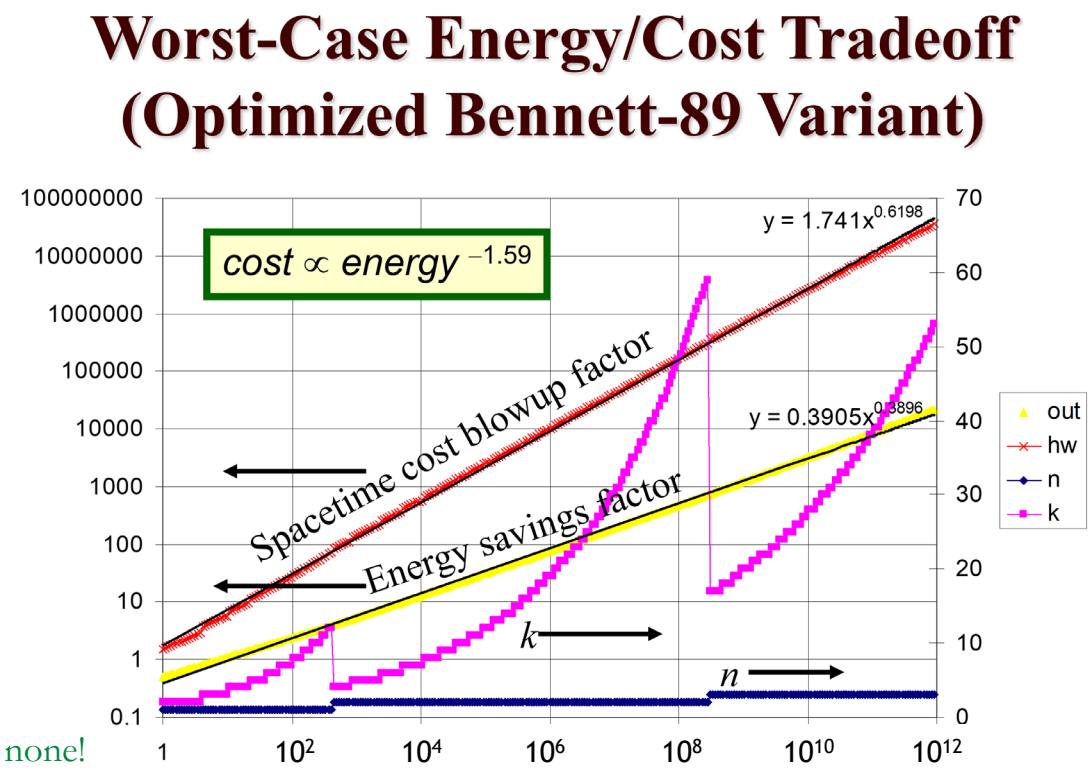
- And, even better general "reversiblization" algorithms may yet be discovered in the future.

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× and mfg. cost/device by 59,000×.
  - Ambitious but doable!! This gives us **a way forward**, where otherwise there is none!



3. Key Scientific and Engineering Challenges for Reversible Computing

The Reversible Computing Scaling Path: Challenges and Opportunities

Major Scientific/Engineering Challenges for RC (Overview)

Physics challenges:

- Deriving fundamental limits of energy dissipation as a function of key physical parameters (Shukla @ Brown)
- Clarifying the *asymptotic* scaling behavior of synchronous and asynchronous reversible machines (Earley @ Cambridge)
- Searching for novel (*e.g.* quantum) mechanisms to suppress dissipation at finite scales

Engineering challenges:

- Methods to systematically increase the effective quality factor of energy-recovering driving mechanisms for synchronous reversible machines
- Further explore the potential engineering realizations of the asynchronous ballistic reversible paradigm
- Extension of design automation tools and methodologies for reversible design

Workforce development challenges:

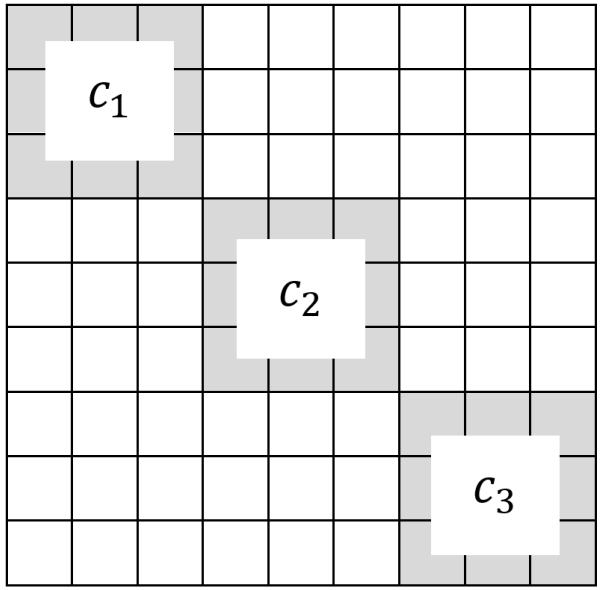
- Educational tools & materials needed to train/retrain the engineering workforce to work in this unfamiliar technological paradigm

