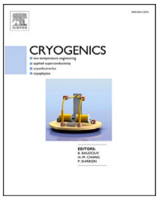
Cryogenics 146 (2025) 104013

Contents lists available at ScienceDirect

Cryogenics

journal homepage: www.elsevier.com/locate/cryogenics



Heat load efficiency in multi-temperature cryogenic computing systems ☆

Nurzhan Zhuldassov *, Rassul Bairamkulov, Eby G. Friedman

Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627, United States of America

ARTICLE INFO

Keywords:
Thermal optimization
CryoCMOS
SFQ
Cloud computing
Graph theory

ABSTRACT

Heterogeneous cryogenic computing systems often incorporate a variety of technologies, each functioning at different temperatures. The chosen operating temperature of these components significantly influences the overall power dissipation, heat load, and system performance. Existing design methodologies for managing cryogenic systems with multiple temperature zones often overlook thermal variations within these zones, the interconnect between different zones, and are restricted to the temperature within a single zone. A comprehensive framework designed to enhance the efficiency of heterogeneous computing systems operating under cryogenic conditions is presented in this paper. Utilizing a graph theoretic approach, the framework is used to evaluate the influence of operating temperatures on both delay and power consumption. Thermal interactions among different system components are also considered, enabling a more precise estimate of the power requirements and local thermal load. The methodology is applied to two case studies related to cryogenic cloud computing systems. The objective is to minimize overall system-wide power consumption while satisfying specific performance criteria and considering the impact of heat load on the cooling infrastructure.

1. Introduction

The escalating demand for high performance computing (HPC) in recent years, propelled by the rise of computationally demanding applications such as cloud computing, has introduced several challenges. These challenges include energy efficiency, thermal management, and system performance. Data centers, which form the backbone of HPC systems, consume significant energy, ranging from tens to hundreds of megawatts [1]. The global annual energy consumption for HPC is estimated at approximately 200 TWh and is expected to quadruple by 2030 [2].

To sustain this rapid expansion, alternative computational technologies are being explored. Cryogenic technologies have emerged as a promising solution, capable of substantially reducing power consumption in large scale computing systems, including the energy associated with refrigeration [3,4]. Cryogenic technology can enhance communication bandwidths, increase data transmission rates, improve response speeds, enhance reliability, and reduce losses in telecommunication systems [4–7]. Significant performance improvements have been observed in semiconductor devices, such as MOSFETs, MESFETs, HEMTs, and CMOS [6], and in optoelectronic and electro-optic components as well as in metal conductors such as aluminum and copper [8,9]. At extreme cryogenic temperatures, such as liquid helium (4 K), certain materials, such as niobium (Nb), can achieve superconductivity [10]. Combining electronics with other cryogenic technologies can justify the cost and complexity of cryogenic cooling, leading to hybrid systems with enhanced performance and reliability. Improved performance and reliability can lead to wider acceptance and distribution of cryogenic technologies in a large variety of industrial and commercial applications.

The cost effectiveness of cryogenic technology however requires evaluation. While the cooling efficiency at smaller scale remains low, at larger scales, such as helium liquefaction plants, the exergy efficiency can reach 23.4% [11]. The cooling capacity of cryogenic technologies operating at 4 K may however be insufficient for efficient heat dissipation [12]. Furthermore, the available cooling power varies at different temperatures. As described in [7], the available cooling power per kilowatt input power differs across temperatures. Consequently, certain circuits may benefit from operating at lower temperatures, whereas other circuitry would achieve satisfactory performance at higher temperatures. The power consumption of cooling systems in data centers currently accounts for 42% of the total power consumption [13]. In-

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

* Corresponding author.
E-mail address: nzhuldas@ur.rochester.edu (N. Zhuldassov).

https://doi.org/10.1016/j.cryogenics.2024.104013
Received 11 September 2024; Received in revised form 26 December 2024; Accepted 29 December 2024

Available online 6 January 2025

creased efficiency at cryogenic temperatures combined with the higher cooling efficiency of large scale refrigeration would result in higher power efficiency and better performance as compared to cloud computing centers operating at room temperature.

Optimizing the temperature of each subsystem within a computing system enhances overall performance, power efficiency, and heat management. For instance, changing the temperature of a subsystem within a cryogenic system supports the integration of different technologies and functionalities at distinct temperature zones within a cryocooler. This strategic placement can lead to a reduction in refrigeration costs. Raising the temperature of a subsystem may however lead to increased latency and power dissipation. Furthermore, the refrigeration characteristics of adjacent subsystems, particularly those operating at lower temperatures, can be affected if these nearby systems are not adequately thermally isolated. The heat emanating from a subsystem operating at a higher temperature could inadvertently affect the cooling efficiency of neighboring lower temperature systems, particularly if these systems are interconnected.

Multiple research efforts have focused on placing different technologies and functions across different temperature stages within a refrigeration system [14,15]. For instance, a hybrid temperature system integrated in the Sumitomo SRDK-101DP-11C cryocooler is described in [14]. This system features a 4 K stage for low temperature superconductive circuits and a 60 K stage for higher temperature semiconductor circuits, such as analog filters and low noise amplifiers, while other electronics are maintained at room temperature (RT). These studies, however, often overlook the full spectrum of temperatures accessible within a given stage. In this system, for example, although the second stage is set to 60 K, the actual temperature range can accommodate temperatures between 60 K to 80 K. Such insights highlight the importance of considering the entire range of feasible temperatures to optimize the placement and operation of different stages within a cryogenic system.

Exploiting the full range of temperatures available in each stage of a cryocooler can enhance the performance of the overall computing system. A methodology is proposed to determine the operating temperature of each component within a cryogenic system to minimize the total power consumption while ensuring that the performance achieves a target objective. The particular thermal characteristics of each stage within a cryocooler is considered in the optimization process, producing a more efficient distribution of the components based on thermal and power requirements.

In [7], the temperature variability within each stage is exploited. The dependence of the power consumption of each cooler stage on the temperature of a neighboring stage is, however, not considered. A computing unit can be placed inside a refrigerator equipped with multiple cooling stages; for example, three stages, where the temperature of stage 1 is 150 K, stage 2 is 70 K, and stage 3 is 4 K. In contrast, if the same computer is placed inside a refrigerator with 120 K at stage 1, 60 K at stage 2, and 4 K at stage 3, the performance and power consumption will be different. The advancement described here considers, for example, the variation in power consumption of the third stage of a cryocooler that is dependent on the temperature of the first and second stages. Thus, this methodology improves the operation of the overall system rather than individually optimizing the operation within each temperature zone.

