Exceptional service in the national interest

Sandia National Laboratories

Feasible demonstration of ultra-low-power adiabatic CMOS for cubesat applications using LC ladder resonators

Michael Frank
Sandia National Laboratories

Tenth Workshop on Fault-Tolerant Spaceborne Computing
Employing New Technologies
Albuquerque, NM
June 1, 2017

Approved for Unclassified Unlimited Release
SAND2017-5650 C

Abstract

Small space platforms such as cubesats are typically highly constrained in the power available for on-board computation, limiting the scope of achievable missions. Unfortunately, conventional approaches to low-power computing in CMOS are limited in their energy efficiency, because they still follow the conventional *irreversible* computing paradigm, in which digital signals are destructively overwritten on every clock cycle, dissipating the associated CV^2 signal energy to heat. In an alternative approach called *reversible computing*, which can be implemented in rad-hard CMOS, we can *adiabatically* transform digital signals from old states to new ones with almost no dissipation of signal energy, instead recovering almost all of the signal energy and reusing it in subsequent operations. At relatively low (MHz scale) frequencies, this approach can yield orders-of-magnitude gains in power-limited parallel performance compared to more conventional approaches to low-power CMOS. In this paper, we propose a feasible near-term demonstration of reversible adiabatic CMOS at attojoule-per-operation energy scales, using custom LC ladder resonators integrated in-package with the logic IC to achieve high-quality energy recovery.

Talk Outline

- Motivation
  - More power-efficient computing for small spacecraft (nanosats, etc.)
- Background
  - Practical limits of irreversible CMOS
  - Thermodynamic limits of computing—a short tutorial
  - Reversible computing
    - The only long-term sustainable path forward!
- Reversible computing in adiabatic CMOS
  - Basic principles
  - Early proof-of-concept chips
  - 2LAL (two-level adiabatic logic)
  - LC ladder resonators
- Conclusion
  - Towards a demonstration of ultra-low-power reversible CMOS

3

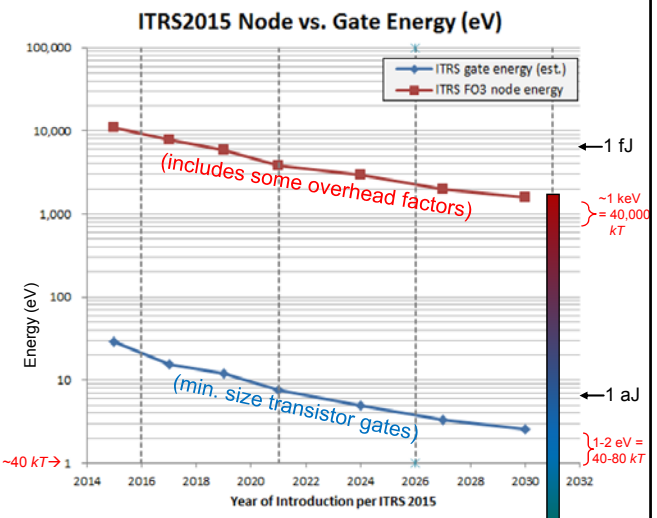
Motivation: Energy Efficiency for Onboard Computing in Spacecraft

- Power efficiency of high-bandwidth downlinks is limited by fundamental communication theory considerations…
  - Majority of downlink power misses the receiver and is wasted
- Thus, it would be desirable to do more processing onboard, if we can find ways to do this within a given power budget…
  - Allows us to save available downlink bandwidth to convey numerous compactly-encoded, higher-level, mission-relevant results extracted from raw sensor data by onboard processing
    - This would then allow us to expand the scope of achievable missions
- Also, even for a fixed mission, if the power requirements for the desired computation can be reduced, this could potentially allow the size of the entire spacecraft to be scaled down…
  - Power supplies, solar panels/radiators, chassis geometry can be scaled
  - Mass of entire spacecraft and fuel requirements can be scaled
  - Total construction and launch costs can be substantially reduced

4

Energy limits for conventional technology are not far away!

- Thermal noise in min.-width FET gates leads to channel fluctuations below ~1-2 eV
  - Increases leakage, impairs device performance
- Note: Real transistors are often sized much wider than minimum width, for speed
  - *E.g.*, ~20× min. width
  - Also there is fanout, wire capacitance, *etc.*
- Note: ITRS is aware of the thermal noise issue, and so has minimum gate energy asymptoting to ~2 eV
  - Node energy *follows*, asymptoting to ~1 keV
- Practical conventional circuit architectures can't just magically cross this gap!
  - ∴ Fundamental thermal limits translate to much *larger* practical limits!



Only reversible computing can take us from the end of the CMOS roadmap all the way down to *kT* and below!

