Multi-Temperature Zone Optimization of Cryogenic Systems

Nurzhan Zhuldassov, Rassul Bairamkulov, and Eby G. Friedman

Department of Electrical and Computer Engineering
University of Rochester
Rochester, New York 14627
nzhuldas@ur.rochester.edu

***Abstract*—Heterogeneous computing exploits several disparate technologies within a single system. The different components of a heterogeneous system are often placed within separate temperature zones. Selecting an appropriate operating temperature strongly affects the dissipated power, cooling power (heat load), system performance, and ambient temperature. To this date, no multi-temperature design methodology exists. To overcome this limitation, a framework for thermal optimization of heterogeneous computing systems is presented. The effects of operating temperature on delay and power consumption are characterized based on a graph theoretic representation of the system. In addition, thermal interactions among the components within a system are considered to accurately evaluate the total power consumption and local heat load. In a practical case study, the target temperature of each component within a quantum computing system is determined to minimize the total power under target performance constraints.**

***Index Terms*—Thermal optimization, cryoCMOS, SFQ, quantum computing, quantum-classical computer**

I. Introduction

The demand for high performance computing (HPC) has greatly increased over the past several decades, driven by the rise in computationally intensive, large scale applications, particularly cloud computing. Further advancements in HPC systems require overcoming a large number of challenges, including energy efficiency, thermal management, and system performance. The energy consumption of a typical data center ranges from tens to hundreds of megawatts [1]. The annual global energy consumption for HPC is estimated at 200 TWh, and is expected to increase fourfold by 2030 [2]. Qualitatively different computational technologies are necessary to sustain this rapid growth in computing.

This research is supported in part by the National Science Foundation under Grant No. 2124453 (DISCOVER Expeditions) and Grant No. 2308863, Department of Energy under FOA number DE-FOA-0002950, and by a grant from Qualcomm Corporation.

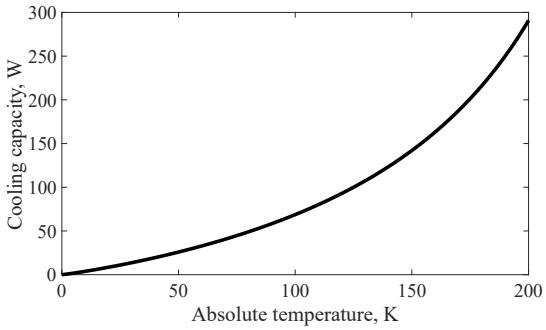

Fig. 1: Available cooling power per 1 kW input power. The data are based on the technical specifications of commercially available cryogenic coolers [6]–[9].

Cryogenic technologies can potentially reduce the power consumption of large scale, stationary computing systems by several orders of magnitude, including the energy cost of the refrigeration [3], [4]. The cooling capacity at 4 K is however often insufficient to efficiently remove the heat generated by the circuitry [5]. Furthermore, as illustrated in Fig. 1, it may be advantageous to place certain circuits at lower temperatures while other circuits are placed at higher temperatures.

The operating temperature also greatly affects the architecture of a heterogeneous computing system. By adjusting the operating temperature of each subsystem, the performance and power of the overall system can be better controlled. Different technologies can be placed at different stages of a cryocooler to reduce overall refrigeration costs. For example, the temperature of a cryogenic CMOS subsystem can be increased. The latency and power dissipation of this subsystem may however also increase. Furthermore, refrigeration of nearby subsystems operating at a lower temperature can be affected if these subsystems are not sufficiently thermally isolated.

An approach where different technologies are placed at different stages of the refrigerator has previously been proposed [10], [11]. A hybrid temperature system exploits multiple stages of a cryocooler; in [10], a Sumitomo SRDK-101DP-11C cryocooler with 4 K and 60 K stages is introduced. Low temperature superconductive

circuits are located at the 4 K stage, higher temperature semiconductor circuits, such as analog filters and low noise amplifiers, are placed at the 60 K stage, and the remaining electronics are placed at room temperature. These studies utilize different stages within a cryocooler, but do not consider the possible range of temperatures within a specific stage. For example, the second stage of the Sumitomo cryocooler in [10] is set to 60 K, while the available temperature range can vary between 60 K to 80 K.

