Silicon Microthermoelectric Coolers for Local Heat Removal in Integrated Circuit Chips

Ruchika Dhawan, Hal Edwards, *Member, IEEE*, and Mark Lee

***Abstract*— Advancements in electronic device fabrication with increasing integration levels have resulted in very high device densities. This has led to higher power dissipation and heat fluxes, increasing integrated circuit (IC) operating temperature. High and nonuniform heat generation degrades device and system performance. Therefore, thermal management to keep ICs within prescribed temperature limits is an important challenge for reliable and economic performance. Cooling techniques, including liquid coolants and air conditioning (AC), have been utilized to remove heat at the package and system level. However, these techniques must overcome high thermal impedances and require complex integration, while global cooling is generally wasteful, inefficient, and expensive. To improve thermal management, we have developed Si microthermoelectric coolers (μTECs) with areas as small $\sim 10^{-5}$ cm^2 that can be integrated on -chip near local hot spots using the standard fabrication processes. While Si μTECs cannot achieve low base temperatures, they can actively pump relatively high heat fluxes directly to a heat sink, thus reducing local temperature increases and allowing targeted rather than global waste heat removal. We demonstrate μTECs that can pump up to 43 W cm^{-2} of locally generated excess heat with no increase in chip temperature.**

***Index Terms*— Integrated circuit (IC) thermal factors, thermoelectric devices, thermoelectrics.**

I. Introduction

With the rapid development of very-large-scale-integrated circuit (VLSIC) technology, feature sizes are now at nanometer dimensions, and the number of components per integrated circuit (IC) chip is increasing at a fast rate. The very high device density and design complexity of VLSICs result in high and nonuniform heat density generation in high-performance Si ICs. This creates hot spots that can significantly degrade device performance and lifetime [1], [2], [3], [4] with heat fluxes of order 100 W cm^{-2} [5], [6], [7], [1], [8], [9].

Thus, the effective removal of high heat fluxes is a fundamental requirement for efficient and reliable operation. Many package-, board-, and system-level cooling strategies are used, including liquid cooling [10], [11], air-cooled heat sinks [12], and air conditioning (AC) [13]. Waste heat is usually dumped to room air, where a global AC system carries it away. With increasingly high heat flux regions, a local heat pump integrated ON-chip may be more efficient than passive package- or board-level cooling [14], [15], but current cooling systems do not scale well to the micrometer dimensions needed for integrated ON-chip cooling.

In 2018, data centers consumed ~1% of worldwide electrical energy, estimated to increase to ~3% by Jones [16]. A large fraction of this usage arises from AC. It has been suggested [14], [15], [17] that integrating microthermoelectric coolers (μTECs) into an IC to pump locally generated heat to a selectively cooled heat sink can significantly reduce AC energy needs. A recent article [18] comprehensively reviews the state-of-the-art μTECs. Temperature reductions of 5–55 K and maximum cooling power densities (i.e., cooling power per unit area at zero temperature difference) exceeding 100 W cm^{-2} are reported from μTECs using high TE figure-of-merit (ZT) materials, usually Bi_2Te_3 or related compounds with $ZT \approx 1$ near 300 K. However, these often require large thermopile (TP) currents of 0.1 to >1 A [19], [20], [21], raising concerns about parasitic ohmic heating in IC applications, where lead resistances may not be negligible. As important, the growth of high ZT materials and processing of such materials into TE devices are generally incompatible with the requirements of complementary metal–oxide–semiconductor (CMOS) silicon IC fabrication, making integration of high ZT μTECs with Si chips very difficult.