4

Entropy in a Nutshell

Basic review + coining some useful terminology

- Define the "surprisingness" or *surprise* $s(x)$ of any event x that has a 1 in m chance of occurring as $s = s(x) = s(m) = \log m$.
  - Call the $m \geq 1$ "improbability;" it can be a non-integer.
    - s is log because the improbabilities of independent surprises multiply.
  - *Indefinite* logarithm; dimensioned in *arbitrary* logarithmic units.
    - Some example units: $\log 2 = 1$ bit; $\log \mathrm{e} = 1$ nat $= k_\mathrm{B}$; $\log 10 = 1$ bel.
- In terms of event's *probability* $p = p(x) = p(m) = 1/m$,

$$s(p) = \log\frac{1}{p} = -\log p\,.$$

- Define event's "*heaviness*" $h = h(x) = h(p)$ (Hopefulness? Horribleness?) as its surprise, weighted by its probability:

$$h(p) = s/m = p \cdot s = p \log m = -p \log p.$$

Improbability $m = 6^2 = 36$

Surprise $s = 2(\log 6)$

Heaviness $h = \frac{s}{m} = \frac{2}{36}\log 6$

- Then for any probability distribution $p(x)$ over any mutually exclusive and exhaustive set of events $X = \{x_1, \ldots, x_n\}$, we have that the ***expected surprise*** $S(X) = \mathrm{E}_p[s(x)]$ and the ***total heaviness*** $H(X) = \sum_{x \in X} h(x)$ associated with that particular set of possible events are the same, and are given by:

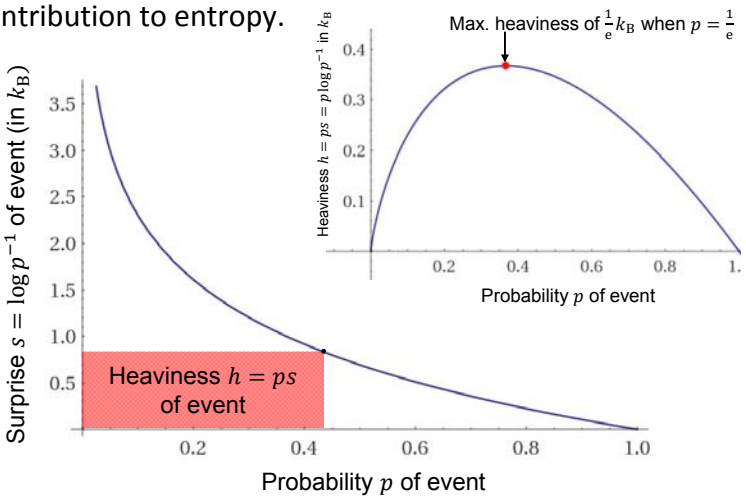
$$S(X) = \sum_{x \in X} p(x) \cdot s(x) = H(X) = -\sum_{x \in X} p(x) \cdot \log p(x)\,.$$

- We call this quantity $H = S$ the *entropy* of the given epistemological situation.
  - By convention, we'll prefer H for "computational" entropy, S for "physical" entropy.

6

Surprise and Heaviness Functions

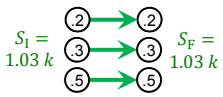
- For an individual state's contribution to entropy.



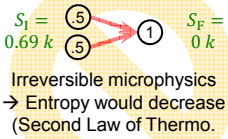
Thermodynamics and Information

- *Physical* entropy quantifies *uncertainty about the detailed microstate of a physical system*.
  - First postulated by Boltzmann (in his H-theorem)
  - Integral to modern physics (Von Neumann entropy)
  - Depends on modeler's state of knowledge (Jaynes)
- The *reversibility (injectivity)* of microphysics underlies the Second Law of Thermodynamics.
  - States cannot merge as they evolve…
  - Thus, entropy of a closed system cannot decrease!
    - Conserved by unitary quantum time-evolution.
  - Entropy can *increase* if we have any uncertainty about the dynamics, or do not track it in detail
- At the most fundamental level, physical information *cannot be destroyed*.
  - Only *reversibly* transformed, and/or transferred between different subsystems…

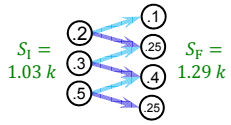
$S[p] = \mathrm{E}_p[\log p^{-1}]$



Bijective microphysics → No "true" entropy change



Irreversible microphysics → Entropy would decrease (Second Law of Thermo. would be violated)



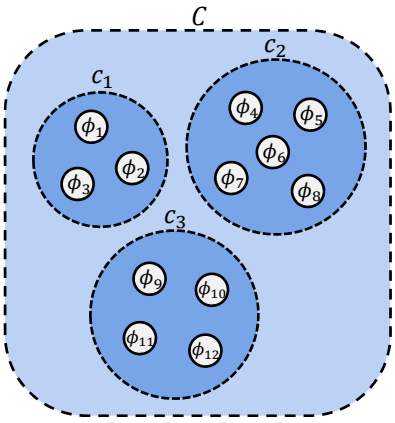
True dynamics uncertain (or not tracked in detail) → Entropy increases

From Physics to Computation

- Thermodynamics and quantum mechanics show that any bounded physical system admits only a finite set $\Phi = \{\phi_1, \ldots, \phi_n\}$ of measurably distinguishable detailed physical states (*microstates*).
  - *E.g.*, Φ could be any orthogonal basis of the system's Hilbert space.
- We can *group* or partition these microstates into subsets c_j of microstates that we consider equivalent to each other for some designated purpose…
  - *e.g.*, representing some specific computational information
- Any probability distribution $p(\phi_i)$ over the physical state space Φ induces a probability distribution over the computational state space (subsystem) $C = \{c_j\}$ as well…

$$P(c_j) = \sum_{\phi_i \in c_j} p(\phi_i) .$$

- This implies a *computational entropy* $H(C)$.

