Cover sheet

Erik P. DeBenedictis, 12/31/2018

This document comprises new "supplementary information" for a previously released document. The new supplementary information starts on the next page. The new material is standalone except for figure 1 and reference 3 of the base document, both reproduced below.

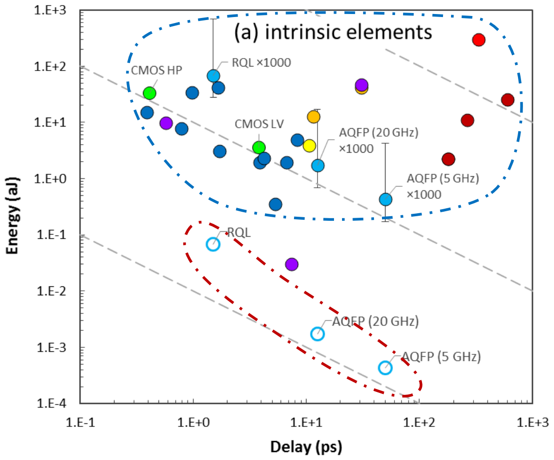

Figure 1. An energy-delay plot of a comprehensive set of logic devices at room temperature, yet including Josephson junctions operating at 4K in three circuits (RQL and AQFP at two speeds). None of the devices stand out at room temperature because only the superconducting ones have refrigeration overhead. However, at 4K they all require refrigeration, causing the superconducting devices to stand out. (The outlier purple dot is a BisFET, which is too immature to be taken as a serious contender.)

References

1. D. E. Nikonov and I. A. Young, "Overview of beyond-CMOS devices and a uniform methodology for their benchmarking," Proc. IEEE, vol. 101, no. 12, pp. 2498–2533, 2013. doi:10.1109/JPROC.2013.2252317

2. V. Anantharam, M. He, K. Natarajan, H. Xie, and M. P. Frank. "Driving fully-adiabatic logic circuits using custom high-Q MEMS resonators," in Proc. Int. Conf. Embedded Systems and Applications and Proc. Int. Conf VLSI (ESA/VLSI). Las Vegas, NV,

3. R. M. Incandela, L. Song, H. Homulle, E. Charbon, A. Vladimirescu, and F. Sebastiano. "Characterization and compact modeling of nanometer CMOS transistors at deep-cryogenic temperatures," IEEE J. Electron Devices Soc., vol. 6, pp/ 996-1006.

4. G. De Simoni, F. Paolucci, P. Solinas, E. Strambini, and F. Giazotto. "Metallic supercurrent field-effect transistor," Nature Nanotechnology, vol. 13, no. 9, pp. 1, 2018. doi:10.1038/ s41565-018-0190-3

5. N. K. Katam, O. A. Mukhanov, and M. Pedram, "Superconducting magnetic field programmable gate array," IEEE Trans. Appl. Supercond., vol. 28, no. 2, pp. 1–12, 2018. doi:10.1109/TASC.2018.2797262

Supplementary Information: FPGA example details

IARPA Cryogenic Computational Complexity project (C3) intended to create an Exascale supercomputer using Josephson junction technology, yet concluded with a million-gate chip containing a 16-bit RQL microprocessor and a few kilobits of memory. C3 was part of a larger effort on future computing methods, which is now considering cryogenic sensor arrays and quantum computer control electronics for subsequent projects (“Super Cables” is an example of such a project).

This section shows how to design with the technology in this article, showing how to add 100 million transistors to the cryogenic chip without significant additional power dissipation. Let’s use RQL and CMOS HP from Figure 1 as a baseline, supplemented by information[3] that will control the performance of 2LAL.

RQL from Figure 1		CMOS HP from Figure 1		2LAL leakage		
E_{RQL} (J)	1.00E-19	E_{CMOS} (J)	4.00E-17	"on" ohms		3,000
$t_{pd,\ RQL}$ (s)	1.25E-12	$t_{pd,\ CMOS}$ (s)	5.00E-13	I_{on}/I_{off}		1.00E+08
$f_{clk,\ RQL}$ (Hz)	1.60E+09	$f_{clk,\ CMOS}$ (Hz)	4.00E+09	Power (1 V, 50% duty, W)		1.67E-12

Assuming the C3 cryogenic chip contains 1 M gates, the superconductor layer will dissipate 160 μW to the 4 K environment. Let’s stipulate that the semiconductor layer dissipates the same power, which the spreadsheet below computes as requiring 1000 gates.

