Journal of Applied Physics PERSPECTIVE pubs.aip.org/aip/jap

Thermal design automation (TDA) for multiscale thermal management of electronics

Cite as: J. Appl. Phys. **138**, 180901 (2025); doi: 10.1063/5.0288828
Submitted: 3 July 2025 · Accepted: 22 October 2025 ·
Published Online: 11 November 2025

Bingyang Cao[a)]

AFFILIATIONS

Key Laboratory for Thermal Science and Power Engineering of Ministry of Education, Department of Engineering Mechanics, Tsinghua University, 100084 Beijing, China

Note: This paper is part of the Special Topic on Thermal Transport in Micro/Nanostructured Materials and Devices.
[a)]**Author to whom correspondence should be addressed:** caoby@tsinghua.edu.cn

ABSTRACT

Modern semiconductor devices face critical thermal management challenges as power densities increase and feature sizes approach deep nanoscale where classical Fourier's heat conduction law breaks down. Traditional chip design approaches that rely primarily on only electrical designs and external cooling solutions are insufficient to address the complex, multiscale nature of thermal transport in advanced integrated circuits. This perspective presents a comprehensive Thermal Design Automation (TDA) framework, as a complementary extension to traditional Electronic Design Automation (EDA) tools, that systematically integrates thermal simulation and management methods across all length scales of semiconductor design. The TDA approach begins at the atomic scale, using first-principles calculations and lattice dynamics simulations to predict intrinsic electron and phonon transport properties. These fundamental properties parameterize phonon Monte Carlo simulations that solve the Boltzmann transport equation to capture non-Fourier heat spreading within transistors, while self-heating effects are simulated by solving fundamental semiconductor device equations. For larger scales, finite element methods and compact thermal models bridge the gap to circuit- and die-level thermal analysis and design, while advanced liquid cooling technologies address chip-level heat dissipation. Through multiscale thermal simulation and design optimization, the TDA framework enables systematic reduction of transistor heat generation, minimization of device and chip thermal resistance, and acceleration of chip design cycles, thereby enhancing overall performance and reliability. This integrated framework addresses the fundamental limitations of existing microscopic and macroscopic thermal simulation tools and establishes a new EDA+TDA paradigm for thermal-aware semiconductor design that can systematically tackle the multiscale thermal bottlenecks limiting further technological advancement in modern electronics.

I. INTRODUCTION

With the continuous advancement of electronic system performance, modern integrated circuits (ICs) are facing increasingly severe thermal management challenges.[1–3] The trends toward higher power densities and greater integration levels have intensified on-chip overheating issues, which degrade system performance, increase power consumption, and compromise long-term reliability.[4–6] Across a wide range of semiconductor platforms—including high-performance integrated circuits,[7] radio-frequency (RF) devices,[8] and power electronics[9]—effective thermal management strategies have become critical bottlenecks limiting further technological advancement.

To implement effective thermal designs and reduce thermal resistance, it is essential to understand the heat transfer chain within semiconductor devices for targeted and systematic thermal optimization. At the transistor level, heat is generated primarily through Joule heating caused by electron–phonon interactions within the active regions.[10–12] As heat spreads throughout the transistor structure, significant thermal spreading resistance develops, creating localized hotspots.[13–15] Also, since transistor feature sizes and hotspot dimensions are comparable to the phonon mean free path (MFP)—the primary heat carriers in semiconductors—Fourier's law of heat conduction breaks down, leading to non-Fourier heat transport behavior that further elevates hotspot temperatures.[16,17] Heat propagation from individual transistors to

the entire die encounters multiple material interfaces due to multilayer epitaxial structures, resulting in thermal boundary resistance (TBR) that impedes heat dissipation.[18–20] Subsequently, heat is transported from the die through the first-level thermal interface material (TIM1) to the chip lid, then through TIM2 to external cooling systems, such as air or liquid cooling, ultimately dissipating to the environment.[21]

While this heat transfer chain applies to single transistors, RF devices and power electronics, such as gallium nitride (GaN), high-electron-mobility transistors (HEMTs), silicon carbide (SiC), and metal–oxide–semiconductor field-effect transistors (MOSFETs), usually contain hundreds of transistors with similar thermal pathways.[22,23] In contrast, advanced ICs comprise billions of transistors on a single die. Modern packaging technologies introduce additional complexity: 2.5D integration connects multiple chiplets through interposers with varying power maps under different operating conditions,[24] while 3D integration vertically stacks multiple dies using through-silicon vias (TSVs).[25] Furthermore, advanced back-side power delivery networks (BSPDNs), where power is supplied from the chip's back side, force thermal transport through back-end-of-line (BEOL) layers, creating more complex thermal transport scenarios compared to traditional front-side power delivery through the substrate.[26,27]

As demonstrated, heat dissipation in chips is a multiscale process encompassing multiple transport stages: heat generation within transistors, thermal spreading across transistor structures, heat conduction through material interfaces, circuit-level thermal transport, and external cooling systems. Each stage can impede overall heat dissipation, making thermal management a systematic bottleneck that requires coordinated thermal design across all hierarchical levels. Traditional chip design workflows have focused primarily on electrical design using electronic design automation (EDA) tools, with thermal considerations limited to external cooling solutions implemented after electrical design and packaging completion.[28] However, as chip power densities continue increasing and thermal transport paths become increasingly complex, internal thermal resistance now constitutes a substantial portion of the total thermal budget, transforming thermal management into a systematic, multiscale challenge.[29,30] Consequently, conventional approaches that rely solely on electrical design and external cooling can no longer adequately address modern thermal constraints. This paradigm shift necessitates the development of a comprehensive electrothermal co-design framework that integrates thermal considerations across all design hierarchies.[31–33]

However, heat transfer in chips involves the transition from nanoscale to macroscale transport regimes and the breakdown of classical Fourier heat conduction law at small scales.[26,34] These complexities render existing commercial thermal simulation software, which relies on macroscopic heat transfer models, inadequate for accurate transistor or chip-level thermal analysis and design. Therefore, a new thermal design toolset and framework is essential to systematically address thermal challenges across all levels of chip design and operation.

To address the aforementioned multiscale thermal challenges, this perspective presents a comprehensive Thermal Design Automation (TDA) framework and toolset that systematically integrates thermal simulation and management methods across all hierarchical levels of semiconductor design, as illustrated in Fig. 1. The TDA framework coherently addresses thermal phenomena from atomic-scale heat generation and phonon transport in transistors to system-level thermal management through a unified computational pipeline.

At the atomic level, the framework employs first-principle calculations to determine lattice energies, stresses, interatomic forces, and electron–phonon coupling (e–ph) parameters.[35] These *ab initio* results train machine learning potentials (MLPs) that accurately describe atomistic interactions for subsequent simulations.[36–39] The computed interatomic forces drive lattice dynamics simulations to predict phonon transmission across material interfaces with realistic atomic structures.[40] Simultaneously, phonon dispersion relations and scattering rates are calculated and integrated with interfacial phonon transmittance data to parameterize phonon Monte Carlo (MC) simulations, which solve the phonon Boltzmann transport equation (BTE) to describe nanoscale thermal transport within individual transistors or transistor arrays.[41,42] The phonon MC method excels at predicting temperature distributions in transistor structures. In conjunction with semiconductor device equations, heat generation rates are calculated based on electron–phonon coupling processes, which capture the microscopic energy transfer from charge carriers to the lattice. This enables comprehensive electro-thermal coupling simulations that accurately capture self-heating effects and thermal behavior at the device level.[43]

While phonon MC effectively addresses thermal transport in systems ranging from tens of nanometers to approximately 10μm, circuit- and die-level thermal analyses require complementary approaches. The framework incorporates finite element method (FEM) and compact thermal model (CTM) based tools to bridge this scale gap.[44] Multiscale coupling between phonon MC and FEM enables comprehensive thermal simulation and design for power electronics, such as GaN HEMTs,[34] while MC-predicted effective thermophysical properties of nanostructures serve as direct inputs for CTM and FEM-based simulations. This hierarchical integration allows CTM tools to incorporate FEM-predicted structural thermal resistances, creating a seamless computational workflow across length scales. At the chip level, the framework addresses the final stage of heat dissipation through high-performance cooling solutions, efficiently transferring heat from the chip to the ambient environment.[45] The integrated multiscale thermal simulation tools provide comprehensive evaluation capabilities for assessing the cooling performance of these systems across chip and transistor levels. Section II provides detailed descriptions of each TDA module, including theoretical foundations, computational implementations, validation studies, and practical applications across the multiscale thermal design hierarchy.

II. FRAMEWORK OF MULTISCALE THERMAL DESIGN AUTOMATION

A. Heat generation and self-heating tunability

As established in the multiscale TDA framework, accurate thermal management begins with understanding and tuning heat generation at the fundamental device level. Heat generation

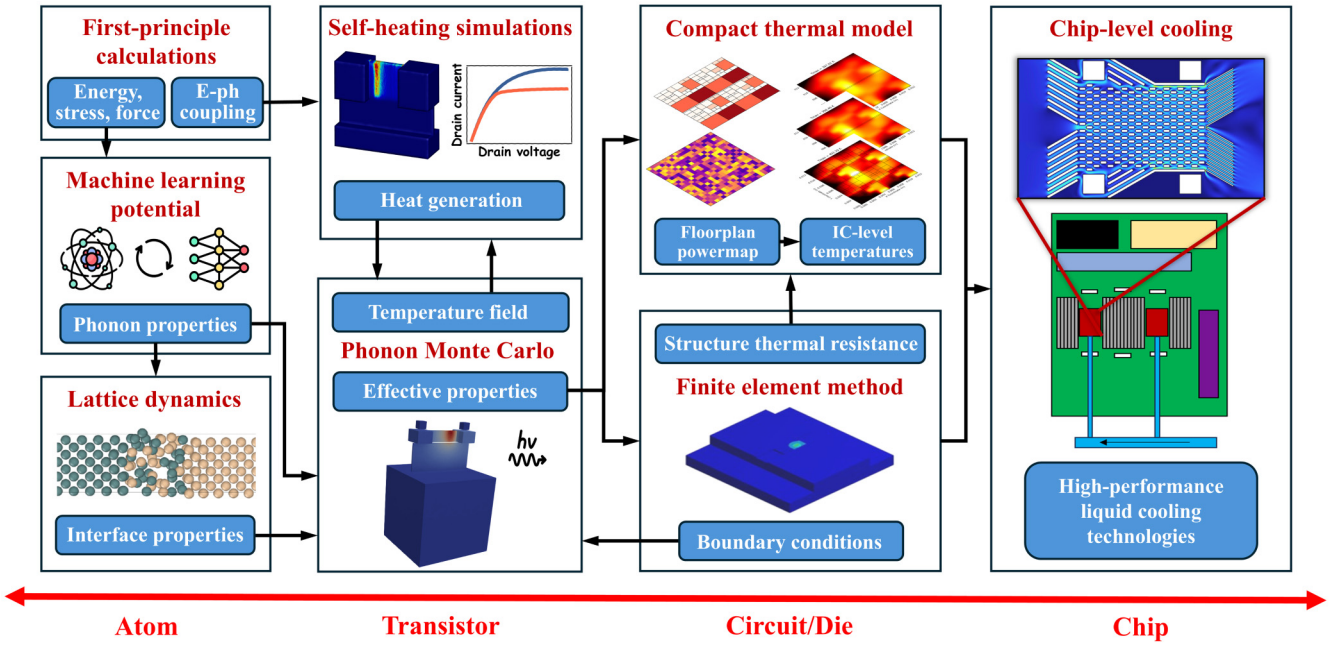

FIG. 1. Multiscale framework of the TDA system for thermal management of electronics.

in transistors originates from microscopic interactions between charge carriers—primarily electrons and holes—and phonons, the dominant heat carriers in semiconductor materials.[12,46] When an external electric field is applied, carriers gain energy and undergo scattering events with phonons during transport. Through these scattering processes, a portion of the carriers' kinetic energy is transferred to the lattice, converting electrical energy into thermal energy and causing localized temperature rises. Non-equilibrium interactions between carriers and both optical and acoustic phonons further intensify local heat accumulation, exacerbating hotspot formation.[15] These hotspots typically concentrate in the channel region of transistors, particularly near the drain side beneath the gate where electric fields are highest.[47]

The resulting heat generation produces self-heating effects that have become a critical physical bottleneck limiting further device scaling. Self-heating in transistors reduces carrier mobility, thereby degrading switching speed and overall device performance. Additionally, elevated temperatures accelerate material degradation, ultimately compromising device reliability and operational lifetime.[48] A detailed understanding of these microscopic heat generation mechanisms provides the foundation for accurate simulation and modeling of self-heating effects in transistors, forming the first critical component of the comprehensive TDA approach.

To analyze the self-heating effect in transistors, the TDA system employs the drift–diffusion model, which solves a set of fundamental semiconductor equations, including the Poisson equation, the carrier continuity equations, and the current transport equations. These equations are expressed as follows:[49]

$$\nabla \cdot (\varepsilon \cdot \nabla \phi) = -q\left(p - n + N_D^+ - N_A^+\right) - \rho_{\mathrm{PE}} - \rho_{\mathrm{trap}}, \tag{1}$$

$$\frac{1}{q}\nabla \cdot \vec{J}_n - \frac{\partial n}{\partial t} = R, \tag{2}$$

$$\frac{1}{q}\nabla \cdot \vec{J}_p + \frac{\partial p}{\partial t} = -R, \tag{3}$$

$$\vec{J}_n = qn\mu_n\vec{E} + qD_n\nabla n, \tag{4}$$

$$\vec{J}_p = qp\mu_p\vec{E} - qD_p\nabla p. \tag{5}$$

Here, ε is the dielectric constant, ϕ is the electrostatic potential, q is the elementary charge, and N_D^+ and N_A^- are the donor and acceptor concentrations in the semiconductor, respectively. ρ_{PE} is the polarization charge density, ρ_{trap} is the trap-induced charge density, and $\vec{J}_n$ and $\vec{J}_p$ are the current densities due to electrons and holes, respectively. n and p are the concentrations of carrier electrons and holes, t is the time, R is the net recombination rate of electrons and holes, μ_n and μ_p are the mobilities of electrons and holes, $\vec{E}$ is the electric field strength, and D_n and D_p are determined by the

Einstein relationship, as shown in Eq. (6),

$$D_n = \frac{kT}{q}\mu_n, \quad D_p = \frac{kT}{q}\mu_p, \tag{6}$$

where T is the temperature inside the device and k is the Boltzmann constant. After solving the electrical equations to obtain the electrical properties, the related parameter can be used to calculate the heat generation distribution inside the device. It is important to note that the physical parameters in the control equations need to use corresponding models to obtain the most realistic physical state. Common models include the Klaassen mobility model, the Shockley–Read–Hall (SRH) recombination model, the velocity saturation model, etc.[50–53]

The reliability of self-heating simulations critically depends on the accurate modeling of heat generation within the device. In the TDA system, heat generation is modeled by incorporating multiple physical mechanisms, and its overall expression is given by[54]

$$H = \left[\frac{|\vec{J}_n|^2}{q\mu_n n} + \frac{|\vec{J}_p|^2}{q\mu_p p}\right] + qR\left(\phi_p - \phi_n + T(P_p - P_n)\right) - T\left(\vec{J}_n \cdot \vec{P}_n + \vec{J}_p \cdot \vec{P}_p\right),$$

where ϕ_n and ϕ_p are the quasi-Fermi levels of electrons and holes, and P_n and P_p represent the absolute thermoelectric powers of electrons and holes, respectively.[55] The first term corresponds to Joule heating, the second term represents net recombination–generation (R-G) heat, and the third term accounts for heat generated by Thomson effects. In most practical simulations, the first two terms—Joule heating and R-G heat—dominate and are sufficient to capture the main thermal effects in the device.[56,57] It is important to note that this expression is derived from the principles of phenomenological irreversible thermodynamics and does not fully capture the microscopic dynamics of carrier–phonon interactions.

