

# Minimizing classical control resources

QRE2022 June 18, 2022

Erik P. DeBenedictis Zettaflops, LLC

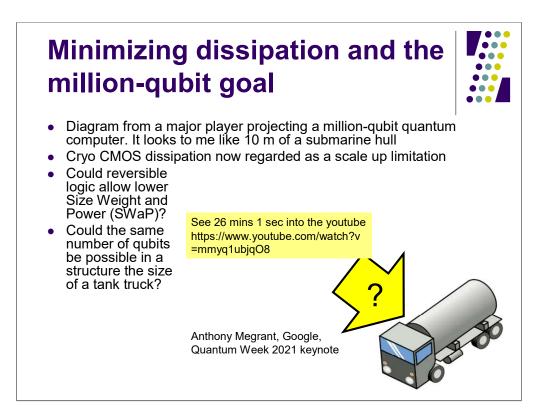


Diagram from Anthony Megrant, Google, Quantum Week 2021 keynote.

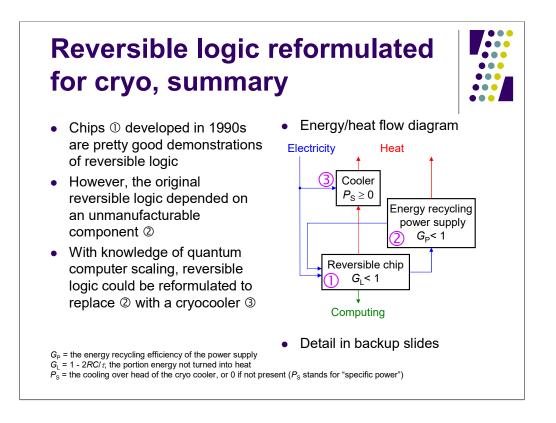
Scaling up quantum computers is a widespread priority. Let us consider a goal of a million qubits.

Currently, each of the million qubits will require a special analog control waveform. If the electronics that creates each waveform is located at room temperature, a million microwave transmission lines will need to cross the cryostat boundary. This would be unwieldy and unscalable.

On the other hand, if the control pulses are generated in the cryostat using cryo CMOS, the large energy of the CMOS will require a large cryostat for heat transfer and a high-capacity cryocooler.

A cryogenic classical logic technology that was more power efficient than cryo CMOS would help.

The diagram includes a cartoon of current plans, which look like 10 m of a submarine, but perhaps a more power efficient cryo technology could reduce system size, weight, and power (SWaP) to the size of a tank truck. Alternatively, the size could stay the same but the structure would have more qubits.



There was a conference in 1981 attended by Feynman, Toffoli, and Fredkin where both quantum and classical reversible computing debuted. This is why we have Toffoli and Fredkin gates in both fields and the CNOT gate is sometimes called the Feynman gate.

Classical reversible logic was studied in the 1990s under DARPA funding (in the USA) and resulted in perfectly satisfactory chips  $\mathbb{O}$ , but a second component of the powertrain, the energy recycling power supply  $\mathbb{Q}$ , was unmanufacturable and caused the field to stall.

It became apparent how quantum computers would scale around 2018. With this information, it was possible to reformulate the reversible logic powertrain so that a cryocooler ③ fills the role unsatisfactorily filled by the energy recycling power supply.

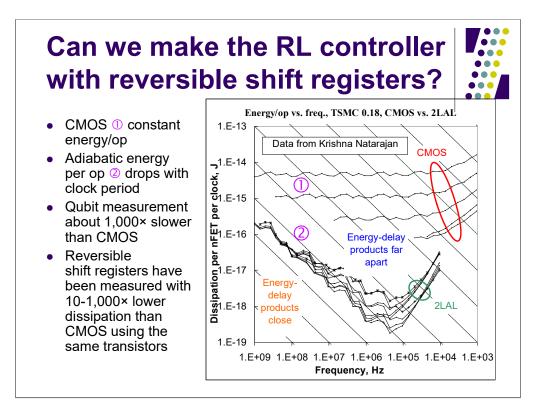
This talk is about the reversible logic as reformulated for cryo.

