Power Throttling in a 3D Integrated Circuit (IC) Dynamic Thermal Simulation

David Geb
david.geb@ansys.com

Subodh Deodhar
subodh.deodhar@ansys.com

Nitin Netake
nitin.netake@ansys.com

Tejas Jeurkar
tejas.jeurkar@ansys.com

Ansys, Inc.

INTRODUCTION AND BACKGROUND

Dynamic thermal management (DTM) generally allows dynamic manipulation of hardware or software to better handle extreme usage scenarios, thereby enabling improved design with lower size, weight, power, and cost (SWaP-C). DTM has received increased interest among the thermal management community, and a DTM Workshop was hosted at DARPA in September, 2023 [1]. DTM can involve passive approaches such as using phase-change materials (PCM), fluids operating near and in the 2-phase regime, and heat pipes. Alternatively, DTM can involve active approaches employing thermostat control, such as fans or blowers, thermoelectric coolers (TECs), and power throttling with, for example, dynamic voltage and frequency scaling (DVFS). The latter approach, power throttling with DVFS and simulation to support optimal placement of on-chip thermal sensors, has been noted as a key thermal simulation challenge for advanced 3D ICs in the Heterogeneous Integration Roadmap (HIR) [2], and will hence be the focus of this paper.

The switching power dissipated by a chip is strongly dependent on voltage and frequency and can be reduced by scaling down these quantities. This leads to DVFS being an effective power management technique that throttles processor power based on live, dynamic conditions [3-4]. Typically, DVFS can be used to conserve power consumption (e.g., for longer battery life in mobile devices), reduce the heat generation on the chip when temperatures are too high, resulting in a decreased cost of the thermal management solution (such as a smaller heat sink) and improved reliability, or reduce noise by allowing fans to run at lower speeds such as in server applications. Power throttling with DVFS has been applied to applications ranging from low-power mobile processors [5] to high-power server chips [6-7], and from single chips to 3D IC systems [8-9].

The focus of this article is on a DTM simulation approach for the case of processor power throttling in a 3D IC using DVFS. Such modeling allows the simulation inputs, such as chip power, to be scaled dynamically during the simulation run based on real-time-simulation temperature data, such as on-chip temperature sensors [10-12]. In other words, if the temperature at a sensor location within the model reaches the threshold, the defined DVFS control logic lowers the voltage and frequency, throttling the dynamic power inputs to the processor on the fly.

Beyond the challenges of integrating DTM into simulation, in parallel, additional challenges are present in chip-package-system (CPS) thermal modeling. For example, various disparate scales need to be integrated into a thermal model [13]: μm to tens of cm, μs to 100s of seconds. Moreover, power can vary widely and change abruptly during the operation of a chip. For example, in the case of a smartphone, the transient power variation can be because of the boot-up sequence, followed by watching a video, talking on the phone or browsing the web. Therefore, it is important to model the spatial and temporal variations in the hotspots accurately. Additionally, the thermal environment can change during the operation of the chip. For example, fan speed, air temperature, and off-chip thermal loading can vary. Therefore, it is of high importance to include these effects accurately. Challenges to effective simulations of throttling can also stem from the lack of relevant material properties (e.g., time-dependent responses of materials and interfaces) and intrinsic and extrinsic sensor errors. 3D ICs present additional modeling challenges due to, for example, the interchip thermal coupling. Further, simulation compute time, particularly for transient simulations presents further challenges. With advanced chip designs having many on-chip thermal sensors in place, DTM simulation needs to overcome these general modeling challenges in addition to the particular challenges related to DTM simulation implementation. The model ultimately should aid

with design, such as optimizing the placement of the on-chip thermal sensors used for DVFS control.

A number of concerns arise when designing a DTM solution. For example, designing cooling solutions for worst-case scenarios can result in costly overdesign. Further, a DTM solution inherently can have an associated performance penalty. These concerns can be minimized with insights obtained from simulations of the DTM solution, such as power throttling with DVFS, in operation. Simulating the DTM solution's performance under various operating scenarios can help assess its effectiveness. Simulation can further guide improvements to the DTM design. Use of efficient DTM strategies becomes imperative to balance between thermal concerns and performance.