Monolithic integration of μTECs with chip-level silicon IC processing is strongly preferable. Si itself has not been considered as a TE material because of its low $ZT \sim 10^{-2}$ near 300 K, leading to an ideal maximum temperature reduction of $-\Delta T_{\text{ideal}} = 1/2ZT_C^2$ of only ~1.5 K [18], [22]. However, if mitigating temperature increases rather than actual refrigeration is the main goal, Si-based μTECs could prove useful as ON-chip integrated heat pumps if

Manuscript received 20 June 2023; revised 5 August 2023; accepted 7 August 2023. Date of publication 21 August 2023; date of current version 22 September 2023. This work was supported in part by the National Science Foundation under Award DMR-2206888. The review of this brief was arranged by Editor E. A. Gutiérrez-D. *(Corresponding author: Mark Lee.)*

Ruchika Dhawan was with the Department of Physics, The University of Texas at Dallas, Richardson, TX 75080 USA. She is now with Texas Instruments, Inc., Dallas, TX 75234 USA (e-mail: r-dhawan@ti.com).

Hal Edwards is with Texas Instruments, Inc., Dallas, TX 75234 USA (e-mail: edwards@ti.com).

Mark Lee is with the Department of Physics, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: marklee@utdallas.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2023.3304277.

Digital Object Identifier 10.1109/TED.2023.3304277

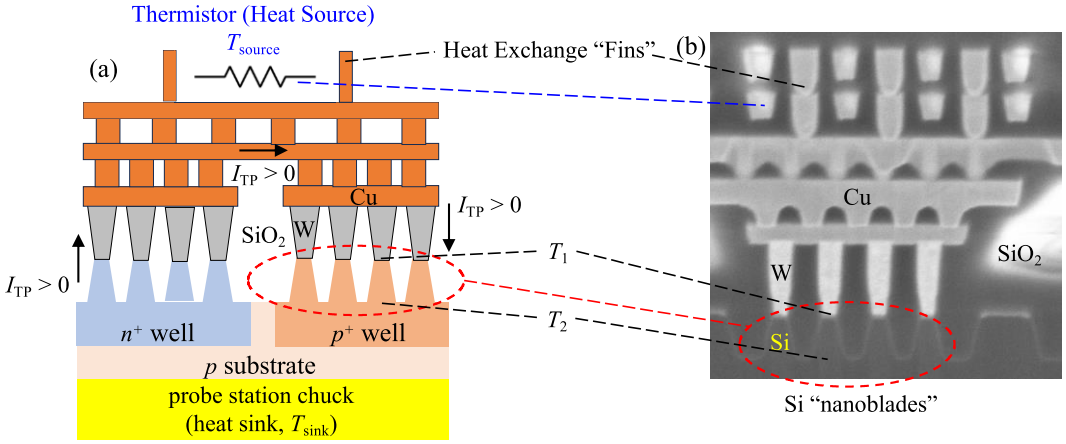

Fig. 1. (a) Illustration of basic thermocouple consisting of *n*- and *p*-type "nanoblades," resistive thermistor as a heat source, and probe station chuck as a heat sink. (b) Cross section SEM image of a set of four Si nanoblades contacted by tungsten plugs. The thermistor is a thin serpentine Cu meander line (winding into and out of the plane of the page) thermally connected to the top of the thermocouple via Cu heat "fins." For scale, the *W* plugs are about 80 nm wide. T_1 and T_2 are the temperatures at the top and bottom, respectively, of the thermocouple elements.

they can achieve maximum cooling power densities of order 100 W cm^{-2}.

In this brief, we report experimental results on Si-based μTECs using $Si_{0.97}Ge_{0.03}$ as the TE material. These μTECs were fabricated using standard "65-nm node" process technology and are thus fully compatible with monolithic VLSIC device integration. While we found that these Si-based μTECs could deliver only modest refrigeration of $-\Delta T_{max} = 0.31$ K, unoptimized prototypes showed maximum cooling power density of 43 W cm^{-2}, a factor of 2–3 away from being practically useful. Potential design improvements to improve cooling performance will be discussed.