This range of available temperatures of each stage within a cryocooler is exploited here to enhance the overall performance of computing systems under a target heat load constraint. A methodology for optimizing the temperature of each component within a cryogenic system is proposed. The total power consumed by the system is minimized while maintaining satisfactory performance. The methodology is validated in a case study requiring cryogenic operation, a quantum computer, a technology which potentially will accelerate a wide range of computing tasks, such as prime factorization, quantum simulation, and complex optimization [12], [13].

The paper is organized as follows: in Section II the problem is formulated, the thermal behavior of the system is discussed, and insight into the organization of cryogenic computing systems are described. An optimization methodology is proposed in Section III. An example case study, a hybrid quantum computing system, optimized using the proposed methodology, is presented in Section IV. Some conclusions are offered in Section V.

II. Background

The efficient integration of cryogenic computing systems requires electronic circuits operating at different temperatures [5]. The primary design objective is to determine a set of temperatures for each of the components at which the total power consumption and/or delay is minimized while satisfying target constraints, which denotes optimal operation.

A methodology is proposed to determine the optimal temperature of the different parts of an electronic cryogenic system. Four steps are performed in the methodology. A graph of the system is initially formulated as step one, and the available range of temperatures for each component within the system is determined. An algorithm to evaluate the set of optimal temperatures, exploiting graph theory [14], is proposed. This algorithm is used to select the paths by a power or delay constraint in the second step. After determining the set of temperatures which satisfies the constraint, a thermal model of the system is generated in the third step, to evaluate the flow of heat (or power) from unit to unit. The heat flow depends upon the thermal conductance between units. The rate of heat flow depends upon the temperature of the connecting wires. In step four, the heat flow can be used to estimate the leakage power; specifically, the power lost from the additional cooling required at lower temperatures due to the flow of heat from the higher temperature components. The net power consumption at a specific set of temperatures therefore includes the leakage power between temperature zones. Optimal operation of the system, considering delay and power constraints and the heat flow among the components, sets the temperature for each component.

The rest of the section is organized as follows. The problem formulation based on graph theory is described in Section II-A. A thermal model of the system is discussed in Section II-B.

A. Formulation of Thermal Optimization Problem

The objective is to determine a suitable operating temperature at each step of the process. Temperature optimization of a process can be described as a directed acyclic multiweighted multigraph $G \coloneqq \langle S, U, W \rangle$. A finite set of states in the process $S = \{S_1, S_2, \ldots, S_n\}$ specifies an instance of the temperature optimization problem. A set of edges is denoted by U and represents a unit performing a computational step. Parallel edges correspond to computing unit i which comprise a subset $U_i \subseteq U$. A typical refrigeration system operates at a specific set of temperatures, such as liquid helium temperature (LHT) or liquid nitrogen temperature (LNT). Index j represents the set of available temperatures, $T = \{T_1, T_2, \ldots, T_j\}$. A unit at different temperatures at each step is represented by $u_{i,j} \in U_i$. Two weights are associated with each edge $u_{i,j}$, $W \coloneqq \langle p, d \rangle \in \mathbb{R}^2_{>0}$, where p and d represent, respectively, the power consumption and delay of a unit at a specific temperature.

A set of operating temperatures corresponding to each computing unit constitutes a path connecting the source to the sink of the process graph. Path π is the collection of specific edges between two endpoints of a process,

$$\pi = (U_1(T_j), U_2(T_j), \ldots, U_i(T_j)) \ . \tag{1}$$

The power consumption of a process is the sum of the power weights along a path, $P(\pi) = p_1 + p_2 + \cdots + p_{n-1}$. The weight of an edge represents the power consumption of a unit. Similarly, the delay of the process is the total cost of the weights, which represents the total weight of the edges along a path, $D(\pi) = d_1 + d_2 + \cdots + d_{n-1}$. Given set U at different temperatures performing a computation among states S, the temperature optimization