IMPLEMENTING POWER THROTTLING IN A THERMAL MODEL

To apply DTM simulation in practice, a python script can be used to define the DTM specifics, and integrated into the thermal model. The python script can include control algorithms for active cooling or dynamic power reduction, depending on live thermal conditions in the model. For example, fan speed can be varied depending on the needs of heated components. In the present case, as discussed, this article will look at processor power throttling using DVFS.

Using a commercially available thermal simulation tool, a python DTM code is applied. The DTM here translates into modulation of power based on temperature sensor feedback. Power modulation is driven by the external python code that gets executed every timestep and gives flexibility for implementing any further complex logic needed, see Figure 1. Further, the python code is generated automatically in this case by a toolkit UI in the simulation tool based on comprehensive user inputs. Figure 1 shows the implementation of a DTM power throttling algorithm into a 3D field solver.
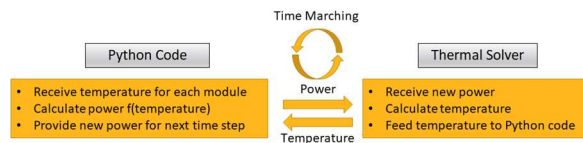

Figure 1: DTM python code applied within a commercially available 3D field thermal solver.

While useful and accurate, DTM with a full-field Computational Fluid Dynamics (CFD) or Finite Element Analysis (FEA) solver solution can take a relatively long time to solve, particularly due to being a transient thermal simulation. Generally, transient simulations take significantly longer than static simulations. Moreover, long sequences of transient power need to be simulated to accurately predict how thermal hotspots change with time. Traditional FEA/CFD-based simulation approaches can potentially run into limitations addressing these needs, and faster transient thermal simulations are generally needed for more effective DTM simulations.

Reduced-Order Models (ROMs) are a potential solution to this problem. ROMs are accurate, compact models from detailed 3D physics simulations, see Figure 2. ROMs can address linear, non-linear, static, or dynamic thermal models with scalar or 3D field output. Various approaches exist to generating ROMs, and generally they can be solved in a fraction of the time required by 3D techniques as well as be integrated into a system model. Thermal ROMs can be integrated with DTM control logic in the same way a full-field 3D thermal solver can.
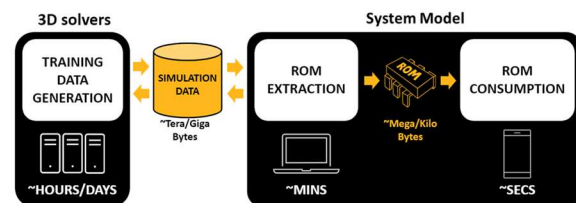

Figure 2: Schematic of a Reduced-Order Model generated from simulation data.

DTM MODEL SETUP

A simplified 3D IC model with a five-tier chip stack and several hotspot locations is created that approximates the specifications listed in [14], see Figure 3. The chip and hot spot sizes are 3cm x 3cm and 2mm x 2mm, respectively. Three hotspots are active per chip, as this was sufficient to demonstrate strong and weak intrachip hot spot thermal coupling. HotSpot 3 and 4 have relatively strong thermal coupling with each other, and weak thermal coupling with HotSpot 1. For simplicity, additional hot spots (e.g., HotSpot 2) took the background heat flux in this example and are therefore not shown. The mesh size for the 3D field model is 19136 cells. A conduction-only, transient model, with ambient and initial temperature set to 20°C was prepared. The simulation stop time is 60s, with time step size of 25ms.

