

Article

Temperature-Conscious Test Scheduling Strategies for 3D Integrated Circuits

Hsiu-Wei Chen , Sz-Chi Ting, Liang-Ying Su and Shih-Hsu Huang * 

Department of Electronic Engineering, Chung Yuan Christian University, Taoyuan 320314, Taiwan; g11402603@cycu.edu.tw (H.-W.C.); s11226138@cycu.edu.tw (S.-C.T.); g11202601@cycu.edu.tw (L.-Y.S.)

* Correspondence: shhuang@cycu.edu.tw; Tel.: +886-3-2654611

Abstract

Vertical thermal coupling in three-dimensional (3D) stacked integrated circuits causes localized hotspots during memory testing—a critical reliability risk that conventional power-constrained schedulers cannot detect. This paper presents a Zone-Based Thermal-Aware Scheduling (ZBTAS) framework that jointly enforces a global power budget and a novel per-column vertical-occupancy constraint derived from the anisotropic thermal resistance of 3D stacks. A transient resistor–capacitor (RC) thermal model with predictive thermal gating is integrated into the scheduling loop to preemptively block test assignments that would violate the junction-temperature limit. The resulting schedule naturally serializes tests within high-density vertical zones, creating temporal gaps that a second-stage schedule-aware BIST grouping algorithm uses to reduce modeled BIST area. Experiments on 100 synthetic HBM-like benchmark instances show that, compared with a power-only baseline that violates the thermal limit in 88% of cases, ZBTAS achieves 100% thermal-limit compliance while reducing the peak temperature by 6.28 °C on average. The schedule-driven serialization enables an average BIST area reduction of 2.17%. These benefits come at the cost of 13.96% additional test time, with the PTG guard margin chosen to balance thermal headroom and throughput while preserving zero observed thermal violations.

Keywords: 3D integrated circuits; built-in self-test; test scheduling; vertical thermal coupling; memory testing; BIST area optimization

1. Introduction

With the approaching physical limits of Moore’s Law, the semiconductor industry has witnessed a paradigm shift towards three-dimensional integrated circuits (3D ICs) as a key solution to sustain performance scaling. By vertically stacking multiple active dies and interconnecting them using through-silicon vias (TSVs), 3D technologies, particularly in high bandwidth memory (HBM) architectures, offer significant advantages, including reduced interconnect latency, smaller form factors, and higher bandwidth density [1].

However, these structural benefits pose severe challenges to thermal management. The vertical stacking of logic and memory layers drastically increases the power density per unit footprint. Furthermore, inter-layer dielectric materials, bonding layers, microbump arrays, and emerging hybrid bonding interfaces affect vertical heat-transfer paths and can introduce package-dependent thermal boundary resistances that complicate heat removal [2,3]. Recent studies on HBM stacks have shown that non-Fourier heat conduction effects can further complicate thermal prediction in stacked HBM structures [4]. This issue is exacerbated during the manufacturing testing phase. Unlike functional operation, testing



Academic Editor: Nicola Lusardi

Received: 15 June 2026

Revised: 7 July 2026

Accepted: 9 July 2026

Published: 11 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

involves elevated switching activity to maximize fault coverage, leading to rapid heat accumulation. If not properly managed, this “test power” can induce localized hotspots and false yield loss from temperature-induced timing violations. Pak et al. [5] have further shown that electromigration and stress/current-density effects in TSV-based 3D ICs should be considered in reliability analysis.

Traditional test scheduling strategies have relied mainly on power constraints as a proxy for thermal safety [6]. The underlying assumption is that keeping the total power consumption below a global budget (e.g., P_{limit}) is sufficient to maintain the chip temperature within a safe range. However, this assumption is fundamentally flawed in the context of 3D ICs due to the phenomenon of vertical thermal coupling. As heat conducts through vertical silicon stacks, simultaneously activating memory modules at the same planar coordinates (x, y) across different layers creates localized vertical heat-accumulation paths, causing temperature spikes even if the global power budget is respected.

Moreover, the design of built-in self-test (BIST) architectures is typically treated as an independent optimization problem, focusing solely on minimizing silicon area overhead. Existing approaches often optimize BIST grouping and scheduling under fixed or power-aware assumptions [7–9]. This decoupling leads to suboptimal trade-offs: schedulers may enforce excessive guard bands that prolong test time to avoid overheating; conversely, aggressive area optimization may force parallel testing that violates localized thermal limits.

To address these limitations, we present a unified co-optimization framework that simultaneously enforces thermal constraints and reduces BIST area overhead. We introduce a zone-based thermal-aware scheduling (ZBTAS) strategy integrated with a predictive thermal gating (PTG) mechanism. Unlike traditional methods, our approach proactively manages vertical concurrency to prevent the formation of hotspots. Furthermore, we demonstrate that this thermally driven strategy creates temporal opportunities for BIST grouping, thereby producing a modest but measurable BIST area co-benefit. The main contributions of this paper are summarized as follows:

- **Zone-Based Vertical Constraint:** We propose a novel scheduling constraint that explicitly limits the number of active modules within vertical columns (zones). This method effectively mitigates thermal superposition in high-density 3D stacked memories.
- **Predictive Thermal Gating:** Instead of relying on static thermal margins, we implement a dynamic look-ahead mechanism that predicts future temperature states based on a resistor–capacitor (RC) thermal model, allowing precise temperature control near the safety limit.
- **Synergistic Area Co-Benefit:** We quantify how thermally induced serialization affects the BIST area model and add a schedule-aware grouping tie-breaker. In the nominal benchmark set, most of the modeled area benefit comes from the safer schedule itself, while the grouping tie-breaker provides a smaller secondary refinement.

Note that this paper considers only post-bond testing. Therefore, the BIST controller can be designed to control memories between different layers. In our experiments, this assumption is applied to both our method and the prior literature.

Experimental results in 100 HBM-like benchmark instances show that the proposed framework reduces the maximum temperature by 6.28 °C compared to a power-only baseline that violates the thermal limit in 88% of cases, achieving 100% thermal compliance while allowing a reduction in the average BIST area of 2.17%.

2. Related Work

The synthesis of BIST architectures and test scheduling for 3D ICs has evolved significantly to address the complexities of vertical integration. The existing literature can be

broadly categorized into three streams: power-constrained BIST optimization, heuristic-based approaches, and thermal-aware test scheduling.

2.1. Power-Constrained BIST Optimization

The foundational work by Ko and Huang [7] formalized the 3D memory BIST synthesis problem as a two-stage process: memory grouping followed by test scheduling. Adopting the session-less scheduling paradigm proposed by Zadegan et al. [8], they demonstrated that allowing tests to start immediately upon resource availability significantly reduces test application time (TAT) compared to session-based methods. More recently, Huang and Huang [9] identified that treating grouping and scheduling as decoupled stages leads to suboptimal area overhead. They proposed a schedule-aware optimization approach where pre-determined scheduling results are utilized to guide the grouping process, achieving more than 10% reduction in BIST area.

However, a critical limitation persists in this lineage of research: the reliance on peak power as a proxy for thermal safety. These methods assume that maintaining aggregate power consumption below a global budget ($P_{total} \leq P_{max}$) is sufficient to prevent overheating. This assumption, while valid for 2D ICs, fails in 3D architectures due to vertical thermal coupling. Because thinned substrates severely restrict lateral heat spreading [10], vertically aligning active modules creates localized hotspot columns that can violate junction temperature limits even if the global power budget is respected.

2.2. Heuristic and Evolutionary Approaches

To handle the NP-hard nature of 3D scheduling, metaheuristic approaches have been widely adopted. Kaibartta et al. [11] proposed a genetic algorithm (GA) to co-optimize test time and hardware overhead (test data registers). Similarly, Chaudhuri et al. [12] utilized graph-theoretic coloring to optimize BIST grouping for inter-layer vias (ILVs) in high-density monolithic 3D (M3D) ICs.

Although these advanced heuristics effectively explore the solution space for their respective targets, applying such time-minimization engines to core or memory scheduling fundamentally inherits the “Power Proxy” flaw. In fact, optimization algorithms such as GA can inadvertently exacerbate thermal risks. By aggressively packing tests to minimize time, the optimizer can select a schedule that stacks high-power modules vertically, leading to a configuration that is mathematically power-safe but physically thermally hazardous. None of the aforementioned works integrates a transient thermal simulation loop or spatial constraints into their optimization engine.

2.3. Thermal-Aware Test Scheduling

Prior work has directly incorporated thermal models into the scheduling loop. Millican and Saluja [13] formulated linear-programming approaches for thermal-aware test scheduling of 3D-stacked integrated circuits. However, the computational complexity of exact optimization grows rapidly with the number of modules and constraints, making it impractical for modern high-density stacks with hundreds of test units. Manna et al. [14] proposed a particle-swarm-optimization-based approach for thermal-aware multifrequency NoC testing, incorporating temperature-related constraints within the search process. More recently, Chatterjee et al. [15] applied machine learning to estimate temperature during frequency-scaled thermal-aware test scheduling for 3D ICs, thereby reducing repeated thermal-simulation overhead.

At the system level, Shen et al. [16] proposed a unified thermal and power management framework for 3D logic-memory systems, demonstrating that decoupled core-memory regulation can lead to suboptimal results. Although this work targets functional-mode reg-

ulation rather than manufacturing test, its core finding, that thermal and power constraints must be jointly managed, strongly supports our co-optimization framework.

Despite their thermal awareness, these thermal-centric methods share a critical blind spot: none of them exploit the scheduling outcome to reduce the BIST hardware area. Their optimization objective is limited to minimizing test time or peak temperature, treating the BIST architecture as a fixed, pre-determined input. In contrast, while recent work has explored schedule-aware grouping under power constraints [9], our ZBTAS framework closes this loop in the thermal domain by demonstrating that a thermally driven schedule inherently creates forced temporal serialization (e.g., to prevent localized hotspots) that can be leveraged by the BIST grouping mechanism to reduce silicon area.

