BEST UTILIZATIONS OF LOW-ENERGY HIGH-SPEED ADVANTAGES OF SUPERCONDUCTING LOGIC

M. Pedram
USC ECE
pedram@usc.edu

http://coldflux.usc.edu/
Talk given at Foundations of Superconducting Digital Logic Workshop

Aug. 1, 2022
Going Faster with Less Energy

- Very high clock frequencies
  - High single-threaded performance through short switching times and fast signal propagation
- Ultra low switching energy
  - Energy related to the generation of the bit information should be made as small as possible

Superconductor circuits meet both of the aforesaid requirements and have become a strong candidate for a beyond-CMOS computing fabric. This technology is known for the high clock frequencies, 2–50 GHz, and the low energy dissipation per logical operation, down to several zJ [1].


Case for Superconductor Circuits

Energy per ALU operation

Neural network inference energy consumption (TOPs/Joule)
Single Flux Quantum (SFQ) Logic

- **Pulse-based logic family**, inherently synchronous, limited fanout drive

- Basic element of RSFQ logic is superconducting loop interrupted by one or more JJs
  - If the loop inductance $L$ is high enough such that $I_C L \geq \Phi_0$, then SFQ can be held in the loop as a persistent current representing logical one, whereas an absence of SFQ means logical zero
  - Magnetic flux ejected from a superconducting loop through a JJ takes the form of **tiny voltage pulses** (mV, ps)

- Building blocks of SFQ circuits
  - Transfer and storage sections
  - Decision-making pairs
Unique Characteristics of SFQ Logic

Different active and passive components
- Two-terminal JJ’s and inductors vs. 3-terminal transistors and capacitors in CMOS

Superconducting interconnect
- Josephson transmission lines (JTLs) and passive transmission lines with small series resistors to prevent flux storage (PTLs) vs. lossy interconnect in CMOS

Different suites of basic cells
- Simple clocked cells (NOT, AND, OR, XOR, DFF, NDRO) and some clockless cells (delay cells, splitters) vs. complex multi-input logic cells in CMOS

The RSFQ main advantage driving its development has been the high speed at very low overall power consumption. Major disadvantage has been the large SQUID cell size severely limiting the integration level. Furthermore, the parallel bias lines carry high currents, and associated magnetic fields must be shielded, which is difficult. The clock network requires a significant overhead in JJs number.
Advantage: Passive Transmission Lines

- Superconducting transmission is dispersion free, i.e., we have **ballistic propagation of an SFQ pulse** [5]
- Speeds up to 1/3 of the speed of light are achieved (**Typical delay: 10 ps/mm**)
- Needs PTL driver and receiver circuits, **Delay scales linearly with PTL length** [6].
- Must be impedance matched to the JJs of driver and receiver circuits. Junctions in the latest fabrication processes allow about 5 to 10 Ω lines, which reduces line width for tighter integration.

**PTLs are a highly appealing feature in RSFQ circuits.**

SFQ Challenges and Potential Solutions

### Technical Challenges and Open Problems

#### Socioeconomic Impacts

<table>
<thead>
<tr>
<th>Economic impact</th>
<th>Climate/environmental impact</th>
<th>Statistical and data analytics methods</th>
</tr>
</thead>
</table>

#### SuperSoCC Design

<table>
<thead>
<tr>
<th>Superconductive MCM</th>
<th>Application optimization and mapping</th>
<th>Software stack development</th>
</tr>
</thead>
</table>

#### Integration and Interfaces

<table>
<thead>
<tr>
<th>Electrical-thermal-mechanical optimization</th>
<th>Passive and active integration technologies</th>
<th>Cables, clocks, and bias lines</th>
</tr>
</thead>
</table>

#### Circuits and Architecture

<table>
<thead>
<tr>
<th>Multi-bit SFQ and temporal logic circuits</th>
<th>All-JJ, 2φ-JJ, and DSFQ logic</th>
<th>Dual clocking, current recycling, and power biasing</th>
</tr>
</thead>
</table>

