REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
10-03-2015	Final Report	29-Oct-2009 - 28-Jan-2017

4. TITLE AND SUBTITLE
Final Report: Design and Demonstration of a 30 GHz 16-bit Superconductor RSFQ Microprocessor

5a. CONTRACT NUMBER
W911NF-10-1-0012

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Mikhail Dorojevets

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
State University of New York (SUNY) at St
Office of Sponsored Programs
Research Foundation Of SUNY
Stony Brook, NY 11794 -3362

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
57342-PH-OC.6

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
The major objective of the project was to design and demonstrate operation of key components of a 30 GHz 16-bit RSFQ processor prototype implemented with the AIST/ISTEC 10 kA/cm sq. fabrication process. Our team has developed complete logical and physical designs of five RSFQ chips using the CONNECT cell library and RSFQ CAD tools developed at the Universities of Yokohama and Nagoya (Japan). The major results are the world's first successful design, fabrication, and demonstration of correct operation of a 20 GHz 8x8-bit parallel carry-save RSFQ multiplier with ~6K JJs, a 16-bit sparse-tree wave-pipelined RSFQ adder with ~10K JJs, and partial

15. SUBJECT TERMS
Superconductor technology, RSFQ, RQL, processor design, arithmetic units, high-performance computing

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Mikhail Dorojevets
UU	UU	UU	UU		19b. TELEPHONE NUMBER 631-632-8611

Standard Form 298 (Rev 8/98)
Prescribed by ANSI Std. Z39.18

Report Title

Final Report: Design and Demonstration of a 30 GHz 16-bit Superconductor RSFQ Microprocessor

ABSTRACT

The major objective of the project was to design and demonstrate operation of key components of a 30 GHz 16-bit RSFQ processor prototype implemented with the AIST/ISTEC 10 kA/cm sq. fabrication process. Our team has developed complete logical and physical designs of five RSFQ chips using the CONNECT cell library and RSFQ CAD tools developed at the Universities of Yokohama and Nagoya (Japan). The major results are the world's first successful design, fabrication, and demonstration of correct operation of a 20 GHz 8x8-bit parallel carry-save RSFQ multiplier with ~6K JJs, a 16-bit sparse-tree wave-pipelined RSFQ adder with ~10K JJs, and partial operation of an 8-bit ALU chip with ~9K JJs. The goal of the second phase of the project was to get detailed understanding of the performance, complexity, and energy efficiency of on-chip storage units implemented with superconductor Reciprocal Quantum Logic (RQL) using our RQL VHDL cell library tuned to the MIT Lincoln Laboratory 10 kA/cm^2 248 nm process. The 8.5 GHz 1-4 Kbit 32-/64-bit multi-ported scratchpad memory, register files, write-through and write-back caches designed with RQL Non-Destructive Read-Out storage cells have the average energy consumption of 3.0-9.5 fJ/bit/operation at room temperature using the cryocooling efficiency of 0.1%.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received		Paper
02/18/2015	5.00	Mikhail Dorojevets, Zuoting Chen, Christopher L. Ayala, Artur K. Kasperek. Towards 32-bit Energy-Efficient Superconductor RQL Processors: The Cell-Level Design and Analysis of Key Processing and On-Chip Storage Units, IEEE Transactions on Applied Superconductivity, (06 2015): 0. doi: 10.1109/TASC.2014.2368354
08/12/2013	4.00	C. L. Ayala, N. Yoshikawa, A. Fujimaki, M. Dorojevets. 8-Bit Asynchronous Sparse-Tree Superconductor RSFQ Arithmetic-Logic Unit With a Rich Set of Operations, IEEE Transactions on Applied Superconductivity, (06 2013): 0. doi: 10.1109/TASC.2012.2229334
08/12/2013	2.00	M. Dorojevets, C. L. Ayala, N. Yoshikawa, A. Fujimaki. 16-Bit Wave-Pipelined Sparse-Tree RSFQ Adder, IEEE Transactions on Applied Superconductivity, (06 2013): 0. doi: 10.1109/TASC.2012.2233846
08/12/2013	3.00	A. K. Kasperek, N. Yoshikawa, A. Fujimaki, M. Dorojevets. 20-GHz 8 x 8-bit Parallel Carry-Save Pipelined RSFQ Multiplier, IEEE Transactions on Applied Superconductivity, (06 2013): 0. doi: 10.1109/TASC.2012.2227648
TOTAL:	**4**	

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

1. Mikhail Dorojevets, Energy-Efficient Superconductor Circuits for High-Performance Computing, Proc. of the 2014 Advanced International High-Performance Workshop (HPC-2014), Cetraro, Italy, July 2014.

2. Mikhail Dorojevets, Processor Design with Superconductor Single-Flux-Quantum Technology, 2012 CMOS Emerging Technologies, Vancouver (BC, Canada), July 17-20, 2012.

Number of Presentations: 1.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received		Paper
08/04/2011	1.00	Mikhail Dorojevets, , Christopher L. Ayala, , Artur K. Kasperek. Data-Flow Microarchitecture for Wide Datapath RSFQ Processors: Design Study, IEEE Transactions on Applied Superconductivity, vol. 21, no. 3, June 2011 (06 2011)
TOTAL:		**1**

Number of Manuscripts:

Books

Received	Book
TOTAL:	

Received	Book Chapter
TOTAL:	

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Artur Kasperek	0.50	
Christopher Ayala	0.60	
Surabhi Garg	0.20	
Swati Shah	0.10	
Subramaniyan Venkatachalam	0.10	
Hao Chen	0.10	
Zuoting Chen	0.60	
Nimit Goel	0.10	
Zhihua Gan	0.05	
FTE Equivalent:	**2.35**	
Total Number:	**9**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Milutin Stanacevic	0.01	
Mikhail Dorojevets	0.25	
FTE Equivalent:	**0.26**	
Total Number:	**2**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):..... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: 0.00

Names of Personnel receiving masters degrees

NAME
Artur K. Kasperek
Christopher L. Ayala
Swati Shah
Nimit Goel
Surabhi Garg
Hao Chen
Zuoting Chen
Subramaniyan Venkatachalam
Total Number: **8**

Names of personnel receiving PHDs

NAME
Artur K. Kasperek
Christopher L. Ayala
Total Number: **2**

Names of other research staff

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

In order to be seriously considered as a competitor to high-performance CMOS processors, superconductor processors need to demonstrate sufficient functionality, complexity, reliability, speed, and energy efficiency in practical designs. To address these challenges, the project had two different tasks discussed below.