As transistor gate lengths approach the electron MFPs, nonequilibrium interactions between electrons and phonons play a critical role in device self-heating effects. Specifically, when the gate length becomes comparable to the electron MFPs, electrons require a finite time to gain energy from the electric field and subsequently transfer it to the phonon. This delay alters the spatial distribution of heat within the device, often shifting the hotspot toward the drain side. Under such conditions, the classical drift–diffusion model becomes inadequate, and more accurate modeling approaches, such as electron MC simulations, are needed to capture the heat generation and hotspot localization.[10,58,59] Furthermore, since electrons primarily interact with optical phonons, whereas thermal conduction is mainly governed by acoustic phonons, the resulting mismatch in energy exchange leads to additional thermal resistance. Phonons at different frequencies acquire varying energies from electrons, complicating thermal transport. To more accurately model self-heating effects in nanoscale devices, first-principles-based methods can be employed to capture the nonequilibrium energy transfer between electrons and phonons.[60] Once the heat generation profile is obtained from electrical simulation, the resulting temperature distribution can be evaluated using either Fourier's law of heat conduction or the phonon BTE, depending on the required accuracy and device scale.[60,61]

While traditional thermal management techniques predominantly focus on lowering the device thermal resistance to improve heat dissipation, the TDA system can begin by addressing the root cause of self-heating—heat generation. Specifically, it emphasizes structural design strategies to suppress heat generation at the source before optimizing thermal conduction. When the carrier drift velocity approaches saturation, a further increase in the electric field primarily enhances carrier scattering, resulting in intensified energy dissipation and thus increased heat generation.[62] As a result, regions with strong electric fields typically correspond to high local heat densities. In field-effect transistors, such hot spots are commonly found near the drain side beneath the gate.[63,64] Based on this correlation, reducing the peak electric field within the transistor channel through structural design can effectively lower the maximum heat generation density and thereby reduce the devices junction temperature.

To demonstrate the effectiveness of structural design in reducing heat generation, Fig. 2 presents an example using an asymmetric slant field plate (FP) to lower the maximum heat generation density in the device.[43] The FP structure, originally introduced in HEMTs to enhance breakdown voltage, has also drawn attention for its potential to alleviate self-heating effects.[65–68] In a conventional AlGaN/GaN HEMT, a gate-connected FP with a slant angle of 6° and a length of 1200 nm is employed, as shown in Fig. 2(a). The FP is composed of the same material as the gate, while the passivation layer beneath it consists of SiN. The structural parameters are based on previously reported experimental data.[69]

The simulated heat generation profile under bias conditions of $V_{GS} = -1$ V and $V_{DS} = 8$ V is presented in Figs. 2(b) and 2(c). Compared to the conventional design, where heat is highly localized near the drain-side gate edge, the introduction of the slant FP results in a more uniform heat distribution across the GaN channel layer. This improvement is attributed to the modified potential profile induced by the FP, which effectively extends the voltage drop region between the gate and the drain. Consequently, the peak electric field and the maximum heat generation density are both significantly reduced. Figures 2(d) and 2(e) show the lateral distribution of heat generation and electric field intensity along the channel beneath the heterojunction interface. The results indicate that the slant FP reduces both the peak electric field and the heat generation density by approximately 50%. Moreover, the width of the high-heat region increases from 100 to 200 nm, further confirming the FP's effectiveness in smoothing the channel potential and mitigating local thermal accumulation.[70]

Figure 2(f) presents the simulated temperature distribution beneath the device channel, comparing results obtained using Fourier's heat conduction law and phonon BTE. With the inclusion of a 1200 nm slant field plate, the maximum temperature decreases from 365.2 to 354.6 K, corresponding to a ≈16.3% reduction in peak temperature rise, thereby demonstrating a notable improvement in thermal performance. The TDA system supports MC simulation to account for non-Fourier heat conduction, which becomes significant when the phonon MFP exceeds the characteristic size of the heat source—as is the case in GaN.[13] In the

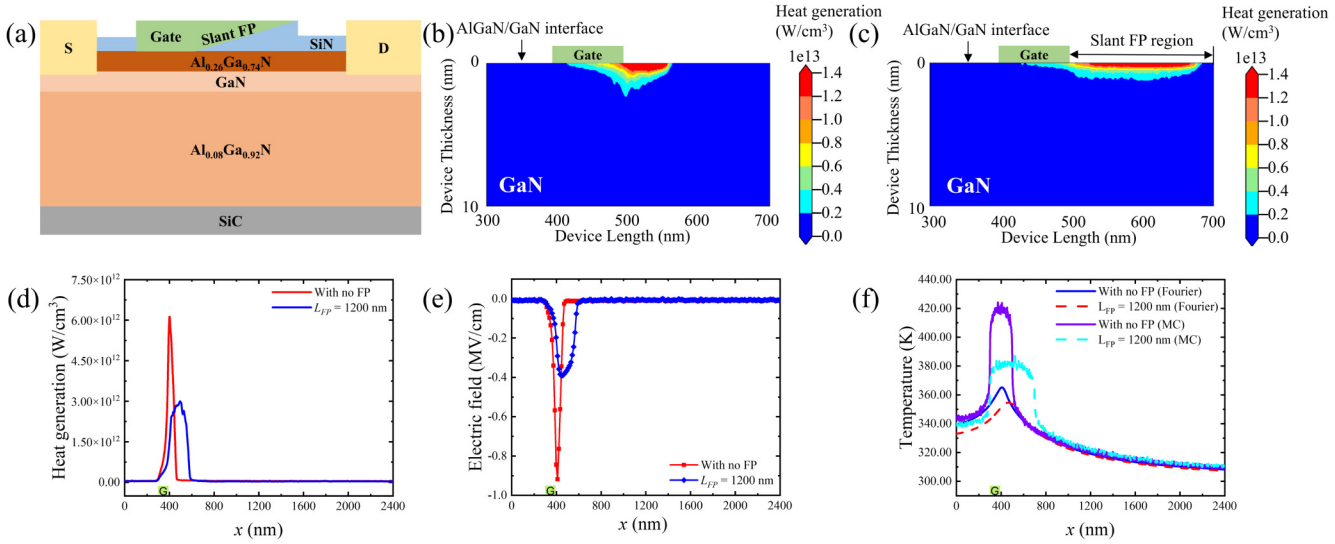

FIG. 2. (a) Structure of an AlGaN/GaN HEMT with a slant field plate. (b) and (c) GaN channel heat generation in conventional and FP-based designs. (d)–(f) Lateral distributions of heat generation, electric field, and temperature.

conventional HEMT, phonon ballistic transport leads to a substantial increase in the hot spot temperature, from 362 K under Fourier's model to 421 K under MC. In contrast, the slant FP design limits this temperature rise to 383 K. This enhanced suppression is attributed to the FP's ability to enlarge the effective heat source region, thereby mitigating phonon ballistic effects. Under non-Fourier conditions, the junction temperature is reduced by 33%, nearly twice the reduction predicted by Fourier's law. These results demonstrate that structural design optimization can effectively reduce heat generation density and mitigate self-heating effects. However, it is important to note that the introduction of FP structures increases parasitic capacitance, which may adversely affect high-frequency device performance. Therefore, future structural optimizations should carefully weigh the trade-off between thermal benefits and potential degradation in electrical performance, to achieve design solutions aligned with specific application requirements.

B. Non-Fourier heat spreading

When heat is generated within a transistor, it initially spreads from localized hotspots across the entire device and eventually into the surrounding chip substrate.[71] This process introduces significant thermal spreading resistance as heat flows from a small, confined region to a larger area. Accurately predicting the resulting temperature field and hotspot temperatures is critical—not only for ensuring the fidelity of electro-thermal co-simulations,[60] which reflect performance degradation in transistors, but also for enabling reliable modeling of device lifetime and failure mechanisms.[16] Also, in power electronics, such as GaN HEMTs, thermal spreading plays a decisive role in device design.[72] Variations in thermal spreading behavior can fundamentally alter design strategies—for instance, the impact of the spreading layer's thickness on device thermal resistance.[73] Understanding and optimizing this effect is essential for developing high-performance, low thermal-resistance transistors.

Thermal spreading analysis based on commercial software typically relies on Fourier's law of heat conduction,

$$\mathbf{q} = -\kappa \nabla T, \tag{7}$$

where $\mathbf{q}$ is the heat flux, T is the temperature, and κ is the intrinsic thermal conductivity—a material property independent of geometry or size.[74] Fourier's law is based on the assumption that heat is transported diffusively, with heat carriers (primarily phonons in semiconductors) undergoing frequent internal scattering events. However, this assumption breaks down when the phonon MFPs are comparable to the characteristic dimensions of the system or the size of the heat source—a common scenario in nanoscale structures or the thermal spreading process in transistors. In such cases, phonons may travel ballistically, traversing the structure with few or no internal scatterings.[75] This leads to pronounced non-Fourier effects, such as boundary temperature jumps and heat flux slip, rendering Fourier's law invalid.[76] Beyond size effects, non-Fourier behavior also emerges in ultra-fast transient thermal processes—such as when a transistor is suddenly switched on—where the timescale of heat conduction becomes comparable to the phonon relaxation time (on the order of picoseconds).[77] Under these conditions, thermal wave effects can be observed.[78] Moreover, in transistors under operation, electrons absorb energy from the electric field and become hot electrons transferring energy to phonons via electron–phonon scattering.[11] High-energy electrons tend to excite high-frequency optical phonons.[79] This selective phonon excitation leads to a highly non-equilibrium phonon population, with optical

phonons reaching much higher temperatures—a phenomenon often referred to as the phonon energy bottleneck, which further contributes to non-Fourier behavior.[80]

Accurately capturing non-Fourier effects is crucial. On one hand, it enables the development of precise models for heat generation and transport, which can be incorporated into reliability analysis and large-scale circuit simulations.[33] On the other hand, it supports advanced thermal design strategies for enhancing transistor performance.[81] To reflect these non-Fourier phenomena, thermal transport must instead be described using the phonon BTE.[82] Rooted in particle dynamics and neglecting the wave nature of heat carriers, the BTE captures ballistic phonon transport, transient thermal wave effects, and inter-carrier coupling, making it well-suited for modeling thermal transport at the nano- to microscale in electronic devices. The general form of the phonon BTE is[82]

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla f = \left(\frac{\partial f}{\partial t}\right)_s + \dot{s}_f, \tag{8}$$

where f is the phonon distribution function, $\mathbf{v}$ is the phonon group velocity, $\left(\frac{\partial f}{\partial t}\right)_s$ represents the scattering term, and $\dot{s}_f$ is the source term. To account for mode-dependent energy non-equilibrium, different phonon modes can be assigned distinct weights in the source term.[80] In practice, the scattering term is often approximated using the relaxation time approximation (RTA),[83]

$$\left(\frac{\partial f}{\partial t}\right)_s = -\frac{f - f^{\mathrm{eq}}}{\tau}, \tag{9}$$

where f^{eq} is the equilibrium phonon distribution and τ is the phonon relaxation time.

To numerically solve the phonon BTE, there are generally two kinds of methods: deterministic and statistical methods.[84] The deterministic methods discretize both spatial and phonon momentum spaces and numerically solve the resulting discrete algebraic equations. The discrete ordinates method (DOM) is the most widely used deterministic approach for solving the phonon BTE.[85] In recent years, several improved schemes have been proposed to enhance its efficiency and accuracy. For example, the Discrete Unified Gas Kinetic Scheme (DUGKS) treats phonon transport and scattering in a unified manner, effectively addressing the slow convergence issue in transient simulations under diffusive conditions.[86] Meanwhile, the synthetic iterative scheme has been proposed to improve convergence in steady-state simulations.[87] Statistical methods primarily involve MC simulations, treating phonon transport as a stochastic process. MC methods rely on tabulated phonon properties and simulate individual phonon movements, making them memory-efficient and particularly advantageous for complex nanoscale structures and varied scattering mechanisms.[82]

Phonon MC methods include traditional phonon number-based methods[88] and energy-based variance-reduced approaches.[89,90] Traditional phonon number-based methods have largely fallen out of favor due to statistical errors and energy conservation issues. Energy-based methods include ensemble MC[89] and tracing (or kinetic) MC approaches.[90] Ensemble MC inherently simulates in transient schemes. In each time step, positive and negative energy phonons need to be newly generated or annihilated in each grid, making it time consuming and poor parallelism.[89] Also, making steady-state computations time-consuming, a long-term average can give the final stable results. Conversely, tracing MC tracks individual phonons directly, allowing efficient transient-state simulations and much more efficient steady-state simulations, which also has a nearly linear parallelism.[91] It should be noted that tracing MC can not only be used to predict the effective thermal conductivities of nanostructures or TBR, but it can also give the temperature field of the whole structures.[13] It can accommodate temperature-dependent phonon properties by iteration, can deliver results consistent with ensemble MC, and significantly reduces computational time—typically to about 1% of the ensemble MC method for steady-state cases since all counts to the phonon packets contribute to the accumulation of the temperature or heat flux, which saves much time.[92]

In recent years, ML methods have been increasingly leveraged to solve the phonon BTE. Among these, physics-informed neural networks (PINNs) have gained particular attention, as they incorporate the governing partial differential equations (PDEs) directly into the loss function. When the underlying PDEs are known, PINNs enable solutions to be learned in a physics-constrained manner without the need for labeled training data. Li *et al.*[93] demonstrated the application of PINNs for efficiently solving the phonon BTE across one- to three-dimensional problems. These approaches have since been extended to address more complex scenarios, including large temperature non-equilibrium conditions,[94] time-dependent problems,[95] and electron—phonon interactions.[96] In parallel, advances in high-performance computing have been explored to further accelerate phonon BTE solvers. For example, Shang *et al.*[97] employed the JAX framework to recast finite-volume updates as graph neural network (GNN)-style message passing, thereby achieving highly parallelized computations on modern GPU architectures.