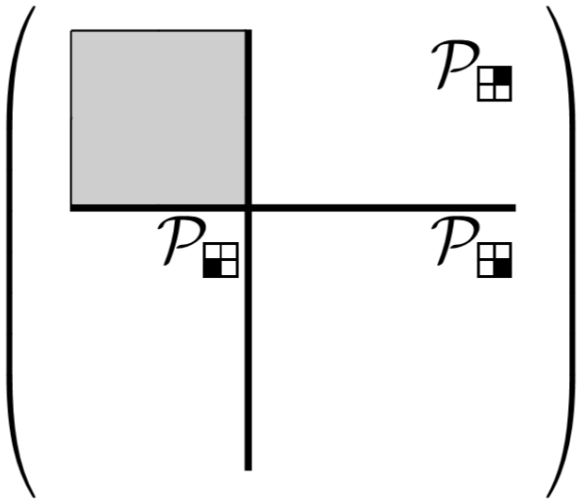
Reversible Operations as Quantum Channels

(Work by K. Shukla, Brown U.
doi:10.3390/e23060701)

BROWN

- Want to characterize dissipation of reversible operations. Do fundamental limits exist? If so, what is the dependence on fundamental parameters?
  - *Most general* limits for practical models: nonequilibrium quantum thermodynamics (NEQT).
  - Unitary evolution: no dissipation, but time to implement (usually) bounded by quantum speed limit (QSL).
  - Dissipation as a function of delay ($D(d)$). In principle, not subject to QSL, but can be used as an initial value of δt for a non-tight bound. Goal: retrieve protocol-based, device-independent expression for $D(d)$.
- Quantum limits: natural framework is representing classical operations as quantum channels.
  - Computational states c_i form equivalence classes over (physical) quantum states $|\psi\rangle$. Permits only coherences between different $|\psi\rangle$ corresponding to the *same* c_i. Thus, each c_i is a single decoherence-free subspace (DFS) of overall Hilbert space.
  - Computation embedded in open system. Information can "leak" into environment, but (we assume) cannot then be recaptured at any future time. Thus, dynamics represented by Markovian (Lindbladian / GKSL) evolution.
  - GKSL with multiple asymptotic states (V. V. Albert *et al.* Phys. Rev. X **6**, 041031 (2016); V. V. Albert, PhD thesis, Yale (2018)): asymptotic states form subspace $\mathsf{As(H)}$ in overall dynamics. $\mathsf{As(H)}$ itself is part of a larger subspace of GKSL evolution; larger subspace (grey in the second image) is where computational information is represented and reversible operations are carried out. This framework provides most general embedding of quantum channel into GKSL dynamics.





Second image from V. V. Albert, Ph.D thesis, Yale U. (2018).

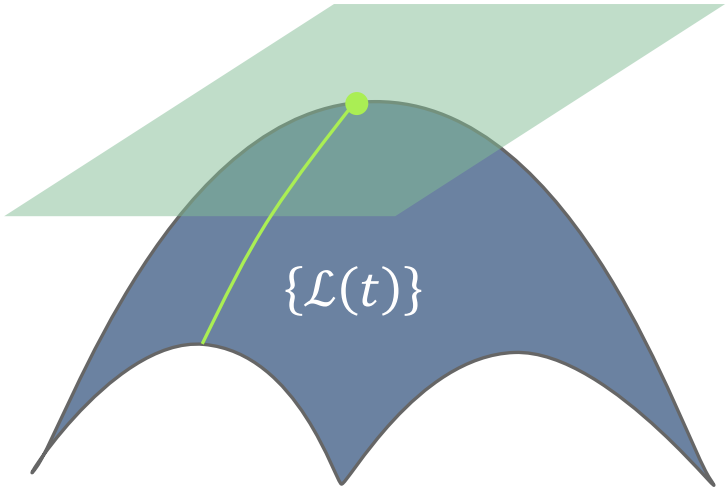
Thermodynamic Quantities and D(d)

(Work by K. Shukla, Brown U. doi:10.3390/e23060701)

BROWN

- Markovian time evolution: (quantum) adiabatic evolution.
  - Direct analogy from quantum adiabatic theorem in (closed system) quantum. States in asymptotic subspace $\mathsf{As}(\mathsf{H})$ evolve over time under adiabatic theorem: full family of all $\mathsf{As}(\mathsf{H})$ is fiber bundle of the manifold describing evolution under GKSL / Lindbladian operator $\mathcal{L}$.
  - As with adiabatic time evolution in quantum, manifold can have nontrivial curvature, parallel transport, and metric tensor properties. V. Albert *et al.* Phys. Rev. X **6**, 041031 (2016) provides a full framework for calculating geometric quantities on manifold by writing down expressions analogous to the Berry curvature $\sigma_{\mu\nu}$ and the quantum geometric tensor $g_{\mu\nu}$.
- Underlying manifold is *manifold of non-equilibrium asymptotic states*. Geometric properties of manifold provide key quantum thermodynamic insights:
  - Geodesics under dissipation metric provide minimal dissipation ("thermodynamic length") for adiabatic evolution of a given Hamiltonian (M. Scandi & M. Perarnau-Llobet, Quantum **3**, 197 (2019)).
  - Geodesics under Fisher information metric provide thermodynamic uncertainty relation (TUR): quantum uncertainty relation between average entropy production rate and average values of currents (G. Guarnieri *et al.*, Phys. Rev. Res. **1**, 033021 (2019)).
- Both of these results rely on manifolds generated by GKSL evolution with a *single* asymptotic state. Results depend *both* on the dimensionality of $\mathsf{As}(\mathsf{H})$ (single asymptotic state vs. several asymptotic states) and the specific metric chosen (dependent on question).
- Steps for $D(d)$: determine thermodynamic length and TUR for multiple asymptotic state setup.