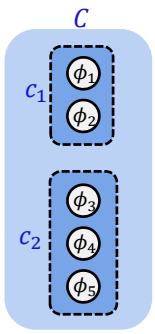


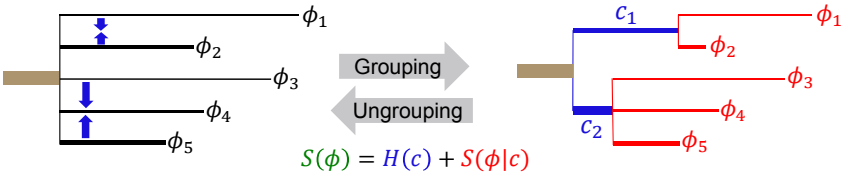
Example of a computational state space C consisting of 3 distinct computational states c_1, c_2, c_3, each defined as a set of equivalent physical states.

9

Visualizing Entropy of Grouped States

- Can represent a hierarchy of events in a tree structure…
  - Branch *thickness* = event probability p.
  - Branch *length* = *incremental surprise* Δs associated w. event,
    - relative to whatever base event it's branching off from.
  - Branch *area* = event's *incremental heaviness* $\Delta h = p\Delta s$
    - contribution to total entropy, in addition to base event.
- *Grouping* events into larger events has these effects:
  - Thicknesses (probs.) of branches combine in parent branch
  - An corresponding part of total length (surprise) of each branch is reassociated to parent (stem) branch.
  - Note: The total heaviness H of all branches and stems (total entropy S) is not affected at all by any grouping/ungrouping!!



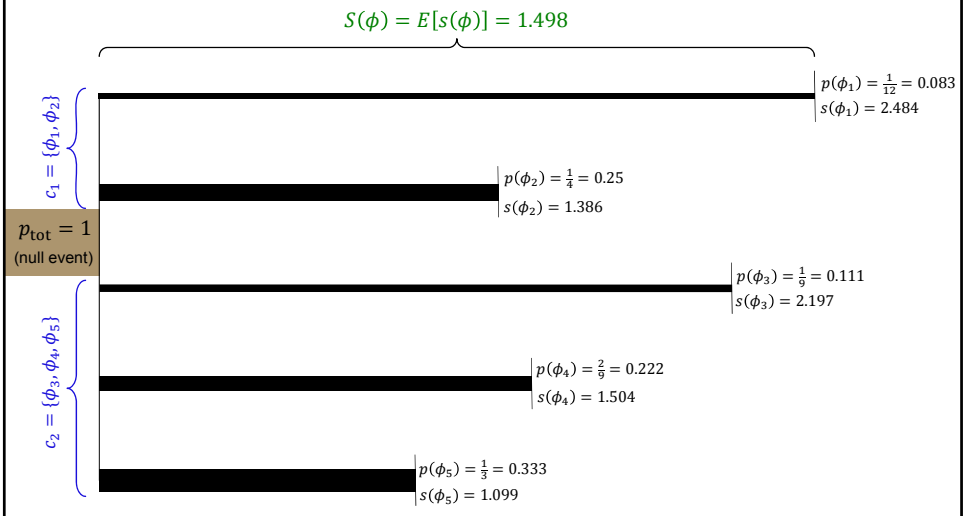


$$S(\phi) = H(c) + S(\phi|c)$$

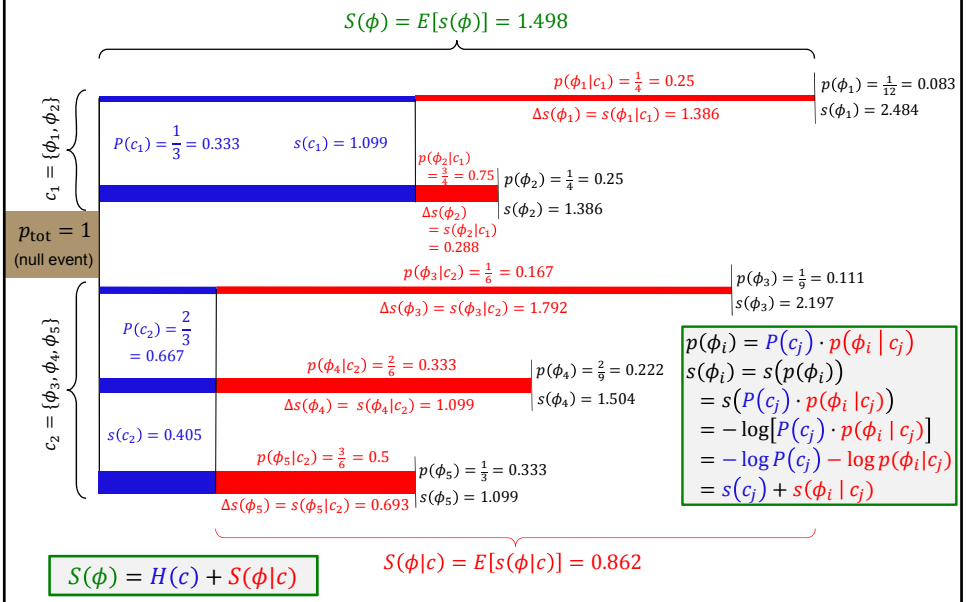
Total system entropy = computational entropy + non-computational entropy