Constant boundary conditions include a Tier 2, Tier 3, and Tier 5 uniform background heat flux of 150,000 W/m^2, and a uniform heat transfer coefficient at the bottom surface of 100 W/m^{2}K, with Tref = 20°C. Tier 2 and Tier 3 both include hot spots in addition to the background heat flux.
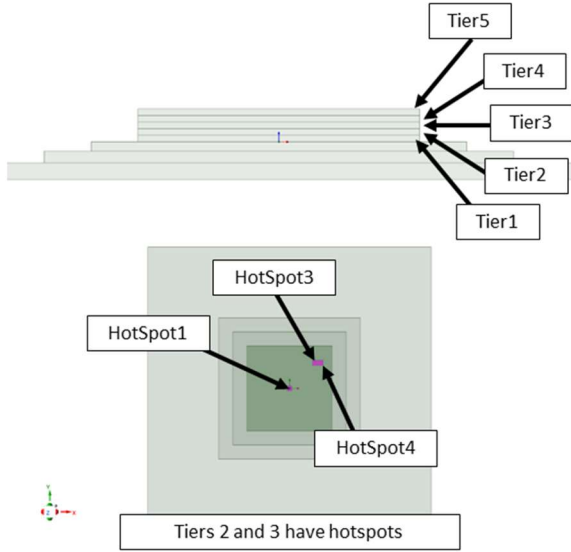
Figure 3: Schematic of 5-tier chip stack, on an interposer, package, and printed circuit board (PCB). Background uniform heat flux is applied to each tier. Tiers 2 and 3 have 2mm x 2mm hot spots, placed as indicated.

Transient boundary conditions, which were set before the simulation began, were each square wave profiles and included the Tier 3 hot spots, the Tier 4 background heat flux and a heat transfer coefficient at the top surface, with Tref = 20°C. These transient boundary conditions can be seen in Table 1 and in Figure 4. The transient hot spot and cooling variations represent changes in chip activities and cooling performance, respectively.

Table 1: Transient boundary conditions' square wave profiles.

	On Value	Phase	On Time	Off Time	Off Value
Tier3 HotSpot 1	5.0E6 W/m² → 20W	2s	2s	8s	2.5E5 W/m² →1W
Tier3 HotSpot 3		6s	1s	9s	
Tier3 HotSpot 4		18s	0.5s	29.5s	
Tier4 Background	1.5E5 W/m² → 135W	0s	33s	4s	7.5E5 W/m² → 675W
Top Cooling	2.5E4 W/m²/C	0s	10s	10s	1.0E4 W/m²/C

In addition to constant, and preset transient boundary conditions, DTM power throttling boundary conditions were also defined. These boundary conditions are defined by different temperature-dependent power profiles for different operating modes (i.e., voltage and frequency). These operating modes are in turn controlled and can be switched depending upon temperature, as DVFS occurs with increasing temperatures. Additionally, a preset On/Off switching profile can be optionally added to these DTM sources. These DTM conditions are defined in a python script, and therefore can include arbitrary complexity in the DVFS algorithm

implemented. Here the Tier 1 Background, see Figure 5, and Tier 2 hot spots 1, 3, and 4, see Figure 6, have DTM power throttling boundary conditions assigned, as shown. Tier 2 hot spots include the same temperature-dependent power profile for each operating mode, while the Tier 1 background source simply has a constant power at each mode.
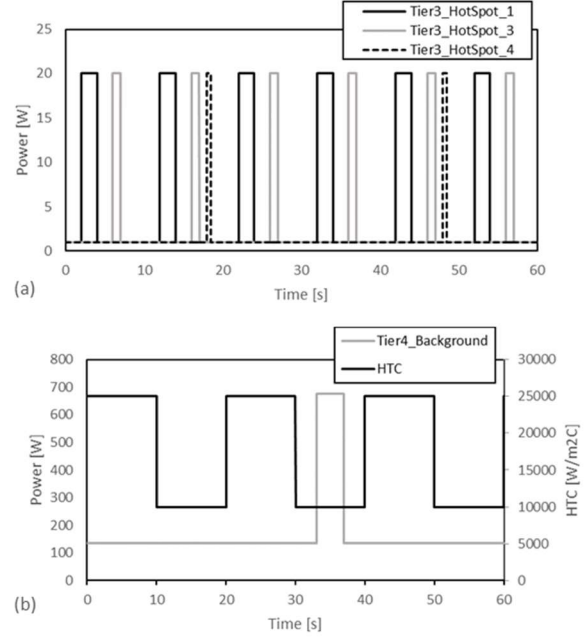