Despite the efficiency improvements enabled by various phonon BTE solution techniques, simulations involving thick substrates—such as GaN HEMTs—remain computationally intensive. Fortunately, ballistic transport near nanoscale heat sources arises due to limited internal scattering and primarily occurs in the vicinity of hotspots, typically within a few phonon MFPs.[79] Beyond this region, phonons undergo sufficient scattering to reach thermal equilibrium, allowing Fourier's law to become valid again. Moreover, ballistic effects caused by phonon-boundary scattering in thin epitaxial layers can be effectively captured using effective thermal conductivities in Fourier-based simulations.[98] As a result, phonon BTE simulations can be confined to regions within a few micrometers of the hotspot.[99] By coupling these local BTE simulations with Fourier-based models through appropriate boundary conditions, a multiscale, device-level thermal analysis can be efficiently and accurately performed.[79]

In the TDA system, we have developed a module centered on a phonon tracing MC framework, integrated with a multiscale MC-FVM hybrid simulation platform. This module supports different kinds of phonon property models, including gray-medium approximations, isotropic dispersion models, and full-band phonon

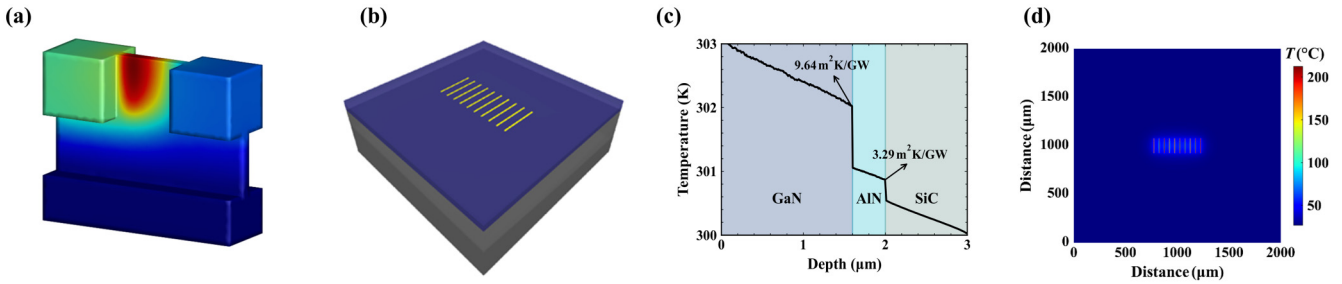

FIG. 3. (a) Electro-thermal co-simulation results for FinFET transistor structures. (b) Schematic of multifinger GaN HEMTs. (c) Temperature distribution profiles and corresponding effective thermal properties in GaN heterostructures. (d) Multiscale thermal analysis predictions of the temperature field distribution in the heat source plane of GaN HEMTs.

dispersion relations imported from first-principles calculations.[42] A comprehensive material database has been constructed to support various semiconductor materials.[38] By decoupling phonon dispersion from relaxation times, the module enables efficient storage of effects, such as temperature variation, isotope scattering, defects, and stress on phonon properties. Additionally, full-band interface phonon information for heterostructures is included in the database. With the aid of ML-driven LD simulations, phonon properties can be computed directly from real atomic configurations.[40] This allows the inclusion of realistic interface conditions, such as interfacial amorphous layers, and other practical factors arising during fabrication. Also, backend processes, including automatic grid interpolation and a model setup, are fully automated, enabling seamless multiscale integration and eliminating the need for cumbersome manual configuration.

As shown in Fig. 3, this TDA module supports a wide range of applications—from fundamental studies of non-Fourier thermal transport in nanostructures, to transistor-level simulations of heat generation and spreading, and device-level multiscale thermal modeling and design. The system can accurately simulate effective thermal conductivity in nanostructures, such as nanoporous materials, confined nanowires, and multilayer interconnects. Given the widespread use of 3D micro/nano interconnects in modern chips—whose dimensions are often comparable to phonon MFPs—the TDA module captures the thermal conductivity degradation due to boundary scattering and provides reliable inputs to circuit-level CTMs or FVM-based simulations for system-level design.[100] Furthermore, by integrating self-heating analysis, the TDA system enables precise prediction of junction temperatures and supports electrothermal co-design, facilitating the design of transistors with low thermal resistance.[43] For high-power devices, such as GaN HEMTs, the module performs multiscale simulations to characterize epitaxial layer properties and interfacial thermal resistances, guiding epitaxial structure design for thermal performance optimization. The TDA module has been applied in 2012 by the Huawei Technologies Co., Ltd. laboratory, where it reduced the design cycle of thermal management structures by 70%. It achieved a 20% reduction in thermal resistance and a 1% improvement in GaN power amplifier efficiency through optimized device structural design.

C. Interfacial phonon transport

In the heat dissipation pathway of electronic devices, heat is initially generated in the device layer and sequentially transferred to the substrate, flange, and heat sink. During this process, TBC plays an important role in the device's overall thermal performance. Due to the electrical properties and design requirements of the device, reducing thermal resistance within bulk materials is extremely challenging. The selection of substrate materials, however, offers more flexibility. Adopting high thermal conductivity substrates does not affect the electrical performance of devices and represents an important approach to improve heat dissipation, and represents an important approach to improve heat dissipation, effectively enhancing power density in electronic devices.

For example, GaN HEMTs with Si substrates demonstrate DC power densities of 4.5 W/mm,[101] while those with SiC substrates (offering higher thermal conductivity) can achieve 15 W/mm,[102] and diamond substrate GaN HEMTs can reach up to 56 W/mm.[103] With diamond possessing the highest thermal conductivity in nature (approximately 2000 W m-1 K-1), it has become the most sought-after substrate material for high-power devices. However, diamond's extreme hardness (Mohs hardness of 10), relatively low thermal expansion coefficient (0.8×10^{-6} K^{-1}), and lattice mismatch make it difficult to form high-quality interfaces with other materials. Another promising high thermal conductivity substrate is SiC, which has relatively lower thermal conductivity but a higher thermal expansion coefficient. The low TBC and high thermal stress between semiconductors and substrates represent major barriers to adopting high thermal conductivity substrates.

Additionally, with device miniaturization and three-dimensional development, the interface density within devices is gradually increasing, making TBC comparable to thin-film thermal resistance. For example, 50 W m^{-2} K−1 of Ga_2O_3/SiC TBC is equivalent to that of 300 nm Ga_2O_3 or 9 μm SiC. Reducing the TBC between wide-bandgap semiconductors and high thermal conductivity substrates will directly lower junction temperatures. When the TBC of Ga_2O_3/diamond increases from 17 to 300 W m^{-2} K^{-1}, the junction temperature rise decreases from 270 to 130 K.[104] Therefore, reducing the TBR between wide-bandgap semiconductors and high thermal conductivity substrates is crucial for addressing chip thermal bottlenecks.

In real-world devices, strain is often a pervasive factor affecting both nanostructures and interfaces. Strained silicon, for instance, has been widely employed to enhance electron mobility. However, strain also influences the phonon and electronic properties of a material, thereby altering its thermal conductivity and TBC.[105,106] For example, in silicon, biaxial tensile strain can reduce thermal conductivity, whereas compressive strain can enhance it. Furthermore, strain at an interface can modulate the TBC by changing the phonon dispersion and group velocities. Therefore, accounting for the effects of strain in thermal design is crucial for accurately predicting and optimizing the thermal performance of devices.

Predicting interfacial thermal transport of realistic interfaces is a challenging task.[107,108] When phonons encounter the interface between two materials, the propagation of phonons is hindered due to the disruption of lattice periodicity by the interface. From a particle perspective, part of the phonon energy is transmitted to the other side of the interface, while another portion is reflected back. The TBC can be calculated using the Landauer formula,[109] which views thermal conduction as a quantum transport process of phonons across the interface,

$$G = \frac{1}{V}\sum_{\mathbf{q},s} \hbar\omega_{\mathbf{q},s}\frac{\partial f_{BE}}{\partial T} v_z(\mathbf{q}, s)\mathscr{T}(\mathbf{q}, s), \tag{10}$$

where V is the unit cell volume, $\mathbf{q}$ is the wave vector, s is the phonon branch index, $\omega_{\mathbf{q},s}$ is the phonon angular frequency, f_{BE} is the Bose–Einstein distribution function, T is the temperature, v_z is the phonon group velocity perpendicular to the interface, and $\mathscr{T}(\mathbf{q}, s)$ is the phonon transmittance. To predict interfacial thermal resistance, physical quantities in the Landauer formula, such as V, $\omega_{\mathbf{q},s}$, and v_z, are relatively easy to obtain. However, calculating $\mathscr{T}$ is extremely challenging. $\mathscr{T}$ is influenced by multiple factors, including the phonon spectra on both sides of the interface, interfacial bonding strength, and interface roughness. At the microscopic level, $\mathscr{T}$ is significantly affected by the atomic morphology of the interface. Determining $\mathscr{T}$ remains the primary challenge in predicting interfacial thermal resistance at micro- and nanoscales.

Two continuum models were developed to represent phonon transmittance at interfaces: the Acoustic Mismatch Model (AMM)[110] and the Diffuse Mismatch Model (DMM).[111,112] The AMM model approximates two media as continuous, homogeneous, and isotropic media, neglecting the details of their lattice structures. The DMM model assumes the opposite extreme, where phonons arriving at an interface undergo complete diffuse scattering, with phonons at different incident angles exhibiting identical scattering characteristics.[113] As macroscopic medium models, AMM and DMM established the theoretical framework for studying thermal transport across solid interfaces and provided explanations for low-temperature interfacial thermal resistance calculations. However, neither model can reflect the relationship between interfacial thermal resistance and interfacial atomic structures.

Several atomistic simulation techniques are currently employed to predict TBC. Molecular Dynamics (MD) simulation, a traditional method for simulating heat transfer, involves tracking the movement of individual atoms within a system. This is achieved by applying Newton's equations of motion and utilizing empirical interatomic potentials. MD simulations are broadly classified based on the presence or absence of non-equilibrium processes, resulting in Non-Equilibrium Molecular Dynamics (NEMD) and Equilibrium Molecular Dynamics.[114–116] NEMD can be further subdivided into Steady-State Molecular Dynamics[117] and Transient Molecular Dynamics.[118] Although MD-based heat flux decomposition techniques can elucidate the contribution of various phonon frequencies to TBC,[119] they are limited in their ability to determine the transmissivity of specific phonon modes. Moreover, MD simulations treat phonons classically, according to Boltzmann statistics, thereby neglecting the quantum mechanical aspects of a phonon distribution. The atomistic Green's function method,[120] on the other hand, accounts for the Bose–Einstein distribution of phonons and calculates the transmission of excitations across an interface using Green's functions. The original AGF method operated within a harmonic framework, but a more recent iteration incorporates anharmonic scattering.[121] While conventional AGF methods provide frequency-dependent phonon transmissivity, more advanced methods have been developed to calculate the interfacial transmissivity of individual phonon modes.[122]

Recent methodological innovations have introduced Lattice Dynamics (LD) as a transformative approach to interface thermal property prediction.[40,123,124] This technique calculates mode-specific phonon transmission coefficients and applies Landauer formalism to determine thermal boundary conductance. Though mathematically analogous to atomic Green's function methodologies, LD extracts superior phonon phase information, yielding more comprehensive characterization of interfacial thermal transport. The computational efficiency advantage is particularly noteworthy—LD requires system dimensions merely matching phonon wavelengths rather than mean free paths, reducing computational expense by approximately three orders of magnitude compared to conventional molecular simulations.

The complexity of interfacial atomic structures necessitates a precise representation of interatomic forces. Traditional approaches, such as empirical potentials, while commonly employed, may lack accuracy and transferability across different materials. In contrast, although DFT methods provide exceptional precision, they are computationally prohibitive for complex systems. MLPs have emerged as a revolutionary bridge between computational efficiency and first-principles accuracy.[125,126] These advanced computational tools maintain quantum mechanical precision while dramatically reducing computational requirements, enabling analysis of complex interfacial structures and large atomic ensembles. MLPs demonstrate remarkable capabilities in calculating the thermal conductivity of materials.[36,37] Recently, MLPs have been integrated with molecular dynamics to calculate TBC at various interfaces.[127–129] An effective interfacial MLP should cover configurations across bulk phases, amorphous structures, crystal/crystal, amorphous/crystal, and amorphous/amorphous interfaces, elemental mixing, temperature ranges, and strain conditions. General-purpose pretrained MLPs (e.g., MACE-MP[130] and MatterSim[131]) can serve as starting points; however, fine-tuning on small, targeted interfacial datasets is often required to achieve higher accuracy. Initial datasets can be constructed from molecular

dynamics trajectories driven by pretrained MLPs and relabeled using DFT under a unified set of parameters; active-learning loops are widely used to identify rare or high-error interfacial features. Validation is multifaceted: in addition to energy/force/stress errors, checking bulk phonon dispersion is standard practice.

In the LD approach, the system is divided into three parts: left lead (L), device (D), and right lead (R), where L and R are semi-infinite, and D has sufficient thickness to contain localization effects. Assuming a system with harmonic interatomic interactions governed by Newton's equations of motion, it can be described by the eigenvalue equation,

$$\begin{pmatrix} H_{\mathrm{LL}} & H_{\mathrm{LD}} & 0 \\ H_{\mathrm{DL}} & H_{\mathrm{DD}} & H_{\mathrm{DR}} \\ 0 & H_{\mathrm{RD}} & H_{\mathrm{RR}} \end{pmatrix} \begin{pmatrix} u_{\mathrm{L}} \\ u_{\mathrm{D}} \\ u_{\mathrm{R}} \end{pmatrix} = \omega^2 \begin{pmatrix} u_{\mathrm{L}} \\ u_{\mathrm{D}} \\ u_{\mathrm{R}} \end{pmatrix}, \tag{11}$$

where H_{LL}, H_{LD}, H_{DL}, H_{DD}, H_{DR}, and H_{RR} are the coupling Hamiltonian matrices of the left lead, device, and right lead, respectively. Due to the infinite length of the leads, H_{LL} and H_{RR} are infinite-dimensional, which complicates a direct solution. Considering that far from the interface, the atom vibrations can be described by a combination of bulk phonon modes, Then u_{L} can be expressed as the sum of the incident and all reflected phonon modes, while u_{R} can be expressed as the sum of all transmitted phonon modes. The eigenvalue equation can be simplified as

$$\begin{aligned} H_{\mathrm{DL}}u_{\mathrm{L}} + H_{\mathrm{DD}}u_{\mathrm{D}} + H_{\mathrm{DR}}u_{\mathrm{R}} &= \omega^2(u_{\mathrm{L}} + u_{\mathrm{D}} + u_{\mathrm{R}}), \\ u_{\mathrm{L}} &= u_{\mathrm{Inc}} + \sum_{\mathrm{Refl}} A_{\mathrm{Refl}}u_{\mathrm{Refl}}, \\ u_{\mathrm{R}} &= \sum_{\mathrm{Trans}} A_{\mathrm{Trans}}u_{\mathrm{Trans}}, \end{aligned} \tag{12}$$

where u_{Inc} is the incident phonon mode, u_{Refl} is the reflected phonon mode, and u_{Trans} is the transmitted phonon mode. This approach leads to a linear system involving a finite number of variables (u_{D}, A_{Refl}, and A_{Trans}). Solving this linear system, we can obtain the amplitude of the reflected and transmitted phonon modes. The transmittance can be calculated as

$$\mathscr{T} = \sum_{\mathrm{Trans}} \frac{v_{z,\mathrm{Inc}}}{v_{z,\mathrm{Trans}}} |A_{\mathrm{Trans}}|^2. \tag{13}$$

The TBC can be calculated using the Landauer formula [Eq. (10)]. In the TDA system, the TBC value is used in the FVM module, while mode-resolved phonon transmittance is used in the MC module.