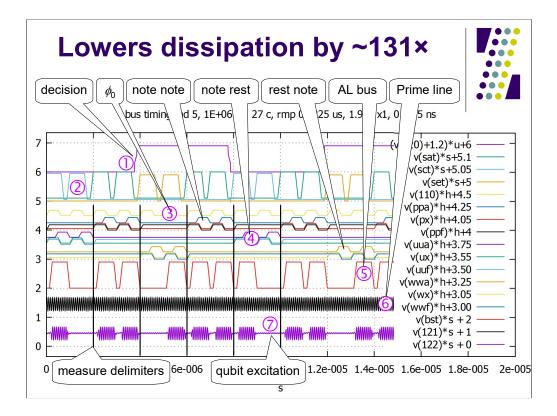
More data in the backup slides.



### Outline

- The speaker devised a reversible logic controller (RL controller) design
  - Schematic + spice simulation
  - Plan to use in real quantum computers, but ...
  - used as an "existence proof" in this talk
- Existence proof of what?
  - No irreversible gates required in cryostat
  - Transistors required in cryostat ∞ source code size (no ∞ stack)
  - Universal architecture in cryostat (.gif/.zip decompressor)
  - Other conclusions in backup material
- Natural metric
  - Irreversible gate-ops + reversible gate-ops + total memory, per quantum operation





The diagram shows a reversible logic controller (RL controller), which is in turn part of the PL/AL architecture. The RL controller becomes an existence proof that lower dissipation is possible. A schematic diagram (see 2<sup>nd</sup> subsequent slide) can be simulated. The simulation shows functional correctness and some information on dissipation.

The overall project included comparing simulated dissipation of the RL controller and a cryo CMOS work alike. This showed a 131× decrease in dissipation.

① Decision line from standard computer at room temperature. When the gate operation sequence reaches a branch point, this signal must be stable at a 0 or 1.

② These are basically state bits that say which of three musical measures are active at a given instant, the three signals are each a different color and overlaid.

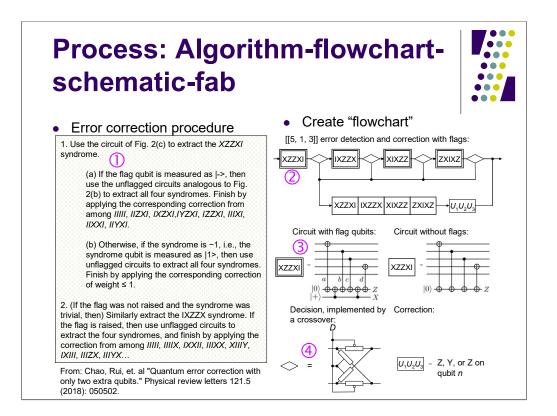
③ Phase 0 clock.

④ Three signals of the data controlled clock for the "note note" music. Note that the lower two lines are the same phase, but stop at a different point. This identifies one of them as the bus enable clock.

© This is the true data wire of the address-line bus, equivalent to the inverse position of the pianist's finger. The bus driver is different for each measure of music, with the signals being note-rest-note-note-rest-note.

<sup>6</sup> Prime line, which is a sine wave for simulation.

⑦ Modulated wave to the qubits.



Background: The Prime-line/Address-line (PL/AL) architecture uses "prime" microwave waveforms created at room temperature for gate operations. These waveforms are distributed to all qubits to be keyed on and off by a control system "pianist." The pianist uses something like music ③ to create the pattern, but a room temperature control computer directs higher level activities like a "playlist" and repeats.