3. Preliminaries and System Model

In this section, we formally define the 3D IC architecture, the thermal constraints, and the BIST area cost model utilized in our co-optimization framework.

3.1. 3D Stacked Architecture and Coordinate System

We consider a core-based 3D SoC comprising L vertically stacked layers. The floor plan of each layer is discretized into a grid of tiles $W \times H$. Let $\mathcal{M} = \{m_1, m_2, \dots, m_N\}$ denote the set of N memory modules to be tested. The physical location of each module m_i is defined by a coordinate tuple (l_i, x_i, y_i) , where $l_i \in \{1, \dots, L\}$ represents the layer index, and (x_i, y_i) denotes the planar grid coordinates. Each module m_i is characterized by its test power consumption p_i and test application time τ_i . For each module m_i , given a scheduled start time S_i , the test end time is $E_i = S_i + \tau_i$.

To address the critical issue of vertical thermal coupling, we define a vertical zone $Z_{x,y}$ as the set of all vertically aligned tiles at the same planar coordinate (x, y) across all layers:

$$Z_{x,y} = \{(l, x, y) \mid l \in \{1, \dots, L\}\} \quad (1)$$

Modules located within the same $Z_{x,y}$ are physically stacked directly on top of each other. Because thinned silicon substrates severely restrict lateral heat spread [10], vertically aligning active modules within these zones leads to rapid localized heat accumulation, a severe form of inter-tier thermal coupling well-documented in the literature [17].

Figure 1 displays the physical stack schematic used by our approach. As shown in Figure 1, the evaluated architecture comprises four vertically aligned memory layers, a shared heat-removal boundary, TSV/microbump heat paths, and BIST control access to memory modules. The highlighted vertical column illustrates the localized hotspot mechanism that motivates the vertical-zone constraint.

3.2. Discrete Thermal Modeling and Equivalent Circuit

The transient thermal behavior of the 3D IC is modeled using a compact resistive-capacitive (RC) network. This methodology, widely adopted in architecture-level thermal simulators such as HotSpot [18] and 3D-ICE [19], provides computational efficiency while maintaining sufficient fidelity for architecture-level exploration based on electro-thermal duality.

Figure 2 displays the equivalent RC thermal network for the discretized 3D stack. Each tile is modeled as a thermal node with injected test power $P_{test}(t)$, effective compact thermal capacitance C_{eff} , lateral resistance R_{lat} , vertical resistance R_{vert} , and a sink/ambient boundary represented by R_{sink} and T_{amb} .

Figures 1 and 2 map the physical stack to the scheduler variables: (x, y) defines a vertical zone, l selects the RC-node layer, active tests inject $P_{test}(t)$, and heat flow is

represented by R_{vert} , R_{lat} , R_{sink} , and T_{amb} . This compact mapping supports repeated PTG look-ahead while retaining the dominant hotspot paths.

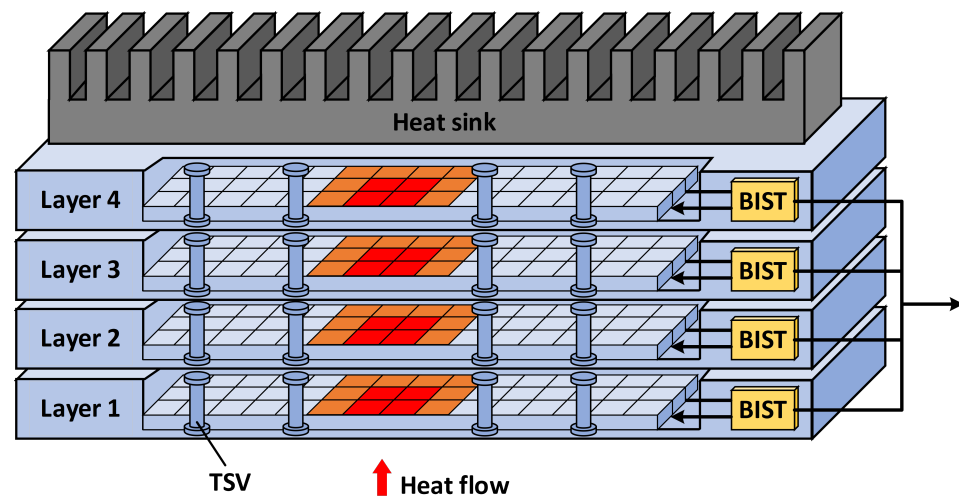


Figure 1. Physical stack schematic used by the ZBTAS approach.

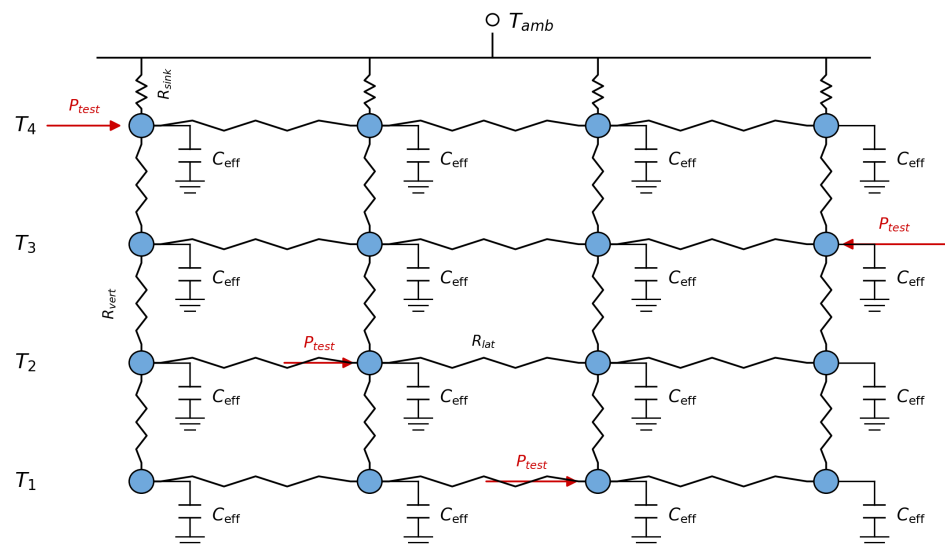


Figure 2. Equivalent RC thermal network for the discretized 3D stack.

3.2.1. Vertical Thermal Resistance Network

The vertical heat transfer path is the most critical component in our analysis. For a given tile at location (x, y) , the total vertical thermal resistance R_{vert} between layer l and $l + 1$ is modeled as:

$$R_{vert} = R_{Si} + R_{interface}^{eff} \quad (2)$$

where R_{Si} represents the resistance of the thinned bulk silicon, and $R_{interface}^{eff}$ represents the thermal interface material and the microbump/inter-tier interface. TSVs, microbumps, and hybrid bonding structures can materially affect inter-layer modeling in HBM-class stacks [20,21]. Reported measurements of underfilled microbump interfaces give area-normalized unit thermal resistances of 8.0–19.0 °C mm²/W for representative pitch values [21]. When normalized to the 4 mm² compact tile used in this paper, the lower end corresponds to approximately 2.0 K/W, i.e., 8.0/4. Therefore, the nominal value $R_{vert} = 2.0$ K/W is used as a compact lower-bound inter-layer scenario parameter rather than as a package-universal measured constant; it is also swept in Section 5.6.

3.2.2. Lateral Thermal Spreading

Lateral heat transfer occurs through a combination of silicon, BEOL/metallization, redistribution structures, and package-level spreading paths rather than through a single homogeneous silicon slab. Although silicon is inherently a good thermal conductor, die thinning and low-conductivity dielectric/interface layers can strongly restrict effective lateral spreading in 3D stacks [10]. HBM-oriented finite-element studies also report a high contrast between silicon and bonding materials, for example silicon at 148 W/mK and non-conductive film (NCF) at 0.2 W/mK, and show that heat can be trapped as DRAM stack height increases [22]. Accordingly, $R_{lat} = 10.0$ K/W is treated here as an effective lateral compact resistance that includes residual spreading through silicon, BEOL/metallization, redistribution structures, and package-level paths. It is not intended to represent a pure thinned-silicon lateral resistance, and it is swept in Section 5.6.

3.2.3. Transient Temperature Update

The transient temperature evolution of a tile at (l, x, y) is governed by the discrete heat conduction equation, solved via the explicit Euler forward method. Let $P_{in}(t)$ denote the active test power injected into the tile at time t , and $P_{out}(t)$ the heat dissipated to adjacent nodes. The update rule is as follows:

$$T_{l,x,y}(t + \Delta t) = T_{l,x,y}(t) + \frac{\Delta t}{C_{eff}} [P_{in}(t) - P_{out}(t)] \quad (3)$$

where C_{eff} is the effective compact thermal capacitance used by the scheduler-level model. In material-level compact RC models, a physical thermal-cell capacitance is derived from the volumetric heat capacity and cell volume, as in HotSpot- and 3D-ICE-style formulations [18,19]; silicon density and heat-capacity values are material properties and have also been tabulated or simulated in the thermophysical literature [23]. However, the value used here is not the silicon-only heat capacity of a thinned die tile. For example, with the present $A_{tile} = 4$ mm² and a 50–100 µm silicon thickness, a silicon-only tile capacitance would be on the order of 3.5×10^{-4} – 7.0×10^{-4} J/K. Therefore, $C_{eff} = 0.5$ J/K is interpreted as an effective transient parameter that lumps the thermal inertia of the compact tile together with BEOL/metallization, bonding/underfill, package/test-fixture, and controlled boundary effects. The same C_{eff} is used for all compared schedulers and is swept in Section 5.6. The dissipation term $P_{out}(t)$ is expanded as:

$$P_{out}(t) = \sum_{k \in \mathcal{N}_{lat}} \frac{T_{l,x,y}(t) - T_k(t)}{R_{lat}} + \sum_{k \in \mathcal{N}_{vert}} \frac{T_{l,x,y}(t) - T_k(t)}{R_{vert}} + \mathbb{I}(l = L) \frac{T_{l,x,y}(t) - T_{amb}}{R_{sink}}, \quad (4)$$

where $\mathcal{N}_{lat}$ denotes the set of up to four planar neighbors sharing the same layer l at coordinates $\{(l, x \pm 1, y), (l, x, y \pm 1)\}$; $\mathcal{N}_{vert}$ denotes the set of vertically adjacent tiles in layers $l - 1$ and $l + 1$ that share the same coordinates (x, y) ; $T_k(t)$ is the temperature of the neighbor tile k at time t ; and $\mathbb{I}(l = L)$ is the indicator function that activates the boundary term only for the top heat-sink layer, where heat is dissipated to the ambient boundary (T_{amb}) through R_{sink} in the single-sided cooling setup shown in Figure 1. The time step Δt is chosen to satisfy the explicit-Euler stability condition $\Delta t \leq C_{eff} / \sum_k G_k$, where G_k denotes the thermal conductance from the current tile to each neighboring node or boundary

sink. With $R_{lat} = 10.0 \text{ K/W}$, $R_{vert} = 2.0 \text{ K/W}$, $R_{sink} = 80.0 \text{ K/W}$, and $C_{eff} = 0.5 \text{ J/K}$, a theoretical worst-case node bounding all possible adjacencies has

$$\sum_k G_k = 4/R_{lat} + 2/R_{vert} + 1/R_{sink} = 1.4125 \text{ W/K}, \quad (5)$$

giving a conservative explicit-Euler stability bound of $\Delta t \leq C_{eff} / \sum_k G_k = 0.354 \text{ s}$. The simulator uses one scheduling cycle of $T_{cycle} = 1.0 \text{ s}$ divided into $n_{sub} = 5$ thermal sub-steps, i.e., $\Delta t = 0.2 \text{ s}$, which is safely below this stability bound. This discrete update function, denoted as $\mathcal{F}_{therm}$, forms the mathematical basis of our predictive thermal gating.

3.3. BIST Area Cost Model

As established earlier, we focus on post-bond testing. Consequently, we assume that one BIST controller may be designed to manage memories located in different layers. The hardware overhead of the BIST architecture is determined jointly by the grouping strategy and the degree of required test concurrency. Following the BIST area model reported in [7], we model the controller area as a linear function of the required testing parallelism.

Let $\mathcal{G} = \{g_1, g_2, \dots, g_K\}$ be a partition of $\mathcal{M}$, where each group g_k is managed by a BIST controller. In [7], the BIST controller area is expressed as $S \times [1 + \alpha(P - 1)]$, where S is the area of a serial BIST controller, P is the number of memories tested in parallel and α is the penalty factor for the area for each additional parallel testing lane. In this work, we extend this formulation to a schedule-aware grouping model by replacing P with $\rho(g_k)$, the maximum temporal overlap required by group g_k . Therefore, the controller area cost for group g_k is modeled as

$$A(g_k) = S_{base} \times [1 + \alpha \cdot (\rho(g_k) - 1)], \quad (6)$$

where S_{base} denotes the area of a serial BIST controller and α is the area penalty coefficient for each additional parallel test lane.

The temporal overlap $\rho(g_k)$ is defined as the maximum number of modules in g_k that are active simultaneously at any discrete time step $t \in \mathcal{T}$:

$$\rho(g_k) = \max_{t \in \mathcal{T}} |\{m_i \in g_k \mid S_i \leq t < E_i\}|. \quad (7)$$

This formulation explicitly captures the coupling between schedule construction and hardware overhead: if all modules in g_k are strictly serially tested, then $\rho(g_k) = 1$, and the controller area reduces to the minimum value S_{base} .

3.4. Problem Formulation

The objective of this work is to determine a start time S_i for each module and a valid grouping partition $\mathcal{G}$ to minimize the total BIST area, subject to a maximum test-time budget, global power limits, and strict thermal safety constraints. Let $\mathcal{T}$ denote the set of discrete time steps during the test application, $Active(t) = \{m_i \in \mathcal{M} \mid S_i \leq t < E_i\}$ the set of modules under test at time step $t \in \mathcal{T}$, T_{max} the maximum allowable test application time (makespan), P_{limit} the maximum aggregate power allowed by the ATE or probe-card constraint, and T_{safe} the junction-temperature safety threshold. The optimization problem is formulated as:

$$\min_{\{S_i\}, \mathcal{G}} \sum_{g_k \in \mathcal{G}} A(g_k) \quad (8)$$

subject to:

$$\max_{m_i \in \mathcal{M}} (S_i + \tau_i) \leq T_{max} \quad (\text{Test Time Budget}) \quad (9)$$

$$\sum_{m_i \in \text{Active}(t)} p_i \leq P_{limit}, \quad \forall t \in \mathcal{T} \quad (\text{Power Constraint}) \quad (10)$$

$$\max_{l,x,y} T_{l,x,y}(t) \leq T_{safe}, \quad \forall t \in \mathcal{T} \quad (\text{Thermal Safety}) \quad (11)$$

This formulation explicitly links the scheduling decision (which dictates $\rho(g_k)$, the overall makespan, and $T_{l,x,y}$) with the area optimization objective, preventing the trivial but impractical solution of infinite serialization.

4. Proposed Methodology

To solve the formulated problem, we propose a unified framework that integrates thermal-aware scheduling with BIST area optimization. The framework operates in two sequential stages: (1) Zone-Based Thermal-Aware Scheduling (ZBTAS), which generates a thermally safe schedule using a transient predictive mechanism; and (2) Schedule-Aware BIST Grouping, which minimizes hardware area by exploiting the natural temporal serialization induced by ZBTAS. We begin with a motivating example that illustrates why power-only scheduling fails in 3D stacks and how our approach addresses this gap.

4.1. Motivating Example

To clarify the core mechanism of the proposed framework, we present a simplified conceptual example, as visually summarized in Figure 3. Consider a 2-layer 3D IC with four memory modules: M_1, M_2 on Layer 1 and M_3, M_4 on Layer 2. Assume M_1 is vertically aligned with M_3 (Zone A), and M_2 is vertically aligned with M_4 (Zone B). Each module consumes 10W, and the global power limit is 25W.

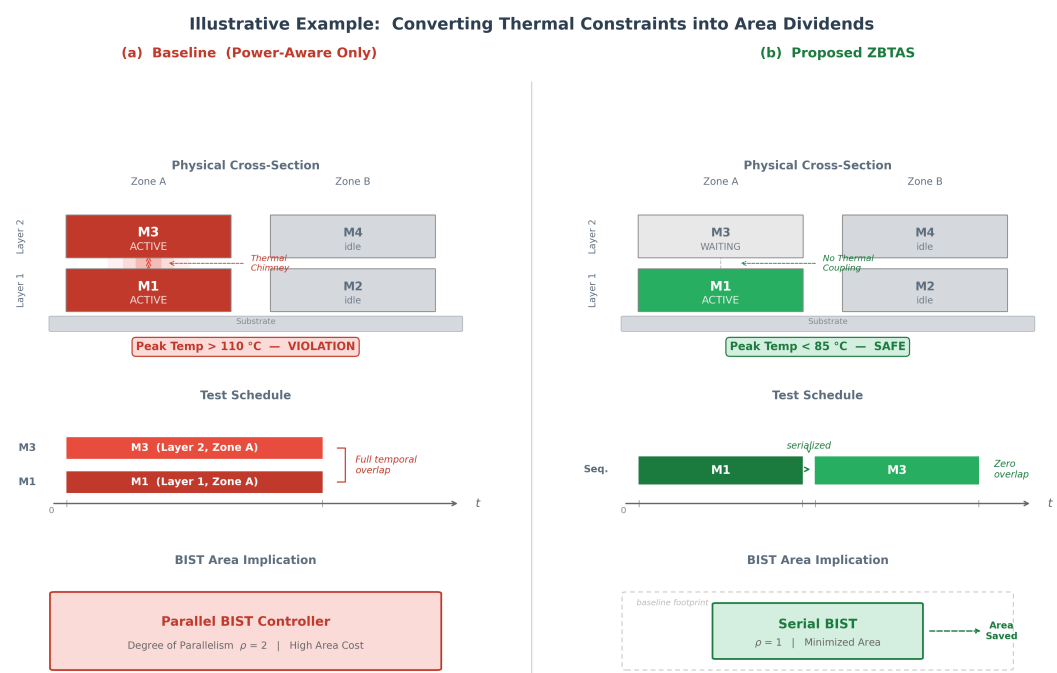


Figure 3. Conceptual illustration of the co-optimization mechanism. (a) The baseline power-aware scheduler inadvertently creates a “thermal chimney,” leading to temperature violations and mandating a costly parallel BIST. (b) The proposed ZBTAS forces temporal serialization, maintaining thermal compliance in the evaluated setting while enabling the use of an area-efficient serial BIST.

4.1.1. Scenario A: Baseline Power-Aware Scheduling

As depicted in Figure 3a, a traditional scheduler evaluates only the global power budget. Since $P(M_1) + P(M_3) = 20W \leq 25W$, it allows M_1 and M_3 to be tested simultaneously.

- **Thermal Outcome:** Because M_1 and M_3 share Vertical Zone A, their active heat superimposes. Local temperature spikes rapidly, creating a vertical hotspot and severely exceeding the safety threshold (e.g., $T_{peak} > 110^\circ\text{C}$).
- **Area Outcome:** The scheduler produces a tightly packed test schedule with full temporal overlap ($M_1 \parallel M_3$). Consequently, the grouping algorithm calculates a degree of parallelism of $\rho = 2$, mandating a larger and more costly parallel BIST controller.