To illustrate the practical application of the methods discussed above, we present a comprehensive analysis of thermal transport in SiC/AlN/GaN heterostructures, which are crucial for high-power and high-frequency electronic devices. Figure 4(a) shows the atomic structure of the SiC/AlN and AlN/GaN heterostructure. Figure 4(b) shows the group velocities for the three materials. The order of group velocities from the largest to smallest is SiC > AlN > GaN. The cutoff frequency of SiC phonons is larger than that of AlN and GaN. All three materials have a phonon bandgap, indicating the separation of acoustic and optical phonons. Figure 4(c) presents the frequency-dependent transmittance across the SiC/AlN and AlN/GaN interfaces using the LD method. Low-frequency acoustic phonons possess higher transmittance.

Phonon density of states overlap is considered to describe the degree of phonon matching between two materials[123] and is believed to correlate positively with TBC. We calculated the phonon density of states for three materials, as shown in Fig. 4(d). Using Eq. (10), the contribution of each phonon mode to the TBC can be calculated. By calculating the contribution of phonons at different frequencies to the TBC, we can compute the spectral TBC, as shown in Fig. 4(e). The spectral TBC exhibits higher peaks in the low-frequency region, especially for the SiC/AlN interface, which is closely related to their larger overlap area of phonon density of states in the low-frequency region. Finally, we calculated the TBC for each interface as a function of temperature, as shown in Fig. 4(f). The TBC of the SiC/AlN interface is much higher than that of the AlN/GaN interface. This can be attributed to the higher group velocities of SiC and AlN phonons and the higher phonon transmittance of the SiC/AlN interface.

Inelastic scattering and higher-order processes open additional transport channels, typically enhancing the TBC when the temperature approaches the Debye temperature, or at disordered/amorphous and highly mismatched interfaces. Two principal theoretical approaches are used to investigate phonon anharmonic scattering. The anharmonic AGF formalism can, in principle, treat anharmonic heat transport across three-dimensional interfaces. However, the complexity and irregularity of real interfaces can make practical applications challenging. Molecular dynamics-based methods (NEMD/EMD) inherently include anharmonicity to all orders, but their predictions are influenced by finite-size effects and by classical statistics that neglect quantum effects. For the system studied here, because the temperature is below the Debye temperature, the interface disorder is limited, and the lattice mismatch is small, elastic predictions reasonably capture the TBC at 300 K. This simplification also enables computation of mode-resolved phonon transmission.

D. Compact thermal modeling

Having established the fundamental mechanisms of heat generation, non-Fourier spreading, and interfacial thermal resistance at the device level, the TDA framework must now address thermal transport at larger scales where multiple devices interact within integrated circuits. At the turn of the century, power density and on-chip temperature rise emerged as critical concerns in very large-scale integration (VLSI) circuits, as studies predicted significant increases in maximum temperatures due to higher interconnect density, increased current density, and enhanced thermal coupling between neighboring devices.[132] Over the past two decades, these thermal challenges have intensified with continued IC technology evolution. According to data from ITRS and Intel, more than 50% of functional units in an 8 nm process multi-processor system-on-chip (SoC) cannot operate at full performance due to thermal reliability concerns.[133]

Contemporary 3D integration technologies, which enable higher integration density and heterogeneous technology integration,[134,135] have further amplified thermal dissipation challenges as

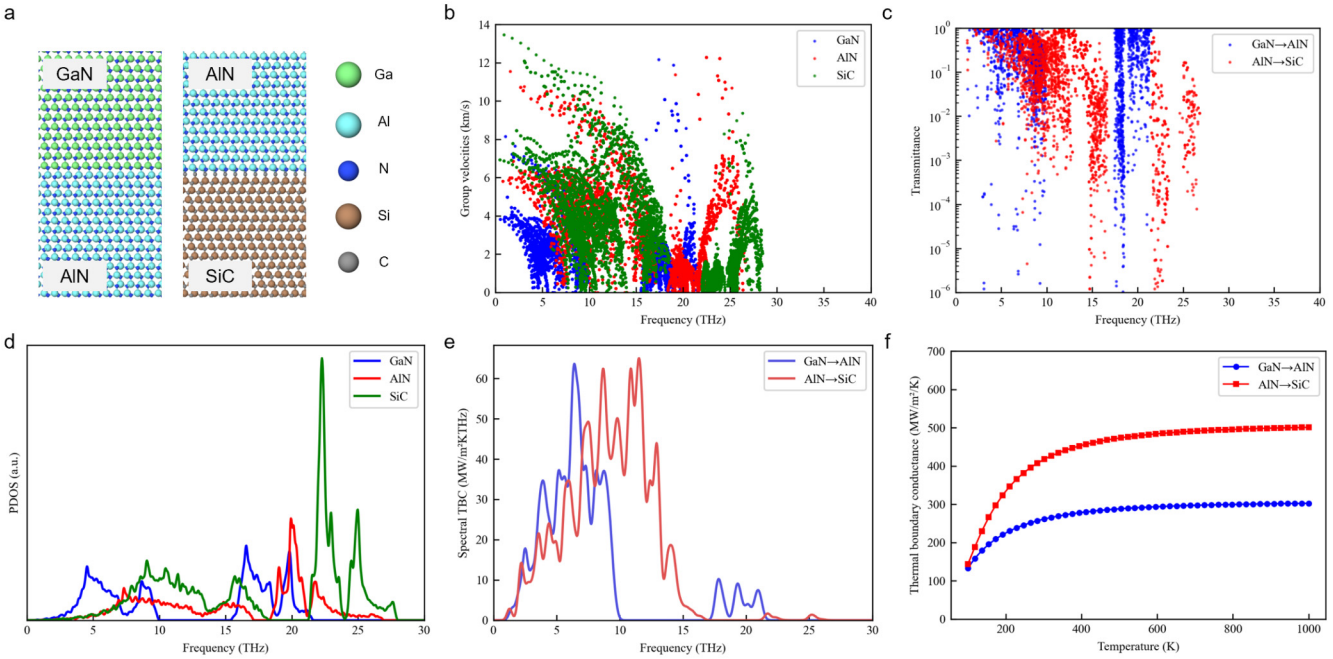

FIG. 4. Illustration of the calculation of TBC. (a) Atomic structure of the SiC/AlN and AlN/GaN heterostructure. (b) Group velocities of GaN, AlN, and SiC. (c) Frequency-dependent transmittance across the SiC/AlN and AlN/GaN interfaces using the LD method. (d) Phonon density of states for GaN, AlN, and SiC. (e) Spectral TBC for the SiC/AlN and AlN/GaN interfaces. (f) TBC for the SiC/AlN and AlN/GaN interfaces as a function of temperature.

multiple layers of simultaneously active devices are vertically stacked.[136] The complex thermal interactions between stacked layers create intricate heat flow patterns that cannot be captured by device-level simulations alone. Moreover, 3D integration has blurred the boundaries between design and manufacturing, driving strong interest in design technology co-optimization (DTCO) and system technology co-optimization (STCO).[137] These developments necessitate circuit-level thermal modeling approaches that can efficiently bridge the gap between detailed device-level physics and system-level thermal management, making compact thermal modeling an essential component of the comprehensive TDA framework.

CTMs are an effective approach to addressing these challenges. CTMs represent the thermal behavior of ICs using lumped RC networks, where heat dissipation is modeled as current sources and heat conduction as resistors and capacitors. This abstraction provides a favorable trade-off, delivering reasonably accurate temperature predictions with minimal computational effort at various levels of abstraction, for both steady-state and transient analyses.[44] Owing to their computational efficiency, CTMs are widely used for architecture-level temperature-aware design[138,139] and for rapid simulation of dynamic thermal management techniques.[140,141] Furthermore, due to the widespread familiarity of SPICE among circuit designers, CTMs have been employed for electro-thermal co-simulation for nearly half a century.[142,143] One of the earliest and most influential CTM solvers, HotSpot, models packaging structures as stacks of homogeneous material layers.[144,145] Since version 6.0, it has supported non-uniform layer properties, making it available for thermal simulations of TSVs.[138,146] Another widely used tool, *3D-ICE*, enables thermal simulations of 3D-ICs with both single-phase and two-phase microchannel cooling.[147] More recently, CTM solvers have been further enhanced to incorporate temperature-dependent and anisotropic effects for improved accuracy in emerging technologies,[148] as well as parallelization for efficiency and EDA integration to support DTCO and STCO.

However, as interconnects continue to shrink and integration density increases, existing CTM solvers face significant challenges in accurately capturing the multiscale thermal behavior of 3D-ICs. The ongoing downscaling of the BEOL, along with the introduction of BPDSN, substantially increases thermal resistance and on-chip temperatures due to micro- and nanoscale effects, such as reduced thermal conductivity, grain boundary scattering, and interface scattering.[149–151] Furthermore, die stacking and monolithic integration exacerbate vertical thermal coupling at these scales.[152,153] While CTM offers faster simulation at the architectural level, it often compromises accuracy due to coarse approximations of local structures.[144,154] As a result, the abstraction level of traditional CTM approaches is frequently insufficient to model the complex thermal behavior of advanced 3D-ICs with high fidelity.

Recently, TDA has developed a heterogeneity-aware CTM framework, an *H3Therm* module, specifically designed to capture micro-nanoscale thermal behavior in IC-level thermal simulations. As illustrated in Fig. 5, *H3Therm* integrates a novel BTE–EMA model for accurate effective thermal conductivity prediction and

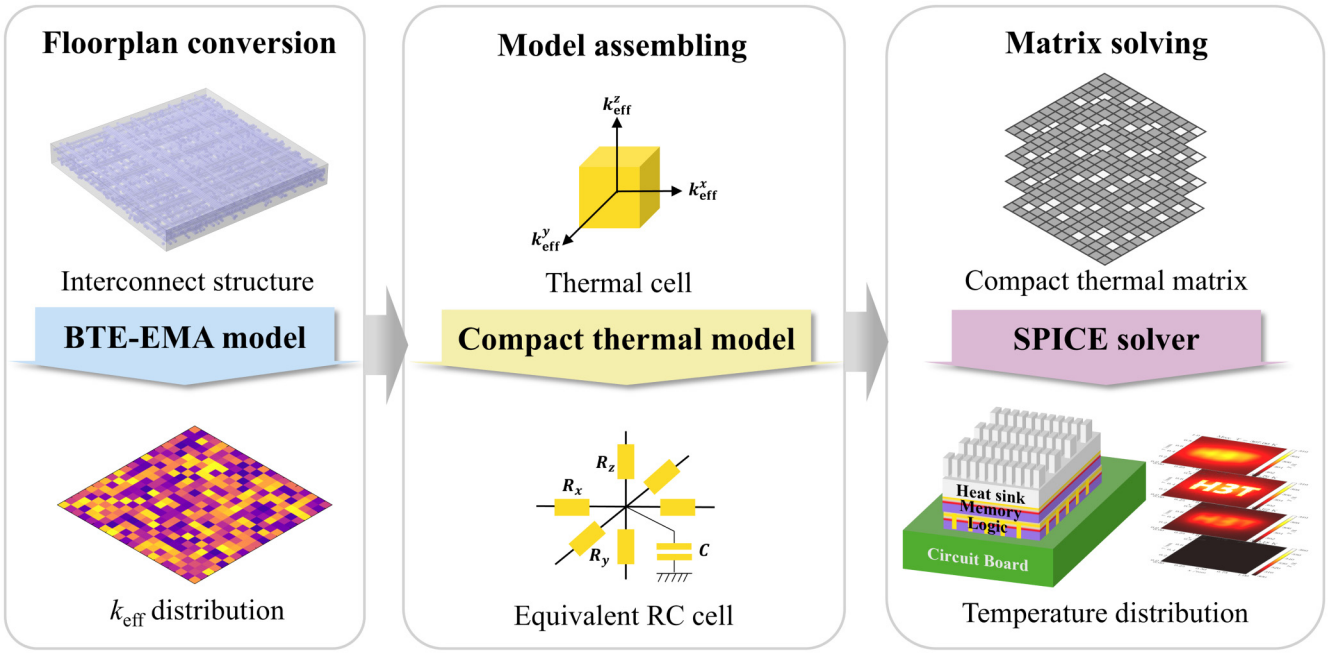

FIG. 5. CTM workflow and thermal-aware design of 3D integrated circuits.

employs an adaptive, power-aware refinement strategy to optimize model granularity. This framework is seamlessly integrated into compact thermal simulators, enabling efficient and accurate thermal simulations for VLSI and 3D-ICs. Building on the EMA theory for ellipsoidal inclusions[155,156] and the BTE approach that accounts for multiple geometric constraints in different directions,[75] we have developed a BTE–EMA approach that accurately predicts the effective thermal conductivity of embedded nanoscale interconnects in 3D-ICs. This model demonstrates excellent agreement with *ab initio* BTE simulation results,[157,158] reliably capturing the effective thermal conductivity of copper nanowires across various dimensions and aspect ratios. To address the variability in interconnect structures and power distributions across ICs, we further propose an adaptive, power-aware refinement method to determine the optimal modeling granularity. The overall framework is integrated into a SPICE-based CTM solver, which preprocesses GDSII files to extract fine-grained thermophysical properties for interconnect layers and performs CTM simulations with adaptive model granularity.

For example, we evaluated the thermal performance of various 3D stack configurations, including face-to-face (F2F), face-to-back (F2B), and back-to-back (B2B) arrangements, using *H3Therm*. Each die has dimensions of $1 \times 1\text{mm}^2$ and dissipates a total power of 1.28 W. About 11.95 % of the units experience a heat flux density of $1000\,\text{W mm}^{-3}$, while the remainder reach up to $10\,\text{W mm}^{-3}$. In the F2F configuration, the BEOL layers are directly bonded, whereas in the F2B and B2B configurations, the substrates incorporate arrays of TSVs with a width of $100\,\mu\text{m}$. The BEOL layers are constructed with random volume fractions of metal and via layers, with a metal width of 10 nm. During preprocessing, the BTE–EMA model is used to map the BEOL layers to effective thermal conductivity distributions. As this model relies on an analytical solution, it incurs minimal computational overhead. The model granularity is then adaptively refined based on the power distribution. Next, the interconnect floorplan files and other simulation parameters are input into the compact thermal solver. After matrix assembly and SPICE-based computation, the temperature distributions of the 3D-IC stacks are generated. The highest temperatures are typically observed in the active layers, and temperature maps are produced for the FEOL layers of both the lower and upper dies. Results show that the maximum temperature rise in the F2F configuration is 70% higher than in the B2B configuration, and the vertical temperature gradients between upper and lower dies differ significantly among the three stacking schemes. The solving time of *H3Therm* is approximately 4.27 s per case, compared to about 12 min for Fourier-BTE coupled simulations. The proposed CTM framework efficiently captures these thermal behaviors and provides valuable guidance for thermal-aware design across different floorplans and workloads.