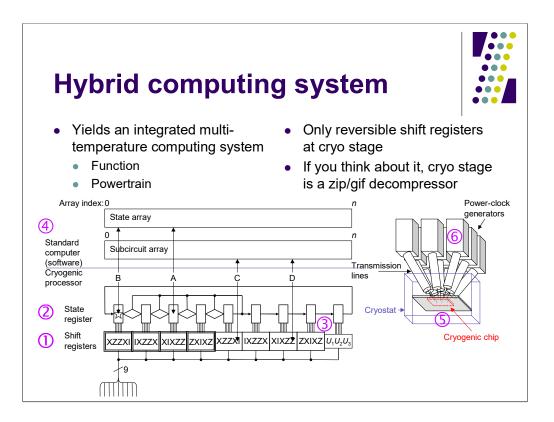
① an example algorithm, specifically quantum error detection and correction for [[5, 1, 3]] code. Task is to apply stimulus to qubits, such as 1 or 2 qubit operations, resets, and measurements.

② the algorithm can be represented by a flowchart, where the boxes represent a quantum gate sequences ③ and the diamonds are decisions ④.

•Flowchart boxes contain a label (XZZXI) to identify a subcircuit sequence. These sequences are like measures in music in that they can only be of a limited length due the need for error correction cycles.

•Decisions are provided by a separate subsystem at room-temperature.

•The correction unit  $U_1U_2U_3$  is a detail not fully described in this slide deck.



The diagram shows a linearized version of the flowchart and its connection to a room temperature computer.

The shift registers with music  $\bigcirc$  are in a row.

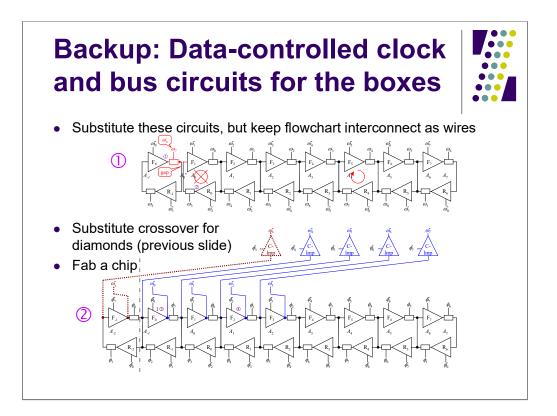
The state registers ② represent the state of the flowchart with a 1 bit, which advances to the right or jumps due to diamonds.

The state machine turns the music on an off via data-controlled clocks ③, one of which is relayed to an "address-line bus" that controls qubits.

A room temperature computer shadows the state of the reversible logic in arrays ④ and uses software to control the diamonds of the flowchart.

On the right:

The reversible computing system also includes a cryo-adiabatic powertrain, comprising a cryogenic chip (6), room temperature power-clock generators (5), and transmission lines into the cryostat.



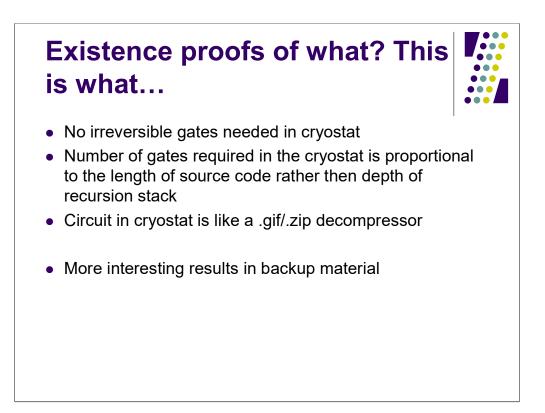
Ngspice simulation code on the Web rigorously defines the circuits.

Circuit ① is a reversible shift register using a static, even-load circuit called Q2LAL. The red annotations allow driving a bus reversibly.

Circuit ② is a data-controlled clock. Adorning the state shift register with the extra circuitry on top creates a clock that runs while the input is a 1. If the input bit is a 0, the clock stops systematically. With the red annotations in ①, some shift register outputs to "tri state."

Each diamond is replaced by a crossover (3 on the second previous slide). An external computer controls the crossover.

Note: There is an optimization that involves a CNOT gate.