I. Design and demonstration of operation of key components of a 30 GHz 16-bit RSFQ processor

We have designed, fabricated, and tested several high-performance RSFQ processing and storage units. The complete designs and physical layouts of all chips have been done at Stony Brook University using the CONNECT cell library and SFQ CAD tools developed at Nagoya and Yokohama Universities (Japan) for the AIST/ISTEC 1.0 μm 10 kA/cm^2 fabrication process.

Testing of all fabricated RSFQ chips was done by the Stony Brook team with assistance from colleagues at Yokohama National University in 2011-2012.

1.1 A 30 GHz 16-bit RSFQ sparse-tree adder overview

We have designed, fabricated, and demonstrated operation of the first 16-bit RSFQ wave-pipelined sparse-tree adder chip with the core complexity of 9941 JJs and the target operation rate of 30 GHz for a 16-bit integer RSFQ processor.

The microarchitecture of the adder has two main features: 1) a use of a technique of asynchronous hybrid wave-pipelined processing developed at SBU, and 2) a prefix sparse-tree carry generate-propagate structure for arithmetic.

In the data-driven asynchronous hybrid wave-pipelining, data waves “self-propagate” through combinational (non-clocked) logic gates without any need for clock signals. The data waves are followed by reset waves that “clean up” the residual logic states of the gates before the next data wave arrival.

The Kogge-Stone adder (KSA) is considered to be the fastest among parallel-prefix adders. However, KSAs have very high complexity and a tremendous amount of wiring congestion because of the need for their prefix trees to provide carries to every individual bit of the adder. In our 16-bit RSFQ adder design, we chose the sparse-tree structure to reduce the number of Josephson junctions (JJs) needed for its implementation without any significant effect on its processing rate. As a side effect, this will also lead to a more energy-efficient design by reducing the total bias current and power consumption.

Our sparse-tree adder has the following three stages: Initialization, Prefix-Tree, and Summation.

The Initialization stage receives two 16-bit data operands A and B to create bitwise Generate (G) and Propagate (P) signals using clocked AND and XOR gates in a co-flow clocking arrangement.

The Prefix-Tree stage consists of Carry-Merge blocks to merge the prefix signals in a logarithmic manner and provide a group carry to each 4-bit summation block. Merging of the prefix signals is implemented with CFFs (resettable Muller C-flip-flop gates based on the Muller C-element) and confluence buffers used as asynchronous OR gates. The first three levels of the sparse tree also perform the ripple-carry addition within each 4-bit group before data arrive at the Summation stage.

The Summation stage computes the final sum with 4-bit carry-skip adders. The lower-half of the adder (bits 7:0) can start the Summation stage early because all appropriate signals are ready. The upper-half of the adder (bits 15:8) must wait until carries for this upper half are calculated by the very last level of the Prefix-Tree stage.

The 16-bit adder was designed for the target operation frequency of 30 GHz with the latency of 352 ps at the bias voltage of 2.5 mV. The 16-bit adder core (without SFQ-to-dc and dc-to-SFQ converters) has 9941 Josephson junctions occupying an area of 8.5 mm^2. The total number of JJs on the adder chip (including test and I/O circuits) is 12,785 and with a total bias current of 1.61 A. The adder chip was fabricated and successfully tested at low frequency for all test patterns with measured bias margins of +9.8%/ - 10.7%.

1.2 20 GHz 8x8-bit parallel carry-save multiplier overview

We have successfully designed, fabricated, and tested the first 8 × 8-bit (by modulo 256) parallel carry-save superconductor RSFQ multiplier with the target frequency of 20 GHz.

When designing our 8 × 8-bit parallel integer multiplier, we had four major targets: high operation frequency of 20 GHz, multiplication time below 500 ps, complexity around 6000 Josephson junctions (JJs), and mostly regular layout employing both local and global connections.

The 8x8-bit (by modulo 28) multiplier operates on two 8-bit input operands and calculates an 8-bit product. The 8x8-bit multiplier has three major components: a partial product generator, partial product compression (reduction) blocks based on 4-to-2

counters, and the final summation block implemented as a ripple-carry adder for the most-significant bits of the product.

The multiplier partial product generator (PPG) consists of 36 partial product (PP) bit generators built with clocked AND gates operating on their multiplicand and multiplier bits. These circuits are organized into three PPG groups, one with 16 and two other with 10 PP generators each. PPs in each PPG group are calculated in parallel, significantly reducing the partial product generation time. Partial products are asynchronously generated and sent to the reduction stage at the internal “hardwired” rate of 80 GHz.

The 8 × 8-bit RSFQ multiplier uses a two-level parallel carry-save reduction tree that significantly reduces the multiplier latency. The 80-GHz carry-save reduction is implemented with asynchronous data-driven wave-pipelined [4:2] compressors (counters) built with toggle flip-flop cells. First, up to 8 PPs in each column are reduced to 4 by two [4:2] compressors working in parallel, each producing 2 PPs. The 4 PPs from the two first-level compressors are merged together with asynchronous confluence buffers and sent ~12.5 ps apart over a single PTL to a second-level [4:2] compressor for that column. Then, the second-level [4:2] compressor will reduce those 4 PPs to 2. The benefits of using this approach are as follows: 1) the O(log2n) PP reduction time, where n is the operand length, and 2) a regular layout with both local and global connections between modules.

