Adaptive Transient Leakage-Aware Linearised Model for Thermal Analysis of 3-D ICs

Chao Zhang
School of Computer Science
The University of Manchester
Manchester, UK
Chao.Zhang@manchester.ac.uk

Milan Mihajlović
School of Computer Science
The University of Manchester
Manchester, UK
milan@cs.man.ac.uk

Vasilis F. Pavlidis
School of Computer Science
The University of Manchester
Manchester, UK
pavlidis@cs.man.ac.uk

***Abstract*—Physics-based models for thermal simulation that involve numerical solution of the heat equation are well placed to accurately capture the heterogeneity of materials and structures in modern 3-D integrated circuits (ICs). The introduction of non-linear effects such as leakage power significantly improves their accuracy. However, this non-linearity increases considerably the complexity and computational time of the analysis. In this paper, we introduce a linearised thermal model by demonstrating that the weak temperature dependence of the specific heat and the thermal conductivity of IC related materials has only minor effect to computed temperature profiles. Thus, these parameters can be considered constant for the operating temperature ranges of modern ICs. The non-linearity in leakage power is approximated by a piecewise linear least square fit and the resulting model is linearised by exact Newton's method. The method is applied to transient thermal analysis with adaptive time step selection, where we demonstrate the importance of applying Newton corrections to obtain the right time step size selection. The resulting method is typically $2-3\times$ faster than a full non-linear method with a global relative error of less than 1%.**

***Index Terms*—Thermal analysis, Leakage power, Finite element method, Adaptive time integration.**

I. Introduction

Thermal characterisation and management are critical issues in the design of modern ICs. The drive towards miniaturisation and increasing package density has increased the contribution of leakage power to the overall power of integrated systems. Leakage power can contribute up to 40% to the total power [1] and strongly depends upon temperature, leading to a positive feedback and potentially to a thermal runaway. Consequently, modern thermal analysis tools for ICs must capture this effect accurately, while maintaining computational efficiency. In addition, they need to be fast, as thermal analysis of ICs forms a critical path in the IC design process, requiring multiple runs over the design cycle.

Thermal analyzers for ICs can broadly be classified into two categories based on the methodology used. These are physical models and electro-thermal duality models. The tools in the former category rely on a numerical solution of the heat equation [2], [3], [4], while the simulators in the latter category rely on the duality between electrical circuits and thermal laws [5], [6]. In general, the duality-based models are faster, but less accurate, and have difficulties to capture heterogeneous features present in modern 3-D ICs. By contrast, physically based models are effective for complex geometries, materials with vastly different thermal properties, and complicated structures encountered in 3-D ICs, such as TSVs [7]. Physical models deploy some discretisation of the heat equation, e.g. by the finite element method (FEM) [8]. In order to improve the computational speed and maintain accuracy, spatio-temporal adaptation is typically considered [3], [4], which assumes both periodic refinement/coarsening of a spatial computational grid and adaptive selection of time steps that closely follow the physics of the problem. Such an approach is implemented in the physical thermal simulator MTA [4], which features both linear and non-linear thermal model, and supports both floorplan and cell level detail.

High computational cost of a full non-linear thermal simulation has prompted the development of linearised techniques for the approximation of non-linearities in the thermal model, including the leakage power. In this context, a global linear approximation of the leakage current is proposed in [9], [10], a piecewise linear approximation (PWL) is considered in [11], [12], and a low-order polynomial approximation in [13]. In [11], the authors propose a continuous PWL approximation leakage current model used in thermal/leakage co-simulation.

The non-linear thermal model linearised by Newton's method and with adaptive time stepping rapidly converges (typically 1-2 iterations, even for loose time integrator tolerances). Thus, if a linearised method is to be competitive in this context, it should require, on average, fewer iterations than the full non-linear method, while maintaining accuracy. In addition, in transient thermal analysis the number and size of the selected time steps should be comparable to both the linear and full non-linear methods.