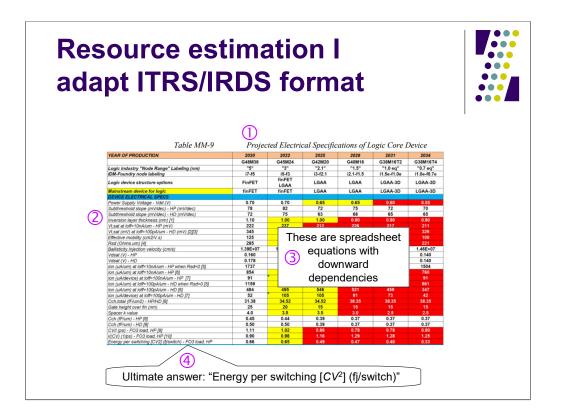
Some additional detail:

First bullet: The design process created the cryogenic part of the control system entirely from reversible gates, so apparently no irreversible gates are required.

Caveats: (a) I have a optimization that uses some CNOT gates. (b) Diamonds are effectively externally controlled Fredkin gates. (c) Irreversible gates may be needed during boot-up.

Second bullet: The design process created the cryogenic part by substituting schematic diagrams for programmatic constructs. This cannot result in a component more than a constant factor larger than the program text (such as an arbitrarily large pushdown stack). (The pushdown stack is at room temperature.)

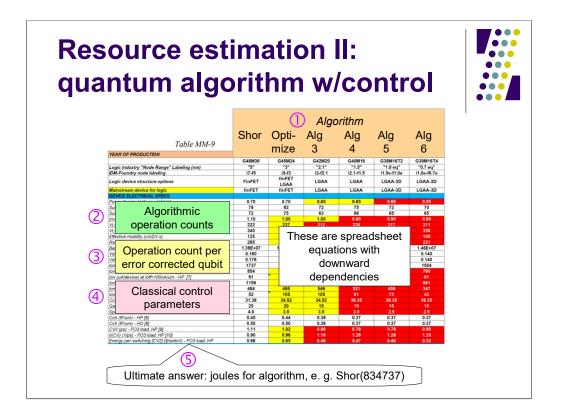
Third bullet: The in-cryostat component is essentially a .gif/.zip decompressor.



The International Technology Roadmap for Semiconductors (ITRS), now the International Roadmap for Devices and Systems (IRDS) may provide an example of how to organize resource estimation.

The iconic colored charts are an Excel spreadsheet. Each column represents a year and is filled with a series of constants and equations with top-to-bottom dependency. The equations compute switching energy for CMOS based on Moore's law assumptions.

Quantum resource estimation is not Moore's law, but perhaps other aspects of the chart are useful.



① The columns are renamed to be algorithms, such as Shor, optimize, etc.

<sup>(2)</sup> Existing quantum resource estimation yields functions for, say, the number of Toffoli magic states required to factor number N. These equations are put into cells to the right of the box.

③ Attention today is focused on logical qubits. This area represents the number of single-qubit resources required for, say, a Toffoli magic state.

 ④ The section to the right of the box includes equations for the number of (ir)reversible gate-ops and total bits (i. e. memory) for each quantum op above.

 $\bigcirc$  The spreadsheet performs the multiplies and adds to compute, for example, the total power and total chip area for the algorithm illustrated based as a function of *N*.



#### Resource model builds on Landauer's minimum dissipation and physics of computation rather than commercial products (e. g. ST Micro 28 nm, 22FFL)

- Developed RL controller out of 100% reversible shift registers
- Simulations suggest >100× dissipation reduction over a cryo CMOS work-alike (but there are many undetermined factors)
- Can prove various things using RL controller as an existence proof:
  No irreversible gates needed in cryostat
  - Resources in cryostat are proportional to source code length
  - Universal cryostat architecture (like .gif/.zip controller)
- Additional data at <a href="https://zettaflops.org/qre-2022">https://zettaflops.org/qre-2022</a>