4.1.2. Scenario B: Proposed ZBTAS

In contrast, as illustrated in Figure 3b, the ZBTAS scheduler evaluates the same candidates using the Vertical Zone Constraint. Although the 25W power check passes, it actively detects that M_1 and M_3 share Zone A.

- **Schedule Action:** The scheduler enforces serialization. It initiates M_1 at time t_0 and stalls M_3 until M_1 is complete and the zone cools.
- **Thermal Outcome:** The peak temperature remains safely within the limit ($<85^\circ\text{C}$) because only one heat source in the pillar is active at any given time, eliminating vertical thermal superposition.
- **Area Outcome:** The grouping algorithm receives a sparse serialized schedule where M_1 and M_3 never overlap. It calculates $\rho = 1$, allowing the assignment of a compact serial BIST controller.

This example illustrates how enforcing thermal safety can create temporal separation that is exploitable by the grouping stage for BIST area reduction.

4.2. Stage 1: Zone-Based Thermal-Aware Scheduling (ZBTAS)

The primary objective of Stage 1 is to determine valid start times S_i for all modules without causing thermal violations. Unlike traditional schedulers that rely solely on global power limits, ZBTAS introduces two critical mechanisms tailored for 3D stacked architectures.

4.2.1. Vertical Zone Constraint

To mitigate rapid heat accumulation along vertically aligned heat paths, we impose a strict spatial constraint on vertical concurrency. For any vertical zone $Z_{x,y}$, the scheduler limits the number of concurrently active modules within $Z_{x,y}$ (recall $Active(t) = \{m_i \in \mathcal{M} \mid S_i \leq t < E_i\}$):

$$\sum_{m_i \in Active(t)} \mathbb{I}(m_i \in Z_{x,y}) \leq 1, \quad \forall x, y \quad (12)$$

where $\mathbb{I}(\cdot)$ is the indicator function. In high-density HBM structures, this constraint ensures that no two modules within the same vertical pillar are tested simultaneously, forcing physical separation between active heat sources.

4.2.2. Predictive Thermal Gating (PTG)

Relying solely on instantaneous temperature measurements is prone to overshoot due to the thermal inertia of silicon. A seemingly safe decision made at time t might inevitably lead to a violation at $t + \delta$. To prevent this, we propose the PTG mechanism.

Before scheduling a candidate module m_c at time t , the scheduler performs a transient look-ahead simulation. It estimates the maximum temperature for a short future horizon of $H = 12$ scheduling cycles assuming that m_c is activated:

$$T_{pred} = \max_{1 \leq k \leq H} \max_{l,x,y} \mathcal{F}_{therm}^{(k)}(T(t), \mathbf{P}^{cand}(k))_{l,x,y} \quad (13)$$

where $\mathcal{F}_{therm}^{(k)}$ applies the discrete thermal update for k future scheduling cycles, including the $n_{sub} = 5$ thermal sub-steps within each cycle. The candidate power map $\mathbf{P}^{cand}(k)$ is spatial rather than scalar:

$$\begin{aligned} P_{l,x,y}^{cand}(k) = & \sum_{\substack{m_i \in Active(t) \\ E_i > t + kT_{cycle}}} p_i \mathbb{I}((l_i, x_i, y_i) = (l, x, y)) \\ & + p_c \mathbb{I}(E_c > t + kT_{cycle}) \mathbb{I}((l_c, x_c, y_c) = (l, x, y)). \end{aligned} \quad (14)$$

Thus, active modules are removed from the predicted map after their expected completion time, and the candidate contributes heat only at its physical tile. Unlike a constant-power look-ahead, which assumes that all currently active modules remain on for the entire horizon and therefore over-estimates the future temperature, this decay-aware spatial formulation yields a tighter prediction. The scheduling guard is set to

$$T_{guard} = T_{safe} - \Delta_{safe}, \quad (15)$$

where $\Delta_{safe} = 5^\circ\text{C}$ in the reported experiments, corresponding to an 80°C effective PTG threshold under the 85°C thermal limit. This guard margin compensates for model discretization and transient uncertainty while avoiding unnecessary over-gating; Section 5.7 reports the guard-margin sweep, including the earlier 17°C conservative setting. If $T_{pred} > T_{guard}$, the module m_c is “gated” (rejected) and remains in the waiting queue, even if the current temperature $T(t)$ is well below the thermal limit. In the evaluated HBM-like synthetic benchmark set, this decay-aware PTG mechanism achieved zero thermal-limit violations for ZBTAS while reducing unnecessary gating relative to a constant-power look-ahead. The general flow of ZBTAS is summarized in Algorithm 1.

4.3. Stage 2: Schedule-Aware Grouping Optimization

Once the thermally safe schedule is generated, the second stage optimizes the BIST controller area. The schedule generated in Stage 1 is kept fixed; Stage 2 only changes the assignment of memories to BIST controller groups and does not modify the start or end times of tests. We model this as a greedy clique-partitioning refinement on a compatibility graph $G = (V, E)$, where an edge exists between two modules if their geometric distance satisfies $dist(m_i, m_j) \leq D_{th}$. Thus, the graph encodes the feasibility of physical BIST-sharing, while temporal compatibility is evaluated separately using the fixed Stage-1 schedule.

The second-stage refinement is the Schedule-Aware Sorting Strategy. Unlike traditional heuristics that exclusively prioritize geometric proximity, we incorporate temporal compatibility derived from Stage 1. We enumerate all maximal cliques and order them according to the following criteria:

1. Maximize the size of the clique ($|g_k|$): By selecting larger cliques, we can minimize the total number of BIST controllers.
2. Minimize Parallelism ($\rho(g_k)$): For cliques of the same size, we prioritize those with a lower temporal overlap ρ .
3. Minimize Impact: As a tie-breaker, we select cliques that share fewer vertices with others to preserve optimization freedom.

Algorithm 1 Zone-Based Thermal-Aware Scheduling (ZBTAS)

Input: Set of modules $\mathcal{M}$, Power limit P_{limit} , Thermal limit T_{safe} , guard margin Δ_{safe} , Test time budget T_{max}

Output: Test Schedule $\{S_i, E_i\}$ for all $m_i \in \mathcal{M}$, or FAIL

Initialize thermal map $T \leftarrow T_{amb}$, $Time \leftarrow 0$, $T_{guard} \leftarrow T_{safe} - \Delta_{safe}$;

while $\exists m_i \in \mathcal{M}$ not completed **do**

Update system temperature T using $\mathcal{F}_{therm}$;

Release resources for completed modules;

Identify active zones Z_{active} and current power P_{curr} ;

$Candidates \leftarrow \{m_j \mid m_j \text{ is WAITING}\}$;

// Deadline guard for the makespan constraint

if $\exists m_j \in Candidates$ such that $Time + \tau_j > T_{max}$ **then**

return FAIL;

end

Sort $Candidates$ by Power Descending (Greedy Strategy);

for m_c in $Candidates$ **do**

// Check 1: Global Power Budget

if $P_{curr} + p_c > P_{limit}$ **then**

continue;

end

// Check 2: Vertical Zone Constraint

if $Zone(m_c) \in Z_{active}$ **then**

continue;

end

// Check 3: Predictive Thermal Gating (PTG)

$T_{pred} \leftarrow \text{LookAhead}(T, \text{Active}(Time) \cup \{m_c\}, H)$; // decay-aware

if $T_{pred} > T_{guard}$ **then**

continue;

end

// All checks passed: Schedule Module

$S_c \leftarrow Time$;

$E_c \leftarrow Time + \tau_c$;

$Status(m_c) \leftarrow RUNNING$;

$P_{curr} \leftarrow P_{curr} + p_c$;

$Z_{active} \leftarrow Z_{active} \cup \{Zone(m_c)\}$;

end

$Time \leftarrow Time + 1$;

end

Because ZBTAS tends to avoid concurrent activation in thermally dense zones, the resulting schedule often lowers $\rho(g_k)$ for candidate groups. As a result, the grouping stage can prefer area-efficient serial BIST configurations when they are compatible with the thermally safe schedule.

4.4. Scalability, Portability, and Complexity

The ZBTAS scheduling loop is portable across different stack heights and floorplans because its architecture-dependent inputs are explicit: the layer count L , floorplan dimensions $W \times H$, module placement, test power and duration, RC parameters, cooling boundary, thermal limit T_{safe} , and guard margin Δ_{safe} . When the framework is applied to another 3D

memory stack, these inputs must be recalibrated or validated, while the zone check, PTG look-ahead, and schedule-aware grouping flow remain unchanged.

The scheduler is executed offline as an EDA step before test application. For a generated schedule length of S cycles, a conservative runtime bound for the approach is

$$O(S \cdot N \cdot H_{ptg} \cdot L \cdot W \cdot H), \quad (16)$$

where N is the number of memory modules, H_{ptg} is the PTG look-ahead horizon, and L , W , and H denote the numbers of compact thermal nodes along the three spatial dimensions. This bound assumes that each candidate check may trigger a full-grid thermal look-ahead. The zone and power checks are lighter-weight scans, and implementation-level local-update caching could reduce practical runtime. This offline scheduler runtime is separate from the test application time reported in Section 5.

5. Experimental Results

5.1. Experimental Setup

To validate the effectiveness of the proposed framework, we developed a cycle-accurate thermal simulator based on the discrete RC equivalent-circuit model, adopting the methodology from HotSpot [18]. The simulation targets a high-density 3D stacked memory architecture (e.g., HBM-like), modeled as a 4-layer stack with a 4×4 grid floorplan per layer. The 4-layer stack is used as a compact 4Hi-style baseline, consistent with HBM thermal studies that compare 4Hi, 8Hi, and 12Hi DRAM-layer configurations [22]. Each tile represents a memory bank or a group of banks with an area of approximately 4 mm^2 ; the 4×4 grid is a coarse architecture-level discretization rather than a claim of a specific vendor floorplan.