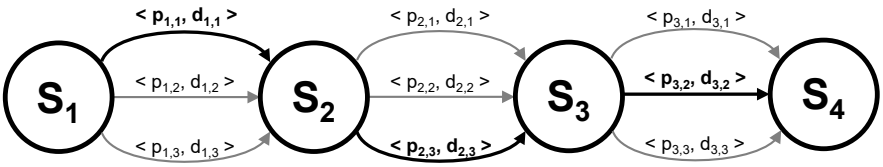

Fig. 2: Example of the temperature design process. The edges between two states describe a unit at different temperatures. An example of a path is shown highlighted in bold. The power consumption of this path is $P(\pi) = p_{1,1} + p_{2,3} + p_{3,2}$. The delay of the path is $D(\pi) = d_{1,1} + d_{2,3} + d_{3,2}$.

problem is to determine a path connecting the source and sink states of a system that minimizes the total power $P(\pi)$ while constraining the total delay of the system, $D(\pi)$, below maximum delay D_{max}.

$$\text{Minimize: } P(\pi), \tag{2}$$

$$\text{subject to: } D(\pi) \leq D_{max}. \tag{3}$$

An example of the process containing three units and four states is depicted in Fig. 2. Note that in this example each unit can operate at three different temperatures, as denoted by the parallel edges between adjacent states. A power and delay are associated with each edge. A path $\pi = (u_{1,1}, u_{2,3}, u_{3,2})$, highlighted in bold in Fig. 2, corresponds to computing units u_1, u_2, and u_3 operating at, respectively, temperature T_1, T_3, and T_2. The total power consumption of the highlighted path is $P(\pi) = p_{1,1} + p_{2,3} + p_{3,2}$. The delay of the highlighted path is $D(\pi) = d_{1,1} + d_{2,3} + d_{3,2} \leq D_{max}$.

B. Thermal Model

Apart from the heat dissipation produced by the units, the power consumed by the path is also due to the leakage power between units. This leakage power is caused by the difference in temperature and is transferred by the connector cables between units [15].

Since the thermal resistance of the cable material varies depending upon the absolute temperature, the thermal resistance can be adjusted to more accurately characterize the flow of heat between units [16], [17]. The thermal resistance linearly, exponentially or logarithmically increases or decreases, depending upon the material type and quality (i.e., purity) of the material [16]–[18]. Specialized cables composed of stainless steel and beryllium copper, such as CryoCoax BCB016, BCB019, and BCB029, are typically utilized in cryogenic applications [19]. The thermal conductivity of these cables exhibits a rising trend with increasing temperature [20]–[22]. The thermal conductivity of beryllium copper can be linearly approximated [22], as depicted in Fig. 3, whereas that of stainless steel can be represented by a dual-line approximation [21].

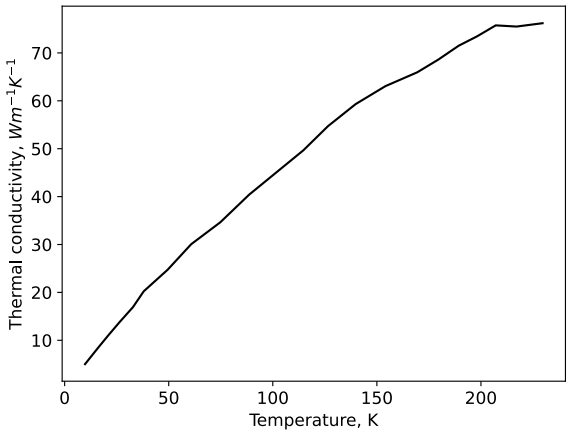

Fig. 3: Thermal conductivity of beryllium copper.

The thermal conductivity of different cables can be used to construct thermal circuits based on an analogy with electrical circuits, as described in [15]. The flow of heat q_T within a system is described by a set of linear expressions,

$$q_T = \begin{array}{c} \\ U_1 \\ U_k \\ U_n \end{array} \begin{array}{c} \begin{array}{ccc} U_1 & U_k & U_n \end{array} \\ \begin{bmatrix} \frac{\Delta T_{1,1}}{R_{1,1}} & \cdots & \frac{\Delta T_{1,n}}{R_{1,n}} \\ \vdots & \ddots & \vdots \\ \frac{\Delta T_{n,1}}{R_{n,1}} & \cdots & \frac{\Delta T_{n,n}}{R_{n,n}} \end{bmatrix} \end{array}. \tag{4}$$