The application of cubic spline interpolation is introduced to estimate the power consumption and performance at different temperatures. The proposed methodology is validated on two case studies of a superconductive cloud computing system. The system utilizes both CMOS and superconductive logic [16] for computation and storage.

The paper is organized as follows: in Section 2, insight into the organization of cryogenic coolers is described, the problem is formulated, and the thermal behavior of the system is discussed. The proposed methodology is presented in Section 3. Two case studies, hybrid superconductive/semiconductor cloud computing systems optimized using the proposed methodology, are described in Section 4. Some conclusions are offered in Section 5.

2. Cryogenic cooling and thermal optimization

The integration of cryogenic computing systems is enhanced when electronic circuits operate at different cryogenic temperatures [12]. The primary goal of the proposed methodology is to identify a set of optimal temperatures for each component, minimizing the total power consumption or delay of the cryogenic system.

The methodology determines the optimal temperature of the different components of an electronic system, as shown in Fig. 1. A graph of the system is initially constructed using power and delay values of each component at different temperatures. Cubic spline interpolation [17] is employed to approximate and interpolate the variable weights within the graph. An algorithm based on graph theory [18] is utilized to determine the set of optimal temperatures. In this algorithm, the power and delay weights in the graph are adjusted after each state based on the temperature from the preceding state, as described in Section 3. To determine the optimal temperature within each zone, the algorithm is executed twice, with the second iteration refining the available temperature range based on the results of the first pass. Once the appropriate temperature set satisfying the constraints is identified, a thermal model of the system is applied to assess the heat flow (or power transfer) between units. This heat flow is influenced by the thermal conductance between the units and the temperature of the interconnecting cables. Furthermore, the leakage power is estimated based on the heat flow; particularly, the power loss due to additional cooling required in the lower temperature components resulting from heat transfer from the higher temperature components. The leakage of heat is contingent on the thermal conductivity of the cables connecting the components. Inclusion of this effect significantly expands the generality of the work described in [7]. The overall power consumption for a given set of temperatures thus includes leakage power between different temperature zones. To achieve optimal system operation, the delay, power, and heat flow among the components should be considered, allowing the temperature of each component (or thermal stage) to be determined.

The rest of this section is organized as follows. Background information describing cryogenic coolers is discussed in Section 2.1. The problem formulation based on a graph theoretic approach is described in Section 2.2. Cubic spline interpolation is summarized in Section 2.3. A thermal model of the system is discussed in Section 2.4.

2.1. Cryogenic coolers

Cryogenic coolers are essential in low temperature applications such as superconductive devices [16], quantum computing, medical systems, and materials research. Several types of cryocoolers exist, such as Gifford-McMahon cryocoolers, pulse-tube cryocoolers, and dilution refrigerators [19]. Dilution refrigerators can achieve tens to hundreds of millikelvins. Achieving these extremely low temperatures requires multiple stages of cooling, each stage being progressively colder than the previous stage. The overall structure of a cryogenic cooler typically consists of multiple temperature chambers, each containing a refrigerant that cools the following stage. For example, the PT415 pulse tube cryocooler from Cryomech [20] has a 50 K stage with a temperature range between 35 K to 77 K, and a 4 K stage with a temperature range between 3 K to 18 K.

The first stage of a cryogenic cooler typically contains both air cooled and water cooled compressors, with temperatures ranging between approximately 120 K to room temperature [21,22]. The next stage usually operates within a temperature range between 50 K to 120 K and is cooled by liquid nitrogen [20–22]. To reach lower temperatures, neon can be used in the stage between 20 K and 50 K [19]. Liquid helium is used as a refrigerant in the next stage to achieve a temperature between 3 K and 5 K [19,21–23]. To reach the lowest temperatures, down to tens to hundreds of millikelvins, ^{3}He/^{4}He dilution refrigerators are used, where Helium-3 and Helium-4 isotopes are mixed [19,23]. Some examples of

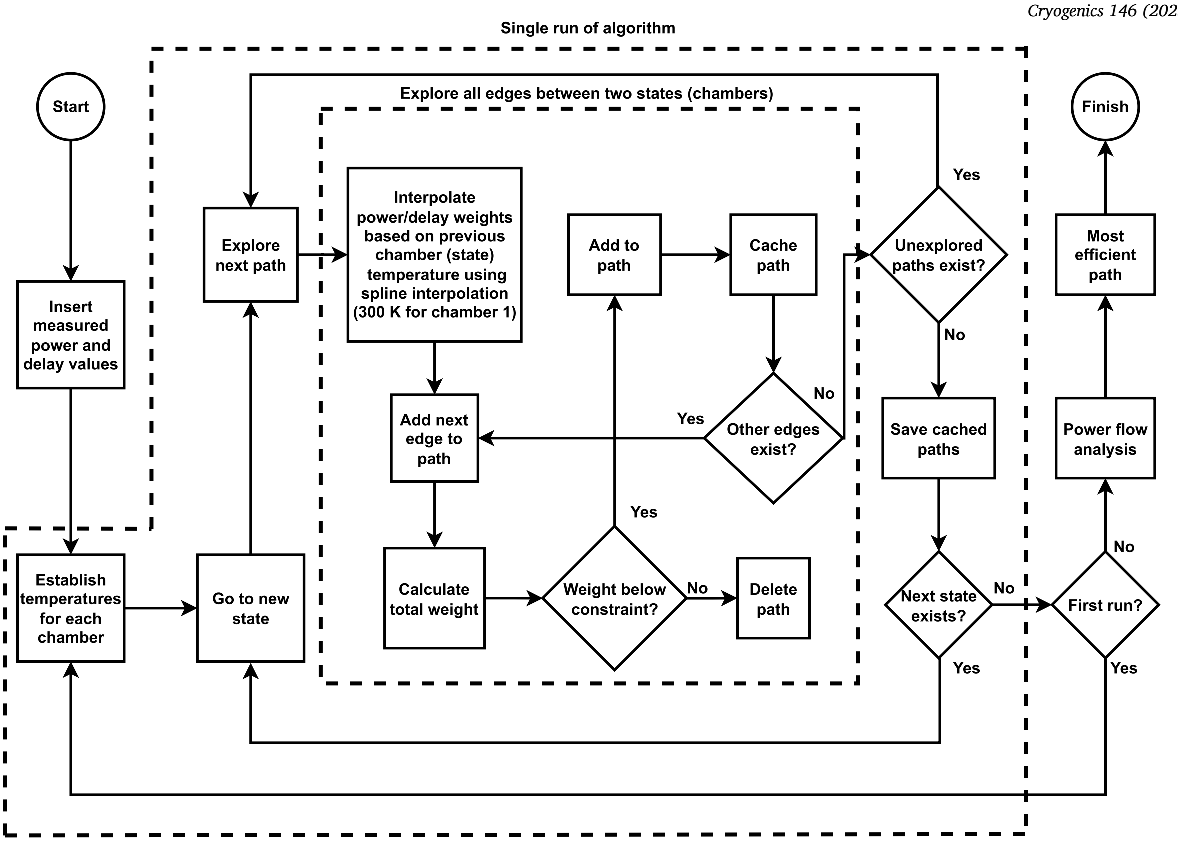


Fig. 1. Flowchart of the proposed methodology.