Physical parameters are treated as compact-model parameters for a constrained wafer-sort environment, as summarized in Tables 1 and 2. The vertical thermal resistance, $R_{vert} = 2.0 \text{ K/W}$, is adopted as a normalized lower-bound compact-model parameter based on measured underfilled microbump interface data reported in [21]. Meanwhile, $R_{lat} = 10.0 \text{ K/W}$ is treated as an effective lateral thermal spreading parameter for the compact model, rather than the physical thermal resistance of thinned silicon. The per-tile sink resistance $R_{sink} = 80.0 \text{ K/W}$ is equivalent to an aggregate 5.0 K/W chip-level single-sided boundary for 16 footprint tiles in parallel. These values define the nominal thermal scenario rather than package-universal measured constants; therefore, Section 5.6 reports a sensitivity sweep of $\pm 25\%$.

The 85°C value is a manufacturing-test scenario bound, not a universal IC limit; it leaves margin for leakage, timing/retention loss, electromigration, and package stress. For example, the thermal analysis of DRAM/HBM presented in [22] adopts 85°C as the DRAM temperature limit. The 15.0 W power limit is likewise used as a constrained benchmark-level ATE/probe-card budget rather than as an industry-wide constant. Because chuck contact, TSV density, material stack, sensor offset, and ATE power profiles vary by product, Section 5.6 sweeps the dominant RC terms and deployment should recalibrate or validate the compact model against the intended package and test fixture.

The proposed framework and baseline algorithms were implemented in Python 3.13. The benchmark set consists of 100 synthetically generated HBM-like instances generated with seeds 0–99. Each nominal instance contains $N = 60$ memory modules with nine vertical “pillar” zones, 3–4 active layers per pillar, pillar module power ranging from 1.1 to 2.0 W , pillar test duration ranging from 50 to 80 cycles, and background modules with $0.3\text{--}0.9 \text{ W}$ power and 15–40 cycle duration. This nominal profile represents a medium density HBM-like workload that captures realistic vertical thermal coupling without artificially

maximizing every pillar. Synthetic benchmarks are adopted because no publicly available 3D memory BIST benchmark suite currently matches the density and layer configuration of modern HBM stacks; creating controlled, parameterized instances allows systematic evaluation across a range of power-map and topology combinations. The parameter choices and aggregate results needed to interpret the study are reported in the manuscript tables, figures, and discussion.

Table 1. Simulation and compact-model parameters.

Parameter	Value	Description
Layer Count (L)	4	Stacked active dies
Grid Size	4×4	Tiles per layer
R_{vert}	2.0 K/W	Effective inter-layer compact resistance (microbump/interface path)
R_{lat}	10.0 K/W	Effective lateral compact resistance (Si/BEOL/package spreading)
R_{sink}	80.0 K/W	Per-tile sink-to-ambient compact boundary
C_{eff}	0.5 J/K	Effective compact thermal capacitance; not silicon-only tile heat capacity
T_{amb}	25.0 °C	Initial and ambient temperature
T_{cycle}	1.0 s	Scheduling-cycle duration
$n_{sub}, \Delta t$	5, 0.2 s	Thermal sub-steps per scheduling cycle
PTG horizon (H)	12 cycles	Look-ahead scheduling horizon
Δ_{safe}	5.0 °C	Selected PTG guard margin (80 °C effective threshold)
D_{th}	3.0	BIST grouping distance threshold
S_{base}, α	0.0089, 0.3	Serial-controller area and parallel-lane penalty
Power Limit	15.0 W	Global ATE/probe-card limit
Thermal Limit	85.0 °C	HBM-like manufacturing-test scenario bound

Table 2. Rationale for simulation and compact-model settings in the evaluated scenario.

Setting Group	Rationale
Vertical compact resistance	Measured underfilled microbump interfaces report 8.0–19.0 °C mm ² /W; normalized by the 4 mm ² tile area, this corresponds to approximately 2.0–4.75 K/W. Therefore, 2.0 K/W is used as a lower-end compact scenario value.
Effective lateral spreading	R_{lat} is treated as an effective lateral spreading resistance that includes silicon, BEOL/metallization, redistribution, and package-level spreading paths. It is not interpreted as a direct pure-silicon geometrical resistance.
Effective transient capacitance	$C_{eff} = 0.5$ J/K is used as an effective transient compact capacitance. A silicon-only 4 mm ² tile with 50–100 µm thickness would be on the order of 10^{-4} – 10^{-3} J/K, so the adopted value is interpreted as a lumped model parameter.
Boundary and thermal limit	$R_{sink} = 80.0$ K/W per tile corresponds to an aggregate 5.0 K/W boundary over 16 footprint tiles. $T_{safe} = 85$ °C is used as a conservative HBM-like manufacturing-test scenario bound, consistent with previous DRAM/HBM thermal discussions that use 85 °C as a DRAM temperature limit.
PTG and grouping settings	$H = 12$, $\Delta_{safe} = 5$ °C, D_{th} , S_{base} , and α are algorithm and area-model settings applied consistently to all compared schedulers.

This paper considers only post-bond testing. Accordingly, the BIST controller is assumed to be capable of controlling memories across different layers. In our experiments, this assumption is applied to both our method and previous work.

5.2. Transient Thermal and Scheduling Analysis

We first analyze the transient thermal behavior of a selected nominal HBM-like instance to demonstrate the necessity of the proposed scheduling constraints. The selected instance (seed 4) contains 60 memory modules distributed across 4 layers on a 4×4 grid, with 9 vertical pillar zones (3–4 active layers per pillar) and module powers ranging from 0.3 to 2.0 W. Figure 4 compares the thermal profiles of the Baseline (Ko’s power-aware scheduler [7]) and the Proposed (ZBTAS) methods.

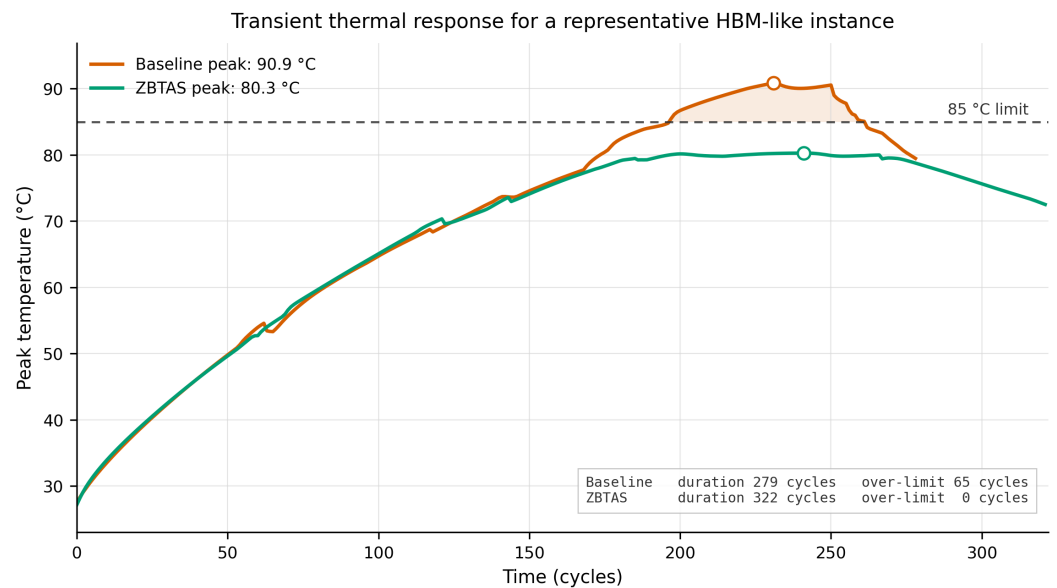


Figure 4. Transient thermal profile comparison for a selected nominal HBM-like instance. The baseline curve (orange) exceeds the thermal limit, peaking at 90.9 °C. The ZBTAS curve (green) throttles test activity via Predictive Thermal Gating (PTG), keeping this illustrative trace at 80.3 °C, below the 85 °C limit.

As observed in Figure 4, the Baseline approach violates the thermal limit, reaching a peak of 90.9 °C. This substantiates our claim that global power constraints are insufficient to prevent localized hotspots in 3D structures. In contrast, the ZBTAS method exhibits a controlled thermal profile. The PTG mechanism predicts potential violations and inserts the necessary cooling intervals, maintaining the peak temperature at 80.3 °C, below the 85 °C safety threshold.

To understand the root cause of this thermal disparity, we visualize the internal scheduling behavior using power-block Gantt charts. Figure 5 illustrates the baseline approach, while Figure 6 demonstrates the proposed method.

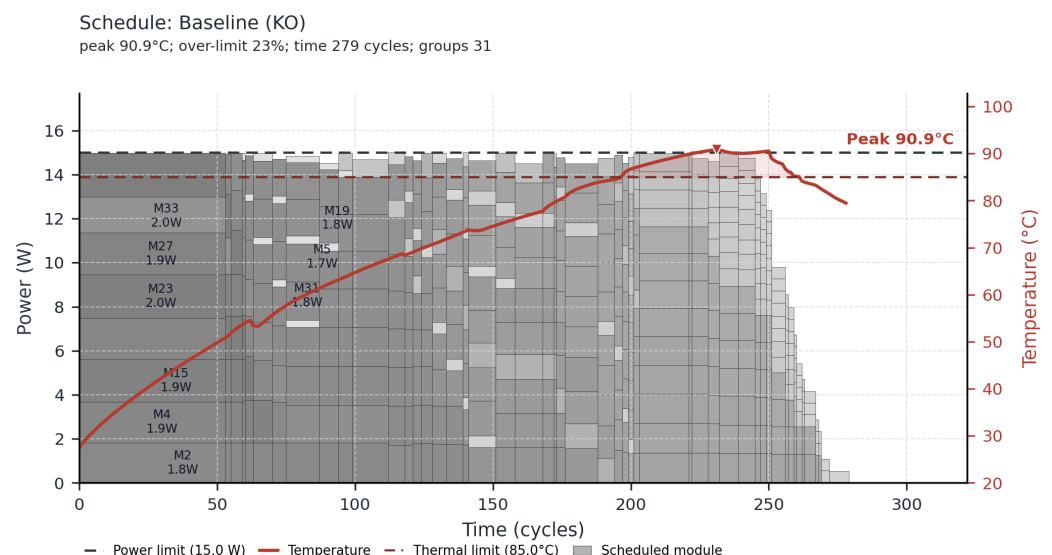


Figure 5. Scheduling execution of the baseline power-aware scheduler. It greedily stacks power blocks, quickly reaching the 15.0 W power limit but causing severe thermal overlap and runaway.

