Thermal Optimization of Hybrid Cryogenic Computing Systems

Nurzhan Zhuldassov, Rassul Bairamkulov, *Member, IEEE*, and Eby G. Friedman, *Life Fellow, IEEE*

***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 multitemperature design methodology exists. To overcome this limitation, a framework for thermal optimization of heterogeneous computing systems is presented in this article. The effects of operating temperature on delay and power consumption are characterized based on a graph representation of the system. In addition, thermal interactions among the components within a system are considered to accurately evaluate the total power consumption and 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*— Cryogenic CMOS, quantum computing, quantum-classical computer, single flux quantum (SFQ), thermal optimization.**

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. 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 dissipate 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, as the temperature is reduced, other circuits are placed at higher temperatures.

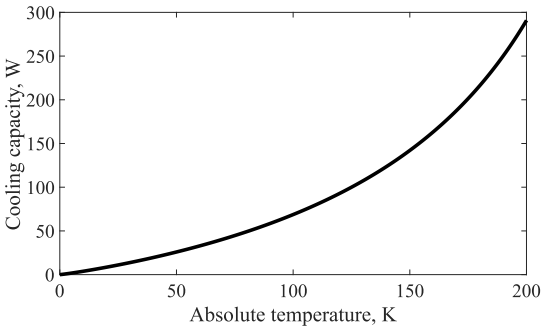


Fig. 1. Available cooling power per 1-kW input power. The data are based on the technical specifications of commercially available cryogenic coolers [6], [7], [8], [9].

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. For example, the temperature of a cryogenic CMOS subsystem can be increased; different technologies can be placed at different stages of a cryocooler to reduce refrigeration costs. The latency and power dissipation of this subsystem, however, may also increase. Furthermore, refrigeration of nearby subsystems operating at a lower temperature can be affected if these subsystems are not 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 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 (LNA), are placed at the 60-K stage, and the remaining electronics are placed at room temperature (RT). These studies use 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 of the second stage of this cryocooler can vary between 60 and 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

Manuscript received 14 August 2022; revised 29 October 2022, 30 December 2022, 27 February 2023, and 15 April 2023; accepted 23 April 2023. Date of publication 23 May 2023; date of current version 25 August 2023. This work was supported in part by the National Science Foundation under Grant 2124453 (Expeditions DISCOVER), in part by Synopsys, and in part by Qualcomm. *(Corresponding author: Nurzhan Zhuldassov.)*

The authors are with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627 USA (e-mail: nzhuldas@ur.rochester.edu).

Digital Object Identifier 10.1109/TVLSI.2023.3271898

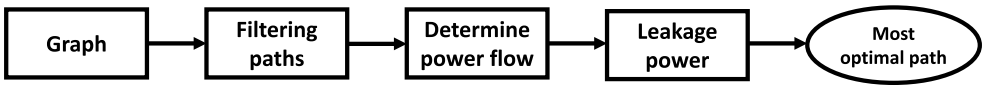


Fig. 2. Flowchart of the proposed methodology. A graph of the system is initially generated. Paths connecting the initial and final states of the process satisfying the constraints are determined. The power flow within the system is subsequently evaluated to determine the total power consumed by the system. The optimal system is based on the total power consumption and delay of each path.

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 acceptable performance. The methodology is confirmed 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 article 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 is described. A proposed methodology is described 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

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

A methodology is proposed to determine the optimal temperature of the different parts of an electronic system. Four steps are performed in the methodology, as shown in Fig. 2. A graph of the system is initially formulated, 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. After determining the set of temperatures which satisfies the constraint, a thermal model of the system is generated to evaluate the flow of heat (or power) from unit to unit. The heat flow depends on the thermal conductance between units. The rate of heat flow depends on the temperature of the connecting wires. Furthermore, the heat flow is 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 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 hybrid quantum computing system as a case study is described in Section II-C.

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 $G := \langle S, U, W \rangle$. 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) in cryogenic CMOS. 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 $u_{i,j}$, $W := \langle p, d \rangle \in \mathbb{R}^2_{>0}$, 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

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

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 delay of a unit 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 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, $D(\pi)$

$$\min\ P(\pi) \tag{2}$$

$$\text{s.t. } D(\pi) \leq D_{\max}. \tag{3}$$

An example of the process containing three units and four states is depicted in Fig. 3. 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, corresponds to computing units u_1–u_3 operating at, respectively, temperatures 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 $D(\pi) = d_{1,1} + d_{2,3} + d_{3,2} \leq D_{\max}$.

