CHALLENGES AND OPPORTUNITIES FOR SUPERCONDUCTIVE COMPUTING

M. Pedram pedram@usc.edu and T. Pinkston tpink@usc.edu
USC ECE

https://discoverexpedition.usc.edu/
Talk given at TAPIA Conference

Sep. 8, 2022
CMOS Scaling and Moore’s Law

- Moore’s law is coming to end
  - Physical scaling limit
  - Thermal design power limit
  - Current delivery limit
- Computing technology advances and economic growth have relied on continuation of Moore’s law
- Huge research and development underway to find compelling beyond-CMOS computing technology
Emerging Beyond CMOS Devices

Emerging Post-CMOS Computing Paradigms

Neuromorphic computing
Superconductive computing
Stochastic computing
Molecular-scale computing
Quantum computing
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

Case for Superconductor Electronics

![Graph showing energy per ALU operation and neural network inference energy consumption (TOPs/Joule)]
Single Flux Quantum Logic

- Bardeen–Cooper–Schrieffer theory: Superconductivity is the condensation of bound pairs of electrons, where these Cooper pairs are connected through electron-phonon interactions.

- Logic circuits based on magnetic flux quantization rather than conductance modulation.

- Allows discrete representation of information in the form of Single Flux Quantum (SFQ).

\[ \Phi_0 = \frac{\hbar}{2e} \approx 2.07 \times 10^{-15} \text{ volt-second} \] (ampere-henry=weber)

\( \hbar \) is the Planck constant, \( e \) is the electron charge.

Only an integer number of flux quanta can exist in a superconductive loop.
**Josephson junction (JJ)** is the nonlinear switching element in superconductive circuits.

- **DC Josephson effect**: Constant DC current bias $I_j < I_C$ results in constant phase difference $\phi$ and zero voltage across JJ.

- **AC Josephson effect**: Constant voltage $V_j$ applied across JJ results in linearly increasing phase difference and current oscillating at $f = 2\pi V_j(t)/\Phi_0$.

- Resistively capacitively shunted junction (RCSJ) model

Most common junction for superconductor electronics is SIS planar junction: Nb/Al-AlOx/Nb @4.2K

Overdamped JJ (small R and C); this is the primary type of JJ.

Superconducting state phase difference $\phi$ exists over JJ, producing supercurrent:

$$I_j(t) = I_C \sin \phi(t) \text{ with } d\phi/dt = 2\pi V_j(t)/\Phi_0$$

where $I_C$ denotes JJ critical current and $\Phi_0 = \hbar/(2e)$

(Josephson equations define **current-phase relationship** and its evolution)
Rapid Single Flux Quantum (RSFQ) 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
Transferring and storing SFQ pulses

Josephson Transmission Line (JTL)

DFF
RSFQ Logic Gates

AND2

OR2

Pedram (2022)
Fabrication Technology

- Existing fabrication processes: ISTEC-AIST, 2.5 kA/cm² (Japan), MIT-LL 10 kA/cm² (USA).

- SIS JJ
- 8 metal layers: M0-M7
- 7 via layers I0-I6 plus special vias C0, C5R, and C5J
- Pad layer: M8
- Resistor layer: R5
- JJ layer: J5
- High kinetic inductance layer: L0
- J5-M6 via: C5J
- R5-M6 via: C5R
- M0-C0 via: C0

Unique Characteristics of SFQ Logic

Unique Features

- 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

Circuit-Level Challenges

- Expensive to implement feedback loops
- Low drive capability
- Path balancing overhead
- Small number of routing layers
- Clocking overhead
- Large bias current requirement
- Sensitivity to stray magnetic fields and flux trapping
- Path balancing overhead

Pedram (2022)
Advantage: Passive Transmission Lines

- Superconducting transmission is dispersion free, i.e., we have **ballistic propagation of an SFQ pulse**
- 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**
- 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

One big question has been how much the energy needed for cooling will increase a superconducting computer’s energy budget. Even so, the power dissipated in a superconducting computer is so small that it remains 100 times more efficient than a comparable silicon computer, even after taking into account the present inefficient cryocooler.
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
1976: High-temperature superconductivity
1987: RSFQ superconducting logic
1990: First superconductive microprocessor
1999: Design tools for RSFQ
2000: Superconducting qubits
2006: Fe-based superconductors

1948: Junction-type transistor
1953: FET theory published
1957: Esaki diode
1959: Fairchild and TI announce IC
1961: Prototyping of MOSFET at Bell Labs
1965: Moore’s law
1967: First DRAM prototyped at IBM
1971: Intel 4004 (2.3K transistors, 740kHz, 10um)
1974: Motorola M6800 (1MHz)
1977: Apple II, Intel 8086 (16-bit)
1978: Intel 8086 (29K transistors, 5MHz)
1980: Flash memory
1983: Nintendo’s video game console
1989: Intel 486 (1.2M transistors, 25MHz, 1um)
1991: Carbon nanotube
1993: Intel Pentium (3.1M transistors, 60MHz, 0.8um)
1994: Ball grid array
2000: 1 Gbit DRAM, chip-scale package
2002: Intel Itanium 2 (221M transistors, 900MHz, 180nm )
2004: Earth Simulator achieves 35.9 TFLOPS
2007: Apple announces the iPhone
2009: Intel 64 Nehalem (781M transistors, 3.33GHz, 45nm)
2011: K computer achieves 10 PFLOPS
2012: Intel 8-core Itanium (3.1B transistors, 2.66GHz, 32nm)
2014: NVIDIA GeForce GT1030, 3D ICs
2016: AlphaGo defeats a human Go master
2021: Fugaku supercomputer achieves 442 PFLOPS