Stack optimization and workload assignment are pivotal in the thermal design of 3D-ICs. Our CTM framework holds significant promise for integration with existing SPICE-based design tools, enabling real-time feedback on IC-level thermal performance and empowering designers to make informed decisions throughout the early design process. Future work will focus on further enhancing the accuracy and efficiency of the CTM framework, as well as expanding its application to thermal-aware design and dynamic thermal management strategies of 3D-ICs. This includes

comprehensive benchmarking across a broader range of architectures, floorplans, and power distributions to assess its generality and robustness. Another important extension is to systematically incorporate the impact of diverse interfaces—such as metal–semiconductor, dielectric—semiconductor, and interconnect—substrate—into the effective compact thermal model. By embedding simplified representations of interfacial resistances and electron–phonon coupling into the CTM framework, we aim to capture their cumulative influence on chip-level heat transport without prohibitive computational cost. This development will enable more accurate and predictive thermal-aware design across a wide range of advanced IC architectures. In addition, ongoing collaborations with leading IC manufacturers aim to obtain real chip data for direct comparison, facilitating further calibration and refinement of the solver and supporting its practical deployment in advanced designs for next-generation ICs.

E. Chip-level thermal management

While the preceding TDA modules address thermal transport from atomic scales through circuit-level analysis, the final stage of heat dissipation requires effective transfer from the chip package to the ambient environment. As established through the multiscale thermal analysis framework, heat generated within individual transistors must ultimately be rejected at the system level to maintain operational temperatures. Contemporary semiconductor devices, driven by continued scaling and performance demands, exhibit exponentially increasing power densities that pose unprecedented challenges to chip-level thermal management. The complete thermal pathway involves heat transfer from the die through TIMs, package substrates, and external cooling systems before final dissipation to the ambient environment.

As the pivotal bridge linking the chip and the heat dissipation module, TIM exerts a direct and critical influence on the thermal conduction efficiency. Traditional polymer-based composite materials encompass diverse types, including thermal conductive greases, thermal conductive gels, thermal conductive adhesives, thermal conductive pads, and thermal conductive phase-change materials. Research has revealed that as the thickness of the TIM diminishes, the proportion of the interfacial thermal resistance within the total thermal resistance will increase correspondingly.[159] For TIMs with high thermal conductivity, such as solder, the proportion of the interfacial thermal resistance is even more pronounced. An ideal TIM should exhibit both high thermal conductivity and low Young's modulus to mitigate the interfacial thermal resistance. Recent state-of-the-art research reveals that the synergistic doping technology of liquid metal and solid fillers can enhance the thermal conductivity of the material by an order of magnitude. This innovative approach enables the transition of the heat carrier from phonon conduction to electron conduction, and elevates the binding force between materials from van der Waals interaction to metallic bonding, for the first time achieving a substantial reduction in the interfacial thermal resistance.[159–161] This dual enhancement achieves unprecedented reductions in interfacial thermal resistance while maintaining the mechanical flexibility necessary for accommodating thermal stress deformation during chip operation.

Based on distinct energy input methods, the thermal design approaches of heat sinks employing modern cooling technologies can be primarily categorized into two types: passive cooling and active cooling. These two cooling methods exhibit notable disparities in terms of thermodynamic principles and engineering implementation. This section undertakes an in-depth investigation of the current mainstream chip cooling technologies as shown in Fig. 6.

1. Air cooling and semiconductor refrigeration

In the realm of thermal management solutions, air cooling, being the most fundamental and representative approach, currently serves as the commonly employed heat dissipation means for chips. This is primarily attributed to its notable advantages of high reliability and low cost. A passive cooling system transfers the heat generated by the chip to the heat sink via the heat conduction mechanism. Subsequently, heat dissipation is accomplished through natural air convection and the process of thermal radiation. Its typical architecture mainly comprises two crucial components: a finned heat sink and a high-thermal-conductivity material. In a typical air cooling system, enhanced heat transfer technologies are often intricately integrated with the design of metal fins and the application of heat pipes[162] or vapor chambers. This passive cooling method boasts numerous advantages, including zero power consumption, noiseless operation, and minimal maintenance requirements. Nevertheless, owing to the constraint of the natural convection heat flux limit, this method is solely applicable to the application scenarios of low-power chips. Forced air cooling technology represents an active heat dissipation approach. It creates a directional air flow via a fan, thereby triggering forced convection on the surface of the heat sink. To attain a high-heat dissipation capacity, a large fin surface area and a high air velocity generated by the fan are indispensable. Inevitably, this gives rise to the issue of noise pollution.

Thermoelectric cooling (TEC), an alternative efficient active cooling technology founded on the Peltier effect, demonstrates a rapid response ability and high-heat dissipation efficiency. It is capable of achieving two-way temperature control and precise temperature regulation, thus offering a reliable solution for maintaining a stable thermal environment across diverse applications. By implementing a collaborative strategy encompassing thermoelectric cooling, phase-change materials, and liquid cold-plates,[163] the cooling rate and latent heat recovery rate can be notably enhanced in high-temperature scenarios.

2. Microchannel cooling

In the context of successive major breakthroughs in 3D packaging and the performance of integrated circuits, the heat flux density of chips has reached the order of kilowatts per square centimeter. Traditional cooling approaches, such as heat sinks and fans, have faced bottlenecks in meeting the stringent demands of advanced packaging technologies. Liquid cooling technology offers significant advantages. The heat storage capacity and thermal conductivity of liquids are substantially higher than those of air, which is of crucial importance for ensuring the reliability, optimizing the performance, and prolonging the service life of chips. Moreover, the application of liquid cooling technology enables more efficient

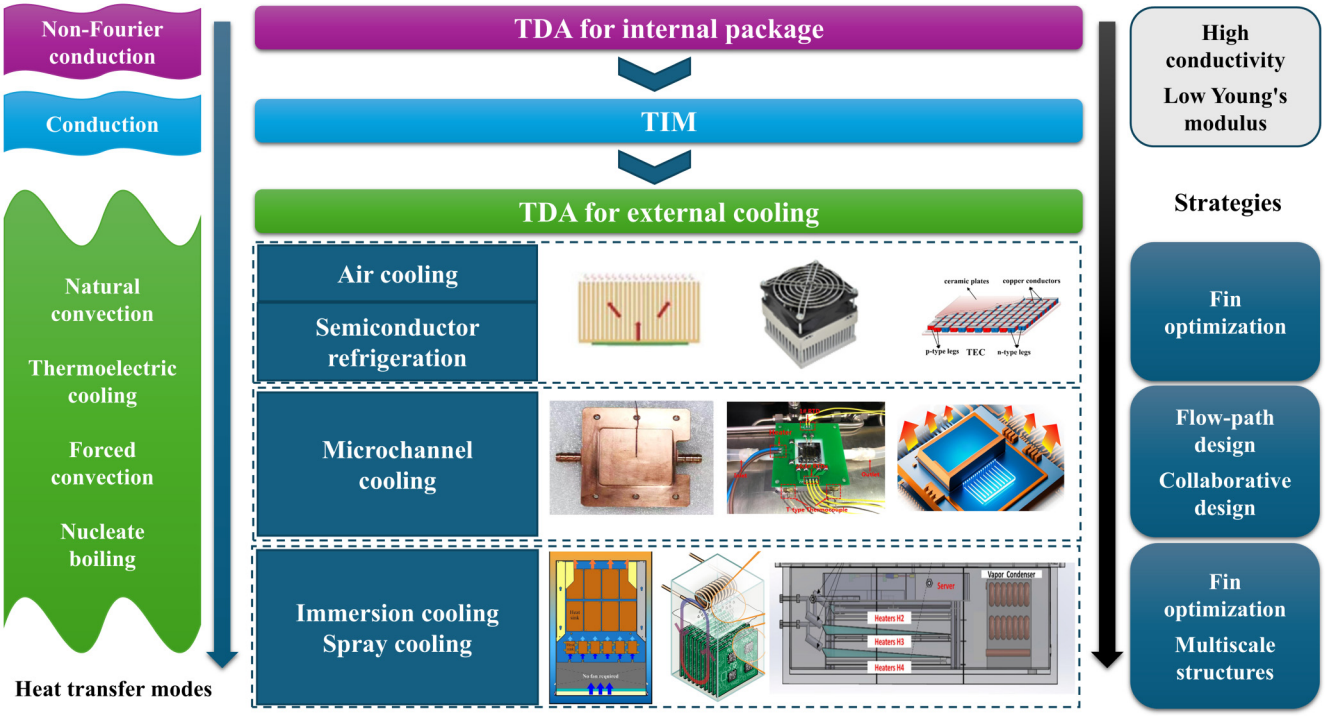

FIG. 6. TDA for external cooling technologies of chips.

utilization of natural cold sources, reduces the frequency of mechanical refrigeration, and, thereby, significantly decreases the energy consumption of the system. Based on whether the cooling medium is in direct contact with the heating element, liquid cooling technology can be primarily classified into direct-contact liquid cooling and indirect-contact liquid cooling. Among these, direct liquid cooling mainly encompasses immersion liquid cooling and spray liquid cooling, whereas indirect liquid cooling is predominantly exemplified by microchannel cooling. At present, globally, microchannel cooling, as the mainstream liquid cooling technology, has witnessed relatively mature development. The key components of a microchannel cooling system consist of a cold-plate, liquid pipes, coolant, and a driving pump. The heat generated by the chip during operation is indirectly transferred to the coolant via the microchannel cold-plate and subsequently removed.

The cooling efficiency and system energy consumption of microchannels are intricately influenced by the flow channel structure. Distinctive microchannel designs have the potential to transcend conventional limitations. Specifically, T-shaped, Y-shaped, wavy, and hybrid microchannel heat sinks have successfully achieved a reduction in pressure drop concurrently with a decrease in thermal resistance.[164] The novel serrated jet microchannels (sJMCs) achieve an ultrahigh-heat flux of 2126 W cm^{-2} under low pressure drop (< 40 kPa),[165,166] with a performance coefficient (COP) reaching 2×10^4. By leveraging the phase-change characteristics of the coolant to augment heat transfer, the heat transfer coefficient of microchannel liquid cooling can be further enhanced. For instance, through the adoption of a capillary-manifold hybrid design and the optimization of the coolant flow channel to intensify the thin-film evaporation effect,[167] the coefficient of performance can surpass 10^5. Moreover, microchannel cooling technology has witnessed a paradigmatic shift from a sole focus on the cooling performance of isolated chips to an emphasis on the innovation of integrated systems. The collaborative design concept of the microchannel cooling system and chip manufacturing, namely, integrating the cooling structure directly into the chip substrate, has inaugurated a new era of three-dimensional thermal electronics integration.[168] This transition, in turn, has given rise to the third-generation thermal management paradigm, which is characterized by the thermal–electrical–mechanical collaborative design.

3. Immersion and spray cooling

Immersion cooling exhibits remarkable advantages at both the packaging and system levels. This can be primarily attributed to the complete immersion of electronic components in the coolant. The heat generated by these components can be directly absorbed by the liquid and then dissipated to the external environment for cooling, thereby enabling the absorption of 100% of the heat on the chip surface. In comparison with the cold-plate solution, immersion cooling not only enables lower power usage effectiveness (PUE) but also has a less pronounced environmental impact.[169]

Immersion cooling is further classified into single-phase immersion and two-phase immersion cooling. The key distinction between the two lies in whether the coolant undergoes a phase change during the heat absorption process. Single-phase immersion cooling mainly enhances convective heat transfer effectively by increasing the surface area. Concerning the issue of enhancing local heat transfer for single-phase immersion-cooled chips, the current main research focus remains on optimizing the geometric parameters of heat sinks. By optimizing the height, thickness, and number of heat sinks, the thermal resistance and the pressure drop under single-phase immersion conditions can be minimized. Phase-change immersion cooling attains efficient heat exchange via enhanced boiling phenomena. This process commences when the local surface temperature surpasses the saturation point of the coolant. Bubble nucleation is triggered on the surface of the multi-scale structures.[170] These gaseous coolant molecules migrate to the condensing surface, release latent heat through the reverse process, and subsequently recycle to complete the entire heat cycle. In an immersion cooling system, the operating components must be in direct and long-term contact with the working fluid. In light of these two requirements, a comprehensive assessment of the thermophysical properties (thermal conductivity, specific heat capacity, and viscosity) and electrochemical compatibility is essential to guarantee the reliability of the system. Spray liquid cooling pertains to the process in which the liquid is atomized into droplets under pressure as it passes through the nozzle. These droplets continuously impinge upon the chip surface, forming a liquid film thereon. This film continuously removes heat via convection, a process whose heat dissipation efficiency is substantially enhanced when phase change occurs.

Overall, the field of chip-level cooling is undergoing a paradigm shift—from isolated performance enhancements toward integrated system-level innovation. The TDA cooling system adopts advanced liquid cooling technologies—including microchannel cooling, as well as single-phase and two-phase immersion cooling—to achieve efficient and scalable thermal management at the chip level. The practical impact of this system is exemplified through successful industry collaborations, most notably with Lenovo, which have led to the development of breakthrough immersion cooling technologies for next-generation server thermal management. Among these TDA-driven innovations are bio-inspired "Flying Fish" heat sink designs that leverage fluid dynamics principles to simultaneously reduce thermal and flow resistance, resulting in a 20% increase in power handling capacity. Additionally, the TDA framework has enabled the creation of advanced dual-circulation phase-change immersion cooling systems that incorporate multiscale composite surfaces to enhance boiling heat transfer. These systems deliver twice the cooling capacity of conventional solutions, with system PUE (Power Usage Effectiveness) as low as 1.04—demonstrating industry-leading energy efficiency and validating the effectiveness of the holistic TDA design methodology.

III. CONCLUSIONS

This perspective has presented a comprehensive TDA framework that systematically addresses the multiscale thermal management challenges for modern electronics. The framework's strength lies in its integration of simulation and electrothermal co-design techniques spanning from atomic-scale physics to system-level cooling solutions. At the atomic level, first-principles calculations and MLPs provide accurate electron and phonon transport properties, while LD simulations predict interfacial phonon transmission using realistic interface structures. These foundational insights inform phonon MC simulations, which capture non-Fourier heat transport phenomena critical to nanoscale devices. Incorporating self-heating effects through coupled electro-thermal simulations further enables accurate device-level thermal predictions and heat generation reductions. In parallel, CTMs and FEMs bridge the gap to circuit- and die-level thermal analysis, supporting hierarchical thermal optimization. At the chip level, the TDA framework addresses the final stage of heat dissipation through advanced solutions, such as high-performance liquid cooling, ensuring efficient heat extraction to the ambient environment.