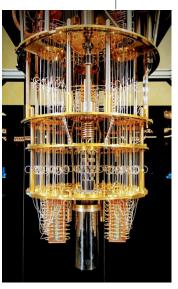
**Conclusions** 

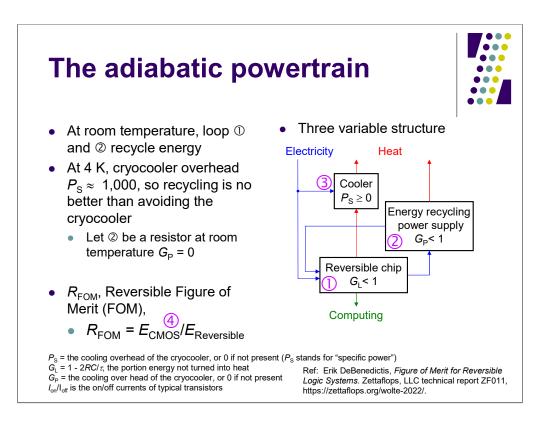


#### **Problem description**

- Quantum computers are a national priority
- The structure on the right does not scale
- For cryogenic qubits, the accepted direction is to compress the data in the cables and use cryogenic electronics to decompress
- *De facto* cryo electronics is cryo CMOS at 4 K, which works about the same as 300 K
- CMOS improving at 2×/decade due to fab, which is not enough for the national priority
- Can we do better?
- Photo by Lars Plougmann, https://www.flickr.com/photos/criminalintent/39660636671, license https://creativecommons.org/licenses/by-sa/2.0/







The next three slides expand on an argument in Erik DeBenedictis, *Figure of Merit for Reversible Logic Systems*. Zettaflops, LLC technical report ZF011 v1, https://zettaflops.org/wolte-2022/.

The next three slides show why room temperature reversible logic could not be productized, but cryogenic reversible logic is different. This diagram applies to room temperature reversible logic if  $P_{\rm S}$  = 0 and cold reversible logic if  $P_{\rm S}$  > 0.  $P_{\rm S}$  ≈ 1,000 for cooling to 4 K.

At room temperature, reversible logic recycles wall-plug energy many times through ① and ② for an efficiency increase of ~1,000×. The cryocooler in a quantum computer ③ has an overhead of  $P_{\rm S} \approx 1,000$ , so an approach that routes energy so it does not go through the cryocooler could give the same benefit.

Energy exits reversible chip ① via two paths, ② (which does not go through the cryocooler) and ③, which is the cryocooler. The approach is to replace the energy recycling power supply with a resistor at room temperature.

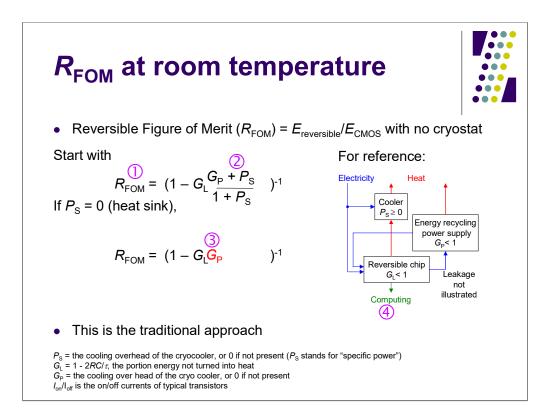
In a quantum computer, the objective is to save energy over cryo CMOS, which can be expressed as maximizing a figure of merit  $R_{\text{FOM}} = E_{\text{CMOS}}/E_{\text{Reversible}}$  ④.

For offline reading:

 $P_{\rm S}$  is "specific power," or the number of watts that must be provided to a cooling system to remove one watt from a cooled environment.

 $G_{L}$  is the sub unity power gain of a reversible chip,  $G_{L}$  has been demonstrated to about 99.9%

 $G_{\rm P}$  is the sub unity power gain of the energy recycling power supply,  $G_{\rm P}$  has been demonstrated to about 95%.