Figure 4: Transient boundary conditions' square wave profiles (a) Tier 3 hot spot powers, (b) Tier 4 background power and top cooling heat transfer coefficient.
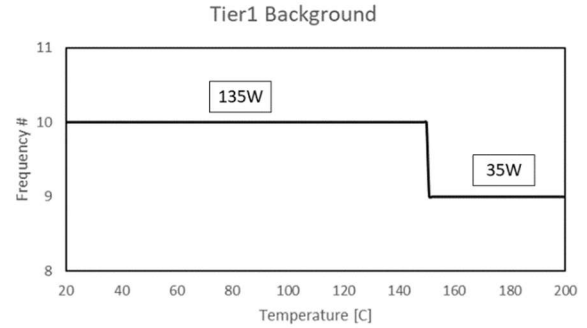

Figure 5: Tier 1 Background Mode # versus Temperature, with power indicated.

In this problem setup, Tier 2 hot spots 1, 3, and 4 have DTM boundary conditions and each are controlled by their own self-temperature sensors. The Tier 1 background boundary condition on the other hand is controlled by both its own self-temperature sensor, as well as Tier 2 hot spots 3 and 4 sensors, therefore having both multiple and remote sensors controlling it. Additionally, each DTM power throttling boundary

condition has an On/Off temperature provision of 65°C/150°C, respectively. Once the Off temperature is reached, the lowest power mode is enforced until the On temperature is reached again.
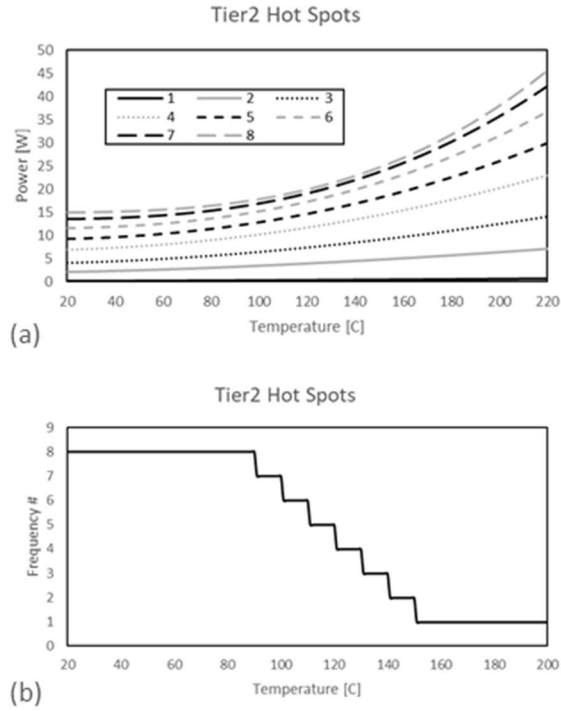

Figure 6: Tier2 Hot Spots (a) Power versus Temperature at Different Modes (b) Frequency # versus Temperature.

ROM GENERATION AND VERIFICATION

To approach the problem setup described, a Linear Parameter Varying (LPV) ROM was generated [12]. The thermal LPV ROM is based on a state-space representation. It processes a set of time-varying inputs, i.e., powers at multiple locations and HTC at the cooling surface, and yields time-varying outputs of temperature at selected locations. The training data necessary for the generation of the LPV ROM was prepared in the 3D field solver model. Here, the ROM inputs include all of the constant, transient, and DTM power throttling sources described above. These sources were excited individually with a constant 200W applied power to generate the time-to-steady-state response. The outputs are temperature sensors at each of the input sources. The top surface heat transfer coefficient was the scheduling parameter, ranging from values between 5000 – 30000 W/m²K, in increments of 5000. The heat transfer coefficient range represents a significant cooling capability that could be provided by a heat sink with forced fan air flow or forced liquid cooling at variable flow rates.