The five least significant bits of the product are calculated during the PP reduction by the [4:2] compressors. The partial products in the three most significant bit columns are reduced to carry-sum pairs and then go through the final summation done by a wave-pipelined ripple-carry adder.

The 8 × 8-bit RSFQ multiplier has the latency of 447 ps at the nominal bias voltage of 2.5 mV and the acceptable DC bias margins at the target frequency of 20 GHz when operating at a slightly higher than nominal bias voltage. The multiplier core (without DC-to-SFQ and SFQ-to-DC converters) is built with 5948 JJs occupying the area of 3.5 mm^2 (2.5 mm × 1.45 mm). It has a bias current of 676 mA.

The multiplier chip was fabricated in December 2011. Despite some challenges due to fabrication process parameter variations and flux trapping, the multiplier chip was successfully tested for the vast majority of test vectors by the Stony Brook designers in February 2012. While multiplier test operations were generated at low frequency, each of these operations was executed at the internal “hardwired” rate of 80 GHz. The fabricated chip operated with the measured DC bias margins of ±5%.

1.3 30 GHz 8-bit ALU block overview

The design of a new 8-bit asynchronous RSFQ ALU was the next step in the development of bit-parallel RSFQ ALUs compared to an 20 GHz 8-bit ALU earlier developed by our team in collaboration with HYPRES, Inc. First, the operation set is significantly extended, including the introduction of a true two’s complement subtract operation. Second, we have developed a sparse-tree design with significantly less complexity of the carry-generate-propagate prefix-tree than in the earlier ALU design.

Our 8-bit ALU has three key microarchitecture and design features: asynchronous wave-pipelined processing, a prefix sparse-tree carry generate-propagate structure, and data pulse counters for implementation of complex operations.
This unit features an extensive set of 8 arithmetic and 12 logical operations. The execution of ALU operations consists of two steps. First, when necessary, one or both operands are inverted, and then operations are performed on these pre-processed data.

List of logical operations:
NOP, AND A, B; OR A, B; XOR A, B; Set All 1s; AND A, ¬B; OR A, ¬B; AND ¬A, B;
OR ¬A, B; XNOR A, B; NOR A, B; NAND A, B.

List of arithmetic operations and their mathematical representations:
1. ADD A, B => A + B
2. ADD ¬A, B => B - A - 1
3. ADD A, ¬B => A - B - 1
4. ADD ¬A, ¬B => -(A + B + 2)
5. INC_ADD A, B => A + B + 1
6. SUB A, B => A - B
7. SUB B, A => B - A
8. INC_ADD ¬A, ¬B => -(A + B + 1)

The entire ALU chip layout including moats, I/O and the test circuit fits within the 4.25 mm × 4.00 mm area with the core circuit with the complexity of 8832 JJs requiring the area of 7.2 mm^2 (2.67 mm × 2.70 mm). It has the simulated DC bias margins of +20%/-16% at the target processing rate of 30 GHz.

Testing of the chip was done at low frequency to check for correct functionality. Testing for some operations, such as XNOR,

showed incorrect results due to the malfunctioning of two gates from the CONNECT cell library. Despite these shortcomings we still verified several other ALU operations, e.g. XOR, ADD, INC_ADD, that did not use those gates. The measured overlapped DC bias margins for these operations were ±1.8%.

1.4 30 GHz 8-bit dual-register block overview

The 8-bit dual-register slice is the component of a register file with two read and one write ports. The slice consists of two 8-bit wide registers (left and right registers that provide left and right operands for ALU) with non-destructive read-out.

The fabricated 2x8-bit register slice has a core complexity of 1,060 JJs. A complete circuit (with test and I/O circuits included) has 1,806 JJs, an area of 1.06 mm^2, and a total bias current of 203 mA.

The register file operation set includes read_left, read_right, write _left, and write_right operations. Before the write operations, the register is reset by reset_left/ reset_right operation depending on which of them will be written with data from ALU. Two registers (one left and one right) can read and one register written each cycle.

The simulation results show that the 8-bit dual-register block design has very wide DC bias margins (greater than ±20%) to operate at the target 30 GHz clock rate.

1.5 Summary of the most important results

The major results of this work are:

1) the world's first successful demonstration of a 16-bit superconductor wave-pipelined RSFQ adder with its complexity of ~10K JJs and target frequency of 30 GHz;
2) the world's first successful demonstration of an 8x8-bit parallel carry-save RSFQ multiplier with its complexity of ~6K JJs and target frequency of 20 GHz;
3) demonstration of partial operation of 8-bit wave-pipelined ALU with its complexity of ~9K JJs and target frequency of 30 GHz.

The results of our work have proven the viability of the hybrid wave-pipelined approach developed by our team at Stony Brook for practical 30 GHz superconductor processor design. Also, the experience of working with the Japanese digital CAD tools, their cell library and fabrication facility has allowed us to clearly understand the benefits and shortcomings of the semi-custom design flow similar to the one used in Japan over a full-custom design flow currently used in the US.

Among those advantages is the separation of the analog (JJ-level) and digital design flows. The latter is based on the standard cell library with cells having their HDL behavioral models and fixed shapes (layouts), providing the opportunity to develop logic synthesis tools for superconductor VLSI circuits in the future.

There were several challenges during this period of work that have significantly influenced the execution of the project.

