Compact Piecewise Linear Model Based Temperature Control of Multi-Core Systems Considering Leakage Power

Hai Wang, *Member, IEEE,* Liwen Hu, Xingxing Guo, Yang Nie, He Tang, *Member, IEEE*

***Abstract*—Temperature control of the new generation integrated multi-core system is challenging. This is because the leakage power, which is significant in modern systems, is nonlinearly related to temperature, resulting in a complex nonlinear control problem in thermal management. In this paper, a new dynamic thermal management (DTM) method with compact piecewise linear (PWL) model based predictive control is proposed to solve the nonlinear control problem. First, a compact PWL thermal model, which takes dynamic power as input, is built by combining multiple local compact linear thermal models expanded at several Taylor expansion points. These local compact linear thermal models are obtained by sampling based model order reduction (MOR) with high accuracy. Their Taylor expansion points are selected by a systematic scheme which exploits the thermal behavior property of the multi-core chips. Based on the compact PWL thermal model, a new predictive control method is proposed to compute the future power recommendation for DTM. By approximating the nonlinearity accurately with the compact PWL thermal model and being equipped with predictive control technique, the new DTM achieves an overall high quality temperature management with smooth and accurate temperature tracking. Experimental results show the new method outperforms the linear model predictive control based method and the echo state network based predictive thermal management method in temperature management quality with lower computing overhead.**

***Index Terms*—Thermal management, leakage power, multi-core, model predictive control.**

I. Introduction

Power density of integrated multi-core systems keeps increasing with technology scaling, causing severe thermal related problems, including system reliability and performance degradation issues [1]. Power budgets, including thermal design power (TDP) [2] and greedy dynamic power (GDP) [3], were provided as the power limit to relieve the high temperature induced reliability problems. In order to dynamically adjust the power of multi-core systems, thermal/power management actions including task migration [4] and dynamic voltage & frequency scaling (DVFS) [5]–[7] were introduced. Then, based on these basic management actions, researchers proposed many dynamic thermal management (DTM) methods to control the temperature of the multi-core system at runtime. For example, a DTM method with fan speed control was proposed in [8] to optimize the overall energy consumption. Liu *et al.* proposed a DTM method which considers the transient temperature behavior of the packaged multi-core system to enhance the temperature control quality [9]. A scheduling based DTM method was invented for heterogeneous multi-core system [10]. Recently, Wang *et al.* introduced a hierarchical DTM method for the many-core system with low computing delay [11].

However, most DTM methods do not consider leakage power, resulting in less accurate thermal management. For high performance systems manufactured using new planar CMOS technology, leakage power, which even accounts for over 50% of the total power consumption, cannot be neglected anymore [12], [13]. Although the recent introduction of FinFET temporally relieved the leakage power problem, the leakage power can still be significant for 7nm FinFET when low/normal VTH transistor is used to achieve high performance [14], and the leakage power ratio may increase again for the future FinFET technology nodes. To make matters worse, leakage power depends on temperature exponentially, forming a positive feedback between power and temperature, which can lead to thermal runaway in the worst case [8], [15]. Therefore, the leakage power induced thermal problem has already become one of the most important limiting factors of multi-core system performance today.

When leakage power is considered, the model in DTM, which connects power and temperature, takes only dynamic power as input and temperature as output. This is because dynamic power is directly controllable by the system (such as by frequency scaling and task scheduling) but leakage power is not. In this work, we call such model as the leakage-aware thermal model or simply thermal model if no confusion will be caused. The leakage-aware thermal model is nonlinear because the nonlinear leakage power is contained inside it.

The major challenge of considering leakage power in DTM lies in building a leakage-aware thermal model which satisfies the following two conditions at the same time: first, it should has high accuracy in leakage modeling; second, it should be easy to integrate with the multi-core DTM methods. However, it is difficult to satisfy both conditions because most DTM methods require a linear thermal model, but the accurate leakage-aware thermal model is inherently nonlinear

This research is supported in part by National Natural Science Foundation of China under grant No. 61974018, in part by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry. *(Corresponding author: Hai Wang.)*

H. Wang, L. Hu, X. Guo, H. Tang are with State Key Laboratory of Electronic Thin Films and Integrated Devices, University of Electronic Science and Technology of China, Chengdu, 610054 China, and also with School of Electronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 610054 China. E-mail: wanghai@uestc.edu.cn (Hai Wang).