The problem of finding the optimal set of temperatures resembles the knapsack problem [15] or multiple knapsack problem [16]. The optimization problem described here, however, is different. In the knapsack problem, the set of items

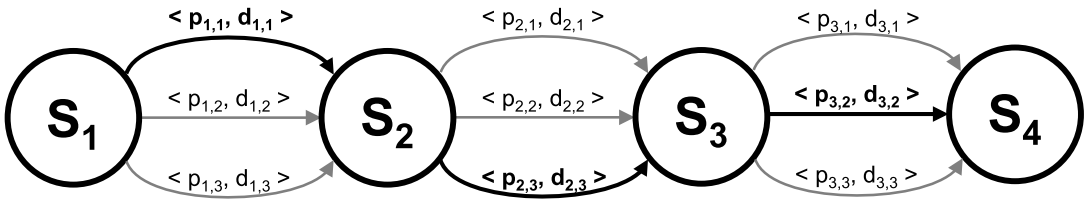


Fig. 3. Example of the temperature design process. The edges between two states describe a unit at different temperatures. An example of a path is shown highlighted in bold. The power consumption of this path is $P(\pi) = p_{1,1} + p_{2,3} + p_{3,2}$. The delay of the path is $D(\pi) = d_{1,1} + d_{2,3}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

maximizing the total value is determined while not exceeding the weight limit. In the multiple knapsack problem, the value of the items in two or more knapsacks is considered. The primary difference is related to the type of units stored in each bag. Since each unit has a unique set of power and delay weights, the knapsack analogy of this problem can be described as a multiple knapsack problem with a unique set of items stored in each knapsack. The cumulative value of each knapsack is maximized while constraining the total weight.

B. Thermal Model

Since computing units are not ideally isolated from each other, any temperature difference between units produces heat flow. The transfer of heat between computing units significantly contributes to the heat load of the cooling system. The choice of path contributes to this transferred power. A refrigerator requires different powers at different temperatures due to the difference between the ambient and operating temperatures [17]. Similarly, a system with a significant difference in temperature between adjacent units likely requires greater refrigeration power when compared with a system with a smaller temperature difference [17]. Newton's law of cooling notes that the rate of temperature change in a body depends on the ambient temperature and body temperature and suggests that additional cooling power is required due to heat leaking from a higher temperature unit to a lower temperature unit. The net cooling power therefore not only depends on the cooling power of each individual unit but also depends on the heat flow between units.

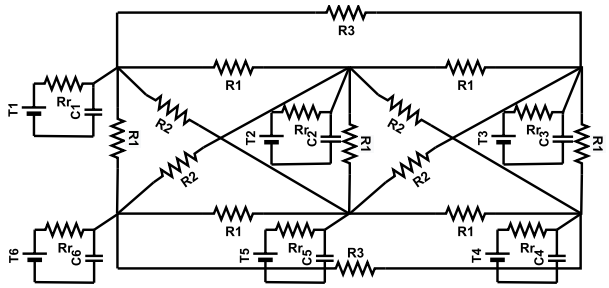
The thermal behavior of a system can be efficiently described based on an analogy with electrical circuits. This analogy between the electrical and thermal processes is summarized in Table I. While electric charge q is transported across a conductor in an electrical circuit, heat Q_T is transported through a thermal conductor. The heat flow (or power) per unit time q_T is analogous to the flow of current i, the electric charge per unit time. Similar to the potential difference ΔV driving an electric current in a conductor, heat is driven by the difference in temperature ΔT. By applying Fourier's law [18], the thermal resistance between units is

$$R_T = \frac{\Delta T}{q_T}. \tag{4}$$

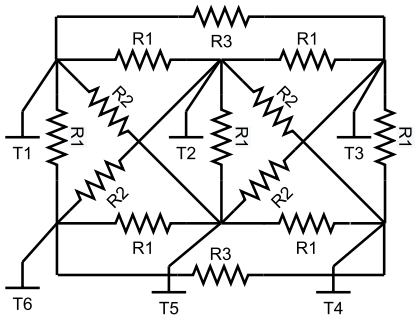
Heat flow within a system is determined for each set of temperatures based on a thermal model. An example of a system containing six CPUs is shown in Fig. 4(a). The CPUs are represented by a thermal capacitance and are labeled as C_n,

TABLE I
THERMAL–ELECTRIC ANALOGY

Thermal parameter	Symbol	Electrical parameter	Symbol
Heat [Joule]	Q_T	Charge [Coulomb]	q
Heat flow [Watt]	q_T	Current [Ampere]	i
Thermal capacitance [J/K]	C_T	Capacitance [Farad]	C
Thermal resistance [K/W]	R_T	Resistance [Ω]	R
Temperature [K]	T	Voltage [Volt]	V