II. Devices and Measurement

The Si-based μTE devices used in this study were designed as TE generators as detailed in [23] and [24] but are capable of Peltier cooling. Briefly summarizing, each thermocouple is composed of $Si_{0.97}Ge_{0.03}$ "nanoblades" (80 nm wide × 350 nm tall × 750 nm long) made using a moat etch from n^+ and p^+ wells (see Fig. 1), where Ge was ion implanted into the surface of a Si wafer. An array of such thermocouples forms a TP with total area for heat flow $A = 48 \times 36$ μm. These devices were shown to output very high power and voltage densities using $\Delta T \sim 20$ K [23], [24]. To test TE cooling, a serpentine Cu film thermistor fabricated ON-chip above each TP is used as a heat source to simulate a local hot spot. The thermistor is thermally connected (but electrically isolated) to the TP top via an interdigitated Cu fin heat exchanger. The backside of the μTEC chip is heat sunk to a probe station chuck actively maintained at $T_{sink} = 300.00$ K.

Fig. 1 shows the (a) illustration and (b) cross section scanning electron microscope image of a device. At the top is the thermistor heat source at temperature T_{source}, connected thermally to the top of the TP at temperature T_1. At the bottom is the probe station chuck acting as the heat sink at temperature T_{sink}, connected thermally to the bottom of the TP at temperature T_2. The TP is electrically connected via n^+- and p^+-well contacts to an external voltage V_{TP} and current I_{TP}, where $V_{TP}I_{TP}$ is the power input needed to pump heat away from the thermistor. Peltier heat flows from source to sink when $I_{TP} > 0$, defined as the direction when electrons in the *n* side and holes in the *p* side drift from top to bottom, as indicated in Fig. 1(a). Peltier cooling of the thermistor only occurs for $I_{TP} > 0$.

The chip with TP and thermistor was placed on a probe station chuck. As illustrated in Fig. 2 (inset), electrical contacts were brought to probe contact pads on the chip surface. Four-wire Kelvin probes were used to measure current–voltage ($I-V$) characteristics of thermistor and TP. All biases were set and measured using an Agilent 4156 parameter analyzer with medium integration time and measurement delay of 100 ms. The thermistor (subscript "th") was voltage biased at V_{th} and current I_{th} measured, giving heater power $Q = V_{th}I_{th}$. Embedded in the chuck was a calibrated thermometer to measure T_{sink}. The temperature difference ΔT of the thermistor relative to the chuck was determined using the procedure described in [24] to calibrate the thermistor's resistive temperature coefficient: $\beta = 0.00269$ K^{-1}, from which $\Delta T = (T_{th} - 300\text{ K}) = (R_{th} - R_{th0})/(\beta R_{th0})$, where $R_{th0} \approx 85$ Ω was the thermistor resistance with the chip at 300 K and $R_{th} = V_{th}/I_{th}$ was determined using positive and negative V_{th} biases to subtract out offsets. For each test condition, R_{th} was measured 40 times and averaged to determine ΔT. In each dataset, the standard deviation in R_{th} was ± 0.003 Ω, corresponding to a ± 0.01 K uncertainty in ΔT.

Fig. 2 shows that Peltier cooling could be observed in these devices with no thermistor heat applied, $Q = 0$, and the TP current biased at $I_{TP} > 0$. Fig. 2 shows ΔT versus I_{TP} across both polarities. The quadratic shape of the curve results from $I_{TP}^2 R_{TP}$ Joule heating, where R_{TP} is the TP resistance, but the offset of the minimum to negative ΔT at $I_{TP} > 0$ is due to Peltier cooling. The yellow shading in Fig. 2 indicates the region where Peltier cooling ($\Delta T < 0$) is strong enough to overcome Joule heating. Cooling occurs for I_{TP} up to 30 mA, beyond which Joule heating overcomes Peltier cooling. The best temperature reduction achieved is $-\Delta T_{max} = 0.31$ K at an optimal current bias near 15 mA. This is less than the 1.0 K reduction reported by Zhang et al. [25] using μTECs made