Y. Nie is with Sichuan Haoxunda Technology Co., Ltd., Chengdu, 610041 China.

as aforementioned. Existing methods make compromises by satisfying only one out of the two conditions. For example, some existing leakage-aware DTM methods [1], [11], [16], [17] use linear models to approximate the original nonlinear leakage model, making it compatible with the traditional linear DTM framework in order to satisfy the second condition. However, they suffer from low accuracy due to the large linear approximation error. On the other hand, DTM with a quadratic polynomial based approximation thermal model was introduced in [18], which has high leakage modeling accuracy. However, since developing an elegant multi-core DTM method for such quadratic polynomial leakage model is difficult, this DTM is developed based on single-core thermal model with scalar quadratic polynomial leakage model, which can only be used for single-core systems [18]. Recently, a leakage-aware DTM using neural network thermal model was proposed in [19]. Being a black-box model based method, it is more suitable for the case where the detailed structural information of the packaged multi-core system is unavailable. It generally has lower accuracy compared with the white-box model based methods due to the absence of the detailed package structural information.

The discussions above reveal that it is difficult to design an accurate leakage-aware predictive DTM method for multi-core systems, especially with the white-box thermal model built accurately from the structural information of the multi-core system. In this work, we resolve this problem by proposing a leakage-aware DTM using compact piecewise linear (PWL) model based predictive control. The major contributions of this work include:

- In order to solve the nonlinear control problem in leakage-aware thermal management, we propose to use a PWL thermal model to approximate the original nonlinear thermal model. With the PWL thermal model, predictive control is enabled for leakage-aware DTM.
- A unified formulation of the PWL thermal model for leakage-aware DTM is derived. Specially, a systematic Taylor expansion point selection scheme is developed to formulate the PWL thermal model by exploiting the thermal behavior property of the integrated multi-core system. The resulted PWL thermal model formulation is concise and elegant. Therefore, it can be integrated into the predictive control framework seamlessly.
- To reduce the runtime and memory overheads of DTM, sampling based model order reduction (MOR) is introduced to reduce the size of the PWL thermal model. Thanks to the sampling based MOR, the resulted compact PWL thermal model achieves both high compression rate and high accuracy.
- We propose the compact PWL thermal model based predictive control framework by integrating the compact PWL thermal model into model predictive control (MPC). Although being a nonlinear control, the compact PWL thermal model based predictive control still retains the concise structure of the traditional linear MPC. By using the new temperature control method, accurate future power recommendations can be computed for the multi-core system.
- We have experimentally compared the new DTM method with traditional DTM using linear thermal model based MPC and the echo state network based predictive thermal management method. Our numerical results show the new method outperforms both methods in thermal management quality with lower computing overhead.

II. Background

In this section, the power models used in this work, including dynamic power model and leakage power model, are introduced first. After that, we briefly review DTM using model predictive control (MPC) and reveal its problem in leakage power consideration.

A. Power modeling

The total power of an integrated multi-core system is composed of dynamic power p_d and leakage power p_s (which is also called static power). In this subsection, we will briefly present the modeling of dynamic power and leakage power.

1) Modeling of dynamic power: Dynamic power is caused by the logic gate switching, whose value depends on the activity of the core. It is expressed as

$$p_d = \alpha V_{dd}^2 f, \tag{1}$$

where V_{dd} and f are the supply voltage and clock frequency of the core, respectively, and α is the activity factor of the core. Dynamic power can be obtained by performance counter based methods, which estimate the activity factor α through performance counts [20].