Based on these observations, we propose a linearised thermal model that includes the non-linear leakage power but neglect the non-linearities in thermal conductance and specific heat due to their weak effect. The leakage current function is approximated by a PWL least square interpolant. The adaptive time integration is done by predictor-corrector schemes [14]. The resulting problem in the corrector needs to be solved by Newton's method both to preserve accuracy and to select appropriate time steps. This is different from the linearised method introduced in other areas [15], where only one iteration per time step suffices to obtain appropriate time steps.

Compared to the full non-linear (NL) method, the resulting method is several times $2-3\times$ faster where the exact

978-3-9819263-2-3/DATE19/©2019 EDAA

gain factor is problem-dependent. The global accuracy of the proposed linearised method is within 1% of the full non-linear simulation. This decrease in execution time in a critical component of the IC design process without compromising accuracy is an improvement over existing techniques.

The rest of the paper is organised as follows. Thermal models based on the NL heat equation and related work are discussed in the Section II. The proposed linearized model is presented in Section III and results on test cases are described in Section IV. Some conclusions are offered in Section V.

II. Background and related work

In this section, we discuss related work in subsection II-A. The physical thermal model involving the non-linear heat equation and our computationally efficient and accurate method are covered in subsections II-B and II-C, respectively.

A. Related work

Different approaches were proposed to include leakage power into thermal models of ICs [10], [11], [16], [17]. The primary objective in each of these methods is to reduce the computational cost of a full NL model while maintaining accuracy. In [16], the authors use Green's function. This leakage model is relatively simple, but introduces severe restrictions to geometries and material properties (in particular, the method does not accommodate efficiently horizontal material heterogeneities frequent in 3-D ICs). In `LightSim` [10], NL leakage power is approximated by a linear function in the temperature interval 40°C–75°C. The accuracy of this model is low and temperature-dependent. In [11], the authors propose an evenly-spaced PWL continuous function for leakage power approximation and tested the method on different ICs. In [18], the authors propose quadratic polynomial interpolation for leakage power. All these works are either considered for steady-state problems or transient problems with fixed time steps.

The nature of 3-D integration considerably increases the size of the system. In order to capture thermal gradients and strong thermal coupling in vertical direction in this context requires that the simulator adapts to the dynamic behaviour of a 3-D system by selecting appropriate time steps. Moreover, the introduction of the NL dependence of leakage power on temperature adds complexity into the system and requires appropriate modification of the solution procedure in order to preserve both accuracy and the time step size selection process. The authors in [11] use simple iteration which has only linear asymptotic convergence (compared to the quadratic convergence of Newton's method) and in [17] inexact Newton's method was used to gain computational efficiency. We demonstrate that this can be achieved by linearising a PWL least square fit with the exact Newton's method.

B. Physics-based thermal model and the FEM discretisation

The heat flow in an IC is governed by the non-linear heat equation:

$$\rho C(\mathbf{x},u)\,\frac{\partial u}{\partial t} - \nabla\cdot(\kappa(\mathbf{x},u)\nabla u(\mathbf{x},t)) = f(\mathbf{x},u,t) \text{ in } \Omega\times[0,T]\,, \tag{1}$$

$$\kappa(\mathbf{x},u)\nabla u\cdot\hat{\mathbf{n}} = \eta(u_a-u) \text{ on } \partial\Omega\times[0,T]\,, \tag{2}$$

$$u(\mathbf{x},0) = u_0 \text{ in } \bar{\Omega} = \Omega\cup\partial\Omega\,. \tag{3}$$

In (1)–(3), the function $u(\mathbf{x},t)$ represents the temperature of a material body $\Omega\subset\mathbb{R}^3$ with the boundary $\partial\Omega$. The physical parameters are the material density ρ $[kg/m^3]$, the temperature-dependent specific heat $C(\mathbf{x},u)$ $[J/kg\,K]$, the temperature-dependent thermal conductivity κ $[W/m\,K]$, and the thermal transmissivity η $[W/m^2K]$. The temperature dependence of the parameters C and κ is a complex function for the silicon over a wide range of temperatures [19], but can be effectively approximated by a linear function (e.g., by the linear least square fit) in the operating temperature range of an IC $(300-400\,K)$.