Baseline						
N_{RQL}		$f_{clk,\ RQL}$ (Hz)		E_{RQL} (J)	$p_{RQL,\ 4\ K}$ (W)	
1,000,000		1.60E+09		1.00E-19	0.000160	
N_{CMOS}	Cplx.	$f_{clk,\ CMOS}$ (Hz)	Clk ratio	E_{CMOS} (J)	$p_{CMOS,\ 4\ K}$ (W)	$p_{Static,\ 4\ K}$ (W)
1,000	1	4.00E+09	1	4E-17	0.000160	n/a

Figure 4 diagrams the result, with superconductors in blue and semiconductors in red. The semiconductor layer is limited by power dissipation, so only the small red area in the lower left contains devices.

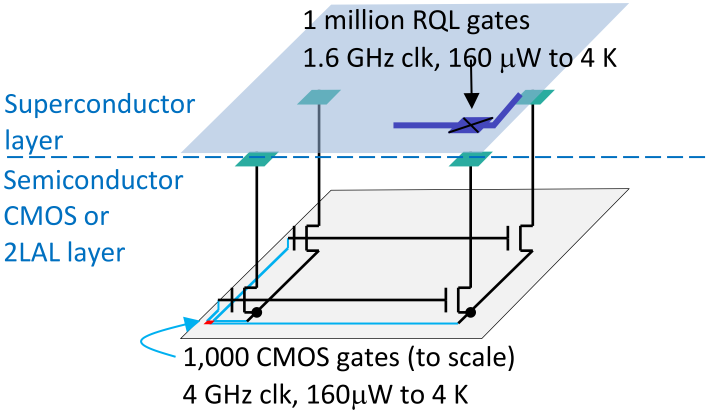
1 million RQL gates
1.6 GHz clk, 160 μW to 4 K
Superconductor layer
Semiconductor CMOS or 2LAL layer
1,000 CMOS gates (to scale)
4 GHz clk, 160μW to 4 K

Figure 4. A thought experiment for a semiconductor-superconductor hybrid FPGA. This is a baseline structure most directly compatible with the gates in Figure 1.

Adiabatic scaling

Now let’s lower the semiconductor layer’s clock rate but exploit the resulting lower power per gate by adding gates so that overall layer power remains the same (leakage power will be negligible in this step). The circuit is first changed from CMOS to 2LAL, which introduces a 10× increase in device count because 2LAL uses more transistors per gate. The two steps each scale down the 4 GHz clock by 10× – to 400 MHz and then to 40 MHz. Each clock frequency reduction cuts dynamic power by the square of the clock rate, or 100×. Our spreadsheet computes the number of gates to maintain power dissipation of 160 μW to the 4 K environment. The superscripts $^{(1)}$, $^{(2)}$, and $^{(3)}$ denote the scaling stages.

Stage 1						
N_{RQL}		$f_{clk,\ RQL}$ (Hz)		E_{RQL} (J)	$p_{RQL,\ 4\ K}$ (W)	
1,000,000		1.60E+09		1.00E-19	0.000160	
$N^{(1)}_{2LAL}$	Cplx.	$f^{(1)}_{clk,\ 2LAL}$ (Hz)	Clk ratio	$E^{(1)}_{2LAL}$ (J)	$p^{(1)}_{2LAL,\ 4\ K}$ (W)	$p^{(1)}_{Static,\ 4\ K}$ (W)
10,000	10	4.00E+08	0.1	4E-18	0.000160	1.67E-08
Stage 2						
N_{RQL}		$f_{clk,\ RQL}$ (Hz)		E_{RQL} (J)	$p_{RQL,\ 4\ K}$ (W)	
1,000,000		1.60E+09		1.00E-19	0.000160	
$N^{(2)}_{2LAL}$	Cplx.	$f^{(2)}_{clk,\ 2LAL}$ (Hz)	Clk ratio	$E^{(2)}_{2LAL}$ (J)	$p^{(2)}_{2LAL,\ 4\ K}$ (W)	$p^{(2)}_{Static,\ 4\ K}$ (W)
1,000,000	10	4.00E+07	0.01	4E-19	0.000160	1.67E-06

Figure 5 diagrams the two stages. The CMOS then 2LAL’s clock drops from 4 GHz to 40 MHz – a pretty substantial but not a fatal drop – but the number of gates increases by an attention-getting 10,000×. There is plenty of chip area available to hold the new gates, as illustrated by the red areas in the diagram.