2) Modeling of leakage power: Unlike dynamic power, leakage power p_s is not directly related to the core's activity. Instead, it depends on the temperature of the chip, and is expressed as [15], [21], [22]

$$p_s = V_{dd} I_{leak}(T_p), \tag{2}$$

where T_p is a scalar representing the temperature at one place of the chip,[1] I_{leak} is the leakage current which is nonlinearly related to temperature.

In this work, we use an n-order polynomial model to accurately model the nonlinear leakage current $I_{leak}(T_p)$ as

$$I_{leak}(T_p) = \alpha_n T_p^n + \alpha_{n-1} T_p^{n-1} + \cdots + \alpha_0. \tag{3}$$

In order to see the accuracy of the leakage model given in (3), Fig. 1 shows an HSPICE simulation result of leakage using $7\,\mathrm{nm}$ PTM-MG FinFET models for high-performance applications ($7\,\mathrm{nm}$ PTM-MG HP NMOS and HP PMOS) provided online at [23], and the leakage computed by the leakage model (with order 3). From the figure, we can see that the leakage model (3) has high accuracy for all common temperatures of multi-core chips.

Finally, the leakage power is accurately modeled by combining equations (2) and (3).

[1] T introduced latter in (4) is a vector representing temperatures at multiple positions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2960414, IEEE Transactions on Industrial Informatics

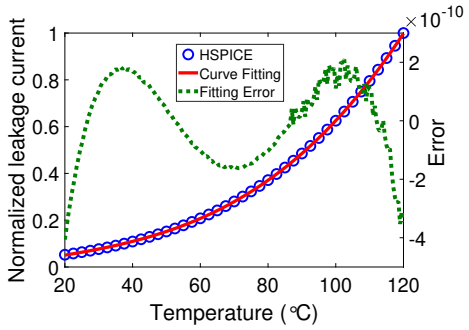


Fig. 1: Comparison of leakage of a PTM-MG 7 nm FinFET from HSPICE simulation with leakage computed using 3-order leakage model (3).

B. Thermal management using model predictive control

In this part, we briefly introduce the MPC based DTM and reveal its difficulty in handling leakage power. For a detailed introduction of MPC based DTM, please refer to [11].

In order to use model predictive control (MPC) in DTM, a thermal model of the multi-core system should be built first. For an l-core system with m total thermal nodes, we can get its thermal model as [3], [11], [15], [17]

$$\begin{aligned} GT(t) + C\frac{\mathrm{d}T(t)}{\mathrm{d}t} &= BP(T,t), \\ Y(t) &= LT(t), \end{aligned} \tag{4}$$

where $T(t) \in \mathbb{R}^m$ is the temperature vector (distinguished from scaler T_p), representing temperatures at m places of the chip and package; $G \in \mathbb{R}^{m\times m}$ and $C \in \mathbb{R}^{m\times m}$ contain equivalent thermal resistance and capacitance information, respectively; $B \in \mathbb{R}^{m\times l}$ contains the power injection topology information; $P(T,t) \in \mathbb{R}^l$ is the power vector of l cores, including both dynamic power vector $P_d(t)$ and leakage power vector $P_s(T,t)$. $Y(t) \in \mathbb{R}^l$ is the output temperatures of l cores; $L \in \mathbb{R}^{l\times m}$ is the output selection matrix which selects the l core temperatures from $T(t)$.

In order to be used in computer, the thermal model (4) is discretized for a given time step h as [24]

$$\begin{aligned} T(k+1) = AT(k) + DP_d(k) \\ + \int_0^h e^{-(h-\tau)C^{-1}G}C^{-1}BP_s(T,\tau)\,\mathrm{d}\tau, \end{aligned} \tag{5}$$

with

$$A = e^{-hC^{-1}G}, \quad D = \int_0^h e^{-(h-\tau)C^{-1}G}C^{-1}B\,\mathrm{d}\tau,$$

where k is the time in discrete form.[2] Note that $A \in \mathbb{R}^{m\times m}$ and $D \in \mathbb{R}^{m\times l}$ are *constant* matrices which are computed *offline* for a given time step h [24].

By using the thermal model (5), MPC calculates the future power recommendation P_d to track a user defined temperature, with the following steps.