(a)



(b)

Fig. 4. Thermal model of a system with six CPUs. (a) Complex model and (b) simplified model. The complex model includes four types of temperature interactions: R_1–R_3, and refrigerator resistance R_r. The simplified thermal model includes only three types of temperature interactions: R_1–R_3.

where n is the CPU number. Similar to a voltage source, the refrigerators are the heat source for the thermal capacitors and represent the voltage source in a thermal circuit. Since a refrigerator cannot instantly regulate the temperature of a CPU, the refrigerators are connected through thermal resistance R_r. Similarly, the CPUs are connected by thermal resistances. In this example, three types of connections among the CPUs exist, corresponding to three thermal resistances, R_1–R_3. The connection between adjacent units is R_1, diagonally adjacent units are connected through thermal resistance R_2, and the

relationship between units separated by one other unit is represented by R_3. The temperature of each unit in this example is maintained at T_i. The flow of power (or heat) between units is based on an estimate of the currents within the circuit.

Since CPUs are often located in different refrigerators (or chambers), the thermal conduction between the CPU and the refrigerator is much higher than the conduction with another CPU (since $R_r \ll R_1, R_2, R_3$). The thermal resistance between the refrigerator and the CPUs (R_r) is therefore assumed to be negligible. The thermal model of the system shown in Fig. 4(a) can therefore be simplified to Fig. 4(b).

The flow of heat within the system is described by a system of linear expressions. A thermal resistance grid between the units is initially described as

$$R_T = \begin{array}{c} \\ U_1 \\ U_k \\ U_n \end{array} \begin{array}{c} \begin{array}{ccc} U_1 & U_k & U_n \end{array} \\ \begin{bmatrix} R_{1,1} & \cdots & R_{1,n} \\ \vdots & \ddots & \vdots \\ R_{n,1} & \cdots & R_{n,n} \end{bmatrix} \end{array} \tag{5}$$

where the thermal resistance between units U_k and U_n is represented by $R_{k,n}$. If the units are thermally isolated, i.e., the units are not thermally connected, the thermal resistance is assumed infinite. The thermal resistances across the diagonal of the matrix are also equated to infinity, as each diagonal element represents a relationship of each unit to itself. Since the temperature of each unit is extracted from a graph within the proposed algorithm, the matrix of temperature differences (or thermal potential differences) can be established as

$$\Delta T = \begin{array}{c} \\ U_1 \\ U_k \\ U_n \end{array} \begin{array}{c} \begin{array}{ccc} U_1 & U_k & U_n \end{array} \\ \begin{bmatrix} \Delta T_{1,1} & \cdots & \Delta T_{1,n} \\ \vdots & \ddots & \vdots \\ \Delta T_{n,1} & \cdots & \Delta T_{n,n} \end{bmatrix} \end{array}. \tag{6}$$

Each element, representing a temperature difference between units, corresponds to an index. For example, $\Delta T_{k,n}$ represents the temperature difference between units k and n. As noted, the diagonal elements are each equal to zero. The power flow q_T between each unit can therefore be evaluated by elementwise matrix division using the electrical analogy, $q_T = (\Delta T / R_T)$, as

$$q_T = \begin{array}{c} \\ U_1 \\ U_k \\ U_n \end{array} \begin{array}{c} \begin{array}{ccc} U_1 & U_k & U_n \end{array} \\ \begin{bmatrix} \frac{\Delta T_{1,1}}{R_{1,1}} & \cdots & \frac{\Delta T_{1,n}}{R_{1,n}} \\ \vdots & \ddots & \vdots \\ \frac{\Delta T_{n,1}}{R_{n,1}} & \cdots & \frac{\Delta T_{n,n}}{R_{n,n}} \end{bmatrix} \end{array}. \tag{7}$$

The power flowing to or from each unit can be evaluated by summing all of the elements along each row as

$$\Delta P = q_T \, \mathbf{1}_n \tag{8}$$

where

$$\mathbf{1}_n = [1, \ldots, 1]^\intercal. \tag{9}$$

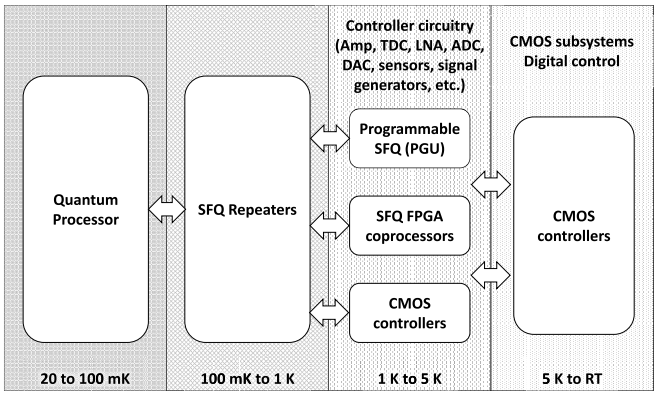


Fig. 5. Block diagram of the hybrid quantum-classical computing system. The system consists of a quantum processor, SFQ repeaters, and controllers, each located at different temperatures. The controllers can be CMOS or SFQ.