10

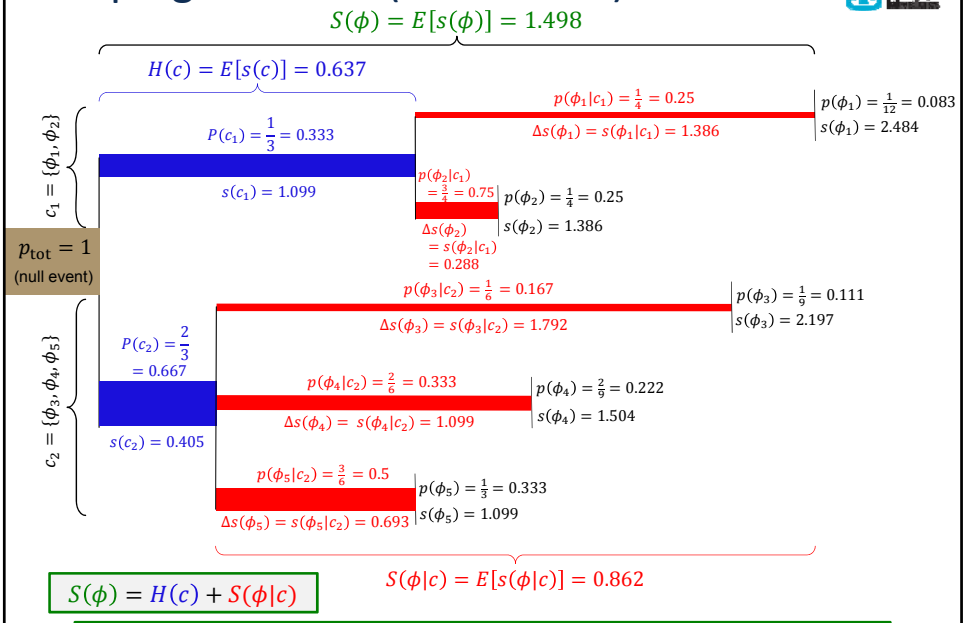
Grouping of States (slide 1 of 3)



Grouping of States (slide 2 of 3)



Grouping of States (slide 3 of 3)



$S(\phi) = H(c) + S(\phi|c)$

Total system entropy = computational entropy + non-computational entropy

13

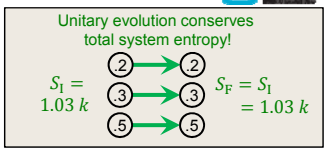
Proof of Landauer's Limit

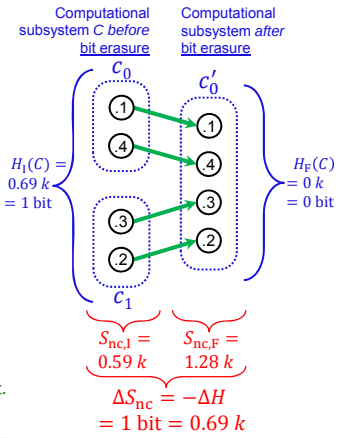
- We've seen that the total system entropy $S(\phi)$ for a given closed system cannot decrease at all…
  - So, what happens if we merge two *computational* states?
- Underlying probability distributions remain the same!
  - Only the identities of the physical states ϕ, and their groupings into computational states, can be changing
- Merging two computational states implies, removing a conceptual partition between groups of physical states
  - Same as the "ungrouping" operation we saw earlier
- The computational contribution $H(C)$ to the total entropy $S(\phi)$ *cannot* simply vanish from existence…
  - Thus, it can only be *ejected* from the **computational** state into the ***non*-computational** state
- We define *non-computational entropy* as:

  $$S_{\text{nc}}(\phi) = S(\phi|C) = S(\phi) - H(C).$$

  - So, the *change* in $S_{\text{nc}}(\phi)$ from a merge operation is thus:

    $$\Delta S_{\text{nc}}(\phi) = \Delta S(\phi|C) = -\Delta H(C).$$
- To extent that "non-computational" = "uncontrolled,"
  - the extra non-computational entropy must end up in some thermal environment at some temperature T
    - We must thus emit at least heat $\Delta Q = T\Delta S$ to that environment. If $\Delta S = 1$ b, then $\Delta Q = kT \ln 2$.





∴ **Landauer Limit**: $E_{\text{diss}} \geq kT \ln 2$ per bit lost. ■

14

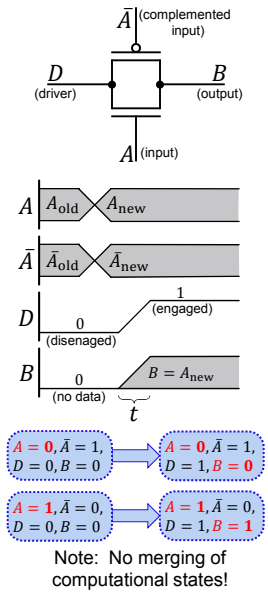
Why Reversible Computing?