The Baseline scheduler (Figure 5) operates greedily, stacking as many test tasks as possible until the power limit of 15.0 W is reached. Consequently, modules located within the same vertical pillar are tested concurrently, creating severe vertical thermal accumulation.

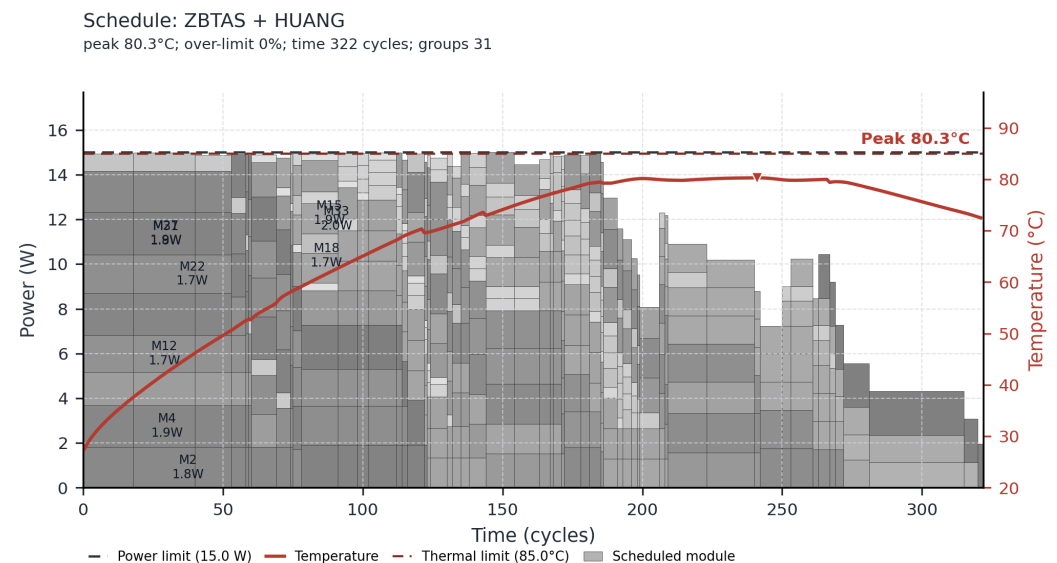


Figure 6. Scheduling execution of the proposed ZBTAS. It forces temporal separation of vertically aligned modules, resulting in a thermally safe but elongated schedule.

In contrast, ZBTAS (Figure 6) proactively serializes modules sharing the same vertical coordinates, resulting in a sparser power distribution. This spatial–temporal separation directly neutralizes the threat of thermal runaway, albeit at the cost of extending the overall test duration.

5.3. Statistical Analysis of Area and Thermal Compliance

To evaluate the batch behavior of the framework, we conducted an experiment in 100 random HBM-like instances under the nominal thermal parameters. Figure 7 presents the trade-off analysis between Peak Temperature and Total BIST Area.

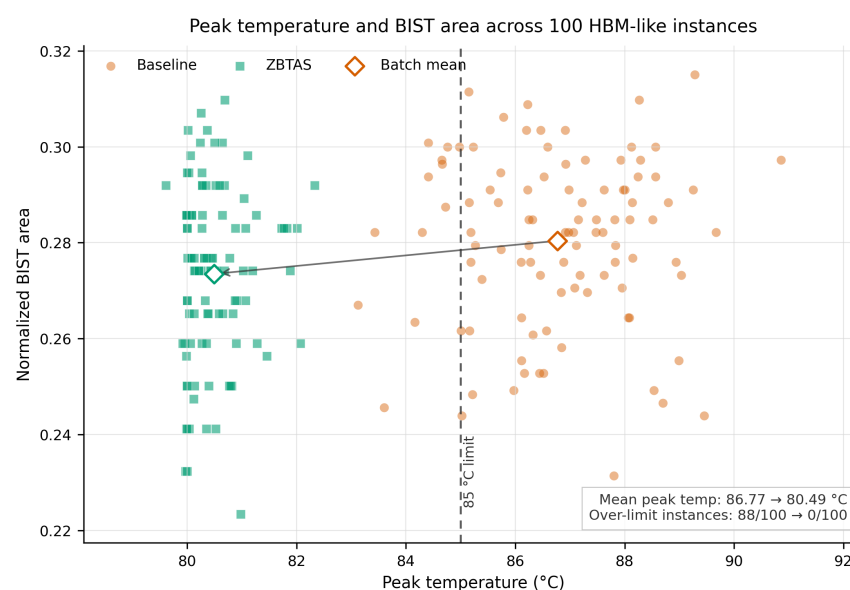


Figure 7. Statistical trade-off analysis ($n = 100$). The proposed method (green squares) consistently achieves 100% thermal safety while simultaneously reducing the BIST area footprint compared with the baseline (orange circles).

The scatter plot reveals a clear separation between the two methods. The Baseline violates the thermal limit in 88 out of 100 instances, confirming that power-only scheduling is insufficient for the large majority (88%) of evaluated HBM-like configurations. In contrast, all ZBTAS results are confined within the safe region, achieving 100% thermal compliance throughout the nominal batch.

In addition to thermal compliance, the proposed method achieves an average reduction in BIST area of 2.17%. This architectural dividend comes directly from the serialization effect of ZBTAS. By forcing tests to be temporally disjoint under the thermal constraint, ZBTAS inherently reduces the parallelism requirement (ρ) of the subsequent grouping stage, allowing the algorithm to assign more memory modules to area-efficient serial BIST controllers. Table 3 summarizes the quantitative metrics.

Table 3. Summary of experimental results (average of 100 nominal HBM-like instances).

Metric	Baseline	Proposed	Improvement
Peak Temperature (°C)	86.8	80.5	−6.3 °C
Thermal Violation Rate	88%	0%	−88%
Normalized BIST Area	0.2795	0.2735	−2.17%
Average Test Time (cycles)	273	311	+13.96%

All comparisons use matched synthetic instances with identical seeds for the baseline and proposed methods. Table 4 reports confidence intervals per-instance results, while the paired Δ column is computed from matched per-seed differences.

Table 4. Statistical reporting for the main nominal comparison ($n = 100$ matched instances; cells report mean and 95% CI).

Metric	Baseline	ZBTAS	Paired Δ
Peak Temp. (°C)	86.77 [86.47, 87.07]	80.49 [80.38, 80.60]	−6.28 [−6.58, −5.98]
Norm. BIST Area	0.2795 [0.2761, 0.2829]	0.2735 [0.2699, 0.2770]	−0.0061 [−0.0067, −0.0054]
Test Time (cycles)	272.7 [270.1, 275.3]	310.8 [306.4, 315.1]	+38.1 [+34.1, +42.1]

5.4. Sensitivity to Workload Intensity

The test-time overhead introduced by ZBTAS depends on the thermal severity of the workload. To characterize this relationship, we additionally evaluate a high-density stress benchmark (12 fully occupied vertical pillars, module powers up to 2.5 W) using the same scheduler and thermal parameters. Table 5 compares the two profiles.

Table 5. Impact of workload intensity on scheduling overhead (nominal: $n = 100$; dense stress check: $n = 30$; violations shown as baseline/ZBTAS).

Profile	Baseline Peak (°C)	ZBTAS Peak (°C)	Δ Temp (°C)	Δ Area	Time Overhead	Violations
Nominal	86.8	80.5	−6.3	−2.17%	+13.96%	88/100/0/100
Dense	104.3	79.3	−25.0	−1.54%	+46.01%	30/30/0/ 30

The dense profile, which maximizes vertical thermal coupling, yields 100% baseline violations and requires 46.01% additional test time for thermal compliance. In contrast, the nominal profile requires only 13.96% overhead while still eliminating 88% baseline violations. This confirms that the overhead scales with the thermal severity of each individual instance and workload profile rather than being a fixed penalty.

5.5. Ablation Study

To quantify the individual contribution of each proposed mechanism, we perform an ablation study in which individual mechanisms are selectively disabled, with the full configuration used as the reference. We evaluate five variants: (1) Power-Only, the power-constrained scheduler of Ko [7]; (2) Zone-Only, which adds the vertical zone constraint but disables predictive thermal gating; (3) PTG-Only, which enables predictive thermal gating but removes the zone constraint; (4) ZBTAS without schedule-aware grouping, which uses the full thermal-aware scheduler but retains the baseline grouping order; and (5) Full ZBTAS, the complete proposed framework. Table 6 summarizes the results averaged over 100 nominal HBM-like instances.

Table 6. Ablation study results (average over 100 nominal HBM-like instances; peak in °C, time in cycles).