Since the thermal resistance of a material varies depending upon the absolute temperature, the thermal resistance can be adjusted to more accurately characterize the heat flow between units [19], [20]. The thermal conductivity of the cryogenic cables, made from materials such as stainless steel, niobium–titanium alloy, and beryllium copper, exhibits a rising trend with increasing temperature [21], [22], [23]. The thermal conductivity of beryllium copper can be linearly approximated [23], whereas that of stainless steel can be represented by a dual-line approximation [22].

C. Control of Quantum Computing Systems

Qubit control requires readout and generation of electronic and optical signals [5]. The readout measures the resonant frequency of a resonator in the case of transmons [24] or the impedance of a charge detector in the case of spin qubits [25]. These measurements are used to evaluate extremely low-noise signals, as a quantum qubit state is highly volatile. Due to this volatility, the controller is also responsible for quantum error correction [26]. A controller consists of control and readout circuitry, service blocks, such as voltage, current, and frequency references [27], and a digital controller [28], as shown in Fig. 5. The service blocks consist of phase locked loop (PLL) oscillators, LNA, analog-to-digital converter (ADC), digital-to-analog converter (DAC), and other circuitry.

The primary cryogenic technologies for the control circuitry of a quantum computer include semiconductor-based cryogenic CMOS, superconductive rapid single flux quantum (RSFQ) [29], and adiabatic quantum flux parametron (AQFP) [29]. Circuits based on cryogenic CMOS technology dissipate less power when compared with RT CMOS, while delivering faster performance. CMOS has been reported to operate at temperatures ranging from 100 mK to RT [30]. At cryogenic temperatures, a MOSFET exhibits enhanced physical properties such as higher transient currents, negligible leakage currents, and increased subthreshold slope [31].

The niobium-based superconductive electronic systems require operation at a temperature below approximately 5 K (and frequently cooled to 4.2 K, helium boiling temperature) [32]. RSFQ and AQFP, two prominent superconductive technologies, have been steadily maturing over recent

years [29]. The RSFQ systems exhibit 100–1000 times less energy per computation as compared with CMOS, while operating at frequencies of hundreds of gigahertz. Systems based on AQFP technology dissipate even less power per computation, several orders of magnitude less than RSFQ, albeit at slower speeds [33]. Current manufacturing technology accommodates more than 6000 Josephson junctions (JJ)/mm^2 [34]. With the development of electronic design automation tools for superconductive electronics [35], [36], single flux quantum (SFQ)-based VLSI complexity systems are currently under development [29]. An 8-bit superconductive microprocessor operating at a frequency of 80 GHz has successfully been fabricated [37].

Despite these promising features, each of these technologies faces a significant challenge in the development of large-scale systems. For example, SFQ memory is challenging due to the poor scaling of the inductors required in SFQ logic and memory. Cryogenic CMOS circuits, in turn, generate significant heat during operation, reducing the cooling efficiency. These issues can be overcome by combining multiple technologies within a single integrated system. The result of computing operations performed using an RSFQ controller can, for example, be stored in a CMOS static or dynamic random access memory (RAM) [31] or a magnetic tunnel junction (MTJ)-based magnetic RAM [38].

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 determine the most power-efficient temperature set, while 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. 6. 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

$$D = \begin{array}{c} \\ T_1 \\ T_j \\ T_m \end{array} \begin{array}{c} \begin{array}{ccc} U_1 & U_k & U_n \end{array} \\ \begin{bmatrix} D_{1,1} & \cdots & D_{1,n} \\ \vdots & \ddots & \vdots \\ D_{m,1} & \cdots & D_{m,n} \end{bmatrix} \end{array}. \tag{10}$$

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 with 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 paths from the source to the sink are evaluated and the unwanted paths are discarded, the algorithm proceeds to the next node.