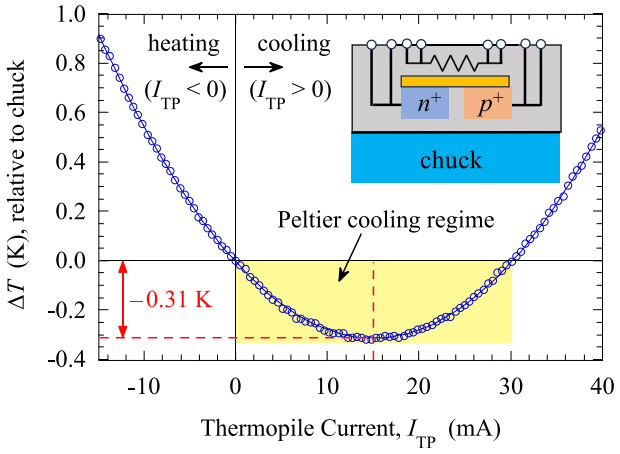

Fig. 2. Temperature difference ΔT between thermistor and chuck as a function of current bias I_{TP} with no heat applied to the thermistor. The chuck is maintained at 300.00 K. Yellow shaded area indicates the region of net Peltier cooling. The maximum temperature reduction $-\Delta T_{max} = 0.31$ K at an optimal $I_{TP} = 15$ mA is indicated. Inset: block schematic of the electrical contact configuration to thermistor and TP, brought out to probe contact pads on the chip surface.

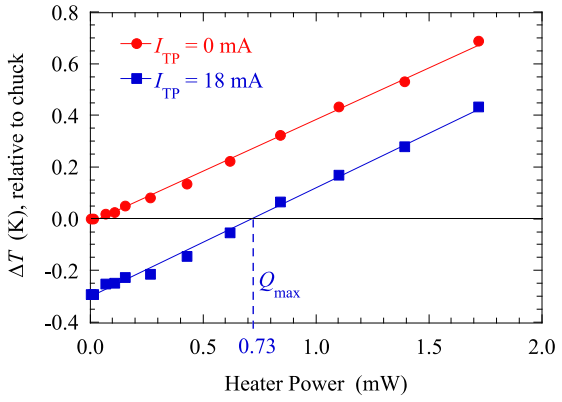

Fig. 3. Temperature difference ΔT between thermistor and chuck as a function of heater power applied to thermistor. The red circles are with the μTEC off: $I_{TP} = 0$ mA. The blue squares are with the μTEC on and biased near its optimal bias: $I_{TP} = 18$ mA. Markers are measured data points and lines are least-square fits to the data. The intersection of $I_{TP} = 18$ mA line on the $\Delta T = 0$ axis gives maximum cooling power Q_{max} of the device.

from surface etched Si mesas and is a factor of ~5 smaller than the ideal limit, most likely due to significant parasitic thermal contact resistances between the TP and the heat source and sink.

Fig. 3 shows the measured Peltier heat pumped from the thermistor to the chuck using a constant I_{TP} to determine the cooling power. With the μTEC off ($I_{TP} = 0$ mA), applying $Q > 0$ to the thermistor always results in $\Delta T > 0$ as expected. However, when the μTEC is biased near its optimal current ($I_{TP} = 18$ mA here), the ΔT versus Q line is offset down with negative intercept on the y-axis giving $-\Delta T_{max}$. The thermistor temperature remains lower than the heat sink until a maximum cooling power $Q_{max} = 0.73$ mW is applied. Normalizing this to area gives a maximum cooling power density of 43 W cm^{-2}. This means the μTEC can remove 43 W cm^{-2} of heat flux generated in addition to its own operating power before local chip temperature rises above that of the heat sink.

The results shown in Figs. 2 and 3 were measured on two separate but nominally identical devices. Both devices

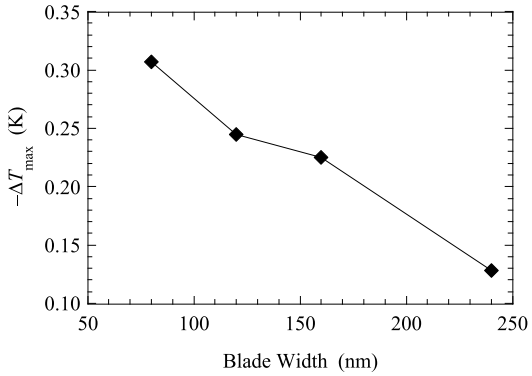

Fig. 4. Maximum temperature reduction versus nanoblade width w in four μTEC variants with the same geometry and material parameters except for w. Cumulative TP area for heat/charge flow was kept constant to keep R_{TP} the same. The lines simply connect data points.

gave the same $-\Delta T_{max}$ to within 0.01 K, Q_{max} to within 0.03 mW, and had the same optimum I_{TP} to within a few milliamperes of each other. This demonstrates good device-to-device reproducibility.