The power density function f $[W/m^3]$ has two components, i.e. $f = f_D(\mathbf{x},t) + f_L(\mathbf{x},u,t)$, where f_D is the dynamic (active) power and f_L is the leakage power. The leakage power has several components of which the subthreshold has a strong dependence on temperature given by

$$f_L = \alpha_L u^2\exp(\beta_L/(u-u_b))\,, \tag{4}$$

where α_L $[W/m^3]$, β_L and the base temperature u_b are technology-dependent parameters [20].

The heat equation (1–3) is discretised in space by standard Galerkin finite element method [8]. The domain Ω is subdivided into a set of non-overlapping cuboid elements that are aligned to internal boundaries between different components of an IC. The procedure is described in detail in [8] and leads to a system of non-linear initial value problems (IVPs)

$$M\dot{\bar{u}} + A(\bar{u})\bar{u} = \bar{f}(\bar{u})\,, \tag{5}$$

where $M, A\in\mathbb{R}^{n\times n}$ are the mass and stiffness matrices [8], $\bar{f}\in\mathbb{R}^n$ is the discrete power vector, and the dot notation stands for the time derivative. The semi-discrete problem has dimension n equal to the number of nodes in the mesh.

C. Adaptive time integration

The system of IVPs (5) can be solved by a number of methods, which discretise the time domain $[0,T]$ into discrete time points $0 = t_0 < t_1 < \cdots < t_k = T$ which can either be equidistant or not. The distribution of time integration points affects the accuracy of computed solutions and computational efficiency and can be done either statically (*a priori*) or dynamically (adaptively), by constructing an error control procedure [8], [14]. We employ a computationally efficient implementation of adaptive implicit methods, where the step selection is governed by computable estimates of the local truncation error (LTE) [8]. This is, in practice, achieved

through a predictor-corrector scheme. We consider the BDF2 method as the corrector [14] (p. 715)

$$\bar{u}_{n+1} - \frac{(1+w_n)^2}{1+2w_n}\,\bar{u}_n + \frac{w_n^2}{1+2w_n}\,\bar{u}_{n-1} = \Delta t_{n+1}\frac{1+w_n}{1+2w_n}\,\dot{\bar{u}}_{n+1}\,, \tag{6}$$

where $w_n = \frac{\Delta t_{n+1}}{\Delta t_n}$, $\Delta t_{n+1} = t_{n+1} - t_n$, $\Delta t_n = t_n - t_{n-1}$. We pair (6) with a second-order explicit method, the explicit mid-point rule (or the leapfrog method) [14]

$$\bar{u}^p_{n+1} = \bar{u}_n + (1+w_{n-1})\Delta t_n \dot{\bar{u}}_n - w^2_{n-1}(\bar{u}_n - \bar{u}_{n-1}) \tag{7}$$

to obtain computable estimates of the LTE $\bar{d}_n$, described by

$$\bar{d}_n \approx C_n(\bar{u}_{n+1} - \bar{u}^p_{n+1})\,, \tag{8}$$

where $C_n = C_n(\Delta t_{n-1}, \Delta t_n, \Delta t_{n+1})$ (see [14] for full detail). With (8), the next time step is estimated as

$$\Delta t_{n+1} = \Delta t_n (\epsilon_t / \|\bar{d}_n\|)^{1/3}\,, \tag{9}$$

where ϵ_t is a user-specified tolerance. Having in mind that the quality of the selected time steps relies on the accuracy of the LTE estimates, the leakage power approximation needs to be implemented consistently (i.e. the same type of the approximation should be used both in (6) and (7)).

III. The proposed method

We propose a linearised thermal model with leakage power that can be applied in an adaptive setting. The model is based on two observations: 1) the non-linearity in thermal coefficients C and κ in (1) is weak and these parameters are considered constant in the operating temperature range of modern ICs, and 2) if leakage power is approximated by a piecewise least square linear function, it is essential to resolve the non-linearities in the model accurately to obtain the right selection of the time step sizes.