As illustrated in Fig. 7, the proposed TDA framework represents a paradigm shift from the conventional EDA-centric research and development workflow to an EDA+TDA codesign methodology that embeds multiscale internal thermal management from the earliest stages of system design. In the traditional EDA workflow [Fig. 7(a)], thermal concerns are primarily addressed at the packaging sign-off stage, where external thermal management techniques—such as heat sinks or TIMs—are employed to dissipate heat. Once these external strategies fail to meet thermal performance requirements, iterations are triggered by failures or performance bottlenecks detected post-fabrication or during packaging, resulting in extended design cycles and suboptimal thermal behavior. Indeed, as internal thermal resistance becomes increasingly critical, this workflow grows progressively inefficient. Although some thermal tools have been integrated into modern EDA environments, they remain fundamentally constrained in capturing the non-Fourier and multiscale nature of heat transport in advanced

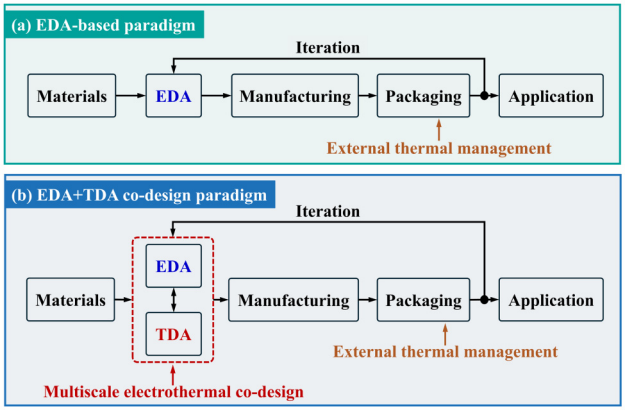

FIG. 7. Comparison of research and development paradigms in thermal-aware electronic design. (a) Traditional EDA-based workflow emphasizes external thermal management at the packaging level. (b) The proposed EDA+TDA codesign paradigm integrates multiscale internal electrothermal co-design into early design stages, enabling iterative optimization between EDA and manufacturing processes for improved thermal performance.

devices. These tools can neither adequately support thermal design at the transistor or interfacial level nor can they account for multi-scale effects across the entire chip.

In contrast, the updated EDA+TDA codesign workflow [Fig. 7(b)] introduces a TDA layer that spans physical scales and is tightly coupled with EDA. This multiscale internal thermal management infrastructure includes interface engineering for reduced TBR, non-Fourier electrothermal simulations for self-heating control, rapid circuit-level thermal optimization for die layout tuning, advanced chip-level cooling strategies, etc. These capabilities enable predictive and scale-aware thermal optimization throughout the design flow—much before manufacturing or packaging. The TDA module facilitates bidirectional iteration with both the materials and manufacturing stages, allowing thermal feedback to shape device architectures, layout strategies, and even fabrication processes. This proactive approach reduces heat generation, lowers thermal resistance across the system, and embeds thermal-awareness into early design decisions. By overcoming the limitations of traditional toolsets and introducing a physically grounded, multiscale-capable methodology, the TDA framework enhances thermal prediction, minimizes late-stage revisions, and accelerates development cycles. Looking ahead, future efforts will focus on expanding realistic electrothermal material databases, leveraging AI and high-performance computing to accelerate multiscale simulations, and establishing an open ecosystem that seamlessly integrates TDA with existing EDA workflows.

ACKNOWLEDGMENTS

This work was supported, in part, by the National Natural Science Foundation of China under Grant Nos. 52425601, 52327809, and 52250273; the National Key Research and Development Program of China under Grant No. 2023YFB4404104; and the Beijing Natural Science Foundation under Grant No. L233022.

AUTHOR DECLARATIONS

Conflict of Interest

The author has no conflicts to disclose.

Author Contributions

Bingyang Cao: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available within the article.

REFERENCES

[1]Z. Wang, R. Dong, R. Ye, S. S. K. Singh, S. Wu, and C. Chen, "A review of thermal performance of 3D stacked chips," Int. J. Heat Mass Transfer **235**, 126212 (2024).
[2]S. S. Salvi and A. Jain, "A review of recent research on heat transfer in three-dimensional integrated circuits (3-D ICs)," IEEE Trans. Compon. Packag. Manuf. Technol. **11**(5), 802–821 (2021).
[3]R. J. Warzoha, A. A. Wilson, B. F. Donovan, N. Donmezer, A. Giri, P. E. Hopkins, S. Choi, D. Pahinkar, J. Shi, S. Graham *et al.*, "Applications and impacts of nanoscale thermal transport in electronics packaging," J. Electron. Packag. **143**(2), 020804 (2021).
[4]S. Tang, J. Chen, Y. B. Hu, C. Yu, H. Lu, S. Zhang, and K. Xiong, "Brief overview of the impact of thermal stress on the reliability of through silicon via: Analysis, characterization, and enhancement," Mater. Sci. Semicond. Process. **183**, 108745 (2024).
[5]W. Zhou, X. Zhong, and K. Sheng, "High temperature stability and the performance degradation of SIC MOSFETs," IEEE Trans. Power Electron. **29**(5), 2329–2337 (2014).
[6]O. Semenov, A. Vassighi, and M. Sachdev, "Impact of self-heating effect on long-term reliability and performance degradation in CMOS circuits," IEEE Trans. Device Mater. Reliab. **6**(1), 17–27 (2006).
[7]Ç. Köroğlu and E. Pop, "High thermal conductivity insulators for thermal management in 3D integrated circuits," IEEE Electron Device Lett. **44**(3), 496–499 (2023).
[8]D. Shoemaker, M. Malakoutian, B. Chatterjee, Y. Song, S. Kim, B. M. Foley, S. Graham, C. D. Nordquist, S. Chowdhury, and S. Choi, "Diamond-incorporated flip-chip integration for thermal management of GaN and ultra-wide bandgap RF power amplifiers," IEEE Trans. Compon. Packag. Manuf. Technol. **11**(8), 1177–1186 (2021).
[9]Y. Qin, B. Albano, J. Spencer, J. S. Lundh, B. Wang, C. Buttay, M. J. Tadjer, C. DiMarino, and Y. Zhang, "Thermal management and packaging of wide and ultra-wide bandgap power devices: A review and perspective," J. Phys. D: Appl. Phys. **56**(9), 093001 (2023).
[10]E. Pop, R. W. Dutton, and K. E. Goodson, "Analytic band Monte Carlo model for electron transport in Si including acoustic and optical phonon dispersion," J. Appl. Phys. **96**(9), 4998–5005 (2004).
[11]K. Raleva, D. Vasileska, S. M. Goodnick, and M. Nedjalkov, "Modeling thermal effects in nanodevices," IEEE Trans. Electron Devices **55**(6), 1306–1316 (2008).
[12]E. Pop, S. Sinha, and K. E. Goodson, "Heat generation and transport in nanometer-scale transistors," Proc. IEEE **94**(8), 1587–1601 (2006).
[13]Y.-C. Hua, H.-L. Li, and B.-Y. Cao, "Thermal spreading resistance in ballistic-diffusive regime for GaN HEMTs," IEEE Trans. Electron Devices **66**(8), 3296–3301 (2019).
[14]S. Sinha, E. Pop, R. W. Dutton, and K. E. Goodson, "Non-equilibrium phonon distributions in sub-100 nm silicon transistors," J. Heat Transfer **128**(7), 638–647 (2006).
[15]E. Pop, "Energy dissipation and transport in nanoscale devices," Nano Res. **3**, 147–169 (2010).
[16]Y. Wang, K. Cheung, A. Oates, and P. Mason, "Ballistic phonon enhanced NBTI," in *2007 IEEE International Reliability Physics Symposium Proceedings. 45th Annual* (IEEE, 2007), pp. 258–263.
[17]R. Yang, G. Chen, M. Laroche, and Y. Taur, "Simulation of nanoscale multi-dimensional transient heat conduction problems using ballistic-diffusive equations and phonon Boltzmann equation," J. Heat Transfer **127**(3), 298–306 (2005).
[18]A. Sarua, H. Ji, K. P. Hilton, D. J. Wallis, M. J. Uren, T. A. M. T. Martin, and M. Kuball, "Thermal boundary resistance between GaN and substrate in AlGaN/GaN electronic devices," IEEE Trans. Electron Devices **54**(12), 3152–3158 (2007).
[19]J. Chen, G. Zhang, and B. Li, "Thermal contact resistance across nanoscale silicon dioxide and silicon interface," J. Appl. Phys. **112**(6), 064319 (2012).

[20]T. Zhan, M. Xu, Z. Cao, C. Zheng, H. Kurita, F. Narita, Y.-J. Wu, Y. Xu, H. Wang, M. Song *et al.*, "Effects of thermal boundary resistance on thermal management of gallium-nitride-based semiconductor devices: A review," Micromachines **14**(11), 2076 (2023).
[21]M. Zhou, L. Li, F. Hou, G. He, and J. Fan, "Thermal modeling of a chiplet-based packaging with a 2.5-D through-silicon via interposer," IEEE Trans. Compon. Packag. Manuf. Technol. **12**(6), 956–963 (2022).
[22]R. Pearson, B. Chatterjee, S. Kim, S. Graham, A. Rattner, and S. Choi, "Guidelines for reduced-order thermal modeling of multifinger GaN HEMTs," J. Electron. Packag. **142**(2), 021012 (2020).
[23]X. Deng, B. Zhang, and Z. Li, "Electro-thermal analytical model and simulation of the self-heating effects in multi-finger 4H-SiC power MESFETs," Semicond. Sci. Technol. **22**(12), 1339 (2007).
[24]X. Ma, Q. Xu, C. Wang, H. Cao, J. Liu, D. Zhang, and Z. Li, "An electrical–thermal co-simulation model of chiplet heterogeneous integration systems," IEEE Trans. Very Large Scale Integr. Syst. **32**(10), 1769–1781 (2024).
[25]Y. Chen, V. SaOnkatali, S. Mishra, J. Ryckaert, J. Myers, and D. Biswas, "Thermal implications in scaling high-performance server 3D chiplet-based 2.5D SoC from FinFET to nanosheet," in *2024 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)* (IEEE, 2024), pp. 45–50.
[26]B. Vermeersch, S. Mishra, M. Brunion, O. Zografos, M. Lofrano, H. Oprins, J. Myers, Z. Tokei, and G. Hellings, "Multiscale thermal impact of BSPDN: SoC hotspot challenges and partial mitigation," in *2024 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2024), pp. 1–4.
[27]T. Takagi, T. Ninomiya, M. Niwa, S. Obara, T. Momose, Y. Shimogaki, M. Nomura, H. Fujioka, M. Mori, and T. Kuroda, "High thermal conductivity AlN films for advanced 3D chiplets," in *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)* (IEEE, 2024), pp. 1–2.
[28]Y.-C. Hua, Y. Shen, Z.-L. Tang, D.-S. Tang, X. Ran, and B.-Y. Cao, "Near-junction thermal managements of electronics," Adv. Heat Transfer **56**, 355–434 (2023).
[29]D.-S. Tang and B.-Y. Cao, "Phonon thermal transport and its tunability in GaN for near-junction thermal management of electronics: A review," Int. J. Heat Mass Transfer **200**, 123497 (2023).
[30]T. Kim, C. Song, S. I. Park, S. H. Lee, B. J. Lee, and J. Cho, "Modeling and analyzing near-junction thermal transport in high-heat-flux GaN devices heterogeneously integrated with diamond," Int. Commun. Heat Mass Transfer **143**, 106682 (2023).
[31]A. J. Gabourie, C. A. Polanco, C. J. McClellan, H. Su, M. Malakoutian, Ç. Köroğlu, S. Chowdhury, D. Donadio, and E. Pop, "AI-accelerated atoms-to-circuits thermal simulation pipeline for integrated circuit design" in *2024 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2024), pp. 1–4.
[32]M. A. Stettler, S. M. Cea, S. Hasan, L. Jiang, P. H. Keys, C. D. Landon, P. Marepalli, D. Pantuso, and C. E. Weber, "Industrial TCAD: Modeling atoms to chips," IEEE Trans. Electron Devices **68**(11), 5350–5357 (2021).
[33]Z. Hassan, N. Allec, L. Shang, R. P. Dick, V. Venkatraman, and R. Yang, "Multiscale thermal analysis for nanometer-scale integrated circuits," IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **28**(6), 860–873 (2009).
[34]H.-L. Li, Y. Shen, Y.-C. Hua, S. L. Sobolev, and B.-Y. Cao, "Hybrid Monte Carlo-diffusion studies of modeling self-heating in ballistic-diffusive regime for gallium nitride HEMTs," J. Electron. Packag. **145**(1), 011203 (2023).
[35]J. Hafner, "*Ab-initio* simulations of materials using VASP: Density-functional theory and beyond," J. Comput. Chem. **29**(13), 2044–2078 (2008).
[36]Y. Liu, H. Liang, L. Yang, G. Yang, H. Yang, S. Song, Z. Mei, G. Csányi, and B. Cao, "Unraveling thermal transport correlated with atomistic structures in amorphous gallium oxide via machine learning combined with experiments," Adv. Mater. **35**(24), 2210873 (2023).
[37]Y.-B. Liu, J.-Y. Yang, G.-M. Xin, L.-H. Liu, G. Csányi, and B.-Y. Cao, "Machine learning interatomic potential developed for molecular simulations on thermal properties of β-Ga_2O_3," J. Chem. Phys. **153**(14), 144501 (2020).
[38]G. Yang, Y.-B. Liu, L. Yang, and B.-Y. Cao, "Machine-learned atomic cluster expansion potentials for fast and quantum-accurate thermal simulations of wurtzite AlN," J. Appl. Phys. **135**(8), 085105 (2024).
[39]L. Guo, Y. Liu, L. Yang, and B. Cao, "Lattice dynamics modeling of thermal transport in solids using machine-learned atomic cluster expansion potentials: A tutorial," J. Appl. Phys. **137**(8), 081101 (2025).
[40]H.-A. Yang and B.-Y. Cao, "Mode-resolved phonon transmittance using lattice dynamics: Robust algorithm and statistical characteristics," J. Appl. Phys. **134**(15), 155302 (2023).
[41]W. Li, J. Carrete, N. A. Katcho, and N. Mingo, "ShengBTE: A solver of the Boltzmann transport equation for phonons," Comput. Phys. Commun. **185**(6), 1747–1758 (2014).
[42]Y. Shen, H.-A. Yang, and B.-Y. Cao, "Near-junction phonon thermal spreading in GaN HEMTs: A comparative study of simulation techniques by full-band phonon Monte Carlo method," Int. J. Heat Mass Transfer **211**, 124284 (2023).
[43]Z.-L. Tang, Y. Shen, and B.-Y. Cao, "Modulating self-heating effects in GaN HEMTs using slant field plate," IEEE Trans. Electron Devices **72**, 1907–1911 (2025).
[44]M. Sabry, "Compact thermal models for electronic systems," IEEE Trans. Compon. Packag. Technol. **26**(1), 179–185 (2003).
[45]M. Yang, M.-T. Li, Y.-C. Hua, W. Wang, and B.-Y. Cao, "Experimental study on single-phase hybrid microchannel cooling using HFE-7100 for liquid-cooled chips," Int. J. Heat Mass Transfer **160**, 120230 (2020).
[46]J. A. Spencer, A. L. Mock, and Y. Zhang, "Heating issues in wide-bandgap semiconductor devices," in *Thermal Management of Gallium Nitride Electronics* (Elsevier, 2022), pp. 1–19.
[47]Y. Shen, X.-S. Chen, Y.-C. Hua, H.-L. Li, L. Wei, and B.-Y. Cao, "Bias dependence of non-Fourier heat spreading in GaN HEMTs," IEEE Trans. Electron Devices **70**(2), 409–417 (2023).
[48]M. Meneghini, C. D. Santi, I. Abid, M. Buffolo, M. Cioni, R. A. Khadar, L. Nela, N. Zagni, A. Chini, F. Medjdoub *et al.*, "GaN-based power devices: Physics, reliability, and perspectives," J. Appl. Phys. **130**(18), 181101 (2021).
[49]S. Selberherr, *Analysis and Simulation of Semiconductor Devices* (Springer Science & Business Media, 1984).
[50]D. B. M. Klaassen, "A unified mobility model for device simulation—I. Model equations and concentration dependence," Solid-State Electron. **35**(7), 953–959 (1992).
[51]R. N. Hall, "Electron-hole recombination in germanium," Phys. Rev. **87**(2), 387 (1952).
[52]W. T. R. W. Shockley and W. T. Read, Jr., "Statistics of the recombinations of holes and electrons," Phys. Rev. **87**(5), 835 (1952).
[53]W. Hänsch, T. Vogelsang, R. Kircher, and M. Orlowski, "Carrier transport near the Si/SiO_2 interface of a MOSFET," Solid-State Electron. **32**(10), 839–849 (1989).
[54]G. K. Wachutka, "Rigorous thermodynamic treatment of heat generation and conduction in semiconductor device modeling," IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **9**(11), 1141–1149 (1990).
[55]H. B. Callen, *Thermodynamics and an Introduction to Thermostatistics* (John Wiley& Sons, 1980), 2.
[56]O. Tornblad, U. Lindefelt, and B. Breitholtz, "Heat generation in Si bipolar power devices: The relative importance of various contributions," Solid-State Electron. **39**(10), 1463–1471 (1996).
[57]Z.-L. Tang and B.-Y. Cao, "Heat generation mechanisms of self-heating effects in SOI-MOS," IEEE J. Electron Devices Soc. **12**, 350–358 (2024).
[58]P. Zhao, S.-H. Zhao, Y.-D. He, and G. Du, "A comparative study of self-heating effects in 3 nm node GAAFETs and FinFETs," in *2022 IEEE 16th International Conference on Solid-State & Integrated Circuit Technology (ICSICT)* (IEEE, 2022), pp. 1–3.
[59]B. Hu, Z. Wang, K. Xu, and D. Tang, "Hotspot and nonequilibrium thermal transport in AlGaN/GaN FinFET: A coupled electron-phonon Monte Carlo simulation study," Int. J. Heat Mass Transfer **241**, 126679 (2025).
[60]Y. Sheng, S. Wang, Y. Hu, J. Xu, Z. Ji, and H. Bao, "Integrating first-principles-based non-Fourier thermal analysis into nanoscale device simulation," IEEE Trans. Electron Devices **71**(3), 1769–1775 (2024).
[61]K. Raleva, D. Vasileska, A. Hossain, S.-K. Yoo, and S. M. Goodnick, "Study of self-heating effects in SOI and conventional MOSFETs with