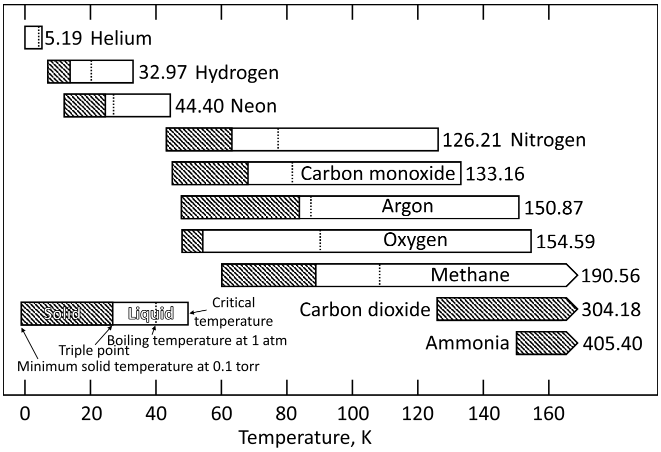


Fig. 2. Commonly used coolants and relevant operating temperatures [19,24–26]. The critical temperature is the maximum temperature above which the substance cannot exist in a liquid state regardless of pressure.

commonly used coolants and related operating temperatures are noted in Fig. 2 [19,24–26].

2.2. Formulation of thermal optimization problem

The objective is to choose the ideal operating temperature at each stage of the process. The optimization process for choosing a zone specific temperature can be conceptualized as a directed acyclic multi-weighted multigraph, with one variable weight and one constant weight at each edge $G := \langle S, C, W\rangle$. Here, $S = S_1, S_2, \ldots, S_n$ is a finite set of states defining a specific instance of the temperature optimization problem. The edges, denoted by C, represent chambers within a refrigeration unit, each containing one or more computing units. For each chamber i, subset $C_i \subseteq C$ corresponds to the parallel edges. Typical refrigeration systems operate at specific temperatures, such as liquid helium temperature (LHT) or liquid nitrogen temperature (LNT), represented by the set of available temperatures $T = T_1, T_2, \ldots, T_j$. Each chamber, operating at a different temperature during each step, is denoted by $c_{i,j} \in C_i$. $W_{i,j} := \langle p(T_{i-1,j}), d\rangle \in \mathbb{R}^2_{>0}$, where $p(T_{i-1,j})$ and d represent, respectively, the power consumption and delay of a unit at a specific temperature. The power consumption is a variable weight dependent on the previous edge weight along the chosen path.

Path π in this process graph, representing a set of operating temperatures for each unit, links the source to the sink. Path π is described as

$$\pi = \left(C_1(T_j), C_2(T_j), \ldots, C_i(T_j)\right) \, . \tag{1}$$

The power consumption of a process is the sum of the power weights along a path, $P(\pi) = p_1 + p_2 + \cdots + p_{n-1}$. Each weight represents the power consumption of a unit. Similarly, the total delay of a process, $D(\pi) = d_1 + d_2 + \cdots + d_{n-1}$, is the sum of the delay weight of the units along the path. The temperature optimization problem is to determine a path that minimizes the total power $P(\pi)$ while keeping the total delay $D(\pi)$ below a maximum limit D_{max},

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

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

An example with three units and four states is illustrated in Fig. 3. Each unit can operate at three distinct temperatures, indicated by the parallel edges between adjacent states. Paths A and B, highlighted in bold, represent two potential configurations. Path A,

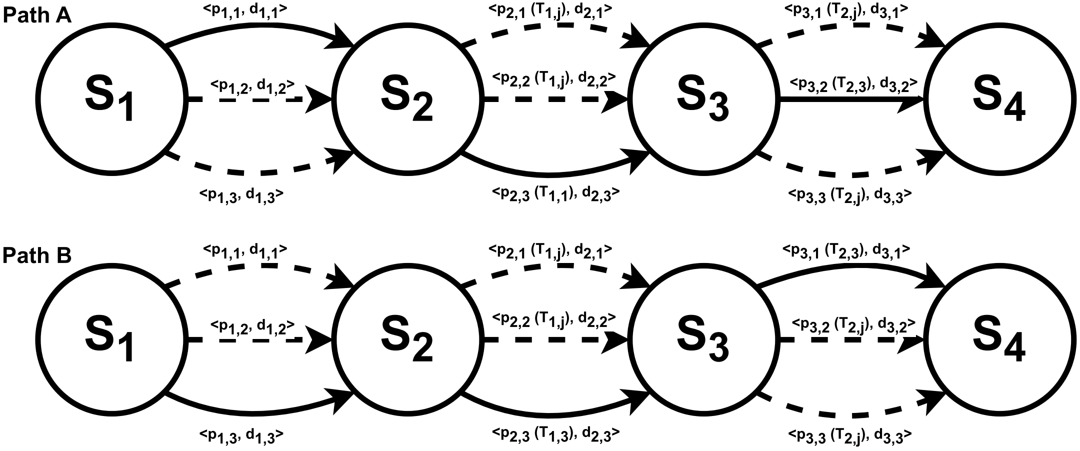


Fig. 3. Example of the temperature design process using a multiweighted multigraph with variable edges. The edges between two states denote a chamber within a refrigerator at different temperatures. Two paths, A and B, are shown highlighted in bold. The power consumption of path A is $P(\pi_A) = p_{1,1} + p_{2,3}(T_{1,1}) + p_{3,2}(T_{2,3})$, and the power consumption of path B is $P(\pi_B) = p_{1,3} + p_{2,3}(T_{1,3}) + p_{3,1}(T_{2,3})$. The first index of the variables represents the device number, while the second index represents the position of the temperature within the list of available temperatures for that specific temperature zone. Note that although the edge between states S_2 and S_3 is the same in both paths, the power weight of the edge is different. The power consumption is determined by the temperature of the previous chamber, which is denoted by $T_{1,1}$ in the edge between states S_2 and S_3 in path A, and by $T_{1,3}$ in path B. The delay of path A and path B is, respectively, $D(\pi_A) = d_{1,1} + d_{2,3} + d_{3,2}$ and $D(\pi_B) = d_{1,3} + d_{2,3} + d_{3,1}$.