1) The AIST chip foundry was damaged during the earthquake in March 2011. After more than a six-month operation recovery period, the foundry resumed fabrication by 2012 but the yield, chip quality, and throughput were found to be low due to some unpredictable variations in the fabrication process parameters.
2) As a result of this, the number of new chips allocated to our Stony Brook team in 2012 was limited to one.
3) The foundry decided to stop fabrication in the summer of 2012 in order to upgrade their clean room equipment.

All of this has forced the Stony Brook team to focus on the design and demonstration of key processor units rather than a complete 16-bit RSFQ processor.

The results were reported at ASC 2012 and published in the IEEE Transactions of Applied Superconductivity [1-3].

II. Cell-level design and analysis of key storage components of a superconductor processor implemented with energy-efficient Reciprocal Quantum Logic

During a new phase of the project started in 2013, the design focus shifted from canonical RSFQ logic with large static power consumption to Reciprocal Quantum Logic (RQL) that removed bias resistors used in RSFQ logic, thus eliminating the very significant static power dissipation associated with these resistors.

Our layout-aware RQL design process includes complete cell-level design and approximate physical layout of the circuits followed by the VHDL simulation, verification, and energy profiling using our RQL VHDL cell library. The Stony Brook VHDL RQL cell library specifies the dynamic energy consumption, latency, JJ complexity, and approximate sizes of individual cells based on the input received from the JJ-level RQL designers. A similar approach has been successfully used earlier for RSFQ

chip design in our joint work with HYPRES, Inc. on the development of several 20 GHz RSFQ chips [4-6].

Our focus was on key 32-/64-bit RQL storage components that are expected to be placed alongside RQL functional units, namely: small dual-ported local (aka scratchpad) memories, multiported register files (RF) with 2 read and 1 write ports, and the (closest to a processor) L1 instruction and data caches.

To meet both performance and energy efficiency targets, such as high rate, adequate bandwidth (bits/s), small latency and low energy for both read and write operations, these storage structures were designed with RQL Non-Destructive Read-Out (NDRO) single-bit storage cells.

A 1-read 1-write RQL NDRO cell has set & reset input signals used for write operations, read and data-out for read operations. It is built with three cells, namely a Set/Reset gate, a connection cell, and AnotB gate (7 JJs total). For register files, we designed and used a new 2-read 1-write NDRO2 cell that includes two additional connection cells and another AnotB gate (13 JJs total).

The design of RQL NDRO-based memory arrays is deeply influenced by some intrinsic features of any SFQ logic with pulse-based propagation of signals over transmission lines. There are no tri-state buffers like the ones used in CMOS memories to connect cells to (vertical) bitlines. Vertical data-out bitlines in RQL NDRO-based storage have to be implemented with OR gates and connection cells that merge data-out signals along bit columns.

Horizontal wordlines and any other signals with their fan-out higher than 1 have to be implemented with RQL connection cells arranged in binary tree-like circuit structures providing the required fan-out increase. Unavoidably, all of this leads to a high number of JJs per bit and low storage density (bits/mm^2) for RQL NDRO-based memory, thus limiting its use to small (few Kbits) on-chip storage structures.

The principal techniques used in the 32-/64-bit RQL storage design were pipelining, predecoding with separate predecoders for all read and write addresses, final stage decoders placed in the middle of a memory/register cell arrays, 8-word data slices (sets), and hierarchical bitlines.

2.1 Summary of the most important results

The major results of the work were as follows:

1) Complete cell-level design of RQL 1-4 Kbit 32-/64-bit pipelined storage units, such as multi-ported memory, register files, write-through (WT) and write-back (WB) instruction and data caches using RQL Non-Destructive Read-Out (NDRO) storage cells using SBU RQL VHDL cell library tuned to the MIT Lincoln Laboratory 10 kA/cm^2 248 nm process with 10 Nb metal layers and the minimum JJ critical current of 38 μA.

2) Analysis of the performance, complexity, and energy efficiency of the designed storage units and their scaling trends by breaking down their latency, JJ complexity, and energy consumption across key memory components, such as decoders, writelines, readlines, data-in and data-out bitlines.

The key technical results are:

1) Clock rate: 8.5 GHz (118 ps clock cycle time);
2) JJ complexity range: from 42,512 JJs for a dual-ported 1 Kbit 32x32-bit memory to 253,918 JJs for a triple-ported 4 Kbit 64x64-bit register file;
3) Read access latency range: from 205 ps for a dual-ported 1 Kbit 32x32-bit memory to 338 ps for a 4 Kbit 64-bit write-back cache;
4) Write access latency range: from 236 ps for a dual-ported 1 Kbit 32x32-bit memory to 469 ps for a 4 Kbit 64-bit write-though and write-back caches;
5) Energy efficiency range: 3.0-9.5 fJ/bit/operation at room temperature using the cryocooling efficiency of 0.1%.

In summary, the cumulative energy at room temperature to fetch a 32-bit instruction from a 2 Kbit WT instruction cache, access four 64-bit operands of a double-precision floating-point multiply-add (DFMA) operation from a four-port 4 Kbit register file, transfer 256 bits over distance of 1–20 mm, and execute a DFMA operation will be ~5.5 pJ, meaning ~2.75 KW per DP PFLOPS at room temperature for the 248 nm fabrication processes using the cryocooling efficiency of 0.1%.

While these results are very promising, more work is needed to evaluate the contribution of the energy costs of instruction scheduling and off-chip main memory access to the energy efficiency of RQL computing across a whole system.

The results were reported at ASC 2014 and published in the IEEE Transactions of Applied Superconductivity [7].

The work on the RQL processor design that was planned to continue for three years has been severely hampered by insufficient funding. Less than one third of the funds approved for this phase of work were received. Due to a lack of funds, our work on energy-efficient RQL processors under this grant had to be stopped in January 2015.