The power flowing to or from each unit is summed along each row,

$$\Delta P = q_T \, \mathbf{1}_n, \tag{5}$$

where

$$\mathbf{1}_n = [1, \ldots, 1]^\intercal. \tag{6}$$

III. Optimization Setup

In the first step of the methodology, a graph of the system is generated, as described in Section II-A. Any path connecting the initial stage of process S_1 to the final stage S_n determines the power consumption and delay of the system. The optimization problem is to choose the most power efficient temperature set, while

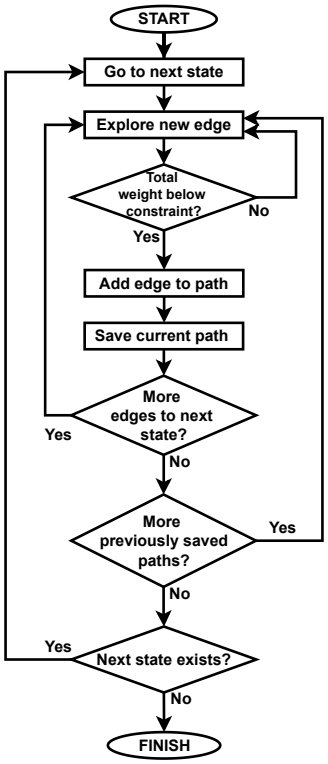

Fig. 4: Flowchart of the proposed algorithm.

ensuring the delay of the system is below constraint D_{max}. A flowchart of the algorithm to determine all of the paths within the graph satisfying the delay constraint is illustrated in Fig. 4. The algorithm requires a matrix of delays D as an input, where entry $D_{i,j}$ denotes the delay of unit i at temperature T_j,

$$D = \begin{array}{c} \\ T_1 \\ T_j \\ T_m \end{array} \begin{array}{c} \begin{array}{ccc} U_1 & U_k & U_n \end{array} \\ \begin{bmatrix} D_{1,1} & \cdots & D_{1,n} \\ \vdots & \ddots & \vdots \\ D_{m,1} & \cdots & D_{m,n} \end{bmatrix} \end{array}. \tag{7}$$

The proposed algorithm is based on breadth-first search traversal of the process graph, starting from the source node. During the traversal, delay D of the partial path is compared to delay constraint D_{max}. If delay D is greater than D_{max}, the algorithm explores the next edge. Partial paths satisfying the delay constraint are recorded and the traversal continues. Upon completing the traversal process, a new path to the current node is treated as an input, and all of the edges are once again explored. After all of the paths from the source to sink are evaluated and the unwanted paths are discarded, the algorithm proceeds to the next node.

Two techniques are used to reduce memory usage and computational runtime. First, by determining the paths satisfying both constraints, power P_{max} and delay D_{max}, those paths not satisfying both constraints are removed earlier in the process. Second, the algorithm is run twice, increasing the precision of the temperature range in the following step. This procedure reduces the runtime while maintaining the same level of precision. For the case of four chambers within a refrigeration system, the procedure reduces the complexity of the graph from $O(n^4)$ to $O(2n^2)$. For example, for a system with four chambers and one hundred possible temperatures, 100^4 possible paths exist. By performing the graph optimization step twice with ten possible temperatures and increasing the precision in the next step, the same result is achieved with the total number of explored paths, $2 * 10^4$.

The algorithm determines all possible paths from source to sink that satisfy the delay constraint. The power flow between each unit is evaluated in the next step of the algorithm to determine the total power consumption of the path, as described in Section II-B. Finally, the optimal temperature set is the set of temperatures consuming the least power.