Superconductor technology

Semiconductor technology

180nm Bulk
130nm
90nm
65nm
45nm
32nm
22nm
16nm
14nm FinFET
5nm
3nm
ColdFlux: Design Methodologies and Tools
qPALACE: ColdFlux Frontend Design Tools
DISCoVER Expedition

USC Viterbi-Led Team Wins $15 Million NSF Expeditions in Computing Award, One of Only Two Announced This Year

Amy Blumenthal | April 22, 2022

Massoud Pedram will be PI for multi-university DISCoVER team focused on green computing, superconductor electronics

Design & Integration of Superconductive Computation For Ventures beyond Exascale Realization
DISCoVER Expedition Goals

- Achieve 100X energy efficiency compared to CMOS at same performance level (including cryogenic cooling)
- Enable very large scale integration (VLSI) of SCE
- Demonstrate superconductive system of cryogenic computing cores (SuperSoCC)
- Realize compute-intensive applications
- Develop technology and distribute knowledge with emphasis on enhancing collaboration and inclusion, fostering innovation and technology transfer, and broadening participation
- Train new generation of diverse workforce for beyond-CMOS world
- Facilitate broader societal adoption
DISCoVER Expedition Team

Ivan Petrovich Nevirkovets
Julia Albright
Mark Bocko
Nobuyuki Yoshikawa
Oleg Mukhanov
Christopher Ayala
Roman Sobolewski
Selcuk Kose
Timothy Pinkston
Yanzhi Wang

Massoud Pedram
Expediton Director

Lead: Eby Friedman
Circuits and architecture
Also Expedition Associate Director

Lead: Murali Annavaram
SuperSoCC design

Lead: Yingying Fan
Socioeconomic impacts

Lead: Grace Xing
Devices and materials

Lead: Michael Hamilton
Integration and interfaces
Vision

• Develop and demonstrate superconductor electronics (SCE) and superconductive computing technology
• Innovate in technology, circuit design, and architectures with targeted application studies and system demonstrations
• Greatly reduce energy requirements of national computing infrastructure
Vision

- Create and promote culture of diversity, equity, and inclusion
- Rich program of student training and participation at all levels
- Instructional modules, new courses, distributive classroom, and MS degree program
- Extensive outreach and BPC activities utilizing NSF i-hubs and national STEM ecosystems
### Challenges and Proposed Solutions

#### Technical Challenges and Open Problems

<table>
<thead>
<tr>
<th>Technical Challenges and Open Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Languages, compilers, and runtimes</td>
</tr>
<tr>
<td>Architecture and multi-threading</td>
</tr>
<tr>
<td>Interfaces across temperature zones and technologies</td>
</tr>
<tr>
<td>Current delivery, clocking, and synchronization</td>
</tr>
<tr>
<td>Dense on-chip memory</td>
</tr>
<tr>
<td>Stray electromagnetic fields</td>
</tr>
<tr>
<td>Physical scaling and integration</td>
</tr>
</tbody>
</table>

#### Socioeconomic Impacts

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

#### SuperSoCC Design

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

#### Integration and Interfaces

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

#### Circuits and Architecture

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

#### Devices and Materials

<table>
<thead>
<tr>
<th>Devices and Materials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nanowire devices e.g., nTron, QPSJ</td>
</tr>
<tr>
<td>Bistable magnetic JJs</td>
</tr>
<tr>
<td>Superconducting ferromagnetic transistors</td>
</tr>
</tbody>
</table>
Research Program

- SuperSoCC design
- Integration and interfaces
- Circuits and architecture
- Devices and materials
- Socioeconomic impacts
  - Industry and application requirements
  - Performance, cost, and energy models
  - Socioeconomic impact models and policy recommendations
- Logic families and custom building blocks
  - Interface and I/O specs
  - Electrical and thermal specs
  - HW/SW mapping and performance specs
- Models and devices
  - Device performance specs
Supercomputing System of Cryogenic Cores (SuperSoCC)
Computing Applications of Superconductor Electronics

• The problem has usually been the cost of the superconducting versus non-superconducting solutions.
• 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.
Neural Network Acceleration

\[
A_{1,0} = W_{0,0} \cdot x_0 + W_{0,3} \cdot x_3 + W_{0,5} \cdot x_5
\]

16 Input Features

Layer 1 weights

 Activation Data Buffer

Weight Buffer

weights
activations
Intermediate results
Splitter

MUL
ADD
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).

**Traditional Encryption**

**Fully Homomorphism Encryption**

**Ring Modular Multiplication**

**NTT: Number Theoretic Transform**

Data in uncontrolled/unsafe state

Data in controlled/safe state
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*.

Error Correction and Control of Quantum Computers

Transformative Concept Framework

- Superconductor electronics poised to disruptively impact digital integrated systems and high-performance computing
- This DISCoVER Expedition, guided by our team’s expertise and vast prior experience, will develop and demonstrate computing systems built from superconductive electronics
- Produce a new generation of well-trained, diverse engineers in post-CMOS technologies