When all the points in the finite element mesh that lie in active layers have temperatures that belong to a single line segment in the PWL approximation of leakage power, then the linearisation procedure converges in precisely one step, as the model is linear. If the leakage current is sampled from different PWL segments, the problem is non-linear, as the coefficients of different linear segments are temperature-dependent, and multiple linearisation iterations are required to determine accurately the temperature. This requirement introduces a tradeoff between the accuracy of the model (*the higher the number of linear segments in PWL approximation, the better the accuracy*) and computational efficiency (*the higher the number of segments, the more non-linear the model*).

In an adaptive time setting, the appropriate choice of time steps leads to a rapid convergence of Newton's method (for a full NL problem it takes typically 1-2 iterations). We demonstrate in Section IV that the proposed method with small to moderate number of PWL segments used to approximate the leakage current exploits the accuracy/complexity tradeoff, thereby converging in a smaller number of linearisation steps compared to the full non-linear method without compromising accuracy. The simplified form of the Jacobian matrix in the linearised case offers computationally efficient assembly (only the matrix rows associated with the nodes lying in active layers of an IC need to be modified at each Newton's step) compared to the full PWL model.

IV. Numerical Results

The proposed methodology is implemented and optimised for speed within the MTA package [4] and evaluated on two circuit floorplans derived from the benchmark suite [21]. All experiments are run on a PC with an Intel i7 4790 processor and 32 GB DRAM. The first circuit comprises two tiers with a 4-core processor as tier 0 and cache memory as tier 1. The second circuit includes three tiers with two memory tiers on top of the processor tier. Both systems have a heat sink with dimension $5\times$ larger than the circuit with 11 fins mounted on the top. We use the same maximum power densities as in [21] and perform a study in which one core and the corresponding memory module(s) are active all the time, two cores and their memory modules are active 75% of time, with randomly switching on and off and one core is inactive (i.e. consumes only leakage power).

A. Non-linearities in C and κ

In this experiment, we demonstrate that the non-linearity of thermal coefficients C and κ in (1) affect negligibly the accuracy of thermal simulation. To this end, we compare optimised implementations of the full non-linear model (**NL**), where C and κ are considered linearly dependent on temperature (i.e. $C = C_0(1+\alpha_c u)$ and $\kappa = \kappa_0(1+\alpha_\kappa u)$) with the approximate non-linear (**NLA**), where $\alpha_c = \alpha_\kappa = 0$, and a linear model (**L**). In both NL and NLA, leakage power is modeled exactly, while in the linear case we take the leakage current at 300 K). The experiment is performed for the two-tier circuit, with spatial discretisation containing $n = 181,243$ nodes and fixed time steps $\Delta t = 0.5\,s$. The global solution error is

$$e = \max_{t_k} \frac{\int_\Omega |u_* - u_{NL}|\,d\Omega}{\int_\Omega u_{NL}\,d\Omega} \cdot 100\,[\%]\,, \tag{10}$$

where u_{NL} is the computed NL solution and u_* is either u_{NLA} or u_L and the maximum is taken over all time steps t_k. The global relative error for the linear case is 22.45%, while for the linearised model is 0.41%. These figures suggest that neglecting the non-linear dependence of C and κ does not affect computed thermal profiles significantly (the error is $< 1\%$). By contrast, the non-linearity in f_L has a profound effect to the computed temperature.

B. Piecewise linear model for f_L

In this experiment, we firstly evaluate the impact of the accuracy in the corrector to the time step selection in the adaptive setting. We fix the number of segments in the linearised model to 5, integrate over the time interval $[0, 200]$ with the LTE parameter $\epsilon_t = 10^{-3}$ and compare the number of time steps N_t, the average number of Newton steps per time step $\bar{N}_n$, and the execution time T_{exe} between the NL, NLA1 (the

NLA model with 1 Newton step), NLAm (the NLA model with multiple Newton steps, required to achieve the convergence to the relative tolerance $\epsilon_N = 10^{-8}$), and L models. The results are listed in Table I.