IV. Quantum Computing Case Study

A quantum computer is a combination of a quantum processor and an electronic controller [5]. A quantum processor uses quantum bits (qubits) to perform operations. Qubits operate at extremely low temperatures; typically, a few millikelvins [23]. An electronic controller reads out the signal and controls the quantum processor [5]. Existing quantum computers utilize classical electronic controllers operating at room temperature (RT) [12]. This approach, however, is challenging and expensive, as the number of qubits is expected to reach thousands and millions [12]. Establishing individual connections between millions of qubits and the controller circuitry operating at room temperature is infeasible due to the read complexity, cost, and signal performance of the interconnect [5], [12], [24]. It has therefore been suggested to utilize a classical CMOS electronic controller operating at cryogenic temperatures [5] or a SFQ controller operating below 4 K [25]–[27], which can be placed closer to the quantum processor.

The proposed algorithm is used to determine the set of optimal temperatures for a hybrid superconductive quantum-classical computing system, as adapted from [26]. While it is possible to operate most of the controller at temperatures below 4 K (i.e., SFQ), the cooling capacity at these temperatures is often insufficient to efficiently remove the heat generated by the controller. Partitioning the controller into higher and lower temperature domains may be more efficient. The proposed algorithm described herein is used as a case study, to determine the set of optimal temperatures for an exemplifying hybrid superconductive quantum-classical computing system, as shown in Fig. 5. The quantum computing system consists

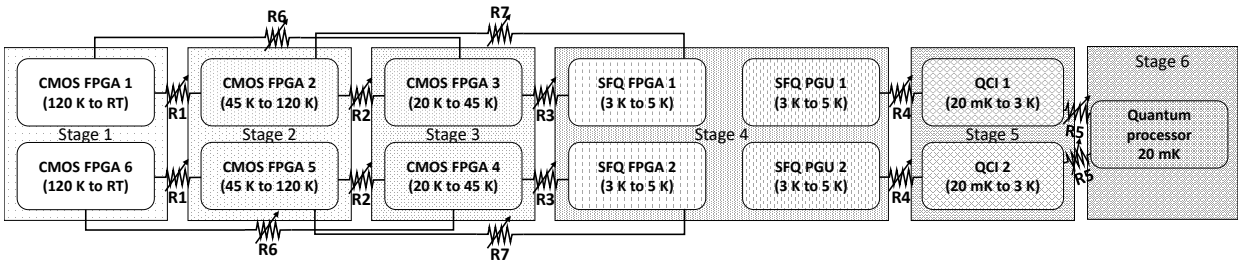

Fig. 5: Simplified thermal-electrical circuit model of a hybrid quantum computer. The thermal relationship between the units is represented by the rheostats, which describe the dependence of the thermal resistance on temperature.

of thirteen devices: six CMOS FPGAs for readout and control, four SFQ FPGAs, two SFQ pulse generating systems, two SFQ quantum-classical interface (QCI) integrated circuits, and a quantum processor. The CMOS circuits operate at a temperature ranging from 20 K to RT, SFQ circuits operate at a temperature ranging from 3 K to 5 K, and SFQ QCI circuits operate over a temperature range from 20 mK to 3 K.

A delay and power at each temperature are assigned to each unit. These numbers are assumed to be the mean value of the delay and power of each unit during operation and are randomly generated assuming an exponential distribution over different temperatures. The delay and power for each unit are listed in Table I. The total power consumption includes the power consumed by the refrigerators.

Table I: Delay D_i and power P_i of each computing unit in a hybrid quantum computer. The delay and power are at the highest possible operating temperature of each unit.

Computing Unit	**Delay** D, [fs]	**Power** P, [W]	**Computing Unit**	**Delay** D, [fs]	**Power** P, [W]
CMOS FPGA1	550	9.0	SFQ FPGA1	6	0.38
CMOS FPGA6	450	7.0	SFQ FPGA2	8	0.39
CMOS FPGA2	70	4.5	SFQ PGU1	3	0.33
CMOS FPGA5	80	5.5	SFQ PGU2	3	0.40
CMOS FPGA3	53	2.2	QCI1	0.6	0.25
CMOS FPGA4	47	1.8	QCI2	0.4	0.25

Any thermal interactions between the units are set by the interconnects between the units and the proximity of the units to each other. The interconnects between the SFQ integrated circuits and the QCI and between the QCI and the SFQ coprocessor are established via superconductive low heat loads and low crosstalk superconductive ribbon cables [26]. These connections maintain accurate timing and reliable transmission of the SFQ pulses. These connections and the nonideality of the refrigerators produce a thermal conductance between units. A simplified thermal-electrical circuit model of the system is illustrated in Fig. 5. Ten different thermal resistances between the units are assumed. The value of the thermal resistances at 4 K is listed in Table II.