First, at current time k, we denote the future dynamic power trajectory (which is unknown and needs to be computed in the end) into the future N_c steps (where N_c is called the *control horizon* in MPC) as

$$\mathcal{P}_d = [P_d(k)^T, P_d(k+1)^T, \ldots, P_d(k+N_c-1)^T]^T. \tag{6}$$

Then, the prediction of core temperatures is written as

$$\mathcal{Y} = [Y(k+1)^T, Y(k+2)^T, \ldots, Y(k+N_p)^T]^T, \tag{7}$$

where N_p is called the *prediction horizon* (with $N_p > N_c$) in MPC and $Y(k+j)$ is the predicted temperatures at time $k+j$ using the information of current time k.

Corresponding to (7), the target temperature vector $Y_g \in \mathbb{R}^l$ is written in a vector trajectory as

$$\mathcal{Y}_g = [Y_g^T, Y_g^T, \ldots, Y_g^T]^T. \tag{8}$$

The objective of the MPC is to bring the predicted output temperature $\mathcal{Y}$ as close as possible to the target temperature $\mathcal{Y}_g$ by adjusting the dynamic power $\mathcal{P}_d$, which is equivalent to minimizing the following cost function

$$\mathcal{J} = (\mathcal{Y}_g - \mathcal{Y})^T(\mathcal{Y}_g - \mathcal{Y}). \tag{9}$$

Please note that $\mathcal{Y}$ is a function of $\mathcal{P}_d$.

Next, optimization is performed to find the $\mathcal{P}_d$ which minimizes (9). However, because there is an integral with the nonlinear P_s in thermal model (5), we cannot express $\mathcal{Y}$ using $\mathcal{P}_d$ (which is the only controllable variable) as the sole variable. Therefore, the optimization problem (9) cannot be solved to find the future dynamic power recommendation, meaning predictive control cannot be directly used for the leakage-aware thermal management.

III. Leakage-aware temperature control using compact piecewise linear model

In this section, we present the new leakage-aware DTM method using compact PWL model based predictive control. The basic idea is to first build a compact PWL thermal model by combining multiple local compact linear thermal models expanded at several Taylor expansion points, and then formulate the predictive control framework based on this compact PWL thermal model. Specifically, we first show how to build the local linear thermal model using Taylor expansion in Section III-A and how to reduce it into a compact local linear thermal model using sampling based MOR in Section III-B. Then, how to formulate the compact PWL thermal model by combining the compact linear thermal models with a systematic Taylor expansion points selection scheme is given in Section III-C. Finally, we present the new predictive control framework based on the compact PWL thermal model in Section III-D.

A. Building local linear thermal model using Taylor expansion

Before presenting the PWL methods, we first show the formulation of the local linear thermal model (at a Taylor expansion point) which will be used in PWL approximation.

First, we can get a local linear leakage power model by performing Taylor expansion on the original nonlinear model (2), (3), expressed in matrix-vector form as

$$P_s = \widehat{P} + \widehat{H}T, \tag{10}$$

[2]We use k to represent the discrete time, and t to represent the continuous time. $k+1$ is equivalent to $t+h$, with h as the discretization time step.

where $\widehat{P} \in \mathbb{R}^l$ is a constant vector not associated with temperature T, and $\widehat{H} \in \mathbb{R}^{l\times m}$ is a constant matrix containing the first order derivative information from Taylor expansion. Due to the page limitation, please refer to [15] for the detailed derivation of (10).

Then, by integrating (10) into (4) and letting $\widehat{G} = G - B\widehat{H}$, we obtain a local linear thermal model as

$$\begin{aligned} \widehat{G}T(t) + C\frac{\mathrm{d}T(t)}{\mathrm{d}t} &= B(P_d(t) + \widehat{P}), \\ Y(t) &= LT(t). \end{aligned} \tag{11}$$

B. Compact local linear thermal model formulation via sampling based MOR

The local linear thermal model built above is large in size, which leads to large computing overhead in DTM with PWL thermal model composed of multiple such local models. Therefore, we introduce a sampling based model order reduction (MOR) technique to build a compact local linear thermal model, which then leads to a compact PWL thermal model.