Hilbert space $\mathsf{As}(\mathsf{H})$ generated by $\{\mathcal{L}(t)\}$



From line element $\mathrm{d}s^2$ on manifold surface:

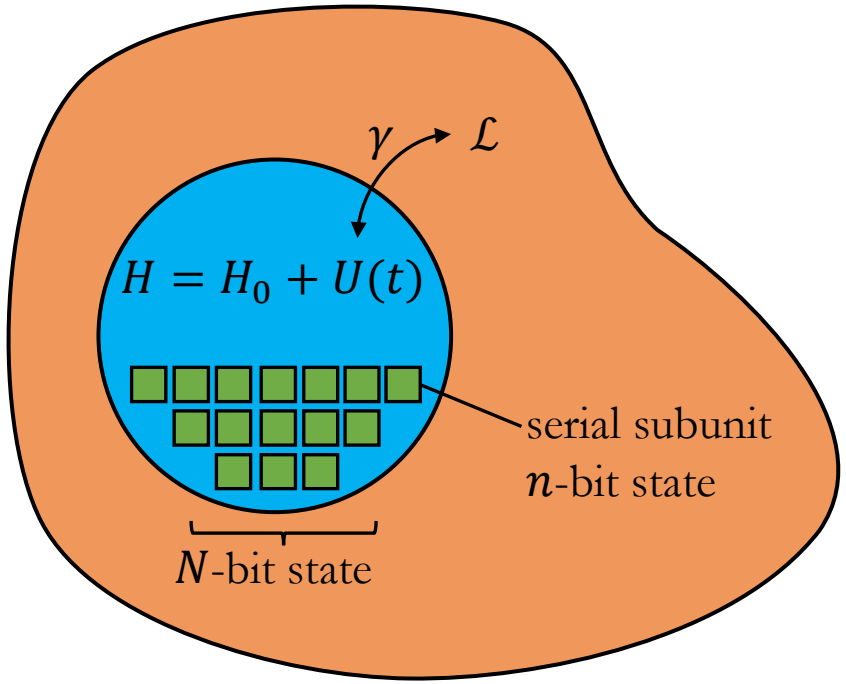
$$\mathrm{d}s^2 = \|\psi(\lambda + \mathrm{d}\lambda) - \psi(\lambda)\|^2$$
$$= \left(\gamma_{\mu\nu} + i\sigma_{\mu\nu}\right) \mathrm{d}\lambda^\mu \, \mathrm{d}\lambda^\nu$$

Berry connection is $A_\mu := \langle n(\lambda) | \partial_\mu | n(\lambda) \rangle$, where $|n(\lambda)\rangle$ is instantaneous eigenstate at λ (parameter corresponding to a time in adiabatic evolution).

- $\sigma_{\mu\nu} = \partial_\mu A_\nu - \partial_\mu A_\nu$: Berry curvature.
- $g_{\mu\nu} = \gamma_{\mu\nu} - A_\mu \wedge A_\nu$: geometric tensor.

Asymptotic Adiabaticity (slide 1 of 2)

(Work by H. Earley, Cambridge U., doi:10.17863/cam.79111)



Setup (per subunit)

- $\mathcal{L}$ is the Lindbladian, γ is the coupling strength
- P is the projector of the correct computational subspace, $1 - P$ is the projector onto the invalid subspace
- ρ is the density, $\rho = P\rho$ at the beginning of each reset cycle
- H_0 is the Ballistic computational Hamiltonian
- $U(t)$ is the (time-dependent) perturbation due to the system's temperature/unconstrained dofs

Key result: Adiabatic scaling is universally maximal for RC, for asymptotically 'large' computers, running for asymptotically long time periods, in the presence of environmental thermal coupling

- 'Large': threshold at point of saturating ability to dissipate entropy from the system (which scales with convex boundary)
- Thermal coupling leads to deviation of state from computational subspace, need to 'reset' to ensure continued correct operation *(Note: resets could possibly be just by continuous environment-inducted superselection)*

Proof in two parts: two regimes of correction frequency

- sufficiently fast resets allow for exploitation of **Quantum Zeno Effect**
- slow resets give normal Fermi golden rule transitions; can also specialize proof to various classical cases such as abstract chemical reaction networks or general Lagrangian descriptions