$\pi_A = (C_{1,1}, C_{2,3}, C_{3,2})$, and Path B, $\pi_B = (C_{1,3}, C_{2,3}, C_{3,1})$, correspond to different temperature settings for the chambers. The power consumption and delay of these paths are based on the respective weight of the edges along the path.

2.3. Cubic spline interpolation

The cubic spline interpolation technique is used here [17] to interpolate the power consumption of a unit within a cooler chamber. An interpolating function is used, as it is intractable to manually define a graph with fixed weights for all possible combinations of temperatures. Cubic spline interpolation is particularly useful in this application, as the cubic curvature and slope of the resulting interpolation are similar to the exponential relationship between temperature and power consumption [4]. Other linear interpolation techniques [27], such as Lagrange polynomial and Newton polynomial, are less suitable due to higher errors, as depicted in Fig. 4

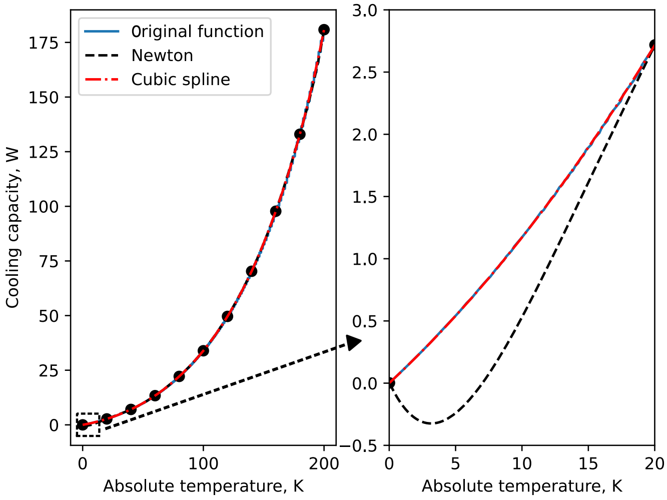
For a given set of $n+1$ data points (x_i, y_i) with no repeating x_i and $a = x_0 < x_1 < \cdots < x_n = b$, spline $S(x)$ is a function satisfying

1. $S(x) \in C^2[a,b]$;
2. On each $[x_{i-1}, x_i], S(x)$ is a polynomial of degree 3, where $i = 1, \ldots, n$;
3. $S(x) = y_i, \forall i = 0, 1, \ldots, n$.

Interpolated function $S(x)$ becomes

$$S(x) = \begin{cases} C_1(x), & x_0 \le x \le x_1 \\ \cdots & \\ C_i(x), & x_{i-1} \le x \le x_i \\ \cdots & \\ C_n(x), & x_{n-1} \le x \le x_n, \end{cases} \tag{4}$$

where each $C_i = a_i + b_i x + c_i x^2 + d_i x^3$ is a cubic function with $d_i \neq 0$ for $i = 1, \ldots, n$. An example comparing cubic spline interpolation with Newton polynomial interpolation is shown in Fig. 4. Interpolation techniques are used to approximate the available cooling power per kilowatt of input power [7]. Note that in this case cubic spline interpolation exhibits less error than Newton polynomial interpolation.



(a)

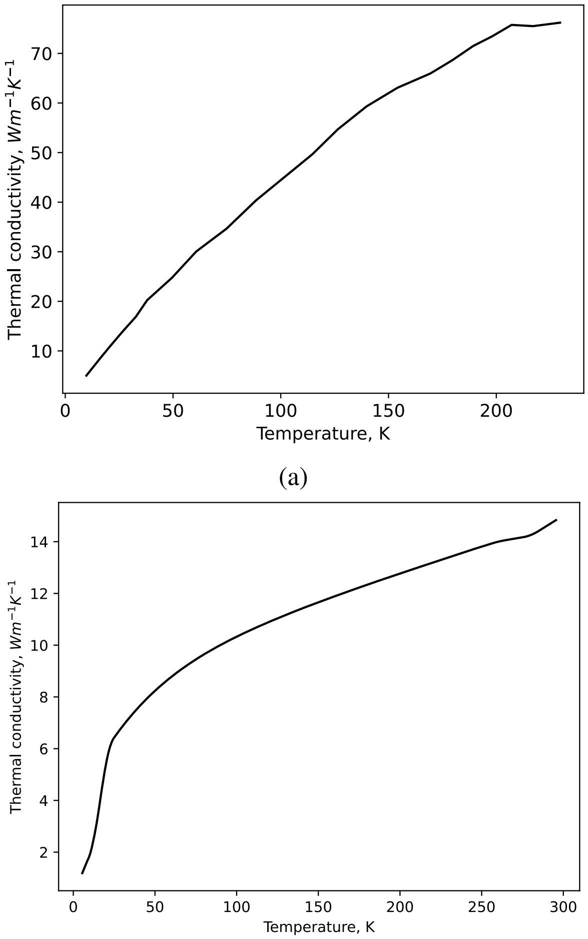
(b)

Fig. 4. Techniques to interpolate the available cooling power per kW input power, (a) the 20 mK to 200 K range, and (b) the 20 mK to 20 K range. Cubic spline interpolation closely resembles the original curve as compared to Newton polynomial interpolation.

2.4. Thermal model

In addition to the heat generated by the units, the power consumed along a path also includes thermal leakage from temperature differences between chambers transferred through the connecting cables [7]. As the thermal resistance of the cable materials changes with temperature, adjustments are made for more precise heat flow characterization between units [28,29]. The thermal resistance changes based on the material type, quality, and purity, and can change either linearly, exponentially, or logarithmically [28–30].

For cryogenic applications, specialized cables such as CryoCoax BCB016, BCB019, and BCB029, composed of stainless steel and beryllium copper, are employed [31]. These materials demonstrate increasing thermal conductivity with rising temperature [8,9,32]. While the thermal conductivity of beryllium copper can be linearly approximated [8], the conductivity of stainless steel is best represented through a dual-line approximation [9], as depicted in Fig. 5.