References:

1. Dorojevets, M., Kasperek A. K, Yoshikawa, N., Fujimaki, A. 20-GHz 8x8-bit Parallel Carry-Save Pipelined RSFQ Multiplier, IEEE Trans. on Applied Superconductivity, vol. 23, no.3, 1300104, June 2013.
2. Dorojevets, M., Ayala, C. L., Yoshikawa, N., Fujimaki, A. 16-Bit Wave-Pipelined Sparse-Tree RSFQ Adder, IEEE Trans. on Applied Superconductivity, vol. 23, no.3, 1700605, June 2013.
3. Dorojevets, M., Ayala, C. L., Yoshikawa, N., Fujimaki, A. 8-Bit Asynchronous Sparse-Tree Superconductor RSFQ Arithmetic-Logic Unit With a Rich Set of Operations, IEEE Trans. on Applied Superconductivity, vol. 23, no.3, 1700104, June 2013.
4. Filippov, T., Dorojevets, M., Sahu, A., Kirichenko, A., Ayala, C., and Mukhanov, O. "8-bit asynchronous wave-pipelined RSFQ arithmetic-logic unit," IEEE Trans. on Applied Superconductivity, vol. 21, no. 3, pp. 847-851, Jun. 2011.
5. Filippov, T. V., Sahu, A., Kirichenko, A. F., Vernik, I. V., Dorojevets, M., Ayala, C. L., and Mukhanov, O. A. "20 GHz operation of an asynchronous wave-pipelined RSFQ arithmetic-logic unit," Physics Procedia, vol. 36, pp. 59-65, 2012.
6. Kirichenko, A. F., Sahu, A., Filippov, T. V., Mukhanov, O. A., A. V. Dotsenko, A. V., Dorojevets, M., Kasperek, A. K. "Demonstration of an 8×8-bit RSFQ multi-port register file, in Proc. IEEE Intl. Superconductive Electronics Conf. (ISEC-2013), Cambridge, MA, July 7-11, 2013.
7. Dorojevets, M., Chen, Z., Ayala, C. L., and Kasperek, A. K. "Towards 32-bit energy-efficient superconductor RQL processors: The cell-level design and analysis of key processing and on-chip storage units," IEEE Trans. on Applied Superconductivity, vol. 25, no.3, pp.1-8, June 2015.

Technology Transfer

Final Report

Design and Demonstration of a 30 GHz 16-bit Superconductor RSFQ Microprocessor

March 10, 2015

Prof. Mikhail Dorojevets

Dept. of Electrical and Computer Engineering

Stony Brook University, Stony Brook, NY 11794-2350

Abstract. The major objective of the project was to design and demonstrate operation of key components of a 30 GHz 16-bit RSFQ processor prototype implemented with the AIST/ISTEC 10 kA/cm^2 fabrication process. Our team has developed complete logical and physical designs of five RSFQ chips using the CONNECT cell library and RSFQ CAD tools developed at the Universities of Yokohama and Nagoya (Japan). The major results are the world's first successful design, fabrication, and demonstration of correct operation of a 20 GHz 8x8-bit parallel carry-save RSFQ multiplier with ~6K JJs, a 16-bit sparse-tree wave-pipelined RSFQ adder with ~10K JJs, and partial operation of an 8-bit ALU chip with ~9K JJs. The goal of the second phase of the project was to get detailed understanding of the performance, complexity, and energy efficiency of on-chip storage units implemented with superconductor Reciprocal Quantum Logic (RQL) using our RQL VHDL cell library tuned to the MIT Lincoln Laboratory 10 kA/cm^2 248 nm process. The 8.5 GHz 1-4 Kbit 32-/64-bit multi-ported scratchpad memory, register files, write-through and write-back caches designed with RQL Non-Destructive Read-Out storage cells have the average energy consumption of 3.0-9.5 fJ/bit/operation at room temperature using the cryocooling efficiency of 0.1%.

In order to be seriously considered as a competitor to high-performance CMOS processors, superconductor processors need to demonstrate sufficient functionality, complexity, reliability, speed, and energy efficiency in practical designs. To address these challenges, the project had two different tasks discussed below.

I. Design and demonstration of operation of key components of a 30 GHz 16-bit RSFQ processor

We have designed, fabricated, and tested several high-performance RSFQ processing and storage units. The complete designs and physical layouts of all chips have been done at Stony Brook University using the CONNECT cell library and SFQ CAD tools developed at Nagoya and Yokohama Universities (Japan) for the AIST/ISTEC 1.0 μm 10 kA/cm^2 fabrication process (Japan).

Testing of all fabricated RSFQ chips was done by the Stony Brook team with assistance from colleagues at Yokohama National University in 2011-2012.

1.1 A 30 GHz 16-bit RSFQ sparse-tree adder overview

We have designed, fabricated, and demonstrated operation of the first 16-bit RSFQ wave-pipelined sparse-tree adder chip with the core complexity of 9941 JJs and the target operation rate of 30 GHz for a 16-bit integer RSFQ processor.

The microarchitecture of the adder has two main features: 1) a use of a technique of asynchronous hybrid wave-pipelined processing developed at SBU, and 2) a prefix sparse-tree carry generate-propagate structure for arithmetic.

In the data-driven asynchronous hybrid wave-pipelining, data waves "self-propagate" through combinational (non-clocked) logic gates without any need for clock signals. The data waves are followed by reset waves that "clean up" the residual logic states of the gates before the next data wave arrival.

The Kogge-Stone adder (KSA) is considered to be the fastest among parallel-prefix adders. However, KSAs have very high complexity and a tremendous amount of wiring congestion because of the need for their prefix trees to provide carries to every individual bit of the adder. In our 16-bit RSFQ adder design, we chose the sparse-tree structure to reduce the number of Josephson junctions (JJs) needed for its implementation without any significant effect on its processing rate. As a side effect, this will also lead to a more energy-efficient design by reducing the total bias current and power consumption.