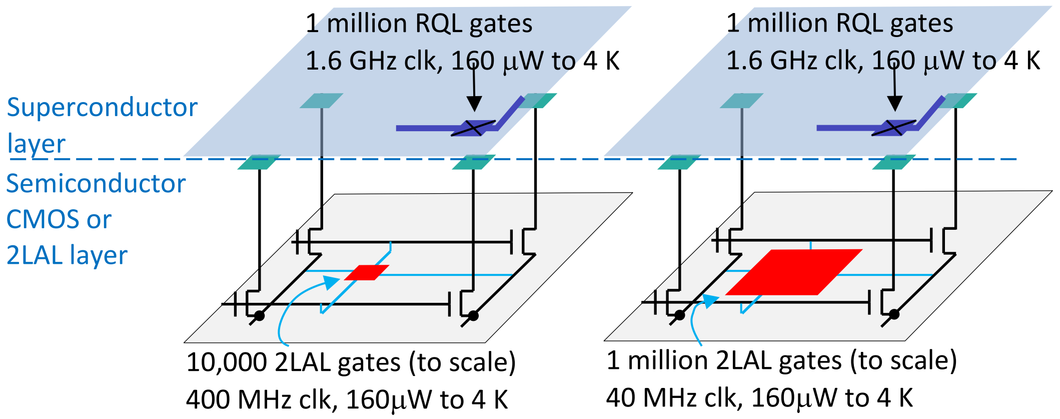
1 million RQL gates
1.6 GHz clk, 160 μW to 4 K
1 million RQL gates
1.6 GHz clk, 160 μW to 4 K
Superconductor layer
Semiconductor CMOS or 2LAL layer
10,000 2LAL gates (to scale)
400 MHz clk, 160μW to 4 K
1 million 2LAL gates (to scale)
40 MHz clk, 160μW to 4 K

Figure 5. Scale up sequence where the semiconductor portion grows by 100× at the price of becoming 10× slower, keeping power the same due to adiabatic scaling.

Let’s pause for a minute to understand an important fact about historical context. The interest in quantum computers has accelerated in the last few years, shining a spotlight on cryogenic CMOS. We’ve stated that cryogenic MOSFETs have an I_{on}/I_{off} ratio of 10^8 at 4 K,[3] but this data is recent. In our opinion, the community’s knowledge of reversible and adiabatic computing has not considered cryogenic operation, so the community would have used the room temperature figure for I_{on}/I_{off}, which is two orders of magnitude lower, or 10^6. If $I_{on}/I_{off} \approx 10^6$, we'd be done with scaling at this stage because the leakage current ($p^{(2)}_{\text{Static, 4K}}$ in the spreadsheet) would go from a negligible 1.67 µW to 167 µW at 4 K – which is about the same as the dynamic power.

Stopping here leads to a pretty useless result. We would have doubled the number of gates, but half the gates are very slow and the system is more complex.

Final scaling step

Given recent attention to cryogenic CMOS, we now realize I_{on}/I_{off} is 10^8 at 4 K and we see the need an opportunity to take another scaling step, per the spreadsheet below.

Stage 3						
N_{RQL}		$f_{clk,\ RQL}$ (Hz)		E_{RQL} (J)	$p_{RQL,\ 4\ K}$ (W)	
1,000,000		1.60E+09		1.00E-19	0.000160	
$N^{(3)}_{2LAL}$	Cplx.	$f^{(3)}_{clk,\ 2LAL}$ (Hz)	Clk ratio	$E^{(3)}_{2LAL}$ (J)	$p^{(3)}_{2LAL,\ 4\ K}$ (W)	$p^{(3)}_{Static,\ 4\ K}$ (W)
100,000,000	10	4.00E+06	0.001	4E-20	0.000160	0.000167

This leads to the diagram in Figure 6, where the semiconductor layer has a 4 MHz clock and 100 million 2LAL gates. This is the last scaling step given our assumptions: the CMOS layer is full and 2LAL leakage is now significant.