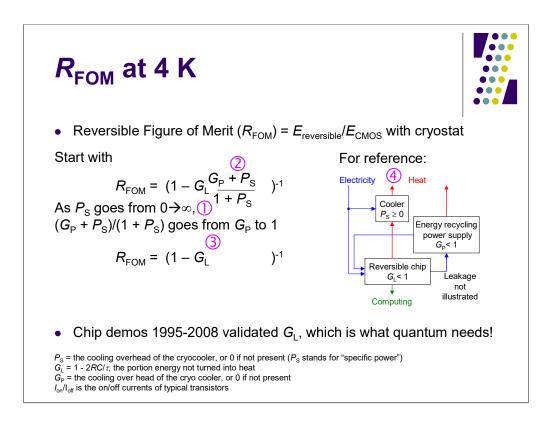
Let us consider room temperature operation where  $P_{\rm S}$  = 0.

 $R_{\text{FOM}}$  (1) is CMOS energy divided by reversible circuit energy. Equation (1) is not very hard to work out, but the derivation is not in this slide deck.

 $R_{\rm FOM}$  has a term 2 that degenerates into familiar forms based on  $P_{\rm S}$ .

Substituting  $P_{\rm S}$  = 0 would be correct for a room temperature system cooled by a heat sink. Simple algebra yields  $R_{\rm FOM}$  = 1/(1 -  $G_{\rm L}G_{\rm P}$ ) ②, as expected. This is the recycling equation that has governed the field for decades ④.

For room temperature operation, the only thing that is important is the product by  $G_LG_P$ , which is why all slide decks and papers have two sections, one for  $G_L$ , the chip's efficiency (which is touted), and  $G_P$ , the recycling efficiency (which is not as good). As a result, reversible logic has been presented as a long-term research direction.

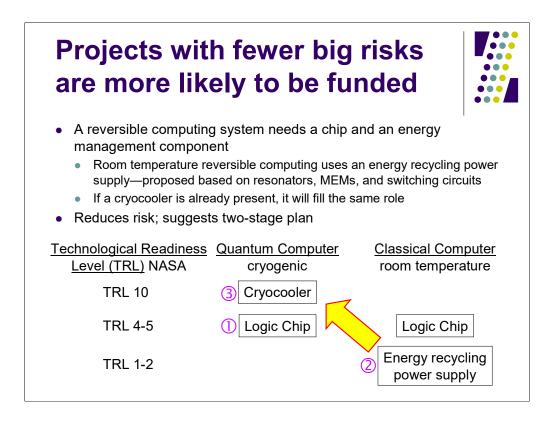


As the temperature of the reversible chip goes down, the cooling overhead  $P_{\rm S}$  goes up. 4 K is often used as a baseline for quantum computer electronics, where  $P_{\rm S} \approx 1,000$ . However,  $P_{\rm S} = \infty$  leads to the same conclusions for this slide deck, so we will use this value. Substituting  $P_{\rm S} = \infty$  will cause the problematic term  $G_{\rm P}$  from the previous slide to "disappear" ③. Explanation:

If  $P_{\rm S}$  is very large ①,  $G_{\rm P} + P_{\rm S} \approx P_{\rm S}$  and  $1 + P_{\rm S} \approx P_{\rm S}$ , so  $R_{\rm FOM}$  ② becomes  $(1 - G_{\rm L})^{-1}$  ③, which his like erasing the problematic term  $G_{\rm P}$  from the last slide.

Interestingly, the demo chips built in the 1994-2008 demonstrated the key checkpoint for reversible logic quantum computer control – but the experimenters did not know this at the time because quantum computers were not well enough defined to know the requirements for the cryogenic classical logic.

Additional detail:  $P_{\rm S}$  does not become infinite in a quantum computer, but is about 1,000 at 4 K. However, the "crossover point" is where  $P_{\rm S}$  is about equal to the slowdown factor. The slowdown factor in a quantum computer is about the ratio of CMOS speed to quantum measurement speed, which is about 1,000 as well. So we are OK, but close to the limit.