These devices can heat or cool a chip depending on the sign of I_{TP}. Thus, such a μTEC could be used as a temperature regulator. ON-chip temperature control is especially important for circuits that rely on ON-chip temperature sensitive metrology references to achieve desired accuracy [26].

III. Potential for Improving Performance

A question of interest is how to improve cooling performance of these Si μTECs. Most obviously, reducing parasitic thermal impedances Θ_1 (between heat source and top of TP) and Θ_2 (between bottom of TP and heat sink) is known to boost performance [27]. When $\Theta_1 + \Theta_2 > 0$, the maximum $-\Delta T$ achievable is reduced from the ideal case by the thermal contact factor $\Theta_{TP}/(\Theta_{TP} + \Theta_1 + \Theta_2)$, where Θ_{TP} is the thermal impedance of the TP itself. Since our measured $-\Delta T_{max} \approx 0.2$ ($-\Delta T_{ideal}$), this suggests that $(\Theta_1 + \Theta_2) \approx 4\Theta_{TP}$, consistent with the conclusions of a previous device model [24]. Thus, a factor of up to 5× cooling performance increase could be achieved by reducing $\Theta_1 + \Theta_2$ through a better heat exchanger design and use of higher thermal conductivity dielectric spacers between source and TP and between TP and heat sink.

In addition, reducing TE element dimensions to below the phonon mean free path (~200 nm at 300 K) [28] has been shown to improve TE device performance [29]. We explored the effect of nanoblade width using a set of four device variants with different widths $w = 80$, 120, 160, and 240 nm. These devices used the same geometry and material characteristics, except that the number of nanoblades in each TP was inversely proportional to w in order to keep total TP cross-sectional area, and hence R_{TP}, constant. Fig. 4 shows that while $R_{TP} \approx 5.2\ \Omega$ remained essentially constant (varying by $\leq 3\%$), as w decreases from 240 to 80 nm $-\Delta T_{max}$ improves by a factor of 2.3× and extrapolates toward continuing improvement at smaller w.

The dopant concentration used in these nanoblades was about 4×10^{18} cm^{-3}, an order-of-magnitude smaller than the densities typically used in semiconductor TE devices.

Modeling of Si TE generators [30] suggests that, including parasitic thermal impedances and electrical resistances, an increase in device level Z of 30%–50% could be expected upon increasing the TE element dopant density from that used to $\sim 2 \times 10^{19}$ cm^{-3}.

Finally, the μTECs used in this work used a fill factor (ratio of TP area occupied by TE elements to total TP area, which includes dielectric filler) of only $\sim$2%. As previously stated, these devices were designed as TE generators, and the use of small fill factors preserves an optimal ΔT needed to maximize power generation [31]. For heat pumping applications however, the aim is to reduce ΔT rather than maintain it, so it may make sense to increase the fill factor up to the 20%–50% level typically found in Bi_2Te_3-based TE devices [19], [21].

Data Availability

Data shown in this brief are available in the Harvard Dataverse at: https://doi.org/10.7910/DVN/WDXFHW.

Acknowledgment

The design and fabrication of TE devices was supported by Texas Instruments (TI), Inc., Dallas, TX, USA. The authors would like to thank Archana Venugopal (TI) and Orlando Lazaro (TI) for useful discussions and ideas and Hari Panthi [University of Texas at Dallas (UTD)] for helping with measurements.

References

[1] Z. Zhang, X. Wang, and Y. Yan, "A review of the state-of-the-art in electronic cooling," *e-Prime-Adv. Electr. Eng., Electron. Energy*, vol. 1, Jan. 2021, Art. no. 100009, doi: 10.1016/j.prime.2021.100009.