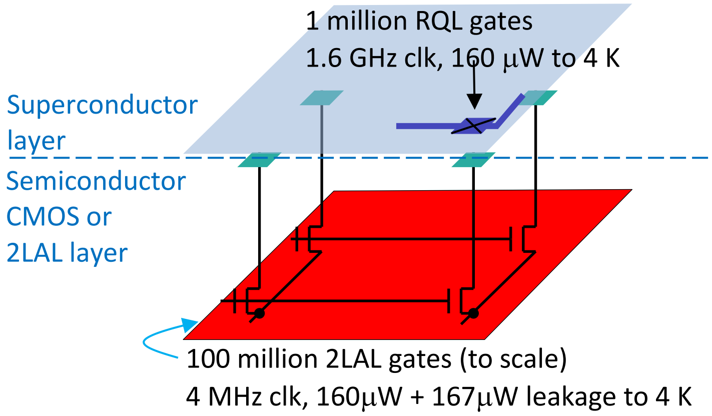
1 million RQL gates
1.6 GHz clk, 160 µW to 4 K
Superconductor layer
Semiconductor CMOS or 2LAL layer
100 million 2LAL gates (to scale)
4 MHz clk, 160µW + 167µW leakage to 4 K

Figure 6. Final scaling step with 100 2LAL gates per RQL gate, yet increasing power by only 3×.

What do with do with it?

We now have a tradeoff that may be exploitable by system designers. Each of the original million fast, low power Josephson junction-based RQL gates are now accompanied by 100 2LAL gates of equally low power. The new gates run slowly, but there are a lot of them. Here are some ideas on what to do:

We used RQL and "CMOS HP" as technology examples in this article, due to their being listed in Figure 1. There are many other logic families that will have the same qualitative behavior (such as ERSFQ and SCRL), but at different densities, speeds, and energy levels. There is also interest in improving both semiconductor density and the I_{on}/I_{off} ratio, so the analysis in this article may be worth redoing later on.

Both quantum computers and cryogenic sensor arrays may benefit from real-time reconfiguration. For example, one part of a quantum computer algorithm may want to operate with a 5-bit, 7-bit, or surface quantum error correction code. The best support for a quantum error correction code may be through logic in the control electronics. If the FPGA-like structure described in the article were used in lieu of a cryogenic ASIC, the code change could be accomplished without new hardware.

The IARPA C3 program includes a substantial effort on design tools. Some of this design tool effort could be avoided if the community settled on one or just a few FPGA-like designs that would be, essentially, hand designed, where application specific functions would be done through an FPGA configuration tool.

The paragraph above also applies to "subroutines" in quantum algorithms. Some quantum algorithms could quite reasonably do a lot of 8-bit quantum integer additions whereas others may need 150-bit additions. In fact, a single algorithm might do quantum integer arithmetic of different bit sizes or over different algebraic groups. The best performance for quantum addition may require specific logic in the control electronics. This FPGA-like structure could be configured for *k*-bit quantum integer addition, with the FPGA reloaded for different *k*'s on the fly.

While the article suggested the semiconductor layer could be used for FPGA-like configuration, there is no reason to limit it this way. The 100 million 2LAL gates could be organized into a memory – either random access or a specific access pattern like a shift register. The shift registers may be useful for storing waveforms in a quantum computer.

The semiconductor layer could be used as an I/O buffer for signals from room temperature. The example in this supplementary information transfers 1 million bits in parallel at 4 MHz, for a bandwidth of 4 Tb/sec. This is 100× the highest rate we're aware of from room temperature.

It would be straightforward to have multiple clock rates on chips like those shown in figures 4-6. Each adiabatic clock involves 4 phases (typically) and uses true and complement signals – or 8 wires from the cryogenic environment to the clock generation equipment at room temperatures. Wires in and out of the cryogenic environment are expensive, but adding an 8 extra wires for good reason is certainly worth considering.

This article was written using 4 K as the only cryogenic temperature, but the ideas could apply to other temperatures as well. In general, larger $I_{\mathrm{on}}/I_{\mathrm{off}}$ ratios enable more extreme adiabatic clock rate scaling and hence allow use at lower temperatures than 4 K. There is also research on creating cryo CMOS that is functional at lower supply voltages (typically by lowering the threshold). Lower supply voltage would lead to more energy efficient semiconductor computation at lower temperatures, affecting architectural tradeoffs.

This article discussed interfacing between semiconductors and superconductors via a superconducting FET. This is not the only approach, in fact a wide semiconductor FET will pass or block a SFQ pulse if designed correctly.