The algorithm includes a power weight attached to the edges. By determining the paths satisfying both the power $P_{\max}$ and delay $D_{\max}$ constraints, the memory usage and computational runtime are significantly reduced. The savings in memory and computational time are more significant if the possible path branch is removed earlier in the process. For example, for a process with ten units and ten possible temperatures, 10^{10} possible paths exist before the constraint check is performed. If all of the paths which start with $\pi = (U_1(T_1), \ldots)$ are removed in the first loop of the algorithm, 10^9 possible paths are removed, reducing the traversal time. The result of the algorithm is all possible paths from the source to the sink 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.

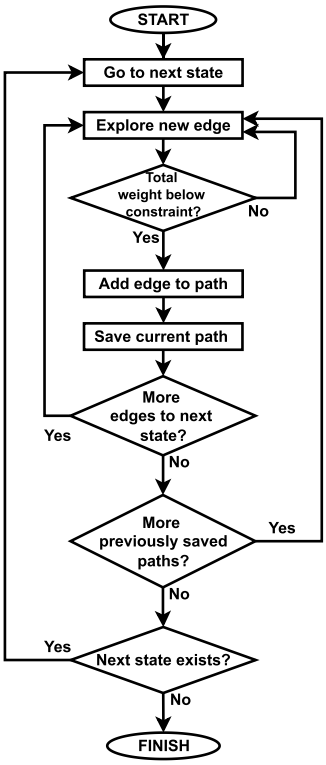


Fig. 6. Flowchart of the proposed algorithm.

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 [39]. An electronic controller reads out the signal and controls the quantum processor [5]. The existing quantum computers use classical electronic controllers operating at 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 RT is infeasible due to the read complexity, cost, and signal performance of the interconnect [5], [12], [40]. It has therefore been suggested to use a classical CMOS electronic controller operating at cryogenic temperatures [5] or an SFQ controller operating

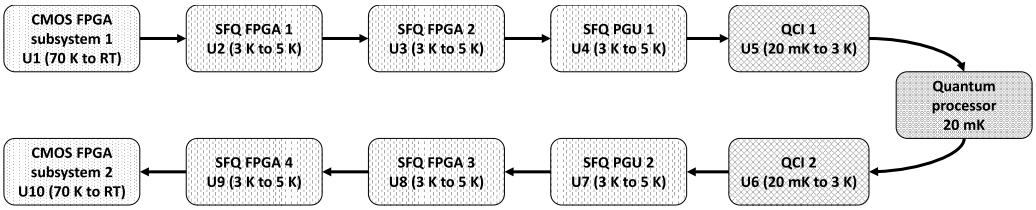


Fig. 7. Flowchart of temperature ranges for each function in a hybrid quantum-classical computing system.

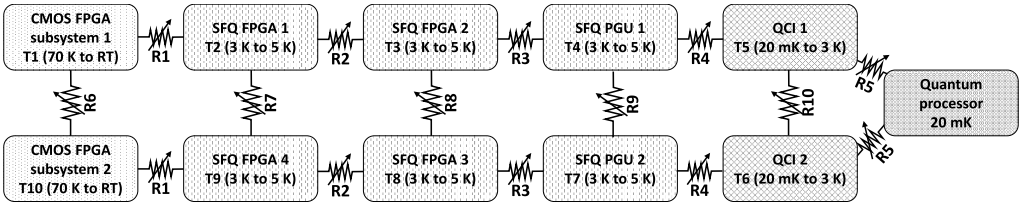


Fig. 8. Simplified thermal–electrical circuit model of the hybrid quantum computer. The thermal relationship between the units is described by the rheostats, which represent the dependence of the thermal resistance on temperature.

below 4 K [29], [41], [42], which can be placed closer to the quantum processor. A block diagram of a hybrid quantum-classical system is depicted in Fig. 5.

While it is possible to operate most of the controller circuitries at temperatures below 4 K, the cooling capacity at these temperatures is often insufficient to efficiently dissipate the heat generated by the controller [5]. Partitioning the controller into higher and lower temperature domains may be more efficient. The proposed algorithm is used to determine the set of optimal temperatures for a hybrid superconductive quantum-classical computing system, as adapted from [41]. The system consists of different readout sensors, multiplexers, demultiplexers, and qubit drivers, placed at temperatures ranging from 20 mK to 4 K [41], [42]. A flowchart of the process is shown in Fig. 7. While the quantum computing system may consist of any number of units [41], the quantum computing system considered here consists of 11 units: two CMOS FPGAs, one for readout and one for control, four SFQ FPGAs, two SFQ pulse generating unit (PGU) systems, two SFQ quantum-classical interface (QCI) integrated circuits, and a quantum processor. Each unit of a quantum computing system is placed within a different temperature domain. The superconductive qubits (quantum processor) are located at 20 mK. The SFQ co-processors for qubit control, error tracking, error correction, readout processing, and execution of the classical portion of the quantum algorithms are placed at temperatures ranging up to 5 K. The digital CMOS FPGA controllers are placed at a temperature ranging from 70 K to RT [41].