- Landauer's Limit is ***absolutely unavoidable*** in any computing scheme based on constantly losing computational information
  - *e.g.*, by erasing it, or (equivalently) destructively overwriting it
- Note: Conventional computers lose information all the time!
  - *Every* active logic gate in a conventional design destructively overwrites its previous output on *every* clock cycle (*e.g.*, billions of times per second)
- Even worse, in practice, erasing a bit dissipates not *just* $kT \ln 2$, but the *entire* logic signal energy associated with that bit!
  - This is still $> 10{,}000\ kT$ even at the very end of the CMOS roadmap!
    - Unlikely to decrease much, given thermal noise and architectural overheads
- The *only* sustainable path forward would be if we increasingly *recover* the signal energies used to encode old bits, and *reuse* almost all of that energy to register newly-computed bits…
  - But, due to Landauer's principle, approaching complete energy recovery *requires us to avoid merging of computational states* (as on prev. slide)
    - Since that would lose information and its associated signal energy!

15

Reversible Computing in Adiabatic CMOS Circuits

- An approach researched since the mid-1980s…
  - MF invented a new scheme in early 2000s (2LAL)
- Here's a simple example of a *reversible copy* operation using a CMOS transmission gate →
  - Semantics: Copy $A \Rightarrow B$, given $B = 0$ initially.
    - Reversible if precondition $B = 0$ is satisfied
  - Boolean AND/OR simply use series/parallel T-gates
- The driving signal D is ramped *gradually* from logic level $0 \rightarrow 1$ over some transition time t…
  - Energy dissipated is CV^2RC/t (to first order)…
    - C = output node capacitance
    - V = logic swing voltage
    - R = resistance of charging path
  - Dissipation approaches 0 in the adiabatic limit…
    - Low speed *and* low leakage through transistors
- This approach could even get below the Landauer limit, given sufficiently low-leakage transistors…
  - Has been empirically demonstrated with resistors

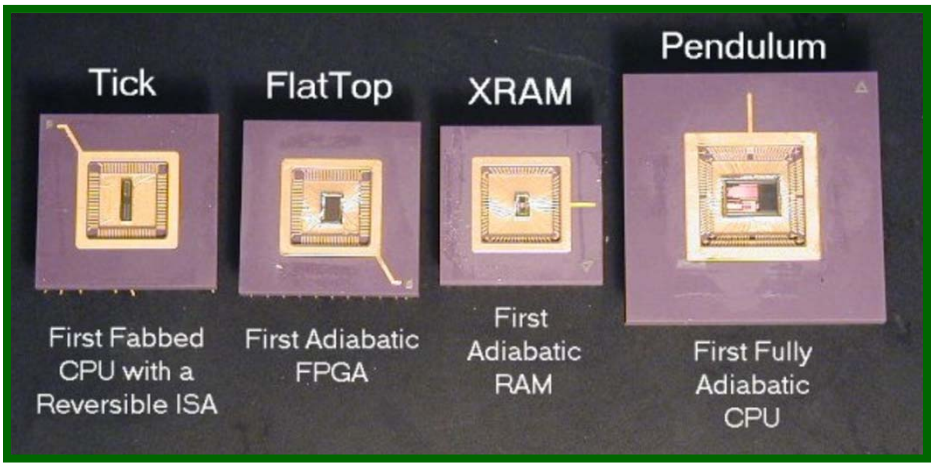


Note: No merging of computational states!

16

Reversible and/or Adiabatic Full-Custom VLSI Chips Designed @ MIT, 1996-1999

By Josie Ammer, Mike Frank, Nicole Love, Scott Rixner, and Carlin Vieri under CS/AI lab members Tom Knight and Norm Margolus.



Circuit Rules for Truly/Fully Adiabatic FET-based Switching

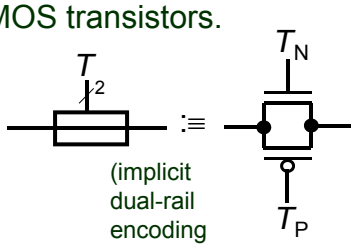
- Avoid passing current through diodes!
  - Crossing the "diode drop" leads to an irreducible dissipation.
- Follow a "dry switching" discipline (in the relay lingo):
  - Never turn on a transistor when $V_{DS} \neq 0$. "No sparks!"
  - Never turn off a transistor when $I_{DS} \neq 0$. "No squelches!"
    - Only exception: If an alternate path for current is available.
- Together these rules imply:
  - The computational function of the circuit must be logically reversible
    - There is no way to erase digital information under these rules!
  - Transitions must be driven by a quasi-trapezoidal waveform
    - It must be generated resonantly, with high Q
- Of course, leakage power must also be kept manageable.
  - Because of this, the optimal design point will not necessarily use the smallest devices that can ever be manufactured!
    - With adiabatics, we can actually achieve lower total dissipation per op (including leakage) and higher aggregate performance (at fixed power) if we back off to using somewhat larger, slower, older-generation devices!

An important rule, that is neglected in almost all of the "adiabatic" circuit literature!

2LAL: 2-level Adiabatic Logic

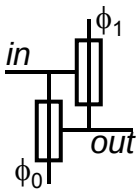
A pipelined fully-adiabatic logic family invented by MF at UF in 2000, implementable using ordinary CMOS transistors.