#### Devices and Materials

<table>
<thead>
<tr>
<th>Nanowire devices e.g., nTron, QPSJ</th>
<th>Bistable magnetic JJs</th>
<th>Superconducting ferromagnetic transistors</th>
</tr>
</thead>
</table>
Story of Two Timelines

1911: Superconductivity discovered
1950: Ginzburg-Landau theory
1957: BCS theory

1962: Josephson effect
1967: Superconducting kinetic inductance

1973: Josephson vortex logic
1974: Superconducting analog microwave electronics
1975: Josephson parametric amplifier

1986: High-temperature superconductivity
1987: RSFQ superconducting logic

1990: First superconductive microprocessor

1999: Design tools for RSFQ
2000: Superconducting qubits

2006: Fe-based superconductors

2000: 1 Gbit DRAM, chip-scale package
2002: Intel Itanium 2 (221M transistors, 900MHz, 180nm)
2002: Earth Simulator achieves 35.9 TFLOPS
2004: Graphene creation experiments
2007: Apple announces the iPhone
2009: Intel 64 Nahalem (781M transistors, 3.33GHz, 45nm)
2011: K computer achieves 10 PFLOPS
2012: Intel 8-core Itanium (31B transistors, 2.66GHz, 32nm)
2014: NVIDIA GeForce GT1030, 3D ICs
2016: AlphaGo defeats a human Go master

2021: Fugaku supercomputer achieves 442 PFLOPS
2021: First AQFP microprocessor (20K jj, 2.5GHz)
2018: AQFP stochastic computing for deep learning
2017: Quantum-classical interface based on SFQ logic
2020: First full version of ColdFlux Tools released

Superconductor technology

Semiconductor technology

Pedram (2022)
IARPA SuperTools Program: ColdFlux Tools

Device level
- Device design
  - FLOOX-SFQ: TCAD
    - Compact SPICE model extraction

Cell/gate level
- Fabrication process
  - XIC: Schematic capture
    - JoSiM: Electrical simulation
      - JoSiM add-on: Margin/yield analyzer and optimizer
  - XIC/Absynth: Cell layout
    - Auron: Layout-vs-Schematic verification
    - InductEx, FFH, TetraHenry: Electrical parameter extraction
    - InductEx, TetraHenry: Compact SPICE model extraction
    - TimEx: HDL model generation

Chip/system level
- Behavioral description
  - qSyn: Logic synthesis
  - qVSim: Logic level simulator
  - qCTS: Clock tree synthesis
  - Timing verification and optimization
    - qPlace, qGDR: Place-and-Route (Layout synthesis)
    - DRC, LVS: Design rule checker, Layout-vs-Schematic
    - qSSTA, qFsim, qPA: STA and functional verification
    - ChipSmith: GDS-II
NSF DISCoVER Expedition

Design & Integration of Superconductive Computation For Ventures beyond Exascale Realization

- **Materials and devices**
  - JJ s with engineered CPRs
  - Self-shunted JJs
  - Non $\Omega$, $\Phi$, $2\Phi$ JJs
  - 3-terminal SC devices
  - SC nanowire devices
  - Epitaxial JJs

- **Architectures and design tools**
  - Compilation/mapping
  - Runtimes
  - Mixed-signal design
  - Synthesis/retiming
  - Dual/multi-phase clocking
  - Power/thermal modeling

- **Circuits and interfaces**
  - Dense memory
  - Multi-threading support
  - Neural networks
  - Temporal logic
  - Ising machines
  - Cross temperature interface

https://discoverexpedition.usc.edu/
SFQ-Based Multi-Function Accelerator
Characteristics of Computing Applications Suitable for Superconductor Electronics