The excitations applied individually to each of the sources were repeated for each value of the heat transfer coefficient scheduling parameter (6 in total here). With the training data generated, the LPV ROM was created with available built-in commercial ROM creation features. The ROM includes 11 power and 1 heat transfer coefficient inputs, as well as 9 temperature sensor outputs. The ROM is then placed in a system modeling solver.

For initial verification, the previous problem setup was considered, with DTM sources turned off. Results from the ROM and 3D field solver were compared. The average error of all 9 temperature sensors, time-averaged over the 60s simulation time, was 1.00%. The solution time for the 3D field model and the ROM for this scenario was 893.52s and 4.76s, respectively, giving a ROM Speedup Factor of 187.71. In this case DTM power throttling sources were not applied, so the next step, as follows, is to include the DTM sources with DVFS power throttling effects.

3D IC THERMAL ROM WITH POWER THROTTLING

The system model that includes the 3D IC thermal ROM can incorporate a DTM component in the form of a python script that defines the DVFS power throttling on the sources, see Figure 7. An available built-in feature can generate the DVFS script based on provided inputs. Here the DTM boundary conditions were applied to Tier 1 Background, and Tier 2 Hot Spots 1, 3, and 4.
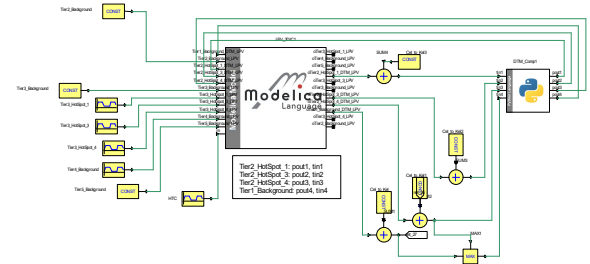

Figure 7: Thermal ROM with DTM schematic in system model.

Results from the ROM with power throttling applied are shown below in Figure 8. Temperatures are compared for a similarly setup 3D field model, with the DTM python script embedded within it. Just as the DTM python script can be embedded within the thermal ROM solution, it can also be embedded within the 3D field solution. This latter aspect, however, is outside the scope of the current article.
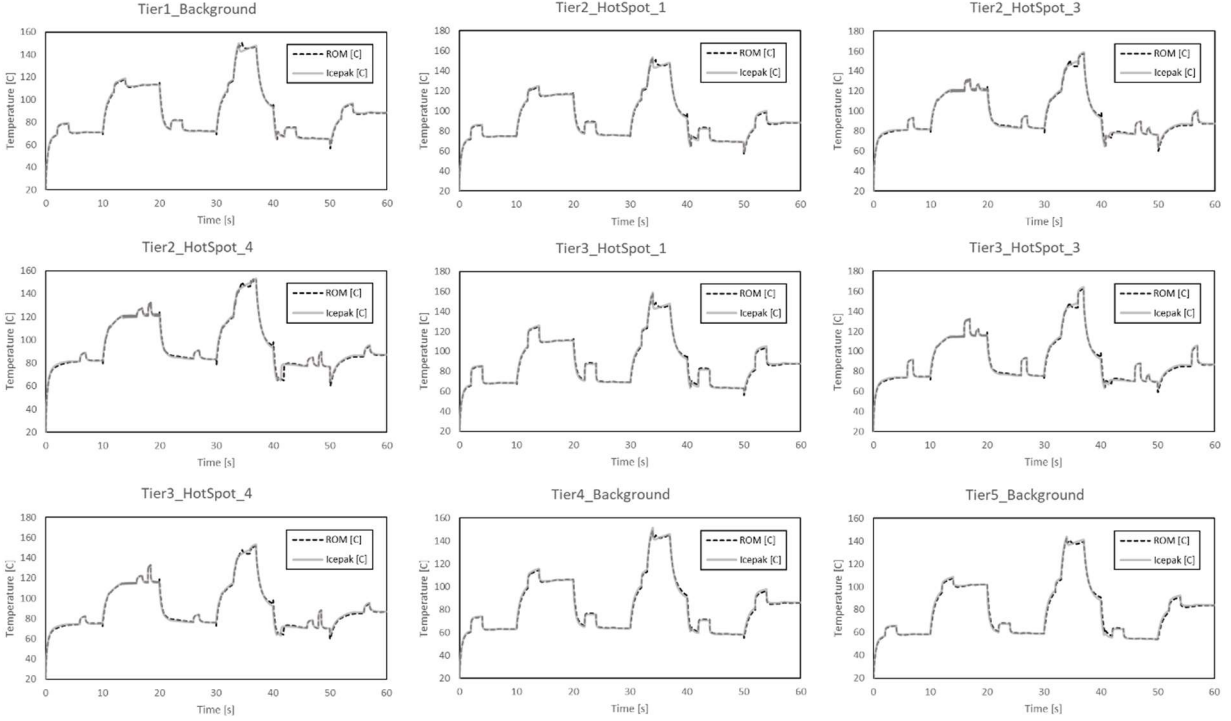
Figure 8: Comparison of DTM with ROM and 3D field solver