MOR has been studied intensively to reduce the computing overhead in control systems. The widely used MOR method in control is truncated balanced realization (TBR) [25], which generates a reduced model with a global error bound over all frequencies. However, TBR's two shortcomings make it unsuitable for reducing the local linear thermal model. First, the TBR process includes solving Lyapunov equations, which is known to be computationally expensive for a large original model like the local linear thermal model [26]. Second, the reduction of local linear thermal model prefers high accuracy in low frequency range rather than TBR's global accuracy over all frequencies. This is because the thermal model is a low pass filter, which only respondes to low frequency inputs (powers) [27], [28], preserving accuracy over frequencies (high frequencies in thermal model) with extremely low frequency response is a waste.

Instead, we introduce a sampling based MOR to reduce the local linear thermal model. The sampling based MOR first solves the original system (11) in frequency domain at n_s frequency sample points $j\omega_1, j\omega_2, \ldots, j\omega_{n_s}$, where the ith solution is

$$z_i = (\widehat{G} + j\omega_i C)^{-1}B. \tag{12}$$

All the solutions are then combined into a matrix

$$Z = [z_1, z_1^*, z_2, z_2^*, \ldots, z_{n_s}, z_{n_s}^*], \tag{13}$$

where z_i^* is the conjugate of z_i.[3] Then, economic singular value decomposition (SVD) is performed on matrix Z, and we take the first q columns of the left singular matrix as U. Finally, we obtain the reduced local linear thermal model by projection with the projection matrix U as

$$\begin{aligned} \widehat{G}_r T_r(t) + C_r\frac{\mathrm{d}T_r(t)}{\mathrm{d}t} &= B_r(P_d(t) + \widehat{P}), \\ Y(t) &= L_r T_r(t), \end{aligned} \tag{14}$$

where $\widehat{G}_r = U^T\widehat{G}U \in \mathbb{R}^{q\times q}$, $C_r = U^T C U \in \mathbb{R}^{q\times q}$, $B_r = U^T B \in \mathbb{R}^{q\times l}$, $L_r = LU \in \mathbb{R}^{l\times q}$.

[3]Please note that adding the conjugate is not necessary for the DC sample point because the solution is real.

It has been proved that the left singular matrix of Z contains original system information at the sampled frequency points [29]. Choosing the first q columns of the left singular matrix as projection matrix U further eliminates the redundant information via principal component analysis (PCA), which balances the accuracy and compactness of the reduced model around the sampled frequency points.

Similar to (5), the local linear thermal model (11) can be discretized into the following *compact local linear thermal model* form but without the integral term in (5) as:

$$\begin{aligned} T_r(t+h) &= \widehat{A}(h)T_r(t) + \widehat{D}(h)P_d + \widehat{D}(h)\widehat{P}, \\ Y(t+h) &= L_r T_r(t+h), \end{aligned} \tag{15}$$

with

$$\widehat{A}(h) = e^{-hC_r^{-1}\widehat{G}_r}, \quad \widehat{D}(h) = \int_0^h e^{-(h-\tau)C_r^{-1}\widehat{G}_r}C_r^{-1}B_r\,\mathrm{d}\tau.$$

C. Compact PWL thermal model formulation

In this part, we formulate the compact PWL thermal model using the reduced local linear thermal model presented in previous parts. The compact PWL thermal model can then be integrated into the predictive control framework for leakage-aware DTM.

1) Taylor expansion thermal points selection scheme for leakage-aware DTM: Although there exists PWL approximation based leakage-aware thermal estimation method [15], it is not straightforward to apply similar PWL approximation to DTM due to the difficulty in Taylor expansion thermal points selection. In thermal estimation problem, the Taylor expansion point can be easily chosen by using the self-estimated temperature or the on-chip thermal sensor temperature [15]. However, DTM will not know the proper Taylor expansion points directly, because its computing target is the future power recommendation, not the temperature. The only things that DTM knows are the current temperature, the target temperature, and also the fact that the temperature prediction trajectory (excited by the unknown future power recommendation to be computed) should be between the two temperatures. In this work, we propose a novel Taylor expansion points selection scheme as the following.