electro-thermal particle-based device simulator," J. Comput. Electron. **11**, 106–117 (2012).
[62]B. J. Baliga, *Fundamentals of Power Semiconductor Devices* (Springer, 2018).
[63]M. Wang, Y. Lv, H. Zhou, Z. Wen, P. Cui, C. Liu, and Z. Lin, "A hybrid simulation technique to investigate bias-dependent electron transport and self-heating in AlGaN/GaN HFETs," IEEE Trans. Electron Devices **70**(10), 5479–5483 (2023).
[64]Y. Liu, Z. Zhou, and X. Liu, "Investigation of self-heating effect on the void embedded SOI MOSFETs," IEEE J. Electron Devices Soc. **13**, 831–837 (2024).
[65]J. Ma and E. Matioli, "2 kV slanted tri-gate GaN-on-Si Schottky barrier diodes with ultra-low leakage current," Appl. Phys. Lett. **112**(5), 052101 (2018).
[66]C. Dundar, D. Kara, and N. Donmezer, "The effects of gate-connected field plates on hotspot temperatures of AlGaN/GaN HEMTs," IEEE Trans. Electron Devices **67**(1), 57–62 (2020).
[67]N. Shi, K. Wang, B. Zhou, J. Weng, and Z. Cheng, "Optimization AlGaN/GaN HEMT with field plate structures," Micromachines **13**(5), 702 (2022).
[68]S. Bordoloi, A. Ray, and G. Trivedi, "Introspection into reliability aspects in AlGaN/GaN HEMTs with gate geometry modification," IEEE Access **9**, 99828–99841 (2021).
[69]J. Wong, K. Shinohara, A. L. Corrion, D. F. Brown, Z. Carlos, A. Williams, Y. Tang, J. F. Robinson, I. Khalaf, H. Fung *et al.*, "Novel asymmetric slant field plate technology for high-speed low-dynamic R_{on} E/D-mode GaN HEMTs," IEEE Electron Device Lett. **38**(1), 95–98 (2017).
[70]J. Ma and E. Matioli, "Slanted tri-gates for high-voltage GaN power devices," IEEE Electron Device Lett. **38**(9), 1305–1308 (2017).
[71]M. Razavi, Y. S. Muzychka, and S. Kocabiyik, "Review of advances in thermal spreading resistance problems," J. Thermophys. Heat Transfer **30**(4), 863–879 (2016).
[72]K. R. Bagnall, Y. S. Muzychka, and E. N. Wang, "Analytical solution for temperature rise in complex multilayer structures with discrete heat sources," IEEE Trans. Compon. Packag. Manuf. Technol. **4**(5), 817–830 (2014).
[73]C. Song, J. Kim, and J. Cho, "The effect of GaN epilayer thickness on the near-junction thermal resistance of GaN-on-diamond devices," Int. J. Heat Mass Transfer **158**, 119992 (2020).
[74]G. Chen, *Nanoscale Energy Transport and Conversion: A Parallel Treatment of Electrons, Molecules, Phonons, and Photons* (Oxford University Press, 2005).
[75]Y.-C. Hua and B.-Y. Cao, "Ballistic-diffusive heat conduction in multiply-constrained nanostructures," Int. J. Therm. Sci. **101**, 126–132 (2016).
[76]Y.-C. Hua and B.-Y. Cao, "The effective thermal conductivity of ballistic–diffusive heat conduction in nanostructures with internal heat source," Int. J. Heat Mass Transfer **92**, 995–1003 (2016).
[77]H. Rezgui, F. Nasri, A. B. H. Ali, and A. A. Guizani, "Analysis of the ultrafast transient heat transport in sub 7-nm SOI FinFETs technology nodes using phonon hydrodynamic equation," IEEE Trans. Electron Devices **68**(1), 10–16 (2021).
[78]D.-S. Tang, Y.-C. Hua, and B.-Y. Cao, "Thermal wave propagation through nanofilms in ballistic-diffusive regime by Monte Carlo simulations," Int. J. Therm. Sci. **109**, 81–89 (2016).
[79]Q. Hao, H. Zhao, and Y. Xiao, "A hybrid simulation technique for electrothermal studies of two-dimensional GaN-on-SiC high electron mobility transistors," J. Appl. Phys. **121**(20), 204501 (2017).
[80]J. Xu, Y. Hu, and H. Bao, "Quantitative analysis of nonequilibrium phonon transport near a nanoscale hotspot," Phys. Rev. Appl. **19**(1), 014007 (2023).
[81]M. Jin, Y. J. Lee, and S. Y. Kim, "Reduction of local thermal effects in FinFETs with a heat-path design methodology," IEEE Electron Device Lett. **42**(4), 461–464 (2021).
[82]H. Bao, J. Chen, X. Gu, and B. Cao, "A review of simulation methods in micro/nanoscale heat conduction," ES Energy Environ. **1**(39), 16–55 (2018).
[83]J. M. Ziman, *Electrons and Phonons: The Theory of Transport Phenomena in Solids* (Oxford University Press, 2001).
[84]S. Mazumder, "Boltzmann transport equation based modeling of phonon heat conduction: Progress and challenges," Annu. Rev. Heat Transfer **24**, 71–130 (2021).
[85]Y. Guo and M. Wang, "Heat transport in two-dimensional materials by directly solving the phonon Boltzmann equation under callaway's dual relaxation model," Phys. Rev. B **96**(13), 134312 (2017).
[86]C. Zhang and Z. Guo, "Discrete unified gas kinetic scheme for multiscale heat transfer with arbitrary temperature difference," Int. J. Heat Mass Transfer **134**, 1127–1136 (2019).
[87]C. Zhang, S. Chen, Z. Guo, and L. Wu, "A fast synthetic iterative scheme for the stationary phonon Boltzmann transport equation," Int. J. Heat Mass Transfer **174**, 121308 (2021).
[88]Q. Hao, G. Chen, and M.-S. Jeng, "Frequency-dependent Monte Carlo simulations of phonon transport in two-dimensional porous silicon with aligned pores," J. Appl. Phys. **106**(11), 114321 (2009).
[89]J.-P. M. Péraud and N. G. Hadjiconstantinou, "Efficient simulation of multidimensional phonon transport using energy-based variance-reduced Monte Carlo formulations," Phys. Rev. B **84**(20), 205331 (2011).
[90]J.-P. M. Péraud and N. G. Hadjiconstantinou, "An alternative approach to efficient simulation of micro/nanoscale phonon transport," Appl. Phys. Lett. **101**(15), 153114 (2012).
[91]D.-S. Tang and B.-Y. Cao, "Ballistic thermal wave propagation along nanowires modeled using phonon Monte Carlo simulations," Appl. Therm. Eng. **117**, 609–616 (2017).
[92]X. Ran and M. Wang, "A steady-state energy-based Monte Carlo method for phonon transport with arbitrary temperature difference," J. Heat Transfer **144**(8), 082502 (2022).
[93]R. Li, E. Lee, and T. Luo, "Physics-informed neural networks for solving multi-scale mode-resolved phonon Boltzmann transport equation," Mater. Today Phys. **19**, 100429 (2021).
[94]R. Li, J.-X. Wang, E. Lee, and T. Luo, "Physics-informed deep learning for solving phonon Boltzmann transport equation with large temperature non-equilibrium," npj Comput. Mater. **8**(1), 29 (2022).
[95]J. Zhou, R. Li, and T. Luo, "Physics-informed neural networks for solving time-dependent mode-resolved phonon Boltzmann transport equation," npj Comput. Mater. **9**(1), 212 (2023).
[96]R. Li, E. Lee, and T. Luo, "Physics-informed deep learning for solving coupled electron and phonon Boltzmann transport equations," Phys. Rev. Appl. **19**(6), 064049 (2023).
[97]W. Shang, J. Zhou, J. P. Panda, Z. Xu, Y. Liu, P. Du, J.-X. Wang, and T. Luo, "JAX-BTE: A GPU-accelerated differentiable solver for phonon Boltzmann transport equations," npj Comput. Mater. **11**(1), 129 (2025).
[98]B. Vermeersch, R. Rodriguez, A. Sibaja-Hernandez, A. Vais, S. Yadav, B. Parvais, and N. Collaert, "Thermal modelling of GaN & InP RF devices with intrinsic account for nanoscale transport effects," in *2022 International Electron Devices Meeting (IEDM)* (IEEE, 2022), pp. 15–23.
[99]Q. Hao, H. B. Zhao, and Y. Xiao, "Multi-length scale thermal simulations of GaN-on-SiC high electron mobility transistors," in *Multiscale Thermal Transport in Energy Systems* (Nova Science Publishers, 2016).
[100]X. Chang, B. Vermeersch, H. Oprins, M. Lofrano, V. Cherman, S. Park, Z. Tokei, and I. De Wolf, "Thermal modeling and analysis of equivalent thermal properties for advanced BEOL stacks," IEEE Trans. Compon. Packag. Manuf. Technol. **15**, 1708–1716 (2025).
[101]S. Martin-Horcajo, A. Wang, M.-F. Romero, M. J. Tadjer, and F. Calle, "Simple and accurate method to estimate channel temperature and thermal resistance in AlGaN/GaN HEMTs," IEEE Trans. Electron Devices **60**(12), 4105–4111 (2013).
[102]J. W. Pomeroy, M. Bernardoni, D. C. Dumka, D. M. Fanning, and M. Kuball, "Low thermal resistance GaN-on-diamond transistors characterized by three-dimensional Raman thermography mapping," Appl. Phys. Lett. **104**(8), 083513 (2014).
[103]M. J. Tadjer, T. J. Anderson, M. G. Ancona, P. E. Raad, P. Komarov, T. Bai, J. C. Gallagher, A. D. Koehler, M. S. Goorsky, D. A. Francis, K. D. Hobart, and