- The problem has usually been the cost of the superconducting versus non-superconducting solutions.
  - Future successes in reducing the cost, size, weight, unreliability, etc. of cryogenic equipment will have a direct and strong impact on how quickly various applications can be commercialized.
- Computing applications: Feedforward pipelined architectures with minimum stalls and pipeline flushes.
- Other applications: Benefit from features such as voltage-controlled oscillation, fast digital-to-analog conversion, operation at ultracold temperatures, true random number generation, etc.
Common Applications

• In electronics, the Josephson effect has enabled devices such as SQUIDs (superconducting quantum interference devices) for the most accurate measurement of magnetic fields, voltage (e.g., the Josephson Voltage Standard), and related electromagnetic quantities.

• Researchers have also experimented with SQUID RF amplifiers, superconducting low-inductance undulating galvanometer (SLUG), SQUID magnetometers (with flux-locked feedback loops), Josephson parametric oscillator, and A/D converters.

• In addition to the magnetic resonance imaging, one of the largest commercial use at present of superconducting electronic devices (HTS) is as filters in cell phone base stations.

• Superconductors are also used for the highest Q resonant cavities, particularly in high-energy particle accelerators.
Ising Hamiltonian solver based on Josephson parametric oscillators

- Adiabatic annealing involves the encoding of an optimization problem using the interactions between classical variables that can take the values $\pm 1$.
- The combinatorial optimization problem is then cast in the form of an all-to-all Ising spin glass model:

$$\mathcal{H} = \sum_{i,j<i} J_{ij} \sigma_i \sigma_j + \sum_i h_i \sigma_i$$

where $\sigma_i \in \{-1, +1\}$ represents the $i$th spin.
- The interaction matrix $J_{ij}$ and the additional local magnetic fields $h_i$ fully parameterize the optimization problem.
- The task of finding the optimal solution amounts to finding the ground state of $\mathcal{H}$. 

Logical bit count: $N = 4$
Fully Homomorphic Computation

- FHE schemes use these two primary parameters to tune the provided security and the supported depth of homomorphic computation.
- Example values of $n$ and $\log_2(q)$ are 4,096 and 1,024, respectively.
- The challenge comes from the difficulty of supporting both a large ring dimension $n$ (which provides comparatively better security) and a large $\log_2(q)$ (which increases the depth of supported computations).

NTT: Number Theoretic Transform

Point-wise Mult.

Ring Modular Multiplication

Total 3 stages
Neural Network Acceleration

Layer 1

\[ A_{1,0} = W_{0,0} \times X_0 + W_{0,3} \times X_3 + W_{0,5} \times X_5 \]
Standard HD classification system are comprised of three components:
1. Input feature extractor mapping input signal to a feature vector
2. HD encoder maps low-dimensional feature vector into HD vector
3. HD classifier maps HD vector into a classification decision

Unfortunately, feature computation is highly application specific, not trained as part of design, and highly compute-intensive thus dominating the HD processing complexity and thus overall system complexity.
Stochastic Computing

One can use stochastic computing to realize key computational blocks of neurons, e.g., *inner product calculation*, pooling, normalization, and *activation function* application.

---

Pedram (2022)

---

Error Correction and Control of Quantum Computers

Massive MIMO systems can generate potentially prohibitive amounts of data due to the large numbers of antennas. With modern parallel, low-rate analog-to-digital converters (ADCs), the bottleneck is often not in the quantization of the received signals but, rather, in the processing of the digitized bits.

To separate the $n$ transmitted signals in the receiver, a commonly used scheme linearly combines the received signals using a set of weights that yields the minimum mean square error (MMSE) between the detected data and the true signal samples.
Conclusion

• Because of the cryogenic burden, new superconducting applications are most likely to achieve early, widespread success in situations where there are clear performance and energy efficiency advantages.

• In spite of much progress in SCE, key challenges exist in way of physical scaling, controlling stray electromagnetic fields, supporting multiple clocks, current recycling, design centering to improve operating margins, designing dense on-chip memory, etc.