[2] A. Bar-Cohen and P. Wang, "Thermal management of on-chip hot spot," *J. Heat Transf.*, vol. 134, no. 5, Mar. 2012, Art. no. 051017, doi: 10.1115/1.4005708.

[3] S. Ogrenci-Memik, *Heat Management in Integrated Circuits: On-Chip and System-Level Monitoring and Cooling*. London, U.K.: Institution of Engineering and Technology, 2016.

[4] H. F. Hamann et al., "Hotspot-limited microprocessors: Direct temperature and power distribution measurements," *IEEE J. Solid-State Circuits*, vol. 42, no. 1, pp. 56–65, Jan. 2007, doi: 10.1109/JSSC.2006.885064.

[5] *International Roadmap For Devices and Systems TM 2020 Edition*. Accessed: Jun. 19, 2023. [Online]. Available: irds.ieee.org/images/files/pdf/2020/2020IRDS_ES.pdf

[6] S. G. Kandlikar, "Review and projections of integrated cooling systems for three-dimensional integrated circuits," *J. Electron. Packag.*, vol. 136, no. 2, Mar. 2014, Art. no. 024001, doi: 10.1115/1.4027175.

[7] J. L. Smoyer and P. M. Norris, "Brief historical perspective in thermal management and the shift toward management at the nanoscale," *Heat Transf. Eng.*, vol. 40, nos. 3–4, pp. 269–282, Feb. 2019, doi: 10.1080/01457632.2018.1426265.

[8] *Intel Core I9-12900K Processor*. Accessed: Jun. 19, 2023. [Online]. Available: ark.intel.com/content/www/us/en/ark/products/134599/intel-core-i912900k-processor-30m-cache-up-to-5-20-ghz.html

[9] B. Ramakrishnan et al., "CPU overclocking: A performance assessment of air, cold plates, and two-phase immersion cooling," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 11, no. 10, pp. 1703–1715, Oct. 2021, doi: 10.1109/TCPMT.2021.3106026.

[10] D. Kearney, T. Hilt, and P. Pham, "A numerical hydrodynamic and thermal characterization of an inter-strata liquid cooling solution for 3D ICs," *Microsyst. Technol.*, vol. 18, no. 2, pp. 225–235, Feb. 2012, doi: 10.1007/s00542-011-1376-x.

[11] A. Sridhar, A. Vincenzi, D. Atienza, and T. Brunschwiler, "3D-ICE: A compact thermal model for early-stage design of liquid-cooled ICs," *IEEE Trans. Comput.*, vol. 63, no. 10, pp. 2576–2589, Oct. 2014, doi: 10.1109/TC.2013.127.

[12] L. Gong, J. Zhao, and S. Huang, "Numerical study on layout of micro-channel heat sink for thermal management of electronic devices," *Appl. Thermal Eng.*, vol. 88, pp. 480–490, Sep. 2015, doi: 10.1016/j.applthermaleng.2014.09.048.

[13] A. Capozzoli and G. Primiceri, "Cooling systems in data centers: State of art and emerging technologies," *Energy Proc.*, vol. 83, pp. 484–493, Dec. 2015, Art. no. 021005, doi: 10.1016/J.EGYPRO.2015.12.168.

[14] O. Sullivan, M. P. Gupta, S. Mukhopadhyay, and S. Kumar, "Array of thermoelectric coolers for on-chip thermal management," *J. Electron. Packag.*, vol. 134, no. 2, Jun. 2012, doi: 10.1115/1.4006141.

[15] G. J. Snyder, M. Soto, R. Alley, D. Koester, and B. Conner, "Hot spot cooling using embedded thermoelectric coolers," in *Proc. 22nd Annu. IEEE Semiconductor Thermal Meas. Manag. Symp.*, Dallas, TX, USA, 2006, pp. 135–143, doi: 10.1109/STHERM.2006.1625219.

[16] N. Jones, "How to stop data centres from gobbling up the world's electricity," *Nature*, vol. 561, no. 7722, pp. 163–166, Sep. 2018, doi: 10.1038/d41586-018-06610-y.