Our sparse-tree adder has the following three stages: Initialization, Prefix-Tree, and Summation.

The Initialization stage receives two 16-bit data operands A and B to create bitwise Generate (G) and Propagate (P) signals using clocked AND and XOR gates in a co-flow clocking arrangement.

The Prefix-Tree stage consists of Carry-Merge blocks to merge the prefix signals in a logarithmic manner and provide a group carry to each 4-bit summation block. Merging of the prefix signals is implemented with CFFs (resettable Muller C-flip-flop gates based on the Muller C-element) and confluence buffers used as asynchronous OR gates. The first three levels of the sparse tree also perform the ripple-carry addition within each 4-bit group before data arrive at the Summation stage.

The Summation stage computes the final sum with 4-bit carry-skip adders. The lower-half of the adder (bits 7:0) can start the Summation stage early because all appropriate signals are ready. The upper-half of the adder (bits 15:8) must wait until carries for this upper half are calculated by the very last level of the Prefix-Tree stage.

The 16-bit adder was designed for the target operation frequency of 30 GHz with the latency of 352 ps at the bias voltage of 2.5 mV. The 16-bit adder core (without SFQ-to-dc and dc-to-SFQ converters) has 9941 Josephson junctions occupying an area of 8.5 mm^2. The total number of JJs on the adder chip (including test and I/O circuits) is 12,785 and with a total bias current of 1.61 A. The adder chip was fabricated and successfully tested at low frequency for all test patterns with measured bias margins of +9.8%/ − 10.7%.

1.2 20 GHz 8x8-bit parallel carry-save multiplier overview

We have successfully designed, fabricated, and tested the first 8 × 8-bit (by modulo 256) parallel carry-save superconductor RSFQ multiplier with the target frequency of 20 GHz.

When designing our 8 × 8-bit parallel integer multiplier, we had four major targets: high operation frequency of 20 GHz, multiplication time below 500 ps, complexity around 6000 Josephson junctions (JJs), and mostly regular layout employing both local and global connections.

The 8x8-bit (by modulo 2^8) multiplier operates on two 8-bit input operands and calculates an 8-bit product. The 8x8-bit multiplier has three major components: a partial product generator, partial product compression (reduction) blocks based on 4-to-2 counters, and the final summation block implemented as a ripple-carry adder for the most-significant bits of the product.

The multiplier partial product generator (PPG) consists of 36 partial product (PP) bit generators built with clocked AND gates operating on their multiplicand and multiplier bits. These circuits are organized into three PPG groups, one with 16 and two other with 10 PP generators each. PPs in each PPG group are calculated in parallel, significantly reducing the partial product generation time. Partial products are asynchronously generated and sent to the reduction stage at the internal "hardwired" rate of 80 GHz.

The 8 × 8-bit RSFQ multiplier uses a two-level parallel carry-save reduction tree that significantly reduces the multiplier latency. The 80-GHz carry-save reduction is implemented with asynchronous data-driven wave-pipelined [4:2] compressors (counters) built with toggle flip-flop cells. First, up to 8 PPs in each column are reduced to 4 by two [4:2] compressors working in parallel, each producing 2 PPs. The 4 PPs from the two first-level compressors are merged together with asynchronous confluence buffers and sent ~12.5 ps apart over a single PTL to a second-level [4:2] compressor for that column. Then, the second-level [4:2] compressor will reduce those 4 PPs to 2. The benefits of using this approach are as follows: 1) the O(log2n) PP reduction time, where n is the operand length, and 2) a regular layout with both local and global connections between modules.

The five least significant bits of the product are calculated during the PP reduction by the [4:2] compressors. The partial products in the three most significant bit columns are reduced to carry-sum pairs and then go through the final summation done by a wave-pipelined ripple-carry adder.

The 8 × 8-bit RSFQ multiplier has the latency of 447 ps at the nominal bias voltage of 2.5 mV and the acceptable DC bias margins at the target frequency of 20 GHz when operating at a slightly higher than nominal bias voltage. The multiplier core (without DC-to-SFQ and SFQ-to-DC converters) is built with 5948 JJs occupying the area of 3.5 mm^2 (2.5 mm × 1.45 mm). It has a bias current of 676 mA.

The multiplier chip was fabricated in December 2011. Despite some challenges due to fabrication process parameter variations and flux trapping, the multiplier chip was successfully tested for the vast majority of test vectors by the Stony Brook designers in February 2012. While multiplier test operations were generated at low frequency, each of these operations was executed at the internal "hardwired" rate of 80 GHz. The fabricated chip operated with the measured DC bias margins of ±5%.

1.3 30 GHz 8-bit ALU block overview

The design of a new 8-bit asynchronous RSFQ ALU was the next step in the development of bit-parallel RSFQ ALUs compared to an 20 GHz 8-bit ALU earlier developed by our team in collaboration with HYPRES, Inc. First, the operation set is significantly extended, including the introduction of a true two's complement subtract operation. Second, we have developed a sparse-tree design with significantly less complexity of the carry-generate-propagate prefix-tree than in the earlier ALU design.

Our 8-bit ALU has three key microarchitecture and design features: asynchronous wave-pipelined processing, a prefix sparse-tree carry generate-propagate structure, and data pulse counters for implementation of complex operations.

This unit features an extensive set of 8 arithmetic and 12 logical operations. The execution of ALU operations consists of two steps. First, when necessary, one or both operands are inverted, and then operations are performed on these pre-processed data.