The ROM and 3D field models with power throttling match closely. The time average error for all sensors is about 1%, as shown in Table 2, leading to verification of the ROM with power throttling approach.

The benefit of applying the ROM with power throttling approach is in the significant speed up observed, compared to using a 3D field solver. In this example the ROM with power throttling solves in 21s, 61 times faster than the 3D field solver, refer to Table 3. The ROM solve time would remain 21s irrespective of how large (i.e., with respect to mesh size) the 3D field model was, so speed up factors could be significantly higher for larger 3D models. The size of the 3D model would impact only the ROM training data generation time, which is a significant one-time, upfront investment to be considered. An additional benefit is that the same 3D IC thermal ROM can be reused for any arbitrary chip power input and power throttling algorithm applied. Therefore, various DTM power throttling algorithms can be quickly tested for the given 3D IC thermal model.

Table 2: Time-average error of temperature sensors

	Error [%]		Error [%]
Tier1_Background	0.76	Tier3_HotSpot_3	1.01
Tier2_HotSpot_1	0.90	Tier3_HotSpot_4	1.00
Tier2_HotSpot_3	0.90	Tier4_Background	1.07
Tier2_HotSpot_4	0.89	Tier5_Background	1.22
Tier3_HotSpot_1	1.05	**Average**	**0.98**

Table 3: Solution times and ROM speedup factor.

	Solution Time [s]	ROM Speedup Factor [-]
3D Field Model	1293.7	61.2
Reduced Order Model	21.2	

CONCLUSIONS

Power throttling in a 3D IC dynamic thermal analysis was considered in this article. This is a key thermal simulation challenge identified in the HIR [2]. A thermal ROM was generated from a 3D field model and a power throttling algorithm was applied to the thermal ROM. The 3D IC thermal ROM with power throttling was verified within 1% against the 3D field model and significant solution speed up was observed. Benefits of the 3D IC thermal ROM for power throttling design and temperature sensor optimization were highlighted.

New approaches have been recently reported for chip thermal modeling that hold potential for further improved DTM simulation [15-16]. These ROM and machine-learning-based methods aim to preserve the fine-grain chip-level detail in 3D ICs (e.g., μm and μs levels) required to capture on-chip hot spots accurately, while simultaneously including electronic thermal environments and cooling systems with high fidelity. The approach described in this article shows

promise for extension to these new and advanced thermal solvers as well.

REFERENCES

[1] *Dynamic Thermal Management (DTM) Workshop,* Arlington, VA: DARPA MTO, 2023.