The delay and power consumption for each unit at different temperatures are required. These values are generated for this case study. Ten different operating temperatures are available for each unit, yielding 10^{10} possible paths in the process graph. The set of available temperatures is generated by linearly spacing the temperature range for each unit within the system. The CMOS circuits operate at a temperature ranging from 70 K to RT, SFQ circuits operate ranging from 3 to 5 K, and SFQ QCI circuits operate over a range from 20 mK to 3 K [41]. A delay and power at each temperature are assigned

TABLE II

DELAY D_i AND POWER P_i OF EACH UNIT U_i IN A HYBRID QUANTUM COMPUTER. THE DELAY AND POWER VALUES ARE AT THE HIGHEST POSSIBLE OPERATING TEMPERATURE OF EACH UNIT

Unit U_i	**Delay** D, [fs]	**Power** P, [W]	**Unit** U_i	**Delay** D, [fs]	**Power** P, [W]
U_1	1,000	5	U_6	6	70
U_2	150	15	U_7	25	40
U_3	100	25	U_8	110	20
U_4	20	50	U_9	130	20
U_5	5	80	U_{10}	800	7

TABLE III

THERMAL RESISTANCE OF THE HYBRID QUANTUM COMPUTER

Resistance	Ω_T [K/W]	**Resistance**	Ω_T [K/W]
R_1	60	R_6	30
R_2	150	R_7	50
R_3	200	R_8	100
R_4	400	R_9	150
R_5	600	R_{10}	300

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

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 [41]. 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

TABLE IV
SET OF TEMPERATURES FOR A HYBRID QUANTUM COMPUTING SYSTEM COMPOSED OF TWO CMOS FPGAS, FOUR SFQ FPGACS, TWO SFQ PGUS, TWO SFQ QCIS, AND A QUANTUM PROCESSOR. THE MOST OPTIMAL (LOWEST POWER) SET IS HIGHLIGHTED IN BOLD

Unit temperature, K										**Delay**	**Power**
U_1 CMOS1	U_2 FPGA1	U_3 FPGA2	U_4 PGU1	U_5 QCI1	U_6 QCI2	U_7 PGU2	U_8 FPGA3	U_9 FPGA4	U_{10} CMOS2	D_i, [fs]	P_i, [W]
70	**3**	**3**	**3.2**	**1.1**	**0.82**	**3**	**3**	**3**	**70**	**240.0**	**2,456**
70	3	3	3.2	0.82	0.82	3	3	3	70	239.2	2,775
70	3	3	3	0.82	0.82	3.2	3	3	70	239.7	3,475
70	3	3	3.2	1.4	0.51	3	3	3	70	239.8	3,698
70	3	3	3.2	1.1	0.51	3	3	3	70	239.2	3,857

these units. A simplified thermal–electrical circuit model of the system is illustrated in Fig. 8. Ten different thermal resistances between the units are assumed. Each unit type is located at different temperatures. Thermal resistances R_1–R_5 represent the thermal interaction between components, while R_6–R_{10} denote the thermal conduction within a temperature domain. Due to the nature of the thermal resistance, which varies with temperature, the thermal resistance is adjusted based on the temperature of the connected components. Since the components primarily operate at cryogenic temperatures, the thermal resistance between components is assumed to linearly decrease with increasing temperature [43]. The value of the thermal resistances at 4 K is listed in Table III.

A set of optimal temperatures are determined using the algorithm described in Section III. The set of temperatures minimizing the total power while satisfying the delay constraint of 0.24 ps is determined. The algorithm is implemented in Python and executed on an Intel Core i7-9750H workstation with 8-GB RAM. The algorithm completes in 499.65 s for this case study. The sets of optimal temperatures, excluding the quantum processor, are listed in Table IV, where the most optimal set is highlighted in bold. The temperature of the CMOS FPGA and SFQ FPGA modules is the same in the top five most optimal temperature sets. The difference in performance is due to the difference in the temperature of the SFQ PGU and QCI modules. The power consumption of the optimal path with a delay constraint of 0.24 ps is 2456 W. Since the total consumption of the cooled components is 95.5 W, most of the power is consumed by the refrigerators operating at cryogenic temperatures.

V. CONCLUSION

Hybrid cryogenic computing systems are an emerging technology motivated primarily by cryogenic HPC 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 the system while maintaining correct functionality and performance.