(a)

(b)

Fig. 5. Thermal conductivity, (a) beryllium copper, and (b) stainless steel.

The thermal conductivity of different cables is considered to construct thermal circuits, analogous to electrical circuits. The heat flow within a system, denoted as q_T, is represented by a set of linear equations,

$$q_T = \begin{array}{cc} & \begin{array}{ccc} U_1 & U_k & U_n \end{array} \\ \begin{array}{c} U_1 \\ U_k \\ U_n \end{array} & \begin{bmatrix} \frac{\Delta T_{1,1}}{R_{1,1}} & \cdots & \frac{\Delta T_{1,n}}{R_{1,n}} \\ \vdots & \ddots & \vdots \\ \frac{\Delta T_{n,1}}{R_{n,1}} & \cdots & \frac{\Delta T_{n,n}}{R_{n,n}} \end{bmatrix} \end{array}. \tag{5}$$

Heat flow is determined by the temperature difference ΔT and thermal resistance R between units. The power transfer to or from each unit is the sum of the heat flow along each row of this matrix,

$$\Delta P = q_T\,\mathbf{1}_n, \tag{6}$$

where $\mathbf{1}_n$ is a column vector of ones,

$$\mathbf{1}_n = [1, \ldots, 1]^{\intercal}. \tag{7}$$

This approach supports a systematic determination of the power flow due to heat being transferred among multiple units within a cryogenic system.

3. Optimization setup

The initial phase of this methodology requires a system graph, as described in Section 2.2. In this graph, each path from the starting state S_1 to the final state S_n affects the power consumption and delay. The optimization problem is to identify the most energy efficient set of temperatures, ensuring that the overall delay of the system remains below a maximum delay D_{max}. A flowchart of the algorithm to identify all of the paths satisfying a delay constraint is illustrated in Fig. 1.

The algorithm utilizes an input delay matrix D, where each element $D_{i,j}$ denotes the delay of unit i at temperature T_j,

$$D = \begin{array}{cc} & \begin{array}{ccc} C_1 & C_k & C_n \end{array} \\ \begin{array}{c} T_1 \\ T_j \\ T_m \end{array} & \begin{bmatrix} D_{1,1} & \cdots & D_{1,n} \\ \vdots & \ddots & \vdots \\ D_{m,1} & \cdots & D_{m,n} \end{bmatrix} \end{array}. \tag{8}$$

Delay values within a temperature range are determined via cubic spline interpolation, as outlined in Section 2.3. The power weights are similarly estimated.

Breadth first search is employed in the algorithm to traverse the process graph from the source node, comparing the delay of each partial path to D_{max}. Paths satisfying the delay constraint are tracked. Upon exploring all of the edges from a node, the algorithm proceeds to the next node. Since the power weights are variable, the weight is reliant on the power weight of the edges preceding the current state. The initial and interpolated power weights are adjusted to obtain new power weights. Newton's law of cooling is applied, based on the previous edge, to achieve the required adjustments.

Memory usage and computational time are improved by a two step approach. First, those paths not satisfying both power and delay constraints are discarded. Second, running the algorithm twice refines the temperature range in subsequent steps, reducing the complexity from $O(n^4)$ to $O(n^2)$. This technique maintains precision while enhancing the computational efficiency.

Once all of the potential paths satisfying the delay limit are identified, the next step evaluates the heat flow between units to determine the total power consumption along each path. This process is described in Section 2.4. The set of optimal temperatures is selected based on the lowest system-wide power dissipation.

The proposed methodology has certain limitations. The approach described in this paper relies on a predetermined number of refrigeration stages and fixed unit configurations. Specifically, the optimal temperature of each stage is determined for a preset number of stages, and units are assigned to each refrigeration chamber in a fixed manner without allowing for flexible grouping of units within a chamber. This fixed assignment restricts the flexibility and efficiency of the temperature management process across multiple units and temperature zones. To address these limitations, future research should focus on optimizing the number of refrigeration stages based on the specific requirements of the system to enable more flexible and efficient temperature management.

4. Case study: cryogenic cloud computing

Cloud computing provides computational, software, and storage services [33]. Cloud computing centers are stationary and typically operate at room temperature. Cloud computing systems can be placed within a cryogenic environment to utilize cryoCMOS and superconductive logic to enhance computational speeds while reducing energy consumption and heat load [6,7,16]. A block diagram of a cloud computing system operating within different temperature zones is depicted in Fig. 6. Note that the system is partitioned into different temperature zones based on the cooling requirements.

In many instances, operating the majority of a system at temperatures below 4 K is feasible. However, at these lower temperatures, the cooling capacity often proves inadequate for effectively dissipating the heat produced by the CMOS components within the system [12]. Consequently, dividing a system into domains with higher and lower temperatures may result in more efficient power and delay characteristics. The proposed algorithm determines the set of optimal temperatures

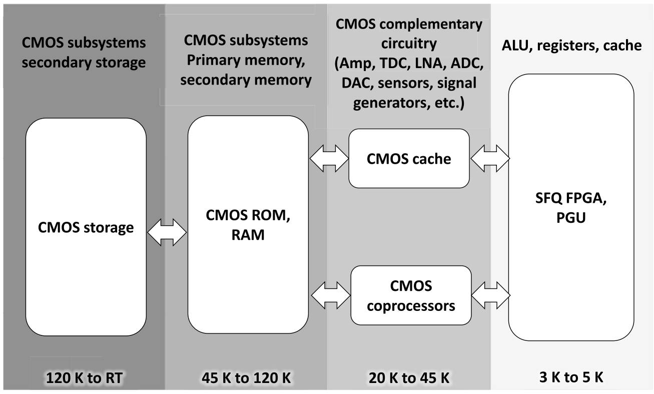

CMOS subsystems secondary storage
CMOS subsystems Primary memory, secondary memory
CMOS complementary circuitry (Amp, TDC, LNA, ADC, DAC, sensors, signal generators, etc.)
ALU, registers, cache
CMOS storage
CMOS ROM, RAM
CMOS cache
CMOS coprocessors
SFQ FPGA, PGU
120 K to RT
45 K to 120 K
20 K to 45 K
3 K to 5 K

Fig. 6. Block diagram of cloud computing system placed at different temperature zones.

Table 1
Delay D_i and power P_i (including refrigeration) of each circuit group within a cryogenic cloud computing system at the maximum feasible operating temperature for each unit.