- Uses transmission gates, symbolized as:

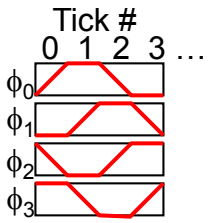


(implicit dual-rail encoding everywhere)

- Basic buffer element:
  - cross-coupled T-gates:
    - needs 8 transistors to buffer 1 dual-rail signal by 1 transition time (tick)

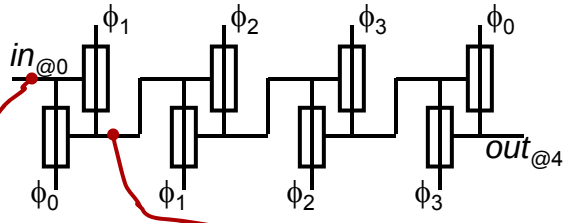


- Only 4 timing signals $\phi_{0\text{-}3}$ are needed. Only 4 ticks per cycle:
  - ϕ_i rises during ticks $t \equiv i \pmod 4$
  - ϕ_i falls during ticks $t \equiv i + 2 \pmod 4$

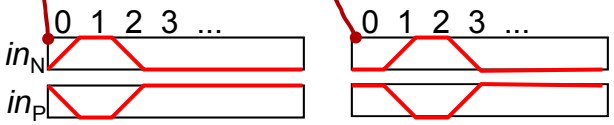


2LAL Shift Register Structure

- 1-tick delay per logic stage:

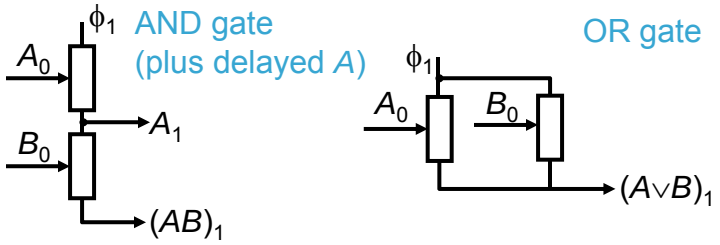


- Logic pulse timing and signal propagation:

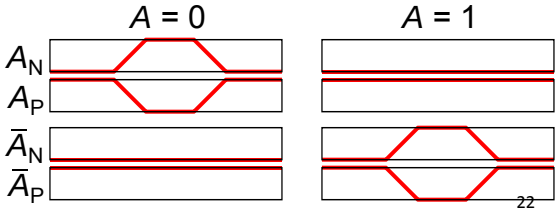


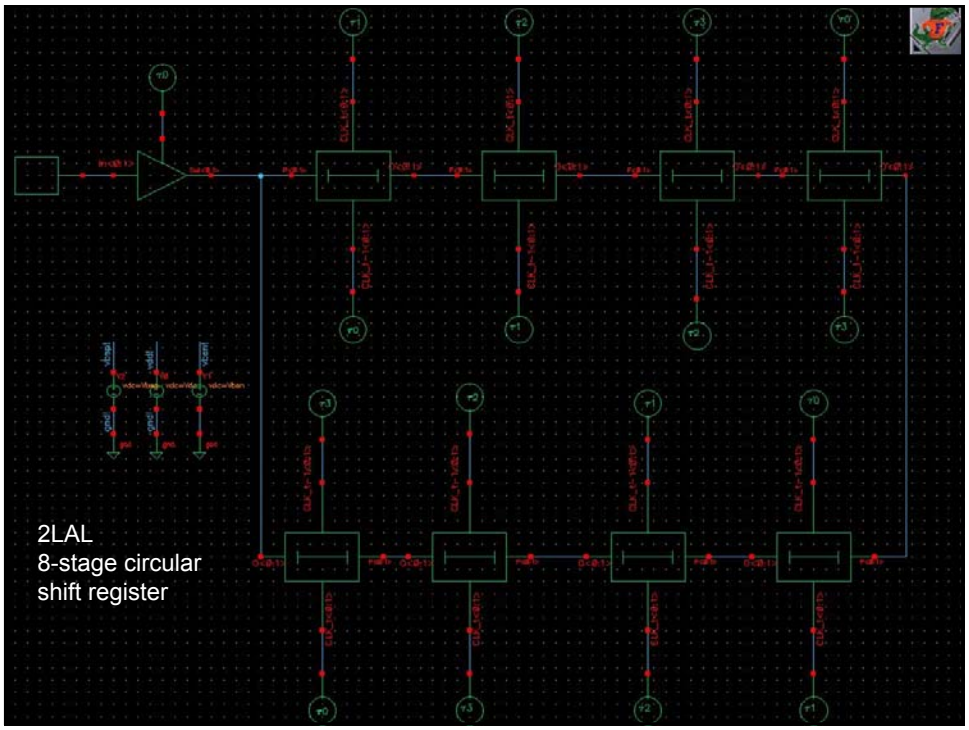
More Complex Logic Functions

- Non-inverting multi-input Boolean functions:

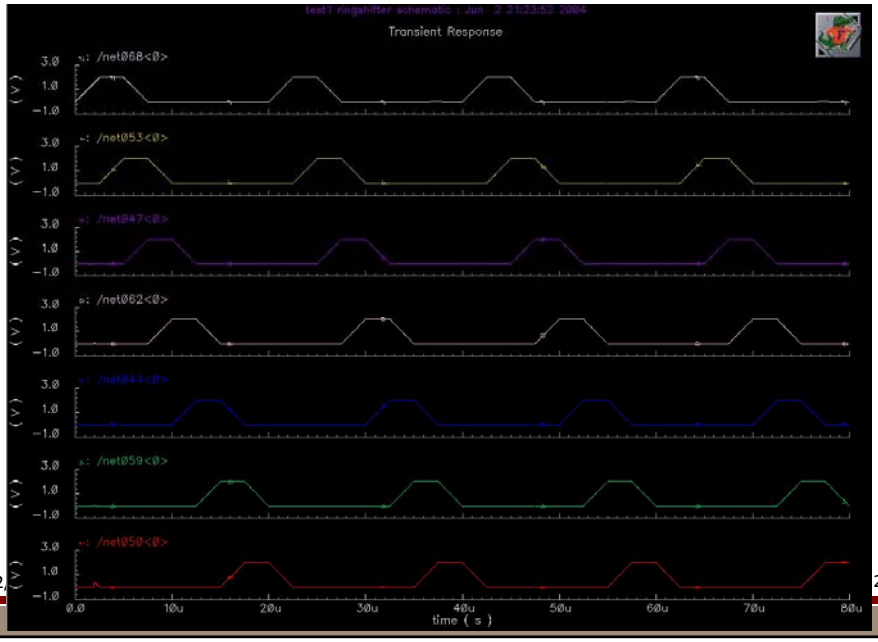


  - Can also complement inputs/outputs and use DeMorgan substitution
- One way to do inverting functions in pipelined logic is to use a quad-rail logic encoding:
  - To invert, just swap the rails!
    - Zero-transistor "inverters."

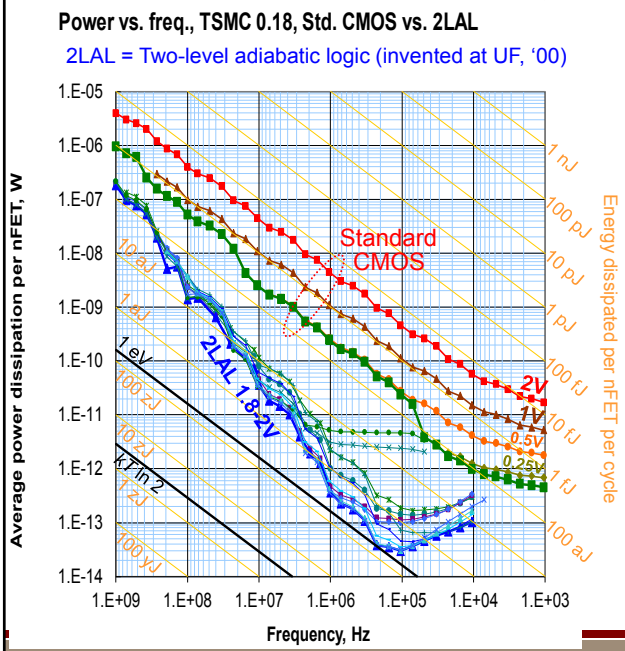




Pulse propagation in 8-stage circuit



Simulation Results (Cadence/Spectre)



- Graph shows per-FET power dissipation vs. frequency
  - in an 8-stage shift register.
- At moderate freqs. (1 MHz),
  - Reversible uses < 1/100th the power of irreversible!
- At ultra-low power levels (1 pW/transistor)
  - Reversible is 100× faster than irreversible!
- Minimum energy dissipation per nFET is **< 1 electron volt**!
  - 500× *lower* dissipation than best irreversible CMOS!
    - 500× *higher* computational energy efficiency!
- Energy transferred per nFET per cycle is still on the order of 10 fJ (100 keV)
  - So, energy recovery efficiency is on the order of 99.999%!
    - Quality factor $Q = 100{,}000$!
      - Note this does not include any of the parasitic losses associated with power supply and clock distribution yet, though

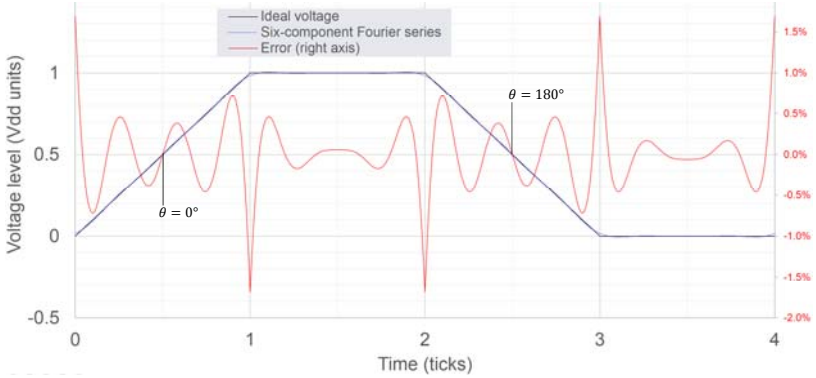
How to generate clock signals?