List of logical operations:

NOP, AND A, B; OR A, B; XOR A, B; Set All 1s; AND A, ¬B; OR A, ¬B; AND ¬A, B; OR ¬A, B; XNOR A, B; NOR A, B; NAND A, B.

List of arithmetic operations and their mathematical representations:

1. ADD A, B ⇒ A + B

2. ADD ¬A, B ⇒ B − A − 1

3. ADD A, ¬B ⇒ A − B − 1

4. ADD ¬A, ¬B ⇒ −(A + B + 2)

5. INC_ADD A, B ⇒ A + B + 1

6. SUB A, B ⇒ A − B

7. SUB B, A ⇒ B − A

8. INC_ADD ¬A, ¬B ⇒ −(A + B + 1)

The entire ALU chip layout including moats, I/O and the test circuit fits within the 4.25 mm × 4.00 mm area with the core circuit with the complexity of 8832 JJs requiring the area of 7.2 mm^2 (2.67 mm × 2.70 mm). It has the simulated DC bias margins of +20%/−16% at the target processing rate of 30 GHz.

Testing of the chip was done at low frequency to check for correct functionality. Testing for some operations, such as XNOR, showed incorrect results due to the malfunctioning of two gates from the CONNECT cell library. Despite these shortcomings we still verified several other ALU operations, e.g. XOR, ADD, INC_ADD, that did not use those gates. The measured overlapped DC bias margins for these operations were ±1.8%.

1.4 30 GHz 8-bit dual-register block overview

The 8-bit dual-register slice is the component of a register file with two read and one write ports. The slice consists of two 8-bit wide registers (left and right registers that provide left and right operands for ALU) with non-destructive read-out.

The fabricated 2x8-bit register slice has a core complexity of 1,060 JJs**.** A complete circuit (with test and I/O circuits included) has 1,806 JJs, an area of 1.06 mm^2, and a total bias current of 203 mA.

The register file operation set includes read_left, read_right, write _left, and write_right operations. Before the write operations, the register is reset by reset_left/ reset_right operation depending on which of them will be written with data from ALU. Two registers (one left and one right) can read and one register written each cycle.

The simulation results show that the 8-bit dual-register block design has very wide DC bias margins (greater than ±20%) to operate at the target 30 GHz clock rate.

1.5 Summary of the most important results

The major results of this work are:

1) the world's first successful demonstration of a 16-bit superconductor wave-pipelined RSFQ adder with its complexity of ~10K JJs and target frequency of 30 GHz;

2) the world's first successful demonstration of an 8x8-bit parallel carry-save RSFQ multiplier with its complexity of ~6K JJs and target frequency of 20 GHz;

3) demonstration of partial operation of 8-bit wave-pipelined ALU with its complexity of ~9K JJs and target frequency of 30 GHz.

The results of our work have proven the viability of the hybrid wave-pipelined approach developed by our team at Stony Brook for practical 30 GHz superconductor processor design. Also, the experience of working with the Japanese digital CAD tools, their cell library and fabrication facility has allowed us to clearly understand the benefits and shortcomings of the semi-custom design flow similar to the one used in Japan over a full-custom design flow currently used in the US.

Among those advantages is the separation of the analog (JJ-level) and digital design flows. The latter is based on the standard cell library with cells having their HDL behavioral models and fixed shapes (layouts), providing the opportunity to develop logic synthesis tools for superconductor VLSI circuits in the future.

There were several challenges during this period of work that have significantly influenced the execution of the project.

1) The AIST chip foundry was damaged during the earthquake in March 2011. After more than a six-month operation recovery period, the foundry resumed fabrication by 2012 but the yield, chip quality, and throughput were found to be low due to some unpredictable variations in the fabrication process parameters.

2) As a result of this, the number of new chips allocated to the SBU team in 2012 was limited to one.

3) The foundry decided to stop fabrication in the summer of 2012 in order to upgrade their clean room equipment.

All of this has forced the SBU team to focus on the design and demonstration of key processor units rather than a complete 16-bit RSFQ processor.

The results were reported at ASC 2012 and published in the IEEE Transactions of Applied Superconductivity [1-3].

II. Cell-level design and analysis of key storage components of a superconductor processor implemented with energy-efficient Reciprocal Quantum Logic

During a new phase of the project started in 2013, the design focus shifted from canonical RSFQ logic with large static power consumption to Reciprocal Quantum Logic (RQL) that removed bias resistors used in RSFQ logic, thus eliminating the very significant static power dissipation associated with these resistors.

Our layout-aware RQL design process includes complete cell-level design and approximate physical layout of the circuits followed by the VHDL simulation, verification, and energy profiling using our RQL VHDL cell library. The Stony Brook VHDL RQL cell library specifies the dynamic energy consumption, latency, JJ complexity, and approximate sizes of individual cells based on the input received from the JJ-level RQL designers. A similar approach has been successfully used earlier for RSFQ chip design in our joint work with HYPRES, Inc. on the development of several 20 GHz RSFQ chips [4-6].

Our focus was on key 32-/64-bit RQL storage components that are expected to be placed alongside RQL functional units, namely: small dual-ported local (aka scratchpad) memories, multiported register files (RF) with 2 read and 1 write ports, and the (closest to a processor) L1 instruction and data caches.

To meet both performance and energy efficiency targets, such as high rate, adequate bandwidth (bits/s), small latency and low energy for both read and write operations, these storage structures were designed with RQL Non-Destructive Read-Out (NDRO) single-bit storage cells.

A 1-read 1-write RQL NDRO cell has set & reset input signals used for write operations, read and data-out for read operations. It is built with three cells, namely a Set/Reset gate, a connection cell, and AnotB gate (7 JJs total). For register files, we designed and used a new 2-read 1-write NDRO2 cell that includes two additional connection cells and another AnotB gate (13 JJs total).