Variant	Peak	Area	Time	Viol.
Power-Only	86.8	0.2795	273	88/100
Zone-Only	83.1	0.2732	324	13/100
PTG-Only	71.3	0.2761	395	0/100
ZBTAS w/o sched.-aware grouping	80.5	0.2737	311	0/100
Full ZBTAS	80.5	0.2735	311	0/100

The results reveal complementary roles of the two scheduling mechanisms. Zone restriction alone reduces the peak temperature by 3.7 °C and reduces the thermal violation rate from 88% to 13%, confirming that vertical occupancy is a major source of overheating. However, zone-only scheduling cannot fully prevent violations because laterally adjacent active zones still accumulate heat through thermal spreading. Predictive thermal gating alone eliminates all thermal violations, confirming that the transient look-ahead is the primary safety mechanism. Adding the zone constraint on top of PTG preserves 0/100 violations, reduces test time from 395 to 311 cycles, and lowers the normalized BIST area from 0.2761 to 0.2735. The comparison between the last two rows shows that, in the nominal benchmark set, the schedule-aware grouping contributes a small additional area reduction (0.2737 to 0.2735) after the thermal schedule has already created most of the serialization benefit.

5.6. Thermal Compact-Parameter Sensitivity

To test whether the thermal-compliance conclusion depends on a single compact RC setting, we sweep each nominal thermal compact parameter by $\pm 25\%$ while holding the others fixed. Each point is evaluated over 30 matched nominal HBM-like instances. Table 7 shows that ZBTAS maintains zero thermal-limit violations for all points tested, although the required cooling overhead increases when heat removal becomes more difficult, especially at $R_{\text{sink}} = 100 \text{ K/W}$.

Table 7. Thermal parameter sensitivity on nominal HBM-like instances (30 seeds per point).

Parameter	Value	Baseline Peak (°C)	ZBTAS Peak (°C)	Time Overhead (%)	ZBTAS Viol.
R_{vert}	1.5	86.9	80.3	13.6	0/30
R_{vert}	2.0 *	87.5	80.5	15.5	0/30
R_{vert}	2.5	88.0	80.6	15.5	0/30
R_{lat}	7.5	86.6	80.3	13.6	0/30
R_{lat}	10.0 *	87.5	80.5	15.5	0/30
R_{lat}	12.5	88.2	80.7	15.5	0/30
R_{sink}	60.0	78.5	75.9	5.2	0/30
R_{sink}	80.0 *	87.5	80.5	15.5	0/30
R_{sink}	100.0	94.3	81.3	31.6	0/30
C_{eff}	0.375	94.4	80.7	23.1	0/30
C_{eff}	0.5 *	87.5	80.5	15.5	0/30
C_{eff}	0.625	81.5	79.2	6.7	0/30

* Nominal value used in the main experiment.

5.7. PTG Margin and Test-Time Trade-Off

The reported configuration uses $T_{safe} = 85\text{ }^{\circ}\text{C}$ and $\Delta_{safe} = 5\text{ }^{\circ}\text{C}$, so PTG rejects candidates above an $80\text{ }^{\circ}\text{C}$ effective threshold. To address the temperature/throughput trade-off, we sweep only Δ_{safe} on the same 100 nominal HBM-like instances; Table 8 reports the diagnostic results.

Table 8. PTG guard-margin sweep on nominal HBM-like instances ($n = 100$).

Effective PTG Threshold	Δ_{safe} ($^{\circ}\text{C}$)	ZBTAS Peak ($^{\circ}\text{C}$)	Time Overhead	Area Reduction	Viol.
68 $^{\circ}\text{C}$	17	70.00	41.08%	2.62%	0/100
70 $^{\circ}\text{C}$	15	71.66	37.38%	2.45%	0/100
75 $^{\circ}\text{C}$	10	76.53	24.29%	2.24%	0/100
80 $^{\circ}\text{C}$ *	5	80.49	13.96%	2.17%	0/100
82 $^{\circ}\text{C}$	3	82.08	11.29%	2.14%	0/100
85 $^{\circ}\text{C}$	0	83.91	6.38%	2.10%	23/100

* Default setting used for the reported configuration.

Relaxing the effective threshold reduces test application time but eventually loses full thermal compliance. We select the $80\text{ }^{\circ}\text{C}$ row because it lowers overhead from 41.08% to 13.96% while keeping 0/100 violations and a $5\text{ }^{\circ}\text{C}$ guard; the $85\text{ }^{\circ}\text{C}$ no-margin row has 23/100 violations. The overhead column measures test application time, not offline scheduler runtime.

6. Discussion

6.1. The Cost of Thermal Compliance: Re-Evaluating Test Time

Our experimental results indicate a test time overhead of 13.96% compared to the baseline at the selected $80\text{ }^{\circ}\text{C}$ effective PTG threshold. Although minimizing test application time is a traditional objective in design-for-test (DfT), we argue that this metric cannot be optimized in a vacuum, ignoring physical safety boundaries.

The baseline's shorter duration is not acceptable under the stated thermal criterion in 88% of the evaluated instances because it exceeds the $85\text{ }^{\circ}\text{C}$ threshold. ZBTAS therefore adds the cooling time needed to safely remove heat under the evaluated compact model. The guard-margin sweep shows that this cost is tunable: $68\text{ }^{\circ}\text{C}$ gives 41.08% overhead, while the selected $80\text{ }^{\circ}\text{C}$ threshold keeps 0/100 violations with 13.96% overhead. This is compliant scheduling, not a claim of global time optimality.

6.2. Comparison with Prior Work

Table 9 compares mechanisms and limitations across representative scheduling and BIST-optimization work, including Chou et al. [6], Ko and Huang [7], Huang and Huang [9], Millican and Saluja [13], Manna et al. [14], Chatterjee et al. [15], and our approach. The studies use different architectures, benchmarks, thermal models, and cost definitions, so the table is qualitative rather than a direct numerical ranking. The power-constrained methods are included as non-thermal baselines rather than as thermal-aware schedulers.

The comparison defines the intended scope: ZBTAS does not claim minimum test time. Its contribution is to couple thermally valid scheduling with schedule-aware BIST grouping, a link that prior thermal-aware schedulers and prior grouping methods usually treat separately.

Table 9. Qualitative comparison with representative prior work.

Work	Thermal/BIST Treatment	Distinction from ZBTAS
Chou et al. [6]	Power-constrained test scheduling; power is used as a proxy for test safety.	No 3D transient hotspot model; vertically aligned hotspot columns cannot be detected.
Ko and Huang [7]	3D memory BIST allocation and scheduling under a global power constraint.	Power compliance is treated as sufficient; localized thermal accumulation is not checked.
Huang and Huang [9]	Schedule-aware BIST grouping based on a supplied schedule.	BIST area is reduced, but the method does not generate a thermally safe schedule.
Millican and Saluja [13]	LP-based thermal-aware scheduling with explicit thermal constraints.	Targets thermal scheduling; constrained optimization complexity can grow with problem size.
Manna et al. [14]	Metaheuristic thermal test scheduling for NoC-based 3D IC testing.	Thermal serialization is not exploited for BIST area reduction.
Chatterjee et al. [15]	ML-assisted thermal scheduling using temperature estimation to reduce simulation cost.	Requires training data and assumes a fixed BIST architecture.
This work	Zone constraint, predictive thermal gating, and schedule-aware BIST grouping.	Couples thermally compliant scheduling with a 2.17% modeled BIST-area reduction at +13.96% test application time.

6.3. Scalability to Future Architectures

The proposed zone-based constraint is directly motivated by deeper 3D stacks. As the number of stacked layers increases, for example from 4-layer to 8- or 12-layer stacks, the cumulative vertical heat path becomes more important, intensifying the “thermal chimney” effect. In such settings, power-only schedulers become less reliable because vertical coupling is not explicit. The exact guard margin and thermal parameters, however, must be recalibrated for each package and heat-sink assumption, as discussed in Section 4.

The complexity bound in Section 4 also clarifies that scalability concerns mainly affect the offline scheduler runtime, not the reported test application time. The present prototype keeps full-grid thermal updates for traceability; local-neighborhood caching could reduce implementation cost in an optimized tool.

6.4. Hardware Implementation Perspectives

The proposed framework serves as an Electronic Design Automation (EDA) software step performed prior to tape-out. It generates a static BIST configuration file and test schedule that is thermally compliant under the evaluated compact model. Consequently, it does not require the integration of on-chip thermal sensors or dynamic throttling control loops during runtime; PTG is executed entirely offline. The reported 2.17% area reduction is therefore a modeled BIST-controller area co-benefit under the area model in Equation (6). Implementation-specific schedule storage, decode logic, and controller-interface costs should still be accounted for when mapping the static schedule into a concrete DfT architecture.

6.5. Limitations and Future Work

Several limitations of this work should be acknowledged. First, our thermal model adopts a simplified lumped RC network with uniform parameters across all tiles. While this captures the dominant vertical coupling behavior, it does not account for spatially varying thermal conductivity caused by non-uniform TSV density or heterogeneous die compositions. Integration with more detailed simulators such as 3D-ICE [19] could improve fidelity at the cost of runtime.

Second, the benchmark circuits used in this study are synthetically generated to emulate HBM-like structures. Although compact thermal parameters are chosen to represent a constrained HBM-like test environment and are analyzed in Section 5.6, validation on industrial floorplans, high-fidelity thermal simulation, or silicon measurements would further strengthen the conclusions.

Third, the strict zone constraint (≤ 1 active module per vertical column) represents the most conservative policy evaluated in this work. Although thermal compliance was achieved in all reported HBM-like synthetic instances, it may impose unnecessary serializa-

tion in columns containing only low-power modules. A parametric relaxation, allowing up to K concurrent modules per zone when total zone power remains below a threshold per-column, could reduce the test time overhead while maintaining safety. Exploring this adaptive zone policy remains a promising direction for future work.

Finally, the current framework treats BIST grouping as a post-scheduling optimization step. A fully integrated approach that jointly solves scheduling and grouping, potentially through reinforcement learning or constraint programming, could yield globally superior solutions and constitutes a natural extension of this research.