[17] J. Sharp, J. Bierschenk, and H. B. Lyon, "Overview of solid-state thermoelectric refrigerators and possible applications to on-chip thermal management: On-chip thermal engineering," *Proc. IEEE*, vol. 94, no. 8, pp. 1602–1612, Aug. 2006, doi: 0.1109/JPROC.2006.879795.

[18] Q. Zhang, K. Deng, L. Wilkens, H. Reith, and K. Nielsch, "Micro-thermoelectric devices," *Nature Electron.*, vol. 5, no. 6, pp. 333–347, Jun. 2022, doi: 10.1038/s41928-022-00776-0.

[19] G. Bulman et al., "Superlattice-based thin-film thermoelectric modules with high cooling fluxes," *Nature Commun.*, vol. 7, no. 1, p. 10302, Jan. 2016, doi: 10.1038/ncomms10302.

[20] I. Chowdhury et al., "On-chip cooling by superlattice-based thin-film thermoelectrics," *Nature Nanotechnol.*, vol. 4, no. 4, pp. 235–238, Apr. 2009, doi: 10.1038/nnano.2008.417.

[21] G. Li et al., "Integrated microthermoelectric coolers with rapid response time and high device reliability," *Nature Electron.*, vol. 1, no. 10, pp. 555–561, Oct. 2018, doi: 10.1038/s41928-018-0148-3.

[22] J. Mao, G. Chen, and Z. Ren, "Thermoelectric cooling materials," *Nature Mater.*, vol. 20, no. 4, pp. 454–461, Apr. 2021, doi: 10.1038/s41563-020-00852-w.

[23] R. Dhawan et al., "$Si_{0.97}Ge_{0.03}$ microelectronic thermoelectric generators with high power and voltage densities," *Nature Commun.*, vol. 11, no. 1, p. 4362, Aug. 2020, doi: 10.1038/s41467-020-18122-3.

[24] G. Hu, H. Edwards, and M. Lee, "Silicon integrated circuit thermoelectric generators with a high specific power generation capacity," *Nature Electron.*, vol. 2, no. 7, pp. 300–306, Jul. 2019, doi: 10.1038/s41928-019-0271-9.

[25] Y. Zhang, G. Zeng, and A. Shakouri, "Silicon microrefrigerator," *IEEE Trans. Compon. Packag. Technol.*, vol. 29, no. 3, pp. 570–576, Sep. 2006, doi: 10.1109/TCAPT.2006.880508.

[26] D. Verma et al., "A design of 10-bit asynchronous SAR ADC with an on-chip bandgap reference voltage generator," *Sensors*, vol. 22, no. 14, p. 5393, Jul. 2022, doi: 10.3390/s22145393.

[27] G. Pennelli, E. Dimaggio, and M. Macucci, "Electrical and thermal optimization of energy-conversion systems based on thermoelectric generators," *Energy*, vol. 240, Feb. 2022, Art. no. 122494, doi: 10.1016/j.energy.2021.122494.

[28] L. Weber and E. Gmelin, "Transport properties of silicon," *Appl. Phys. A Solids Surf.*, vol. 53, no. 2, pp. 136–140, Aug. 1991, doi: 10.1007/BF00323873.

[29] J. M. Sojo-Gordillo et al., "Superior thermoelectric performance of SiGe nanowires epitaxially integrated into thermal micro-harvesters," *Small*, vol. 19, no. 17, Apr. 2023, Art. no. 2206399, doi: 10.1002/smll.202206399.

[30] G. Hu et al., "Scaling of power generation with dopant density in integrated circuit silicon thermoelectric generators," *IEEE Electron Device Lett.*, vol. 40, no. 12, pp. 1917–1920, Dec. 2019, doi: 10.1109/led.2019.2947357.

[31] R. Dhawan, P. Madusanka, G. Hu, K. Maggio, H. Edwards, and M. Lee, "Maximizing performance of microelectronic thermoelectric generators with parasitic thermal and electrical resistances," *IEEE Trans. Electron Devices*, vol. 68, no. 5, pp. 2434–2439, May 2021, doi: 10.1109/ted.2021.3067624.