The design of RQL NDRO-based memory arrays is deeply influenced by some intrinsic features of any SFQ logic with pulse-based propagation of signals over transmission lines. There are no tri-state buffers like the ones used in CMOS memories to connect cells to (vertical) bitlines. Vertical data-out bitlines in RQL NDRO-based storage have to be implemented with OR gates and connection cells that merge data-out signals along bit columns.

Horizontal wordlines and any other signals with their fan-out higher than 1 have to be implemented with RQL connection cells arranged in binary tree-like circuit structures providing the required fan-out increase. Unavoidably, all of this leads to a high number of JJs per bit and low storage density (bits/mm^2) for RQL NDRO-based memory, thus limiting its use to small (few Kbits) on-chip storage structures.

The principal techniques used in the 32-/64-bit RQL storage design were pipelining, predecoding with separate predecoders for all read and write addresses, final stage decoders placed in the middle of a memory/register cell arrays, 8-word data slices (sets), and hierarchical bitlines.

2.1 Summary of the most important results

The major results of the work were as follows:

1) Complete cell-level design of RQL 1-4 Kbit 32-/64-bit pipelined storage units, such as multi-ported memory, register files, write-through (WT) and write-back (WB) instruction and data caches using RQL Non-Destructive Read-Out (NDRO) storage cells using SBU RQL VHDL cell library tuned to the future MIT Lincoln Laboratory 10 kA/cm^2 248 nm process with 10 Nb metal layers and the minimum JJ critical current of 38 μA.

2) Analysis of the performance, complexity, and energy efficiency of the designed storage units and their scaling trends by breaking down their JJ complexity and energy consumption across key memory components, such as decoders, writelines, readlines, data-in and data-out bitlines.

The key technical results are:

1) Clock rate: 8.5 GHz (118 ps clock cycle time);

2) JJ complexity range: from 42,512 JJs for a dual-ported 1 Kbit 32x32-bit memory to 253,918 JJs for a triple-ported 4 Kbit 64x64-bit register file;

3) Read access latency range: from 205 ps for a dual-ported 1 Kbit 32x32-bit memory to 338 ps for a 4 Kbit 64-bit write-back cache;

4) Write access latency range: from 236 ps for a dual-ported 1 Kbit 32x32-bit memory to 469 ps for a 4 Kbit 64-bit write-though and write-back caches;

5) Energy efficiency range: 3.0-9.5 fJ/bit/operation at room temperature using the cryocooling efficiency of 0.1%.

In summary, the cumulative energy at room temperature to fetch a 32-bit instruction from a 2 Kbit WT instruction cache, access four 64- bit operands of a double-precision floating-point multiply-add (DFMA) operation from a four-port 4 Kbit register file, transfer 256 bits over distance of 1–20 mm, and execute a DFMA operation will be ~5.5 pJ, meaning ~2.75 KW per DP PFLOPS at room temperature for the 248 nm fabrication processes using the cryocooling efficiency of 0.1%.

While these results are very promising, more work is needed to evaluate the contribution of the energy costs of instruction scheduling and off-chip main memory access to the energy efficiency of RQL computing across a whole system.

The results were reported at ASC 2014 and published in the IEEE Transactions of Applied Superconductivity [7].

The work on the RQL processor design that was planned to continue for three years has been severely hampered by insufficient funding. Less than one third of the funds approved for this phase of work were received. Due to a lack of funds, our work on energy-efficient RQL processors under this grant had to be stopped in January 2015.

References:

1. Dorojevets, M., Kasperek A. K, Yoshikawa, N., Fujimaki, A. 20-GHz 8x8-bit Parallel Carry-Save Pipelined RSFQ Multiplier, IEEE Trans. on Applied Superconductivity, vol. 23, no.3, 1300104, June 2013.

2. Dorojevets, M., Ayala, C. L., Yoshikawa, N., Fujimaki, A. 16-Bit Wave-Pipelined Sparse-Tree RSFQ Adder, IEEE Trans. on Applied Superconductivity, vol. 23, no.3, 1700605, June 2013.

3. Dorojevets, M., Ayala, C. L., Yoshikawa, N., Fujimaki, A. 8-Bit Asynchronous Sparse-Tree Superconductor RSFQ Arithmetic-Logic Unit With a Rich Set of Operations, IEEE Trans. on Applied Superconductivity, vol. 23, no.3, 1700104, June 2013.

4. Filippov, T., Dorojevets, M., Sahu, A., Kirichenko, A., Ayala, C., and Mukhanov, O. "8-bit asynchronous wave-pipelined RSFQ arithmetic-logic unit," IEEE Trans. on Applied Superconductivity, vol. 21, no. 3, pp. 847-851, Jun. 2011.

5. Filippov, T. V., Sahu, A., Kirichenko, A. F., Vernik, I. V., Dorojevets, M., Ayala, C. L., and Mukhanov, O. A. "20 GHz operation of an asynchronous wave-pipelined RSFQ arithmetic-logic unit," Physics Procedia, vol. 36, pp. 59-65, 2012.

6. Kirichenko, A. F., Sahu, A., Filippov, T. V., Mukhanov, O. A., A. V. Dotsenko, A. V., Dorojevets, M., Kasperek, A. K. "Demonstration of an 8×8-bit RSFQ multi-port register file, in Proc. IEEE Intl. Superconductive Electronics Conf. (ISEC-2013), Cambridge, MA, July 7-11, 2013.

7. Dorojevets, M., Chen, Z., Ayala, C. L., and Kasperek, A. K."Towards 32-bit energy-efficient superconductor RQL processors: The cell-level design and analysis of key processing and on-chip storage units," IEEE Trans. on Applied Superconductivity, vol. 25, no.3, pp.1-8, June 2015.