Refrigerator	**Delay**	**Power**
Stage	D, [ns]	P, [kW]
Stage 1	2,500	32
Stage 2	350	25
Stage 3	150	3
Stage 4	80	1

to minimize energy for a cryogenic cloud computing system. The system is comprised of different components such as multiplexers, demultiplexers, arithmetic logic units, registers, and other computing units, which are strategically placed within different temperature zones ranging from 3 K to 300 K. Depending upon the dissipated heat of the circuits, some of the components are placed in the temperature domain between 3 K to 5 K, cooled by liquid helium. SFQ circuits can be placed in this temperature zone due to significantly lower heat dissipation as compared to CMOS [16]. CMOS circuitry can be partitioned into temperature domains ranging between 20 K to RT. The CMOS circuitry dissipating the least heat is placed within the 20 K to 45 K temperature domain, cooled by liquid or solid neon. Primary and secondary memory, such as RAM, are placed within the next refrigerator chamber, cooled by liquid or solid nitrogen within the 45 K to 120 K temperature domain. The CMOS circuitry consuming the most heat is placed within the 120 K to RT domain chamber, which is cooled by an air and/or water cooled compressor [21,22].

Estimates of the delay and power of each unit at different temperatures are generated for this case study. Four different operating temperatures are available in each chamber. The set of available temperatures is generated by linearly spacing the temperature range for each device within each system. Each computing unit is assigned delay and power values at each temperature, which are treated as the average delay and power. In this case study, power consumption and delay measurements at ten distinct temperatures are used. These measurements are subsequently employed to estimate the power consumption and delay at different temperature intervals through cubic spline interpolation, as described in Section 2.3. The delay and power for each circuit group placed in separate refrigerator chambers are listed in Table 1. The overall power consumption encompasses not only the power dissipated by the computing units but also includes the energy consumed by the refrigeration systems.

Thermal interactions between units within the cryogenic system are governed by the interconnects and spatial proximity. The connections between refrigerator chambers utilize low heat superconductive loads and superconductive ribbon cables with minimal crosstalk for reliable and precise transmission of the logic signals [16]. Other necessary interconnects are established through cryocoax cryogenic cables [31]. These connections, combined with the nonideal cooling efficiency of the refrigerators, produce a thermal conductance between the chambers. A simplified thermal-electrical circuit model of the system is depicted in Fig. 7.

Table 2
Thermal resistances of the cryogenic cloud computing system.

Resistance	Ω_T [K/W]
R_1	60
R_2	150
R_3	200
R_4	310
R_5	450

Table 3
Set of temperatures for a cryogenic cloud computing system. The set of temperatures producing the lowest power consumption is in bold.

Stage temperature, K				**Delay**	**Power**
C_1	C_2	C_3	C_4	D_i, [ns]	P_i, [kW]
180	**53.0**	**20.0**	**4.78**	**797.90**	**225.94**
200	61.4	20.0	4.88	797.75	229.05
200	53.2	22.7	4.90	795.45	229.77
200	69.7	20.0	4.90	798.93	229.83
220	61.4	20.0	4.95	796.20	230.51

The example system includes five distinct thermal resistances influencing the interactions between circuit blocks. Some blocks are situated within the same stage of a refrigerator. Thermal interactions are assumed to only occur between different stages of a refrigerator. Thermal resistances R_1, R_2, and R_3 represent interactions between adjacent stages, R_4 denotes the thermal conduction between stages 1 and 3, and R_5 considers interactions between stages 2 and 4. Since the thermal resistance changes with temperature, the resistance is adjusted based on the temperature of the interconnected components. Since the components operate at cryogenic temperatures, the thermal resistance between these components linearly decreases as the temperature rises [34]. The thermal resistances at 4 K are listed in Table 2.

The proposed algorithm, described in Section 3, identifies the set of temperatures that minimizes the total power consumption while adhering to a maximum delay constraint of 0.8 µs. The algorithm is implemented using Python on an Intel Core i7-9750H workstation equipped with 8 GB RAM. In this case study, the execution time is 0.65 seconds. The sets of optimal temperatures identified by the algorithm are listed in Table 3. The most power efficient set is highlighted in bold. Notably, as the temperature of the higher stage changes, the optimal temperature of the lower stage also changes. Although the first five most optimal sets of temperatures are different, the combination of different temperatures results in similar values of total delay and power. The power consumption of the optimal path, which satisfies the 0.8 µs delay constraint, is 225.94 kilowatts. The total delay of a nonoptimized system with all of the devices placed at the highest operating temperature within each zone is over 3 µs, as listed in Table 1. If all of the devices are placed at the lowest available temperature, as listed in Table 4, the total power consumption will exceed 1.6 MW. The optimization process therefore saves significant power while satisfying the target performance constraints. The methodology has also been applied to an additional cloud computing case study. The total delay of this nonoptimized cloud computing system is 5 µs. The total delay of the optimized system is 1.5 µs, as listed in Table 5. The power consumption of the nonoptimized system is 2,348 kW, over eleven times larger than the optimized system.

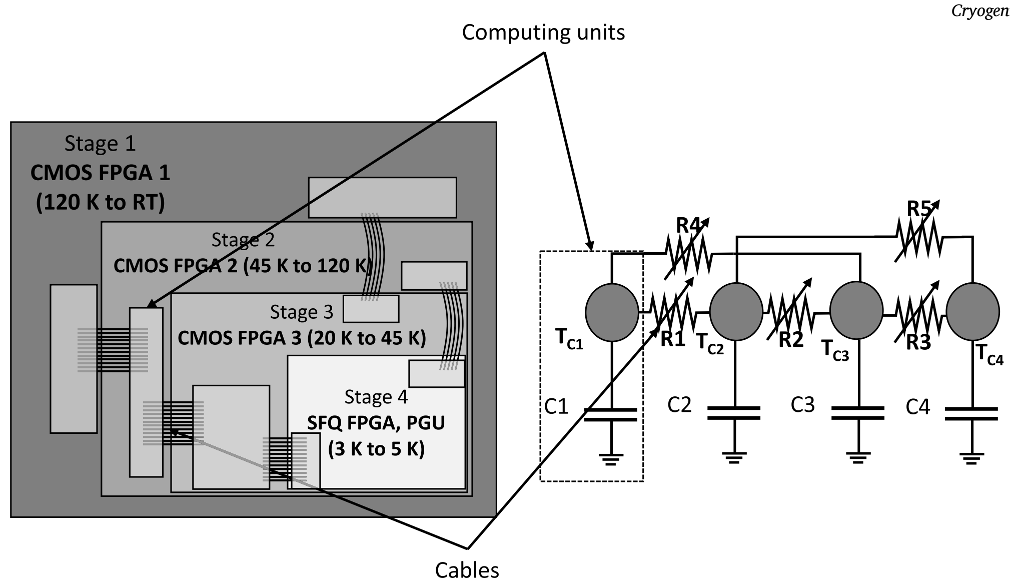


Fig. 7. Thermal-electrical circuit model of a cryogenic cloud computing system. The thermal relationship between chambers due to a temperature difference is represented as a thermal resistor.