7. Conclusions

This paper presented a co-optimization framework addressing the dual challenges of thermal safety and hardware overhead in 3D IC memory testing. Recognizing that vertical thermal coupling invalidates traditional power-aware scheduling, we introduced the Zone-Based Thermal-Aware Scheduling (ZBTAS) algorithm. Equipped with a predictive thermal gating (PTG) mechanism, ZBTAS prevents the formation of vertical hotspots by intelligently serializing vertically aligned test tasks. Furthermore, we demonstrate a schedule-aware BIST grouping strategy that capitalizes on this thermally induced temporal sparsity.

Experimental results in 100 HBM-like benchmark instances validate that the proposed framework achieves full thermal compliance in all tested cases, eliminating thermal violations observed in 88% of the baseline runs and reducing the peak temperature by 6.28 °C on average. Concurrently, the inherent serialization results in a 2.17% average reduction in the modeled BIST controller area, at the cost of 13.96% additional test time. By converting a thermal constraint into an architectural co-benefit under an explicit compact thermal model, this work provides a reproducible scheduling approach for safer testing of next-generation 3D integrated circuits.

Future work will focus on adaptive zone relaxation policies, integration with high-fidelity thermal simulators, and joint scheduling–grouping optimization through reinforcement learning or constraint programming.

Author Contributions: Conceptualization: H.-W.C., S.-C.T. and S.-H.H.; investigation and methodology: H.-W.C. and S.-C.T.; validation and formal analysis: H.-W.C., S.-C.T. and L.-Y.S.; writing—original draft preparation: H.-W.C. and L.-Y.S.; writing—review and editing: H.-W.C. and S.-H.H.; supervision: S.-H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Science and Technology Council, Taiwan, under grant number 112-2221-E-033-050-MY3.

Data Availability Statement: The parameter choices and aggregate results used to support the findings of this study are included in this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kim, J.S.; Oh, C.S.; Lee, H.; Lee, D.; Hwang, H.R.; Hwang, S.; Na, B.; Moon, J.; Kim, J.G.; Park, H.; et al. A 1.2 V 12.8 GB/s 2 Gb Mobile Wide-I/O DRAM with 4×128 I/Os Using TSV-Based Stacking. *IEEE J. Solid-State Circuits* **2012**, *47*, 107–116. <https://doi.org/10.1109/JSSC.2011.2164731>.
2. Lee, S.H.; Kim, S.J.; Lee, J.S.; Rhi, S.H. Thermal Issues Related to Hybrid Bonding of 3D-Stacked High Bandwidth Memory: A Comprehensive Review. *Electronics* **2025**, *14*, 2682. <https://doi.org/10.3390/electronics14132682>.
3. Loi, G.L.; Agrawal, B.; Srivastava, N.; Lin, S.C.; Sherwood, T.; Banerjee, K. A Thermally-Aware Performance Analysis of Vertically Integrated (3-D) Processor-Memory Hierarchy. In Proceedings of the 43rd Annual Design Automation Conference (DAC), San Francisco, CA, USA, 24–28 July 2006; pp. 991–996. <https://doi.org/10.1145/1146909.1147160>.
4. Zhou, Z.; He, Y.; Liu, S.; Yang, L.; Yang, N. Effect of Non-Fourier Heat Transport on Temperature Distribution in High Bandwidth Memory. *IEEE Trans. Electron Devices* **2026**, *73*, 561–568. <https://doi.org/10.1109/TED.2025.3628342>.

5. Pak, J.; Pathak, M.; Lim, S.K.; Pan, D.Z. Modeling of Electromigration in Through-Silicon-Via Based 3D IC. In Proceedings of the 61st IEEE Electronic Components and Technology Conference (ECTC), Lake Buena Vista, FL, USA, 31 May–3 June 2011; pp. 1420–1427. <https://doi.org/10.1109/ECTC.2011.5898698>.
6. Chou, R.M.; Saluja, K.K.; Agrawal, V.D. Scheduling Tests for VLSI Systems Under Power Constraints. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **1997**, *5*, 175–185. <https://doi.org/10.1109/92.585217>.
7. Ko, Y.C.; Huang, S.H. 3D IC Memory BIST Controller Allocation for Test Time Minimization Under Power Constraints. In Proceedings of the 2017 IEEE 26th Asian Test Symposium (ATS), Taipei, Taiwan, 27–30 November 2017; pp. 260–265. <https://doi.org/10.1109/ATS.2017.56>.
8. Zadegan, F.G.; Ingelsson, U.; Asani, G.; Carlsson, G.; Larsson, E. Test Scheduling in an IEEE P1687 Environment with Resource and Power Constraints. In Proceedings of the 2011 IEEE 20th Asian Test Symposium (ATS), New Delhi, India, 20–23 November 2011; pp. 525–531. <https://doi.org/10.1109/ATS.2011.80>.
9. Huang, S.Y.; Huang, S.H. Memory Grouping for the Built-In Self-Test of Three-Dimensional Integrated Circuits. *Electronics* **2024**, *13*, 3759. <https://doi.org/10.3390/electronics13183759>.
10. Melamed, S.; Watanabe, N.; Nemoto, S.; Shimamoto, H.; Kikuchi, K.; Aoyagi, M. Impact of Thinning Stacked Dies on the Thermal Resistance of Bump-Bonded Three-Dimensional Integrated Circuits. *Microelectron. Reliab.* **2016**, *67*, 2–8. <https://doi.org/10.1016/j.microrel.2016.08.015>.
11. Kaibartta, T.; Biswas, G.P.; Pal, A.K.; Das, D.K. A Genetic Algorithm-Based Metaheuristic Approach for Test Cost Optimization of 3D SIC. *IEEE Access* **2021**, *9*, 160987–161002. <https://doi.org/10.1109/ACCESS.2021.3131336>.
12. Chaudhuri, A.; Banerjee, S.; Kim, J.; Lim, S.K.; Chakrabarty, K. Built-In Self-Test of High-Density and Realistic ILV Layouts in Monolithic 3-D ICs. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2023**, *31*, 296–309. <https://doi.org/10.1109/TVLSI.2022.3228850>.
13. Millican, S.K.; Saluja, K.K. Linear Programming Formulations for Thermal-Aware Test Scheduling of 3D-Stacked Integrated Circuits. In Proceedings of the 2012 IEEE 21st Asian Test Symposium (ATS), Niigata, Japan, 19–22 November 2012; pp. 37–42. <https://doi.org/10.1109/ATS.2012.28>.
14. Manna, K.; Sagar, C.S.; Chattopadhyay, S.; Sengupta, I. Thermal-Aware Preemptive Test Scheduling for Network-on-Chip Based 3D ICs. In Proceedings of the 2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), Pittsburgh, PA, USA, 11–13 July 2016; pp. 529–534. <https://doi.org/10.1109/ISVLSI.2016.75>.
15. Chatterjee, S.; Roy, S.K.; Giri, C.; Rahaman, H. Frequency-Scaled Thermal-Aware Test Scheduling for 3D ICs Using Machine Learning Based Temperature Estimation. *Microelectron. J.* **2022**, *128*, 105535. <https://doi.org/10.1016/j.mejo.2022.105535>.
16. Shen, Y.; Schreuders, L.; Pathania, A.; Pimentel, A.D. Thermal Management for 3D-Stacked Systems via Unified Core-Memory Power Regulation. *ACM Trans. Embed. Comput. Syst.* **2023**, *22*, 120:1–120:26. <https://doi.org/10.1145/3608040>.
17. Savidis, I.; Vaisband, B.; Friedman, E.G. Experimental Analysis of Thermal Coupling in 3-D Integrated Circuits. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2015**, *23*, 2077–2089. <https://doi.org/10.1109/TVLSI.2014.2357441>.
18. Skadron, K.; Stan, M.R.; Sankaranarayanan, K.; Huang, W.; Velusamy, S.; Tarjan, D. Temperature-Aware Microarchitecture: Modeling and Implementation. *ACM Trans. Archit. Code Optim.* **2004**, *1*, 94–125. <https://doi.org/10.1145/980152.980157>.
19. Sridhar, A.; Vincenzi, A.; Atienza, D.; Brunschweiler, T. 3D-ICE: A Compact Thermal Model for Early-Stage Design of Liquid-Cooled ICs. *IEEE Trans. Comput.* **2014**, *63*, 2576–2589. <https://doi.org/10.1109/TC.2013.127>.
20. Huang, L.H.; Cheng, Y.Y.; Wu, T.L. Analysis and Optimization of HBM3 PPA for TSV Model With Micro-Bump and Hybrid Bonding. *IEEE Trans. Components Packag. Manuf. Technol.* **2025**, *15*, 22–29. <https://doi.org/10.1109/TCPMT.2024.3483445>.
21. Colgan, E.G.; Andry, P.; Dang, B.; Magerlein, J.H.; Maria, J.; Polastre, R.J.; Wakil, J. Measurement of Microbump Thermal Resistance in 3D Chip Stacks. In Proceedings of the 28th Annual IEEE Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), San Jose, CA, USA, 18–22 March 2012; pp. 1–7. <https://doi.org/10.1109/STHERM.2012.6188818>.
22. Song, J.H.; Yoon, S. Comparative Study of Thermal Dissipation in Increasing DRAM Layers of HBM Using 3D FEA Simulations. *J. Semicond. Technol. Sci.* **2025**, *25*, 142–147. <https://doi.org/10.5573/JSTS.2025.25.2.142>.
23. Endo, R.K.; Fujihara, Y.; Susa, M. Calculation of the Density and Heat Capacity of Silicon by Molecular Dynamics Simulation. *High Temp.–High Press.* **2003**, *35–36*, 505–511. <https://doi.org/10.1068/htjr135>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.