Table II: Thermal resistance of the hybrid quantum computer at 4 K.

Resistance	Ω_T [K/W]	**Resistance**	Ω_T [K/W]
R_1	60	R_5	600
R_2	150	R_6	30
R_3	200	R_7	50
R_4	400		

A set of optimal temperatures for each device is determined using the proposed algorithm. The set of temperatures minimizing the total power while satisfying the delay constraint of 0.135 ps is determined. The algorithm is implemented in Python and executed on an Intel Core i7-9750H workstation with 8 GB RAM. For this case study, the algorithm completes in 0.65 s. Sets of optimal temperatures, excluding the quantum processor, are listed in Table III, where the most optimal set is highlighted in bold. The difference in performance is due to the difference in the temperature of the SFQ FPGA, PGU, and QCI modules. The quantum processor is located in the last stage, chamber 6, operating at 20 mK (see Fig. 5). The power consumption of the optimal path with a delay constraint of 0.135 ps is 258 watts. Most of the power is consumed by the refrigerators operating at cryogenic temperatures.

V. Conclusions

Hybrid cryogenic computing systems are an emerging technology motivated primarily by high performance cloud computing and quantum computing networks. The operating temperature of the circuit components affects the performance, cooling power, and dissipated power. Selecting the appropriate operating temperature is therefore crucial to minimizing the total power dissipated by

Table III: Set of temperatures for a hybrid quantum computing system. The system is composed of six CMOS FPGAs, two SFQ FPGAs, two SFQ PGUs, two SFQ QCIs, and a quantum processor placed in a cryogenic refrigerator with six chambers. The most optimal (lowest power) set is highlighted in bold.

Stage temperature, K						**Delay**	**Power**
C_1	C_2	C_3	C_4	C_5	C_6	D_i, [fs]	P_i, [W]
120	**45**	**20**	**3.94**	**2.65**	**0.02**	**134.99**	**258.76**
120	45	20	3.89	2.72	0.02	134.90	259.23
120	45	20	3.89	2.69	0.02	134.78	259.60
120	45	20	3.94	2.62	0.02	134.87	259.62
120	45	20	3.91	2.69	0.02	134.87	259.79

the system while maintaining correct functionality and performance.

A methodology for the thermal optimization of cryogenic computing systems with multiple temperature zones is presented in this paper. The methodology is validated on a practical case study where the individual temperature of an eleven unit system is optimized. The overall power consumption of a quantum computing system is minimized while satisfying the target delay constraint. A multigraph representation describes the relationship among the temperature, delay, and power of a system. The total cooling power is described by a thermal model of the system, which includes a variable thermal conductance between each unit within the system. The proposed algorithm is applied to a case study, and the temperature of each component that minimizes the total system power dissipation is determined while satisfying target performance constraints.

REFERENCES

[1] P. Sharma *et al.*, "Design and Operational Analysis of a Green Data Center," *IEEE Internet Computing*, vol. 21, no. 4, pp. 16–24, August 2017.

[2] M. Koot and F. Wijnhoven, "Usage Impact on Data Center Electricity Needs: a System Dynamic Forecasting Model," *Applied Energy*, vol. 291, p. 116798, June 2021.

[3] D. S. Holmes, A. L. Ripple, and M. A. Manheimer, "Energy-Efficient Superconducting Computing—Power Budgets and Requirements," *IEEE Transactions on Applied Superconductivity*, vol. 23, no. 3, pp. 1,701,610–1,701,610, June 2013.

[4] N. Zhuldassov and E. G. Friedman, "Temperature–Frequency Boundary of Cryogenic Dynamic Logic," *Microelectronics Journal*, vol. 135, p. 105763, March 2023.

[5] B. Patra *et al.*, "Cryo-CMOS Circuits and Systems for Quantum Computing Applications," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 1, pp. 309–321, September 2017.