Table 4
Delay D_i and power P_i (including refrigeration) of each circuit group within a cryogenic cloud computing system at the minimum feasible operating temperature for each unit.

Refrigerator	Delay	Power
Stage	D, [ns]	P, [kW]
Stage 1	493.7	192
Stage 2	69.7	592
Stage 3	32.2	213
Stage 4	14.4	661

Table 5
Case study of a cryogenic cloud computing system (including refrigeration). The delay limit for this case study is set to 1.5 microseconds. The set of temperatures producing the lowest power consumption is shown in bold.

Stage temperature, K				Delay	Power
C_1	C_2	C_3	C_4	D_i, [ns]	P_i, [kW]
120	**45.0**	**20.0**	**5.0**	**1,454.85**	**198.18**
140	45.0	20.0	4.98	1,450.05	198.84
140	53.2	20.0	4.78	1,445.37	199.51
160	45.0	20.0	4.95	1,440.83	200.19
140	61.4	20.0	4.95	1,436.42	201.48

5. Conclusions

Cloud computing data centers are widely used. These cloud computing systems can exploit cryogenic operation to achieve high performance computing while dissipating significantly lower energy. Operating electronic systems at cryogenic temperatures offers several fundamental benefits: increased carrier mobility in semiconductors, reduced thermal noise, negligible leakage current, superconductive behavior, and reduced electrical resistance. As a result, computing systems exhibit higher operating frequencies, improved reliability, reduced noise, and lower power consumption. The operating temperature profoundly influences the performance, cooling capacity, and power dissipation of the circuit components. It is therefore essential to carefully choose the operating temperatures to reduce the total power dissipation (and heat load) of the overall system while achieving the proper function and enhancing overall performance.

The methodology outlined in this paper focuses on the thermal optimization of cryogenic computing systems operating across different temperature zones. Unlike previous approaches that focus on individually optimizing each subsystem, the proposed methodology considers the operation of the entire system including the mutual dependence between temperature zones. Thermal leakage through the interconnecting cables between separate chambers can be a significant issue. The total cooling power includes this thermal leakage within the thermal model of the system. The total power dissipation of a system is minimized while maintaining the performance within predefined delay constraints.

Two practical case studies are evaluated to validate the methodology in which the temperature of the individual subsystems within a cryogenic cloud computing system, located in a four stage refrigerator, is determined. The temperature, delay, and power of the system are represented by a multigraph with variable power and constant delay weights. This multigraph contains all possible states of the system. The power consumption of an optimized example cryogenic cloud computing system is 225.94 kilowatts with a total delay of 0.8 μs. The optimized system exhibits lower delay as compared to nonoptimized systems operating at higher operating temperatures. Furthermore, although a nonoptimized system operating at the lowest possible temperature is faster, the power consumption of such a system exceeds 1.6 megawatts—seven times greater than the power consumption of the optimized system. Similarly, an eleven times reduction in power is exhibited in the second case study.

CRediT authorship contribution statement

Nurzhan Zhuldassov: Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Rassul Bairamkulov:** Validation. **Eby G. Friedman:** Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Eby G. Friedman reports financial support was provided by National Science Foundation under Grant No. 2124453 (DISCOVER Expeditions) and

Grant No. 2308863. Eby G. Friedman reports financial support was provided by Department of Energy under FOA number DE-FOA-0002950. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] Sharma P, et al. Design and operational analysis of a green data center. IEEE Internet Comput August 2017;21(4):16–24.

[2] Koot M, Wijnhoven F. Usage impact on data center electricity needs: a system dynamic forecasting model. Appl Energy June 2021;291:116798.

[3] Holmes DS, Ripple AL, Manheimer MA. Energy-efficient superconducting computing—power budgets and requirements. IEEE Trans Appl Supercond June 2013;23(3):1701610.

[4] Zhuldassov N, Friedman EG. Temperature–frequency boundary of cryogenic dynamic logic. Microelectron J March 2023;135:105763.

[5] Jha AR. Superconductor technology: applications to microwave, electro-optics, electrical machines, and propulsion systems. John Wiley & Sons; 1998.

[6] Zhuldassov N, Friedman EG. Cryogenic dynamic logic. In: Proceedings of the IEEE international symposium on circuits and systems; October 2020. p. 1–5.

[7] Zhuldassov N, Bairamkulov R, Friedman EG. Thermal optimization of hybrid cryogenic computing systems. IEEE Trans Very Large Scale Integr (VLSI) Syst September 2023;31(9):1339–46.

[8] Simon N, Drexler E, Reed R. Properties of copper and copper alloys at cryogenic temperatures. Final report. Technical Report. Boulder, Colorado: National Institute of Standards and Technology (MSEL); February 1992.

[9] Bradley PE, Radebaugh R. Properties of selected materials at cryogenic temperatures. Natl Inst Stand Technol June 2013;680:1–14.

[10] Jabbari T, Friedman EG. SFQ/DQFP interface circuits. IEEE Trans Appl Supercond August 2023;33(5):1303705.

[11] Pakzad P, Mehrpooya M, Zaitsev A. Investigation of a new energy-efficient cryogenic process configuration for helium extraction and liquefaction. Int J Energy Res June 2021;45(7):355–10377.

[12] Patra B, et al. Cryo-CMOS circuits and systems for quantum computing applications. IEEE J Solid-State Circuits September 2017;53(1):309–21.

[13] Bharany S, et al. A systematic survey on energy-efficient techniques in sustainable cloud computing. Sustainability May 2022;14(10):6256.

[14] Mukhanov OA, et al. Superconductor digital-RF receiver systems. IEICE Trans Electron March 2008;91(3):306–17.

[15] Gupta D, et al. Modular, multi-function digital-RF receiver systems. IEEE Trans Appl Supercond December 2010;21(3):883–90.