Error in computational state given by projection of evolved density onto erroneous subspace

$$\delta p \equiv \mathbb{P}(\text{error} \mid \delta t) = \text{tr}[(1 - P)\rho(t + \delta t)]$$

Expand $\rho(t + \delta t)$ to second order in δt using Lindbladian master equation twice

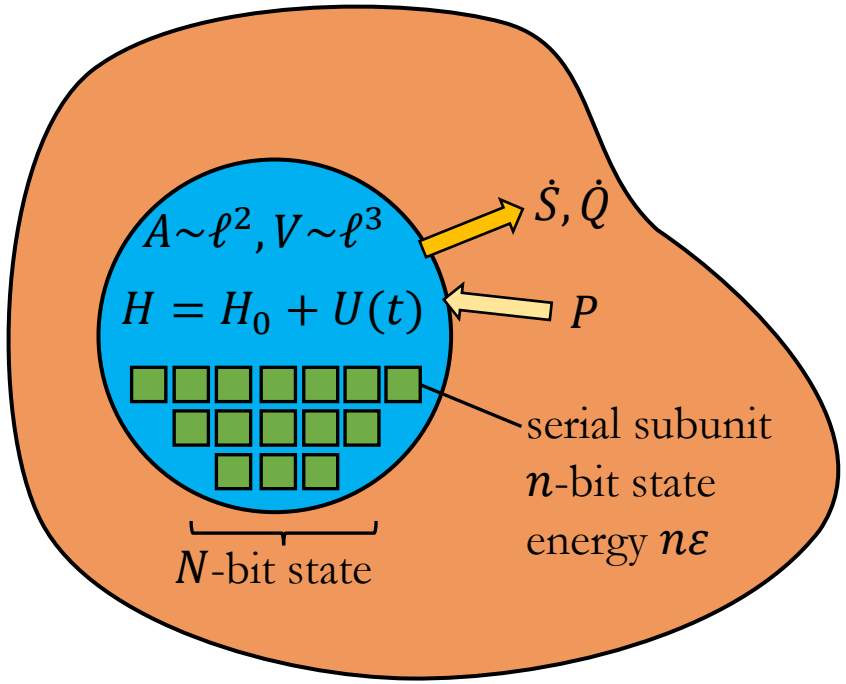
$$\delta p = \delta t^2 \,\text{tr}\left[(1 - P)\left(\frac{1}{\hbar^2} U\rho U + \frac{\gamma}{\delta t}\Sigma_i \mathrm{L}_i \rho \mathrm{L}_i^\dagger\right)\right] + \mathcal{O}(\delta t^3)$$

$$\approx n\left(1 + \frac{\gamma}{\delta t}\frac{A}{V}\right)\delta t^2 \left(\frac{kT}{\hbar}\right)^2 \left(1 - \frac{\text{tr}P}{\text{tr}\mathbb{I}}\right) + \mathcal{O}(\delta t^3)$$

Where the second line comes from assuming U and $\mathcal{L}$ are uncorrelated with ρ and have temperature T. If the strength of the external environment interaction γ is sufficiently weak, $\gamma \lesssim \frac{V}{A}\delta t$, then we get $\delta p = \mathcal{O}(\delta t^2)$, *i.e.*, the Quantum Zeno Effect.

Asymptotic Adiabaticity (slide 2 of 2)

(Work by H. Earley, Cambridge U., doi:10.17863/cam.79111)



Setup

- A is the (convex bounding) surface area
- V is the volume
- P here is input power
- $\dot{S}$, $\dot{Q}$ are rates of generation & dissipation of entropy and heat, respectively

Determine dissipation requirements via entropy increase due to error; let s be local entropy, S total entropy:

$$\delta s = s(t + \delta t) - s(t) \gtrsim \delta p \cdot n(\dots)$$
$$\dot{s} \gtrsim n^2 \delta t(\dots)$$

Let $R_Z = \frac{N}{n}\frac{1}{\delta t}$ be the total measurement/reset ('Zeno') rate, and let $R_C \approx N\varepsilon/h$ (Margolus and Levitin, 1998) be the total computation rate where ε is the energy associated with one bit.

Then $\dot{S} \gtrsim \frac{R_C^2}{R_Z}$ and so $R_C \lesssim \sqrt{\dot{S} R_Z}$ (some constant factors suppressed)

This corresponds to the adiabatic regime, and gives scaling $R_C \lesssim V^{5/6}$ (compare $R_C \lesssim V^{4/6}$ for irreversible computing).