The cryo-reversible powertrain needs two parts, a reversible logic chip ① and an energy handling unit ② or ③.

Extending the energy efficiency of a (room-temperature) reversible microprocessor chip to the wall plug would require an energy recycling power supply ②. Energy recycling power supplies are at at a lower TRL level (see note) than reversible chips, so this would link two research projects to one another, increasing risk.

However, a cryogenic reversible logic quantum control chip would naturally use a cryocooler for the same function. A cryocooler is TRL 10, an off-the-shelf item.

TRL level is "Technology Readiness Level," a NASA term. See https://en.wikipedia.org/wiki/Technology\_readiness\_level

Thus, there less risk for a cryogenic reversible system.

## Existence proofs of what? This is what...



No irreversible gates required in cryostat

- Landauer's minimum is "order of *kT* per irreversible function"
- RL controller has no irreversible gates
- Cryo CMOS design tools lay out mostly irreversible gates
- Caveats: One of my papers has a CNOT for optimization; crossover is an externally controlled Fredkin gate; have data-controlled clocks

Reversible circuit is of bounded and reasonable size

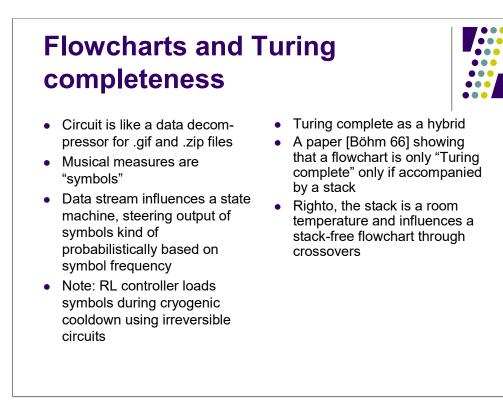
- Reversible gate dissipation = (2 RC/τ) × (½CV<sup>2</sup>)
- For 1 μs quantum measurement, (2 RC/τ) ≈ 1/1,000
- Bennett showed all-reversible computers are possible, but with overhead [Bennett]
- By construction, RL Controller has no infinite stack and data in shift register proportional to source code size

The RL controller circuits forms an existence proof, but of what? It is an existence proof that no irreversible gates are needed in the cryostat.

(right) Landauer identified a minimum dissipation for an irreversible gate, but the minimum does not apply to reversible gates. If the gates are based on transistors, the dissipation is about the same as CMOS at its top clock rate. However, the dissipation decreases linearly with clock period (2 RC/ $\tau$ ).

As stated earlier, qubit measurement is much slower than CMOS, about 1,000× slower.

However, we must address the possibility that reversible logic would require vastly more gates than a CMOS equivalent. This turns out not to be true, as the flowchart will only have a much "music" stored as is in the original algorithms. There is not, for example, a need for a "stack" in the cryostat that contains potentially unbounded storage.



[Böhm 66] Böhm, Corrado, and Giuseppe Jacopini. "Flow diagrams, Turing machines and languages with only two formation rules." *Communications of the ACM* 9.5 (1966): 366-371.

### Landauer's minimum dissipation in a hybrid system

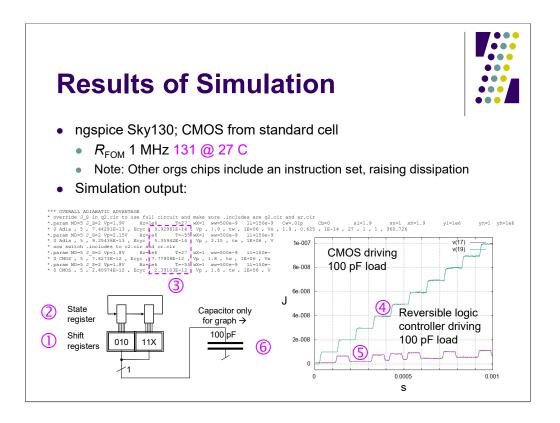


Physicist vs. computer architect

- Landauer's minimum is *kT* per irreversible operation
- In a mixed temperature environment, which *T* do we use?
- Physicist's answer: The *T* of the environment performing the function
- Computer architect's answer: What are the limits of moving the irreversible operations to an environment where *T* is most favorable?