[16] Krylov G, Jabbari T, Friedman EG. Single flux quantum integrated circuit design. Second edition. Springer; 2024.

[17] Dyer SA, Dyer JS. Cubic-spline interpolation: part 1. IEEE Instrum Meas Mag March 2001;4(1):44–6.

[18] Bairamkulov R, Friedman EG. Graphs in VLSI. Springer; 2023.

[19] Jr RG Ross. Refrigeration systems for achieving cryogenic temperatures. In: Proceedings of the low temperature materials and mechanisms. CRC Press; August 2016. p. 127–200.

[20] Cryomech. Technical specifications of Cryomech's Gifford-McMahon cryocoolers. [Online]. Available: https://www.cryomech.com/cryocoolers/gifford-mcmahon-cryocoolers/, 2023.

[21] Lake I. Shore Cryotronics. Technical specifications of ST-500 cryostat, [Online]. Available: https://www.lakeshore.com/products/product-detail/janis/st-500-microscopy-cryostat, 2023.

[22] Group SC. Cryocooler product catalogue of SHI cryogenics group. [Online]. Available: https://www.shicryogenics.com/wp-content/uploads/2020/10/Cryocooler-Product-Catalogue-Letter-06.20-Web.pdf, 2023.

[23] Oy BlueFors. Technical specifications of BlueFors Oy cryocoolers. [Online]. Available: https://bluefors.com/products, 2023.

[24] Çengel YA, Turner RH, Cimbala JM, Kanoglu M. Fundamentals of thermal-fluid sciences, vol. 703. New York: McGraw-Hill; 2001.

[25] Brunner E. Fluid mixtures at high pressures VI. Phase separation and critical phenomena in 18 (n-alkane+ ammonia) and 4 (n-alkane+ methanol) mixtures. J Chem Thermodyn March 1988;20(3):273–97.

[26] Vesovic V, et al. The transport properties of carbon dioxide. J Phys Chem Ref Data May 1990;19(3):763–808.

[27] Atangana A, Araz Sİ. New numerical scheme with Newton polynomial: theory, methods, and applications. Academic Press; 2021.

[28] Lyeo H-K, Cahill DG. Thermal conductance of interfaces between highly dissimilar materials. Phys Rev B April 2006;73(14):144301.

[29] Thornburg D, Thall E, Brous J. A manual of materials for microwave tubes. Technical Report. Radio Corporation of America; January 1961.

[30] Paasschens J, Harmsma S, Van der Toorn R. Dependence of thermal resistance on ambient and actual temperature. In: Proceedings of the bipolar/BiCMOS circuits and technology meeting; September 2004. p. 96–9.

[31] CryoCoax. Technical specifications of CryoCoax's cryogenic cables. [Online]. Available: https://cryocoax.com/cryogenic-cable-and-cable-assemblies/, 2023.

[32] Schmidt C. Simple method to measure the thermal conductivity of technical superconductors, e.g., NbTi. Rev Sci Instrum April 1979;50(4):454–7.

[33] Arvindhan M. Effective motivational factors and comprehensive study of information security and policy challenges. In: System assurances: modeling and management. Elsevier; 2022. p. 531–45.

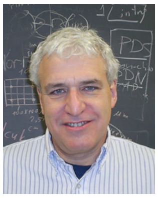
[34] Ho CY, Powell RW, Liley PE. Thermal conductivity of the elements. J Phys Chem Ref Data April 1972;1(2):279–421.

Nurzhan Zhuldassov received his B.S. degree in Electrical and Computer Engineering from Nazarbayev University, Astana, Kazakhstan, in 2018, and his M.S. in 2019, and Ph.D. in 2024, both in Electrical and Computer Engineering at the University of Rochester, Rochester, NY, USA.

In 2022 he was an intern at Google, Sunnyvale, California, and in 2023 he was an intern at Apple, Cupertino, California. His research interests encompass on-chip and package level power delivery networks, power integrity, cryogenic operation of MOSFETs, and optimization of cryogenic electronic circuits.

Rassul Bairamkulov is a Postdoctoral Scholar at the Integrated Systems Laboratory, EPFL, Switzerland. He received his Ph.D. degree in Electrical and Computer Engineering from University of Rochester, New York, in 2022. In 2018 and 2020, he was an intern at Qualcomm, San Diego, California.

Dr. Bairamkulov is a recipient of the Best Paper Award at the 2023 International Conference on ery Large-Scale Integration (VLSI-SoC) as well as the Best Paper Award Nomination at the 2024 Asia-South-Pacific Design Automation Conference (ASP-DAC). His current research interests encompass logic synthesis, power integrity, and electronic design automation for emerging VLSI technologies.



Eby G. Friedman (Life Fellow, IEEE) received the B.S. degree in electrical engineering from the Lafayette College, and the M.S. and Ph.D. degrees in electrical engineering from the University of California at Irvine. He was with Hughes Aircraft Company, from 1979 to 1991, rising to the Manager of the Signal Processing Design and Test Department, where he was responsible for the design and test of high-performance digital and analog ICs. He has been with the Department of Electrical and Computer Engineering, University of Rochester, since 1991, where he is currently a Distinguished Professor and the Director of the High Performance VLSI/IC Design and Analysis Laboratory. He is also a Visiting Professor with the Technion—Israel Institute of Technology, Haifa, Israel. He has authored more than 600 articles and book chapters, holds 29 patents, and has authored or edited 21 books in the fields of high speed and low power CMOS design techniques, 3-D design methodologies, high speed interconnect, superconductive circuits, and the theory and application of synchronous clock and power distribution networks. His current research and teaching interests include high performance synchronous digital and mixed-signal circuit design and analysis with application to high speed portable processors, low power wireless communications, and data centers. Dr. Friedman is a Senior Fulbright Fellow, a Fellow of the NAI, a National Sun Yat-sen University Honorary Chair Professor, and an Inaugural Member of the UC Irvine Engineering Hall of Fame. He was a recipient of the IEEE Circuits and Systems Mac Van Valkenburg Award, the IEEE Circuits and Systems Charles A. Desoer Technical Achievement Award, the University of Rochester Graduate Teaching Award and Lifetime Achievement Award, and the College of Engineering Teaching Excellence Award. He was the Editor-in-Chief and Chair of the Steering Committee of the *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, the Editor-in-Chief of the *Microelectronics Journal*, a Regional Editor of the *Journal of Circuits, Systems and Computers*, an editorial board member for numerous journals, and a program and technical chair for several IEEE conferences.