[6] B. Oy, (2023) Technical specifications of BlueFors Oy cryocoolers, [Online]. Available: https://bluefors.com/products

[7] Sumitomo Heavy Industries, Ltd., (2023) Technical specifications of Sumitomo cryocoolers, [Online]. Available: https://www.shi.co.jp/english/products/machinery/cold/index.html

[8] Sunpower Inc., (2023) Technical specifications of Sunpower cryocoolers, [Online]. Available: https://www.sunpowerinc.com/products/stirling-cryocoolers/

[9] Northrop Grumman, (2023) Technical specifications of Northrop Grumman cryocoolers, [Online]. Available: https://www.northropgrumman.com/space/cryocoolers/

[10] O. A. Mukhanov *et al.*, "Superconductor Digital-RF Receiver Systems," *IEICE Transactions on Electronics*, vol. 91, no. 3, pp. 306–317, March 2008.

[11] D. Gupta *et al.*, "Modular, Multi-Function Digital-RF Receiver Systems," *IEEE Transactions on Applied Superconductivity*, vol. 21, no. 3, pp. 883–890, December 2010.

[12] L. Vandersypen and A. van Leeuwenhoek, "Quantum Computing–the Next Challenge in Circuit and System Design," *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 24–29, February 2017.

[13] A. Montanaro, "Quantum Algorithms: an Overview," *npj Quantum Information*, vol. 2, no. 1, pp. 1–8, January 2016.

[14] R. Bairamkulov and E. G. Friedman, *Graphs in VLSI*, Springer, 2023.

[15] N. Zhuldassov, R. Bairamkulov, and E. G. Friedman, "Thermal Optimization of Hybrid Cryogenic Computing Systems," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 31, no. 9, pp. 1339–1346, September 2023.

[16] H.-K. Lyeo and D. G. Cahill, "Thermal Conductance of Interfaces Between Highly Dissimilar Materials," *Physical Review B*, vol. 73, no. 14, p. 144301, April 2006.

[17] D. Thornburg, E. Thall, and J. Brous, "A Manual of Materials for Microwave Tubes," Radio Corporation of America, Technical Report, January 1961.

[18] J. Paasschens, S. Harmsma, and R. Van der Toorn, "Dependence of Thermal Resistance on Ambient and Actual Temperature," *Proceedings of the Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 96–99, September 2004.

[19] CryoCoax, (2023) Technical Specifications of CryoCoax's Cryogenic Cables, [Online]. Available: https://cryocoax.com/cryogenic-cable-and-cable-assemblies/

[20] C. Schmidt, "Simple Method to Measure the Thermal Conductivity of Technical Superconductors, e.g., NbTi," *Review of Scientific Instruments*, vol. 50, no. 4, pp. 454–457, April 1979.

[21] P. E. Bradley, R. Radebaugh *et al.*, "Properties of Selected Materials at Cryogenic Temperatures," *National Institute of Standards and Technology*, vol. 680, pp. 1–14, June 2013.

[22] N. Simon, E. Drexler, and R. Reed, "Properties of Copper and Copper Alloys at Cryogenic Temperatures. Final Report," National Institute of Standards and Technology (MSEL), Boulder, Colorado, Technical Report, February 1992.

[23] M. H. Devoret, A. Wallraff, and J. M. Martinis, "Superconducting Qubits: A Short Review," *arXiv preprint cond-mat/0411174*, November 2004.

[24] M. Reiher *et al.*, "Elucidating Reaction Mechanisms on Quantum Computers," *Proceedings of the National Academy of Sciences*, vol. 114, no. 29, pp. 7555–7560, July 2017.

[25] G. Krylov and E. G. Friedman, *Single Flux Quantum Integrated Circuit Design*, Springer, 2022.

[26] O. Mukhanov *et al.*, "Scalable Quantum Computing Infrastructure Based on Superconducting Electronics," *Proceedings of the IEEE International Electron Devices Meeting*, pp. 31.2.1–31.2.4, December 2019.

[27] N. K. Katam, O. A. Mukhanov, and M. Pedram, "Superconducting Magnetic Field Programmable Gate Array," *IEEE Transactions on Applied Superconductivity*, vol. 28, no. 2, pp. 1–12, January 2018.