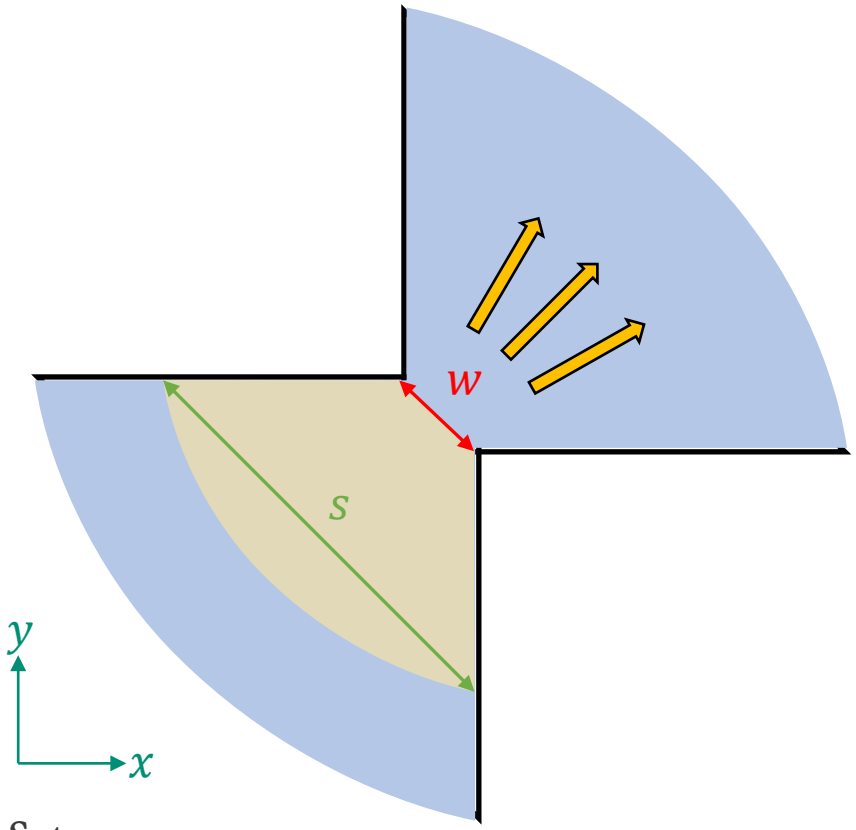
This proof made a number of simplifying assumptions; it can be substantially generalised (Earley, arXiv:2007.03605; paper in preparation)

Can also prove adiabaticity for slow-reset regime, using Fermi transition rates

- Consequences of this regime are that local computation rate scales like $r_C \sim V^{-1/6}$, i.e. local computation asymptotically slow but total computation asymptotically fast
- Compare QZE case where can pick size of subunits to maintain asymptotically fast local computation rate too ($r_C \sim n \cdot V^{-1/6}$)
- QZE allows greater control over quantum state, possibility of QC; Fermi regime implies significantly more decoherence…

Synchronization in asynchronous reversible computers is expensive

(Work by H. Earley, Cambridge U., doi:10.17863/cam.79111)



Setup

- 2D phase geometry corresponds to two asynchronous subunits X and Y evolving independently along their trajectories
- constriction in phase space corresponds to a constraint on their joint config., and a synchronisation/communication event
- w is the width of the constriction
- s is the ~width of the ~steady state distribution

Parallel computers need to intermittently synchronize their state in order to communicate, aggregate & distribute information, etc

- Asynchronous subunits evolve independently except during synchronization events
- Independent evolution leads to dispersion in phase space (*e.g.*, due to Brownian motion, Markovian dynamics, etc; more fundamentally, Heisenberg uncertainty)

Parallel computer state can be represented in phase space

- For a deterministic reversible subunit, each subunit's progress along its configuration trajectory corresponds to a single linear dimension
- Synchronous subunits have a correlated distribution, so synchronization can be performed for free (assuming programs carefully track joint state)
- Asynchronous subunit distribution decorrelates over time; synchronization re-correlates the distribution

Key result: synchronization in asynchronous RC corresponds to an effective **erasure** of information in the joint distribution

- In the limit of low free energy density, corresponding to large system sizes, dispersion is faster than net computational progress; this leads to a steady state approximation at the constriction
- Bounded-above free energy supply means there is a correspondence between time and Landauer-like entropy erasure
- Leak rate $\sim wb^2$ where w is the constriction width and $b \ll 1$ is the 'computational bias', corresponding to the ratio of net computational rate to maximum computational rate
- $s \sim \frac{1}{b}$ is the approximate width of the steady state phase distribution; information erased $\sim \log\left(\frac{s}{w}\right)$
- actual analysis more complicated (see Earley, arXiv:2011.04054)

Consequence: a moderate level of concurrency introduces Landauer-like costs to asynchronous RCs, *reducing* RC advantage

- constant factor advantage still possible, but asymptotic scaling advantage from previous slide is lost unless frequency of 'concurrency-operations' is asymptotically vanishing

Can dissipation scale better than linearly with speed?