A methodology for thermal optimization of cryogenic computing systems with multiple temperature zones is presented in this article. The methodology is validated on a practical case study where the individual temperature of an 11-unit system is optimized. The 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. All possible conditions of the system are represented by this multigraph. 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 the 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, P. Pegus, II, D. Irwin, P. Shenoy, J. Goodhue, and J. Culbert, "Design and operational analysis of a green data center," *IEEE Internet Comput.*, vol. 21, no. 4, pp. 16–24, Aug. 2017.
[2] M. Koot and F. Wijnhoven, "Usage impact on data center electricity needs: A system dynamic forecasting model," *Appl. Energy*, vol. 291, Jun. 2021, Art. no. 116798.
[3] D. S. Holmes, A. L. Ripple, and M. A. Manheimer, "Energy-efficient superconducting computing—Power budgets and requirements," *IEEE Trans. Appl. Supercond.*, vol. 23, no. 3, Jun. 2013, Art. no. 1701610.
[4] N. Zhuldassov and E. G. Friedman, "Temperature–frequency boundary of cryogenic dynamic logic," *Microelectron. J.*, vol. 135, Mar. 2023, Art. no. 105763.
[5] B. Patra et al., "Cryo-CMOS circuits and systems for quantum computing applications," *IEEE J. Solid-State Circuits*, vol. 53, no. 1, pp. 309–321, Jan. 2018.
[6] B. Oy. (2023). *Technical Specifications of BlueFors Oy Cryocoolers*. [Online]. Available: https://bluefors.com/products
[7] Sumitomo Heavy Industries. (2023). *Technical Specifications of Sumitomo Cryocoolers*. [Online]. Available: https://www.shi.co.jp/english/products/machinery/cold/index.html
[8] SunPower. (2023). *Technical Specifications of Sunpower Cryocoolers*. [Online]. Available: https://www.sunpowerinc.com/products/stirling-cryocoolers/cryocoolers-overview
[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 Trans. Electron.*, vol. E91-C, no. 3, pp. 306–317, Mar. 2008.
[11] D. Gupta et al., "Modular, multi-function digital-RF receiver systems," *IEEE Trans. Appl. Supercond.*, vol. 21, no. 3, pp. 883–890, Jun. 2011.
[12] L. Vandersypen and A. van Leeuwenhoek, "Quantum computing—The next challenge in circuit and system design," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 24–29.
[13] A. Montanaro, "Quantum algorithms: An overview," *NPJ Quantum Inf.*, vol. 2, no. 1, pp. 1–8, Jan. 2016.
[14] R. Bairamkulov and E. G. Friedman, *Graphs in VLSI*. Springer, 2023.

[15] G. B. Mathews, "On the partition of numbers," *Proc. London Math. Soc.*, vol. 1, no. 1, pp. 486–490, 1896.

[16] S. Martello and P. Toth, "Solution of the zero-one multiple knapsack problem," *Eur. J. Oper. Res.*, vol. 4, no. 4, pp. 276–283, Apr. 1980.

[17] K. Chowdhury and S. Sarangi, "Performance of cryogenic heat exchangers with heat leak from the surroundings," in *Advances in Cryogenic Engineering*, vol. 29, R. W. Fast, Ed. Boston, MA, USA: Springer, 1984, pp. 273–280.

[18] D. W. Hahn and M. N. Özisik, *Heat Conduction*. Hoboken, NJ, USA: Wiley, 2012.

[19] H.-K. Lyeo and D. G. Cahill, "Thermal conductance of interfaces between highly dissimilar materials," *Phys. Rev. B, Condens. Matter*, vol. 73, no. 14, Apr. 2006, Art. no. 144301.

[20] D. Thornburg, E. Thall, and J. Brous, "A manual of materials for microwave tubes," Radio Corp. Amer., New York, NY, USA, Tech. Rep. 60-325, Jan. 1961.

[21] C. Schmidt, "Simple method to measure the thermal conductivity of technical superconductors, e.g., NbTi," *Rev. Sci. Instrum.*, vol. 50, no. 4, pp. 454–457, Apr. 1979.

[22] P. E. Bradley et al., "Properties of selected materials at cryogenic temperatures," *Nat. Inst. Standards Technol.*, vol. 680, pp. 1–14, Jun. 2013.

[23] N. Simon, E. Drexler, and R. Reed, "Properties of copper and copper alloys at cryogenic temperatures," Nat. Inst. Standards Technol. (MSEL), Boulder, CO, USA, Tech. Rep. PB-92-172766/XAB; NIST/MONO-177, Feb. 1992.