F. J. Kub, "GaN-on-diamond HEMT technology with TAVG = 176 °C at PDC, max = 56 W/mm measured by transient thermoreflectance imaging," IEEE Electron Device Lett. **40**(6), 881–884 (2019).
[104]Z. Cheng, L. Yates, J. Shi, M. J. Tadjer, K. D. Hobart, and S. Graham, "Thermal conductance across β-Ga_2O_3-diamond van der Waals heterogeneous interfaces," APL Mater. **7**(3), 031118 (2019).
[105]Q. Chen, F. J. Medina, S. Wang, and Q. Hao, "In-plane thermal conductivity measurements of Si thin films under a uniaxial tensile strain," J. Appl. Phys. **133**(3), 035103 (2023).
[106]D. Xu, R. Hanus, Y. Xiao, S. Wang, G. J. Snyder, and Q. Hao, "Thermal boundary resistance correlated with strain energy in individual Si film-wafer twist boundaries," Mater. Today Phys. **6**, 53–59 (2018).
[107]A. Giri and P. E. Hopkins, "A review of experimental and computational advances in thermal boundary conductance and nanoscale thermal transport across solid interfaces," Adv. Funct. Mater. **30**(8), 1903857 (2020).
[108]J. Chen, X. Xu, J. Zhou, and B. Li, "Interfacial thermal resistance: Past, present, and future," Rev. Mod. Phys. **94**(2), 025002 (2022).
[109]E. S. Landry and A. J. H. McGaughey, "Thermal boundary resistance predictions from molecular dynamics simulations and theoretical calculations," Phys. Rev. B **80**(16), 165304 (2009).
[110]W. A. Little, "The transport of heat between dissimilar solids at low temperatures," Can. J. Phys. **37**(3), 334–349 (1959).
[111]E. T. Swartz and R. O. Pohl, "Thermal resistance at interfaces," Appl. Phys. Lett. **51**(26), 2200–2202 (1987).
[112]E. T. Swartz and R. O. Pohl, "Thermal boundary resistance," Rev. Mod. Phys. **61**(3), 605–668 (1989).
[113]P. Reddy, K. Castelino, and A. Majumdar, "Diffuse mismatch model of thermal boundary conductance using exact phonon dispersion," Appl. Phys. Lett. **87**(21), 211908 (2005).
[114]A. Rajabpour and S. Volz, "Thermal boundary resistance from mode energy relaxation times: Case study of argon-like crystals by molecular dynamics," J. Appl. Phys. **108**(9), 094324 (2010).
[115]Y. Chalopin, K. Esfarjani, A. Henry, S. Volz, and G. Chen, "Thermal interface conductance in Si/Ge superlattices by equilibrium molecular dynamics," Phys. Rev. B **85**(19), 195302 (2012).
[116]Y. Ni, Y. Chalopin, and S. Volz, "Significant thickness dependence of the thermal resistance between few-layer graphenes," Appl. Phys. Lett. **103**(6), 061906 (2013).
[117]S. Merabia and K. Termentzidis, "Thermal boundary conductance across rough interfaces probed by molecular dynamics," Phys. Rev. B **89**(5), 054309 (2014).
[118]B.-Y. Cao, J.-H. Zou, G.-J. Hu, and G.-X. Cao, "Enhanced thermal transport across multilayer graphene and water by interlayer functionalization," Appl. Phys. Lett. **112**(4), 041603 (2018).
[119]K. Sääskilahti, J. Oksanen, J. Tulkki, and S. Volz, "Role of anharmonic phonon scattering in the spectrally decomposed thermal conductance at planar interfaces," Phys. Rev. B **90**(13), 134312 (2014).
[120]W. Zhang, T. S. Fisher, and N. Mingo, "The atomistic Green's function method: An efficient simulation approach for nanoscale phonon transport," Numer. Heat Transfer, Part B **51**(4), 333–349 (2007).
[121]N. Mingo, "Anharmonic phonon flow through molecular-sized junctions," Phys. Rev. B **74**(12), 125402 (2006).
[122]Z.-Y. Ong and G. Zhang, "Efficient approach for modeling phonon transmission probability in nanoscale interfacial thermal transport," Phys. Rev. B **91**(17), 174302 (2015).
[123]H. Zhao and J. B. Freund, "Lattice-dynamical calculation of phonon scattering at ideal Si–Ge interfaces," J. Appl. Phys. **97**(2), 024903 (2005).
[124]J.-S. Wang, J. Wang, and J. T. Lü, "Quantum thermal transport in nanostructures," Eur. Phys. J. B **62**(4), 381–404 (2008).
[125]V. L. Deringer, M. A. Caro, and G. Csányi, "Machine learning interatomic potentials as emerging tools for materials science," Adv. Mater. **31**(46), 1902765 (2019).
[126]K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C. W. Park, A. Choudhary, A. Agrawal, S. J. L. Billinge, E. Holm, S. P. Ong, and C. Wolverton, "Recent advances and applications of deep learning methods in materials science," npj Comput. Mater. **8**(1), 1–26 (2022).
[127]S. Fujii and A. Seko, "Structure and lattice thermal conductivity of grain boundaries in silicon by using machine learning potential and molecular dynamics," Comput. Mater. Sci. **204**, 111137 (2022).
[128]A. Hashemi, R. Guo, K. Esfarjani, and S. Lee, "*Ab initio* phonon transport across grain boundaries in graphene using machine learning based on small dataset," Phys. Rev. Mater. **6**(4), 044004 (2022).
[129]J. Wu, E. Zhou, A. Huang, H. Zhang, M. Hu, and G. Qin, "Deep-potential enabled multiscale simulation of gallium nitride devices on boron arsenide cooling substrates," Nat. Commun. **15**(1), 2540 (2024).
[130]I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, and G. Csányi, "A foundation model for atomistic materials chemistry," arXiv:2401.00096 (2023).
[131]H. Yang, C. Hu, Y. Zhou, X. Liu, Y. Shi, J. Li, G. Li, and Z. Lu, "Mattersim: A deep learning atomistic model across elements, temperatures and pressures," arXiv:2405.04967 (2024).
[132]K. Banerjee, A. Mehrotra, A. Sangiovanni-Vincentelli, and C. Hu, "On thermal effects in deep sub-micron VLSI interconnects," in *Proceedings of the 36th Annual ACM/IEEE Design Automation Conference* (ACM, New Orleans, LA, 1999), pp. 885–891.
[133]H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," in *Proceedings of the 38th Annual International Symposium on Computer Architecture, ISCA '11* (Association for Computing Machinery, New York, 2011), pp. 365–376.
[134]E. Beyne, "The 3-D interconnect technology landscape," IEEE Des. Test **33**(3), 8–20 (2016).
[135]G. Van der Plas and E. Beyne, "Design and technology solutions for 3D integrated high performance systems," in *2021 Symposium on VLSI Circuits* (IEEE, 2021), pp. 1–2.
[136]W. Y. Woon, S. Vaziri, C. C. Shih, I. Datye, M. Malakoutian, J. Hsu, K. F. Yang, J. R. Huang, T. M. Shen, S. Chowdhury, X. Y. Bao, and S. S. Liao, "Thermal dissipation in stacked devices," in *2023 International Electron Devices Meeting (IEDM)* (IEEE, 2023), pp. 1–4.
[137]S. Pal, A. Mallik, and P. Gupta, "System technology co-optimization for advanced integration," Nat. Rev. Electr. Eng. **1**(9), 1–12 (2024).
[138]J. Meng, K. Kawakami, and A. K. Coskun, "Optimizing energy efficiency of 3-D multicore systems with stacked DRAM under power and thermal constraints," in *DAC Design Automation Conference 2012* (Association for Computing Machinery, 2012), pp. 648–655.
[139]P. Shukla, D. Aguren, T. Burd, A. K. Coskun, and J. Kalamatianos, "Temperature-aware sizing of multi-chip module accelerators for multi-DNN workloads," in *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (IEEE, 2023), pp. 1–6.
[140]A. K. Coskun, J. L. Ayala, D. Atienza, and T. Simunic Rosing, "Modeling and dynamic management of 3D multicore systems with liquid cooling," in *2009 17th IFIP International Conference on Very Large Scale Integration (VLSI-SoC)* (IEEE, Florianopolis, Brazil, 2009), pp. 35–40.
[141]Y. Hua, L. Luo, S. L. Corre, and Y. Fan, "An online learning framework for self-adaptive dynamic thermal modeling of building envelopes," Appl. Therm. Eng. **232**, 121032 (2023).
[142]K. Fukahori and P. R. Gray, "Computer simulation of integrated circuits in the presence of electrothermal interaction," IEEE J. Solid-State Circuits **11**(6), 834–846 (1976).
[143]D. Huang, L. Costero, and D. Atienza, "An evaluation framework for dynamic thermal management strategies in 3D multiProcessor system-on-chip co-design," IEEE Trans. Parallel Distributed Syst. **35**(11), 1–17 (2024).
[144]W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. Stan, "HotSpot: A compact thermal modeling methodology for early-stage VLSI design," IEEE Trans. Very Large Scale Integr. Syst. **14**, 501–513 (2006).
[145]K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, "Temperature-aware microarchitecture," in *Proceedings of the 30th*

Annual International Symposium on Computer Architecture,ISCA '03 (Association for Computing Machinery, New York, 2003), pp. 2–13.
[146]K. Skadron, R. Zhang, and M. Stan, "HotSpot 6.0: Validation, acceleration and extension," Report, University of Virginia, Department of Computer Science, 2015.
[147]A. Sridhar, A. Vincenzi, M. Ruggiero, T. Brunschwiler, and D. Atienza, "3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling," in *2010 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* (IEEE, 2010), pp. 463–470.
[148]C. Wang, Q. Xu, C. Nie, H. Cao, J. Liu, D. Zhang, and Z. Li, "A multiscale anisotropic thermal model of chiplet heterogeneous integration system," IEEE Trans. Very Large Scale Integr. Syst. **32**(1), 178–189 (2024).
[149]F. Xie, R. Chen, and T. Wei, "Thermal mitigation strategy for backside power delivery network," in *2024 IEEE 74th Electronic Components and Technology Conference (ECTC)* (IEEE, 2024), pp. 1485–1492.
[150]L. Wang, F. Xie, J. Liu, T. Liu, L. Peng, Z. Zhang, T. Wei, and R. Chen, "Power and thermal integrity analysis of high performance and low power CPUs at sub-2 nm node designed with various advanced backside PDNs," in *2024 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2024), pp. 1–4.
[151]V. Bjorn, "Multiscale thermal challenges and modelling frameworks for advanced device technologies and systems," in *2024 International Electron Devices Meeting (IEDM)* (IEEE, 2024).
[152]P. Shukla, A. K. Coskun, V. F. Pavlidis, and E. Salman, "An overview of thermal challenges and opportunities for monolithic 3D ICs," in *Proceedings of the 2019 on Great Lakes Symposium on VLSI* (Association for Computing Machinery, New York, 2019), pp. 439–444.
[153]V. Cherman, S. Van Huylenbroeck, M. Lofrano, X. Chang, H. Oprins, M. Gonzalez, G. Van er Plas, G. Beyer, K. J. Rebibis, and E. Beyne, "Thermal, mechanical and reliability assessment of hybrid bonded wafers, bonded at 2.5 μm pitch," in *2020 IEEE 70th Electronic Components and Technology Conference (ECTC)* (IEEE, 2020), pp. 548–553.
[154]Q. Wang, T. Zhu, Y. Lin, R. Wang, and R. Huang, "ATSim3D: Towards accurate thermal simulator for heterogeneous 3D-IC systems considering nonlinear leakage and conductivity," in *2024 2nd International Symposium of Electronics Design Automation (ISEDA)* (IEEE, 2024), pp. 618–623.
[155]W. Ahn, H. Jiang, S. Shin, and M. A. Alam, "A novel synthesis of Rent's rule and effective-media theory predicts FEOL and BEOL reliability of self-heated ICs," in *2016 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2016), pp. 7.1.1–7.1.4.
[156]C.-W. Nan, R. Birringer, D. R. Clarke, and H. Gleiter, "Effective thermal conductivity of particulate composites with interfacial thermal resistance," J. Appl. Phys. **81**(10), 6692–6699 (1997).
[157]N. Stojanovic, D. H. S. Maithripala, J. M. Berg, and M. Holtz, "Thermal conductivity in metallic nanostructures at high temperature: Electrons, phonons, and the Wiedemann-Franz law," Phys. Rev. B **82**(7), 075418 (2010).
[158]X. Chang, H. Oprins, M. Lofrano, B. Vermeersch, I. Ciofi, O. V. Pedreira, Z. Tokei, and I. De Wolf, "Thermal analysis of advanced back-end-of-line structures and the impact of design parameters," in *2022 21st IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm)* (IEEE, 2022), pp. 1–8.
[159]X.-D. Zhang, G. Yang, and B.-Y. Cao, "Bonding-enhanced interfacial thermal transport: Mechanisms, materials, and applications," Adv. Mater. Interfaces **9**(27), 2200078 (2022).
[160]X.-D. Zhang, Z.-T. Zhang, H.-Z. Wang, and B.-Y. Cao, "Thermal interface materials with high thermal conductivity and low Young's modulus using a solid–liquid metal codoping strategy," ACS Appl. Mater. Interfaces **15**(2), 3534–3542 (2023).
[161]X. Zhang, Y. Dong, Y. Du, L. Yang, W. Ma, and B. Cao, "Improving the thermal performance of liquid metal thermal interface materials: The role of intermetallic compounds at the gallium/copper interface," Adv. Mater. Interfaces **12**, 2500041 (2025).
[162]H. Shabgard, M. J. Allen, N. Sharifi, S. P. Benn, A. Faghri, and T. L. Bergman, "Heat pipe heat exchangers and heat sinks: Opportunities, challenges, applications, analysis, and state of the art," Int. J. Heat Mass Transfer **89**, 138–158 (2015).
[163]D. Luo, Z. Wu, L. Jiang, Y. Yan, W.-H. Chen, J. Cao, and B. Cao, "Realizing rapid cooling and latent heat recovery in the thermoelectric-based battery thermal management system at high temperatures," Appl. Energy **370**, 123642 (2024).
[164]Z.-Q. Yu, M.-T. Li, and B.-Y. Cao, "A comprehensive review on microchannel heat sinks for electronics cooling," Int. J. Extreme Manuf. **6**(2), 022005 (2024).
[165]Z. Wu, W. Xiao, and B. Song, "Efficient thermal management of high-power electronics via jet-enhanced HU-type manifold microchannel," Int. J. Heat Mass Transfer **221**, 125113 (2024).
[166]Z. Wu, Z. Jiang, W. Yan, Y. Yang, J. Kang, K. Zheng, W. Bu, W. Wang, and B. Song, "Jet microchannel with sawtooth wall for efficient cooling of high-power electronics," Int. J. Heat Mass Transfer **206**, 123955 (2023).
[167]H. Shi, S. Grall, R. Yanagisawa, L. Jalabert, S. Paul, S. H. Kim, J. L. Viovy, H. Daiguji, and M. Nomura, "Chip cooling with manifold-capillary structures enables 10^5 COP in two-phase systems," Cell Rep. Phys. Sci. **6**(4), 102520 (2025).
[168]R. V. van Erp, R. Soleimanzadeh, L. Nela, G. Kampitsis, and E. Matioli, "Co-designing electronics with microfluidics for more sustainable cooling," Nature **585**(7824), 211–216 (2020).
[169]H. Alissa, T. Nick, A. Raniwala, A. A. Herranz, K. Frost, I. Manousakis, K. Lio, B. Warrier, V. Oruganti, T. J. DiCaprio *et al.*, "Using life cycle assessment to drive innovation for sustainable cool clouds," Nature **641**, 331–338 (2025).
[170]Q. Wang, H. Ren, P. Huang, D.-C. Gao, and Y. Sun, "Multiscale hybrid surface structure modifications for enhanced pool boiling heat transfer: State-of-the-art review," Renew. Sustain. Energy Rev. **208**, 115018 (2025).