Some observations from Pidaparthi & 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

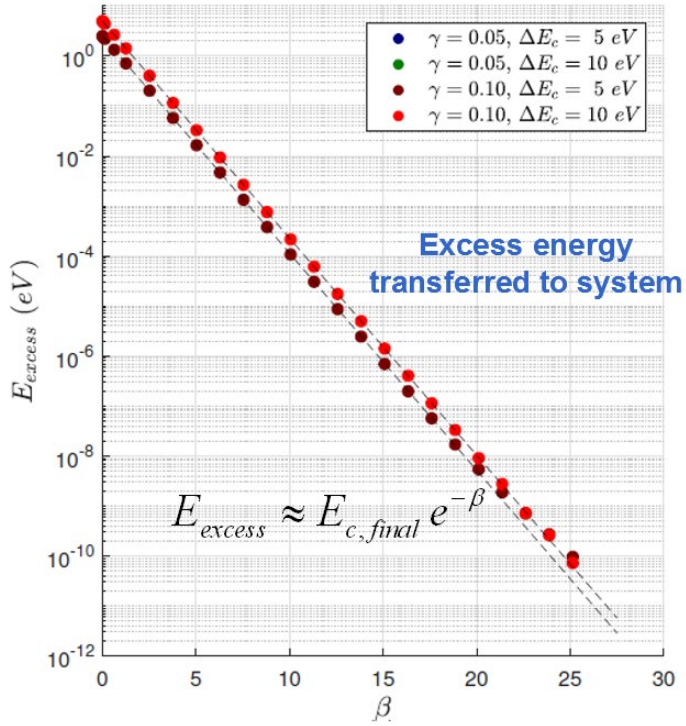
Subhash S. Pidaparthi and **Craig S. Lent** *

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)



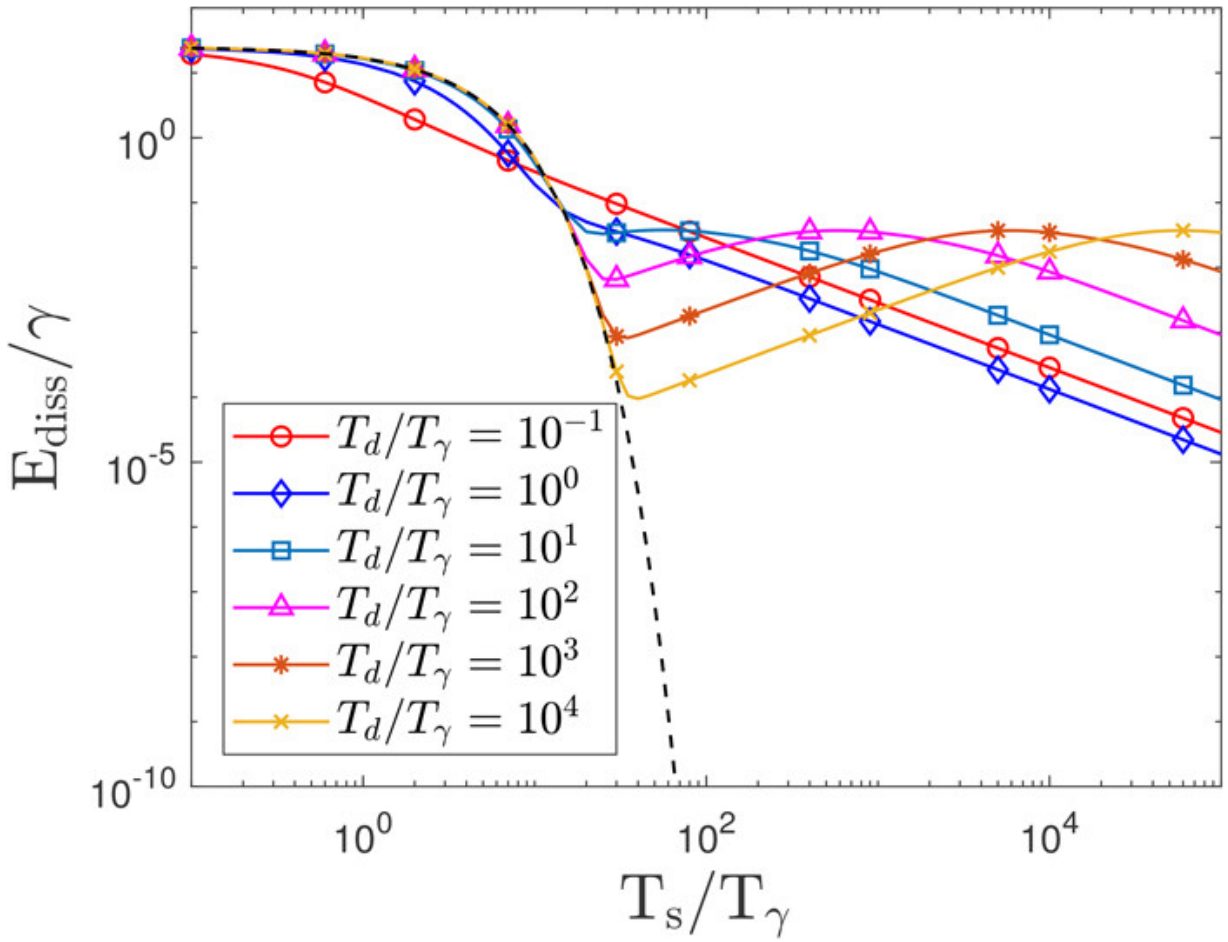


FIG. 10. Dissipated energy of an open system as a function of switching speed for different dissipation time constants. The dashed line is the excess energy of an isolated system. Here, the environmental temperature $k_BT / \gamma = 0.5$.