[2] IEEE EPS, "Heterogeneous Integration Roadmap, Chapter 20: Thermal," 2023.

[3] D. Brooks and M. Martonosi, "Dynamic thermal management for high-performance microprocessors," *Proceedings HPCA Seventh International Symposium on High-Performance Computer Architecture,* pp. 171-182, 2001.

[4] A. Mirtar, S. Dey and A. Raghunathan, "Joint Work and Voltage/Frequency Scaling for Quality-Optimized Dynamic Thermal Management," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 23, no. 6, pp. 1017-1030, 6 2015.

[5] J. M. Kim, Y. G. Kim and S. W. Chung, "Stabilizing CPU Frequency and Voltage for Temperature-Aware DVFS in Mobile Devices," *IEEE Transactions on Computers,* vol. 64, no. 1, pp. 286-292, 1 2015.

[6] W. Huang, M. Allen-Ware, J. B. Carter, E. Elnozahy, H. Hamann, T. Keller, C. Lefurgy, J. Li, K. Rajamani and J. Rubio, "TAPO: Thermal-aware power optimization techniques for servers and data centers," *2011 International Green Computing Conference and Workshops,* pp. 1-8, 7 2011.

[7] G. Xu, "Multi-core server processors thermal analysis," *2017 16th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm),* pp. 416-421, 5 2017.

[8] M. M. Sabry, A. K. Coskun, D. Atienza, T. Š. Rosing and T. Brunschwiler, "Energy-Efficient Multiobjective Thermal Control for Liquid-Cooled 3-D Stacked Architectures," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 30, no. 12, pp. 1883-1896, 12 2011.

[9] R. Roy, S. Das, B. Labbe, R. Mathur and S. Jeloka, "Co-design of Thermal Management with System Architecture and Power Management for 3D ICs," *2022 IEEE 72nd Electronic Components and Technology Conference (ECTC),* pp. 211-220, 5 2022.

[10] S. Krishnaswamy, P. Jain, M. Saeidi, A. Kulkarni, A. Adhiya and J. Harvest, "Fast and accurate thermal analysis of smartphone with dynamic power management using reduced order modeling," *2017 16th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm),* pp. 276-281, 5 2017.

[11] Y. Im, W. Kim, T. An, H. Lee, Y. Cho, J. Yoo, H. Lee, Y. Shin, M. Lee and V. K. Yaddanapudi, "Thermal Sensor Placement based on Meta-Model Enhancing Observability and Controllability," *2020 19th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm),* pp. 776-782, 2020.

[12] Y. Im, G. Jung, M. Lee, A. Gangrade and S. Kim, "Thermal Modeling and Optimization of Mobile Device using modified LPV ROM," *2023 22nd IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm),* pp. 1-8, 5 2023.

[13] N. Chang, S. Pan, K. Srinivasan, Z. Feng, W. Xia, T. Pawlak and D. Geb, "Emerging ADAS Thermal Reliability Needs and Solutions," *IEEE Micro,* vol. 38, no. 1, pp. 66-81, 2018.

[14] DARPA Microsystems Technology Office, "Miniature Integrated Thermal Management Systems for 3D Heterogeneous Integration (Minitherms3D)," 2023.

[15] A. Kumar, N. Chang, D. Geb, H. He, S. Pan, J. Wen, S. Asgari, M. Abarham and C. Ortiz, "ML-based Fast On-Chip Transient Thermal Simulation for Heterogeneous 2.5D/3D IC Designs," *2022 International Symposium on VLSI Design, Automation and Test, VLSI-DAT 2022 - Proceedings,* 2022.

[16] D. Geb, S. Asgari, A. Kumar, J. Wen, N. Chang, S. Pan, M. Abarham, H. He and V. Gandhi, "On-Chip Transient Hot Spot Detection with a Multiscale ROM in 3DIC Designs," *Proceedings - Electronic Components and Technology Conference,* Vols. 2022-May, pp. 221-232, 2022.