[24] E. Jeffrey et al., "Fast accurate state measurement with superconducting qubits," *Phys. Rev. Lett.*, vol. 112, no. 19, May 2014, Art. no. 190504.

[25] J. M. Elzerman, R. Hanson, L. H. W. van Beveren, B. Witkamp, L. M. K. Vandersypen, and L. P. Kouwenhoven, "Single-shot read-out of an individual electron spin in a quantum dot," *Nature*, vol. 430, no. 6998, pp. 431–435, Jul. 2004.

[26] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, "Surface codes: Towards practical large-scale quantum computation," *Phys. Rev. A, Gen. Phys.*, vol. 86, pp. 032324.1–032324.48, Sep. 2012.

[27] L. Song, H. Homulle, E. Charbon, and F. Sebastiano, "Characterization of bipolar transistors for cryogenic temperature sensors in standard CMOS," in *Proc. IEEE SENSORS*, Oct. 2016, pp. 1–3.

[28] I. D. C. Lamb et al., "An FPGA-based instrumentation platform for use at deep cryogenic temperatures," *Rev. Sci. Instrum.*, vol. 87, no. 1, Jan. 2016, Art. no. 014701.

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

[30] R. M. Incandela, L. Song, H. Homulle, E. Charbon, A. Vladimirescu, and F. Sebastiano, "Characterization and compact modeling of nanometer CMOS transistors at deep-cryogenic temperatures," *IEEE J. Electron Devices Soc.*, vol. 6, pp. 996–1006, 2018.

[31] N. Zhuldassov and E. G. Friedman, "Cryogenic dynamic logic," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Oct. 2020, pp. 1–5.

[32] V. Lacquaniti, D. Andreone, N. De Leo, M. Fretto, A. Sosso, and M. Belogolovskii, "Engineering overdamped niobium-based Josephson junctions for operation above 4.2 K," *IEEE Trans. Appl. Supercond.*, vol. 19, no. 3, pp. 234–237, Jun. 2009.

[33] N. Takeuchi, D. Ozawa, Y. Yamanashi, and N. Yoshikawa, "An adiabatic quantum flux parametron as an ultra-low-power logic device," *Supercond. Sci. Technol.*, vol. 26, no. 3, Jan. 2013, Art. no. 035010.

[34] V. K. Semenov, Y. A. Polyakov, and S. K. Tolpygo, "New AC-powered SFQ digital circuits," *IEEE Trans. Appl. Supercond.*, vol. 25, no. 3, pp. 1–7, Jun. 2015.

[35] (2018). *IARPA SuperTools Program*. Accessed: May 2022. [Online]. Available: https://www.iarpa.gov/index.php/research-programs/supertools

[36] G. Krylov, J. Kawa, and E. G. Friedman, "Design automation of superconductive digital circuits: A review," *IEEE Nanotechnol. Mag.*, vol. 15, no. 6, pp. 54–67, Dec. 2021.

[37] Y. Ando, R. Sato, M. Tanaka, K. Takagi, N. Takagi, and A. Fujimaki, "Design and demonstration of an 8-bit bit-serial RSFQ microprocessor: CORE e4," *IEEE Trans. Appl. Supercond.*, vol. 26, no. 5, pp. 1–5, Aug. 2016.

[38] A. G. Qoutb and E. G. Friedman, "MTJ magnetization switching mechanisms for IoT applications," in *Proc. Great Lakes Symp. VLSI*, May 2018, pp. 347–352.

[39] M. H. Devoret, A. Wallraff, and J. M. Martinis, "Superconducting qubits: A short review," Nov. 2004, *arXiv:cond-mat/0411174*.

[40] M. Reiher, N. Wiebe, K. M. Svore, D. Wecker, and M. Troyer, "Elucidating reaction mechanisms on quantum computers," *Proc. Nat. Acad. Sci. USA*, vol. 114, pp. 7555–7560, Jul. 2017.

[41] O. Mukhanov et al., "Scalable quantum computing infrastructure based on superconducting electronics," in *IEDM Tech. Dig.*, Dec. 2019, pp. 31.2.1–31.2.4.

[42] N. K. Katam, O. A. Mukhanov, and M. Pedram, "Superconducting magnetic field programmable gate array," *IEEE Trans. Appl. Supercond.*, vol. 28, no. 2, pp. 1–12, Mar. 2018.

[43] C. Y. Ho, R. W. Powell, and P. E. Liley, "Thermal conductivity of the elements," *J. Phys. Chem. Reference Data*, vol. 1, no. 2, pp. 279–421, Apr. 1972.