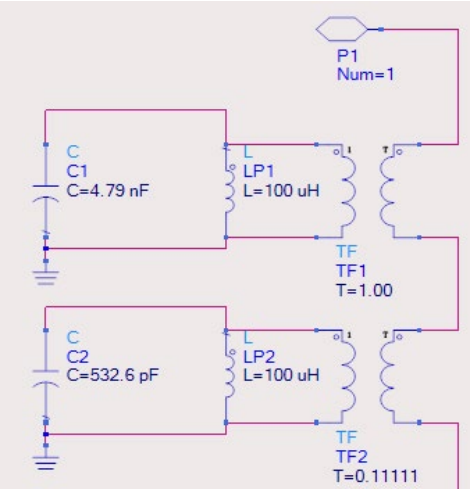
Published in: Subhash S. Pidaparthi; Craig S. Lent; *Journal of Applied Physics* **129,** 024304 (2021)
DOI: 10.1063/5.0033633

Resonator design effort, in progress…

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

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:

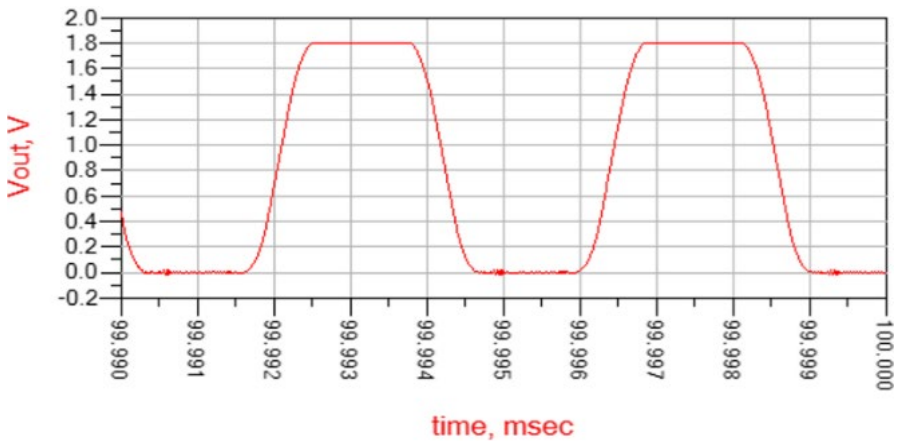
- Initial target operating point: 230 kHz, 1.8V (optimal point for minimum dissipation in the UF study) **(Has been met.)**
  - However, our circuit technique should be adaptable over a wide range of frequencies and voltages.
- Tops and bottoms of trapezoidal wave should be within ≤5% of flatness throughout ¼ clock period. **(Met.)**
- The 10-90% rise/fall time should be between 75 & 100% of its nominal value (80% of 1/4 clock period) **(Met.)**
- Efficiency goals:
  - Quality factor of resonator during unpowered ring-down should be ≥1,000. (**Met.** Simulated value: ~3,000.)
  - Total energy dissipation per cycle during steady-state powered operation should be ≤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 ≤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 patent application has been filed on our resonator design.

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



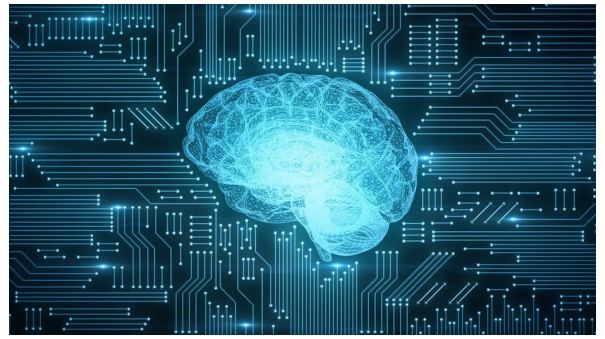
4. Synergies Between Reversible Computing and AI/ML

The Reversible Computing Scaling Path:
Challenges and Opportunities

Ways in Which AI/ML Technology can Help Reversible Computing

Artificial intelligence (AI) and machine learning (ML) technologies can potentially assist in the advancement of reversible computing in the following areas (at least!):

- ***Optimizing*** the design of reversible devices/circuits/architectures to maximize system-level figures of merit (*e.g.*, cost efficiency)
  - Ideally, a co-design optimization should be done spanning multiple layers of the design stack
- ***Discovering*** creative, innovative new designs / design principles for reversible devices, circuits, architectures, and algorithms.
  - And maybe eventually also helping to make useful advances in answering fundamental physics questions?
- More generally, ***accelerating*** design-automation workflows by various methods
  - Existing industry EDA tools for conventional CMOS design already offer a variety of AI-enhanced features
- ***Educating*** the engineering workforce (from students to industry leaders/veterans) to enable them to better understand, work with, design and use reversible computing technologies

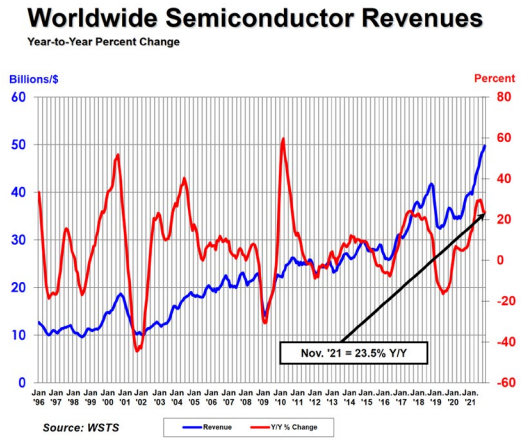


Here's a generic AI/ML graphic that only a marketing person could love

The "reversible computing revolution" would likely cost at least **hundreds of millions of dollars** over **decades** to build out if done *solely* in the traditional, labor-intensive way.