- To achieve a large energy savings, they must be generated resonantly, with a high Q factor.
  - Parasitic losses in clock distribution network must be minimal.
- The waveforms need to have this very nonstandard shape…
  - Not sinusoidal or square-wave, but *trapezoidal*.
    - *Gradual* rise/fall ramps, and flat horizontal wave tops/bottoms.
      - Ramps do not have to be perfectly linear, but slope should be limited.
- A few of the supply techniques that have been considered:
  - Clipped sinusoidal (crystal or LC) oscillators
  - Transmission-line resonators
  - Custom MEMS resonators (various geometries)
- Each of these have issues, and are not close to practical yet
  - Here, we propose an easier approach: LC ladder networks.

Spectrum of Trapezoidal Wave

- Relative to mid-level crossing, waveform is an odd function
  - Spectrum includes only odd harmonics $f, 3f, 5f, \ldots$
- Six-component Fourier series expansion is shown below
  - Maximum offset with $11f$ frequency cutoff is $< 1.7\%$ of V_{dd}

$$v_{f6}(t) = V_{dd}\left[\frac{1}{2} + \frac{4\sqrt{2}}{\pi^2}\left(\sin\theta + \frac{\sin 3\theta}{3^2} - \frac{\sin 5\theta}{5^2} - \frac{\sin 7\theta}{7^2} + \frac{\sin 9\theta}{9^2} + \frac{\sin 11\theta}{11^2}\right)\right]$$



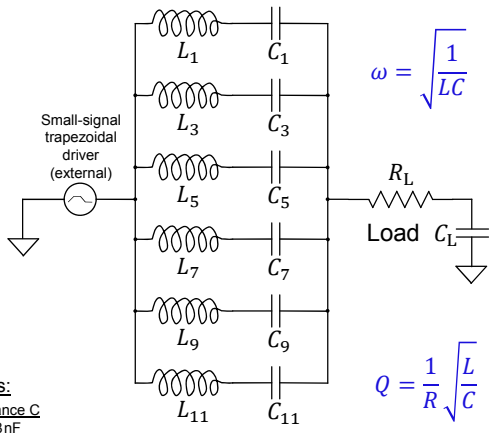
Ladder Resonator Structure

- Can build trapezoidal resonator w. a ladder circuit made of parallel passive bandpass filters, each a sinusoidal LC resonator
  - Each "rung" of ladder passes a different odd multiple of the fundamental clock frequency f
  - Adjust L/C ratio to obtain a target Q value on each path, given parasitic R,C values
- Excite the circuit with a driving signal containing the right distribution of frequency component amplitudes
  - Each frequency component gets amplified by the Q value of its corresponding rung
  - If all rungs are designed to the same target Q, we can just use a trapezoidal driver
- For high Q, clock period must be long compared to the total parasitic RC…
  - Max. possible $Q_n = \frac{1}{2\pi} \cdot \frac{t_{\text{period},n}}{(RC)_{\text{parasitic}}}$

(for $V_{dd} \approx 1.75$ V ↓)

Example values:

harmonic mode (n)	frequency f	component amplitude Va	inductance L	capacitance C
1	230 kHz	1000.00 mV	691.98 nH	691.98 nF
3	690 kHz	111.11 mV	230.66 nH	230.66 nF
5	1150 kHz	-40.00 mV	138.40 nH	138.40 nF
7	1610 kHz	-20.41 mV	98.85 nH	98.85 nF
9	2070 kHz	12.35 mV	76.89 nH	76.89 nF
11	2530 kHz	8.26 mV	62.91 nH	62.91 nF



Ladder Resonator for Odd Harmonics

$\omega = \sqrt{\frac{1}{LC}}$

$Q = \frac{1}{R}\sqrt{\frac{L}{C}}$

28

Design Plan for Demonstration Part

- Select a CMOS fabrication process…
  - Older-generation processes are good, b/c low leakage and rad-hard
- Design a pipelined 2LAL circuit to implement the desired function.
  - To the level of layout and parasitic extraction in the selected process…
    - Minimize the parasitic resistance and capacitance of clock dist. network.
- Identify a target clock frequency that is low enough to obtain the desired energy reuse factor (Q value)
  - This determines the maximum power-limited performance boost that can be achieved compared to conventional irreversible CMOS
- Select a packaging methodology that allows discrete components to be placed as close to the die as possible
  - Ideal: Direct bonding of component leads to pads on chip surface
    - Again, minimize the parasitic resistance/capacitance of joins
- Identify specific COTS inductor and capacitor components for ladder network that maximize the overall Q obtained…
  - Goal: Demonstrate Q values of 10-100×.
- Iteratively refine design as needed…

29

Conclusion

- There is a need for greater energy efficiency in spacecraft
  - Could allow the entire vehicle to be scaled down considerably…
  - Or, afford greater mission scope within a given-size platform
- We can actually ***prove*** from fundamental physics that:
  - The *only* long-term sustainable path to attain ever-better energy efficiency in computing is to use ***reversible computing*** principles.
- The CMOS roadmap will soon run out of steam,
  - and beginning to apply reversible computing principles now can offer near-term benefits, that can be further extended in the future.
- Reversible computing in truly/fully adiabatic CMOS is an approach that could be demonstrated in a short time-frame…
  - LC ladder resonators with die-bonded inductors may be adequate to allow demonstrating 1-2 orders of magnitude energy efficiency gains
- Next step: A detailed design and feasibility study showing the viability of such a demonstration would be highly desirable.

30