TABLE I
NUMBER OF TIME STEPS N_t, AVERAGE NUMBER OF NEWTON STEPS $\bar{N}_n$, EXECUTION TIME T_{exe}, AND THE GLOBAL SOLUTION ERROR (10) FOR DIFFERENT METHODS

Model	N_t	$\bar{N}_n$	T_{exe} $[\times 10^3\ s]$	e [%]
L	142	1.00	1.11	20.45
NLA1	745	1.00	4.00	0.02
NLAm	330	1.05	2.43	0.02
NL	474	1.05	4.16	-

The results suggest that the NLAm model performs well and selects time step sizes that are commensurate to the NL model, but the model NLA1 selects considerably smaller steps, due to the inaccuracy in computing the corrector, ultimately leading to a method that is not competitive. The NLAm model is approximately $2\times$ times faster than the NL model and $2\times$ slower than the linear model but with accuracy close to the NL model. Similar patterns are observed for different number of line segments in the approximation of leakage current.

Finally, we assess the accuracy and computational efficiency of the NLAm method as a function of the number of piecewise linear segments for approximating the leakage current. The experiment is performed for both two-tier and three-tier circuits for the same time interval and other parameters as previously and the results are reported in Table II. The results for L and NL models are given for comparison.

TABLE II
NUMBER OF TIME STEPS N_t, AVERAGE NUMBER OF NEWTON STEPS $\bar{N}_n$ EXECUTION TIME T_{exe}, AND THE GLOBAL SOLUTION ERROR (10) FOR THE NLAM(ℓ) METHOD AS A FUNCTION OF THE NUMBER OF LINE SEGMENTS ℓ.

2-tier	ℓ	1	2	5	L	NL
	N_t	156	143	330	142	474
	$\bar{N}_n$	1.00	1.00	1.05	1.00	1.05
	T_{exe} $[\times 10^3\ s]$	1.33	1.16	2.43	1.11	4.16
	e [%]	10.17	0.45	0.02	20.45	-
3-tier	Method	1	2	4	L	NL
	N_t	211	200	257	218	526
	$\bar{N}_n$	1.00	1.00	1.06	1.00	1.12
	T_{exe} $[\times 10^3\ s]$	2.85	2.75	3.51	2.01	7.56
	e [%]	14.90	1.20	0.01	28.82	-

The performance of the NLAm method depends on the number of linear segments, nonetheless, we obtain temperature profiles with global error smaller than 1% with only a few linear segments. In such cases, the execution times of the NLAm method are significantly shorter than that of the NL method, and comparable to the linear case but with much higher accuracy than the linear model.

V. CONCLUSIONS AND FUTURE WORK

We presented a linearised model for thermal simulation of ICs, based on neglecting the non-linearities in thermal coefficients C and κ and linearising the leakage power model with a PWL least square fit. The obtained discrete problem still requires linearisation by Newton's method, but leads to much shorter execution times than the non-linear model with very good accuracy. The performance is dependent, however, on the number of linear segments and future work involves automatic selection and adaptive change of this number throughout the simulation with the aim of minimising the execution time, whilst maintaining accuracy.

REFERENCES

[1] S. Naffziger *et al.,* "The Implementation of a 2-core, Multi-Threaded Itanium Family Processor," *IEEE J. Solid State Circ.*, 41(1)(2006), 197–209.

[2] P. Li, L. T. Pileggi, M. Asheghi, and R. Chandra, "IC Thermal Simulation and Modeling via Efficient Multigrid-based Approaches," *IEEE Trans. on CAD*, 25(9)(2006), 1763–1776.

[3] Z. Hassan *et al.,* "Full-Spectrum Spatial–Temporal Dynamic Thermal Analysis for Nanometer-Scale Integrated Circuits," *IEEE Trans. on VLSI Syst.*, 19(12)(2011), 2276–2289.