Proposed hybrid multitemperature computing architecture

- Room temperature:
  - Compute the output
  - Compress the output
- Cable:
  - Move compressed string into the cryostat
- Cryo electronics:
  - Decompress the output



The ngspice simulation has two data-controlled clocks ② and two registers ① holding 3 notes of "music" each. The "music" is 010 and 11 or 111.

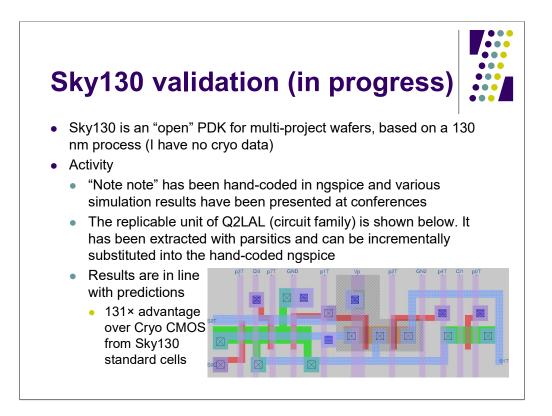
The output of a simulation run is shown above ③. There were four 1 MHz runs in the combinations of CMOS/Adiabatic and 27 C/-55 C. Adiabatic supply voltages were1.9 V (27 C) and 2.05 V (-55 C) and 1.8 V for CMOS. Wire capacitance was .01 pF. The first period of a 20 cycle run has startup effects, so Ecyc is averaged over the remaining 19 cycles.

Within the purple dashed line,  $R_{\text{FOM}} = E_{\text{CMOS}}/E_{\text{Adia}}$  yields 131 at 27 C. The -55 C CMOS simulation did not function.

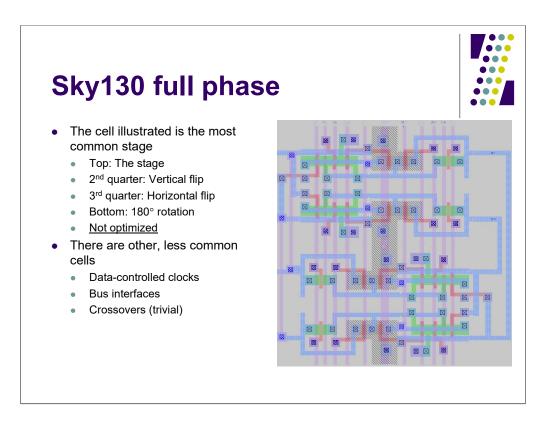
The graph demonstrates functionality from a simulation, with same architecture ① and ②, but with a different transistor model, a 10 V supply, and a 100 pF capacitor ③ on the output bus. Curve ④ is a reference for CMOS, which rises by  $\frac{1}{2}CV^2$  each on the rising edge of the clock. Curve ⑤ is the output.

The was an oversight in the original circuit creating the graph: the second music register was only two stages long instead of three. The missing stage made no functional difference because it was initialized to all 1s. The oversight has been fixed, but the graph on the right appeared in some documents, so the original version is used here for comparisons. For what appears below, assume the second music register holds 111.

The music 010 plays first and alternates with 111. The output (5) represents 3 1/3 "musical measures" of these three notes. The pattern is 010 111 010 1. So (5) is on top of the underlined <u>0</u> in the preceding sequence.



Validation was also performed using Sky130 layout. To assure constraints like even load, the basic replication unit should be create from a single template with reflections and rotations. The reflections and rotations require solving a puzzle on how to organize the flyover wires.



Full stage of Q2LAL. Replication units are flipped v and h and rotated 180°. Not exactly balanced due to the need for crossovers.