- Consider that the existing semiconductor industry spent *hundreds of billions* of $ in recent decades

But, perhaps making RC somewhat commonplace might cost only *tens* of millions & take only about a decade if it could leverage a sufficiently intensive level of assistance from dedicated AI/ML technologists? → May just be wishful thinking, but…



(Area under blue curve = ~$600B)

Ways in Which Reversible Computing can Help AI/ML

Potentially, training and/or inference for machine-learning models *might* be a relatively early application space for reversible computing. Some reasons why it could be a promising area:

- Aggregate energy consumed today by training/operation of large AI models is already substantial.
  - For reasons of sustainability, we'd like to be able to improve the size and capability of AI models w/o increasing resource usage
- Training/inference for common ML models (*e.g.*, artificial neural nets) tends to require relatively simple computational kernels (*e.g.*, matrix processing)
  - Designing a reversible accelerator/coprocessor for such simple kernels could be relatively straightforward (in terms of design labor) compared to hardware for more complex/general-purpose applications. (*E.g.*, an Intel architecture CISC-style CPU is much harder.)
- AI/ML systems are a huge & fast-growing market (est. $8.1B in 2020, forecasted to grow by 34% annually)
  - Increases potential that a large player with deep pockets might be willing to bankroll the required R&D

In addition, in increasing the future growth potential for AI/ML (and digital computing in general), the advent of reversible technologies could help attract increased levels of investment into tech.

- In contrast, without RC, the growth of the tech industry could begin to slow as we run up against the limits of conventional technology

5. Next Steps & Conclusion

The Reversible Computing Future

Some Important Next Steps for Progress in the Reversible Computing Field

Much work is still needed in the following areas:

- More outreach to better inform decision-makers/investors about the potential of this area
  - Talks, books, video lectures…
- Further development of existing CMOS-based technology platforms for RC in the relatively near term
  - Continue development of high-Q trapezoidal resonators, optimize packaging & integration.
  - Re-engineer FET device structures to (more) aggressively minimize leakage.
  - Improve cost-efficiency of densely-packed 3D fabrication processes w. multiple layers of active logic.
- To develop digital circuits & systems of substantial complexity based on RC, we need:
  - Extensions to EDA tools are needed to support reversible circuits & architectures.
  - New RC-based hardware designs (hardware algorithms for functional units, IP blocks, processor designs).
  - (Eventually) reversible programming models/languages & software algorithms.
    - There is substantial work in this area already.
- Longer-term work to improve the energy-delay product of RC implementations (across various temps.):
  - Need to identify practical new classes of RC devices leveraging novel/exotic (quantum-mechanics-based) operating principles.
  - Need to better characterize the fundamental limits of efficiency of RC as a function of various physical timescales of interest.
    - *E.g.*, equilibration, relaxation, fluctuation, decoherence, and switching/interaction timescales are (potentially) all important
  - Further clarification of the very-long-term asymptotic limits of RC scaling.

Conclusion

There will never *not* be a pressing demand for ever more-efficient general digital computing!

- This will remain true *despite* the emergence of a variety of non-digital computing models (*e.g.*, analog, dynamical-systems based, stochastic, quantum) for various specialized applications.

The conventional (non-reversible) paradigm for digital computing is approaching its end-of-life.

- Soon it will no longer be possible to improve its efficiency due to fundamental thermodynamic limits.

Reversible computing offers the *only* physically possible route to continue improving the efficiency of digital computing beyond the limits of the non-reversible paradigm.

- And further, we know of no fundamental limits to the energy-efficiency (and cost-efficiency!) of RC.

Various groups have already demonstrated clear, compelling proofs-of-concept for the implementation of RC in both semiconducting and superconducting technology platforms.

- At this point, there really is nothing fundamental that prevents the further development of RC technology towards eventual commercialization.

Of course, much work remains to be done if we wish to continue improving the efficiency and scale of RC, but no *fundamental* barriers to further ongoing improvement are apparent.

- ∴ RC is a nascent new subfield of ECE that is now quite ripe for significant further development.

Really, the only thing needed at this point is simply **massive levels of new R&D funding** (from government, industry, &/or far-sighted investors).

- IMO, we really need dedicated funding to ramp up to a level of (at least) $10s of M/year in order to make an adequately rapid rate of R&D progress across the entire field if we want to have solutions ready to go by the time the efficiency of non-reversible digital technology totally flatlines…

RC could grow the value of the digital economy by many orders of magnitude.