[4] S. Ladenheim, Y.-C. Chen, M. Mihajlović, and V. F. Pavlidis, "The MTA: An Advanced and Versatile Thermal Simulator for Integrated Systems," *IEEE Trans. on CAD*, 37(12)(2018), 3123–3136.

[5] W. Huang, *et al.,* "HotSpot: A Compact Thermal Modeling Methodology for Early-Stage VLSI Design," *IEEE Trans. on VLSI Syst.*, 14(5)(2006), 501–513.

[6] A. Sridhar, A. Vincenzi, D. Atienza, and T. Brunschwiler, "3D-ICE: A Compact Thermal Model for Early-Stage Design of Liquid-Cooled ICs," *IEEE Trans. on Comp.*, 63(10)(2014), 2576–2589.

[7] V. F. Pavlidis, I. Savidis, and E. G. Friedman, *Three-Dimensional Integrated Circuit Design*, 2nd ed., Morgan Kaufmann Publishers, 2017.

[8] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite Elements and Fast Iterative Solvers: with applications in incompressible fluid dynamics*, 2nd ed., Oxford University Press, 2014.

[9] V. Chaturvedi, H. Huang, and G. Quan, "Leakage Aware Scheduling on Maximum Temperature Minimization for Periodic Hard Real-Time Systems," *Proc. IEEE Conf. on Comput. and Inform. Tech.*, (2010), 855–860.

[10] S. R. Sarangi, G. Ananthanarayanan, and M. Balakrishnan, "Lightsim: A Leakage Aware Ultrafast Temperature Simulator," *Proc. Asia and South Pacific Design Automat. Conf.*, (2014), 855–860.

[11] Y. Liu, R. P. Dick, L. Shang, and H. Yang, "Accurate Temperature-Dependent Integrated Circuit Leakage Power Estimation is Easy," *Proc. IEEE DATE Conf.*, (2007), 1–6.

[12] H. Wang *et al.,* "A Fast Leakage-Aware Full-Chip Transient Thermal Estimation Method," *IEEE Transa. Comput.*, 67(5)(2018), 617–630.

[13] B. Shi and A. Srivastava, "Dynamic Thermal Management Considering Accurate Temperature-Leakage Interdependence," *Encyclopedia of Thermal Packaging: Thermal Packaging Tools*, World Scientific, (2015), 39–50.

[14] P. M. Gresho and R. L. Sani, *Incompressible Flow and the Finite Element Method, Vol. 2: The Navier-Stokes Equations*, John Wiley & Sons, Chichester, UK, 2000.

[15] H.C. Elman, M.D. Mihajlović, D.J. Silvester, "Fast Iterative Solvers for Buoyancy Driven Flow Problems," *J. Comput. Phys.*, 230(10)(2011), 3900-3914.

[16] H. Sultan and S. R. Sarangi, "A Fast Leakage Aware Thermal Simulator for 3D Chips," *Proc. IEEE DATE*, (2017), 1733–1738.

[17] C. Yan, H. Zhu, D. Zhou, and X. Zeng, "An Efficient Leakage-Aware Thermal Simulation Approach for 3D-ICs Using Corrected Linearized Model and Algebraic Multigrid," *Proc. DATE*, (2017), 1207–1212.

[18] H. Su *et al.,* "Full Chip Leakage Estimation Considering Power Supply and Temperature Variations," *Proc. Int. Symp. Low Power Elect. and Design*, (2003), 78–83.

[19] Y. S. Touloukian and C. Ho, *Thermophysical Properties of Matter: Specific Heat - Metallic Elements and Alloys.* New York: IFI/Plenum, 1970, Vol. 4.

[20] A. Chandrakasan, W. Bowhill, and F. Fox, *Design of High-Performance Microprocessor Circuits*, IEEE Press, 2001.

[21] M. M. Sabry, A. Sridhar, and D. Atienza, "ICCAD 2015 Contest in 3D Interlayer Cooling Optimized Network", *Proc. ICCAD*, (2015), 912–915.