First, we define two thermal management cases called *rising case* and *falling case*, depending on the current temperature of the core. We have the falling case if the current temperature is higher than the target temperature. In this case, DTM should lower the core temperature to target temperature for reliability. Otherwise, we have the rising case to boost performance. Here we use the rising case as the illustration example. Please note that DTM for the falling case can be performed in the same way.

Let us denote T_0 as the lowest temperature and T_n as the target temperature of the chip.[4] The operating temperature of the rising case lies between T_0 and T_n. We introduce n potential expansion points in the operating temperature range: $\{T_1, T_2, \ldots, T_n\}$.[5] For simplicity, assume all the potential

[4]Usually, the lowest temperature is set to be the same as or slightly higher than the ambient temperature.

[5]Please note that T_0 is *not* a potential Taylor expansion point.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2960414, IEEE Transactions on Industrial Informatics

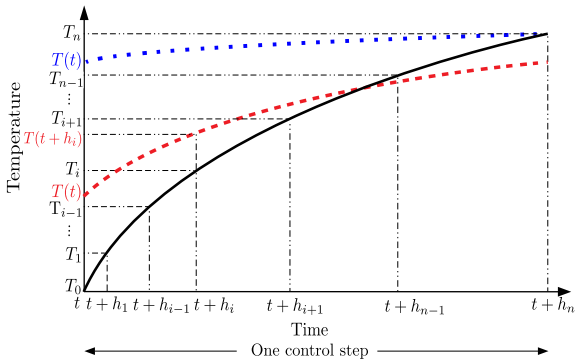


Fig. 2: The sketch map of the PWL method for one control step. $T_1, T_2, \ldots, T_n$ are the potential Taylor expansion points. $t, t+h_1, \ldots, t+h_n$ are the potential local linear model switching time points. The black solid line is the extreme temperature trajectory. The red dashed line is a common temperature trajectory. The blue dot line represents the temperature trajectory which is already very close to the target at time t.

expansion points are uniformly placed in the operating temperature range, i.e., $T_i - T_{i-1} = \frac{T_n - T_0}{n}$ for any integer $i \in [1, n]$, as shown in Fig. 2.

Next, corresponding to the Taylor expansion points, we also need to determine the potential local model switching time points $\{t, t+h_1, \ldots, t+h_n\}$ within one control step. The extreme temperature trajectory in the rising case, which starts from $T(t) = T_0$ and ends at $T(t+h_n) = T_n$ is used to determine these time points. As shown in Fig. 2, the extreme temperature trajectory is the solid black line, and the switching time point $t+h_i$ is chosen as the one which satisfies $T(t+h_i) = T_i$ for this trajectory.

PWL approximation will be performed by using the linear local thermal models constructed using some of these potential Taylor expansion points $\{T_1, T_2, \ldots, T_n\}$ switched only at the corresponding switching time points $\{t, t+h_1, \ldots, t+h_n\}$ as shown in the next part.

2) Compact PWL thermal model for temperature prediction: If the current temperature $T(t)$ lies between T_{i-1} and T_i, the DTM thermal prediction trajectory should look like the red dashed line in Fig. 2 exited by the future power recommendation (which is unknown and need to be computed).[6] For this trajectory, $T_r(t+h_n)$, which is the reduced temperature state of $T(t+h_n)$, can be represented[7] in the following way.

First, $T_r(t+h_i)$, which is the reduced temperature state of $T(t+h_i)$ (shown as red font in Fig. 2), is represented using the compact local linear model (15) expanded at T_i as

$$T_r(t+h_i) = \widehat{A}_i T_r(t) + \widehat{D}_i P_d + \widehat{D}_i \widehat{P}_i, \tag{16}$$

where $\widehat{A}_i = \widehat{A}(h_i)$, $\widehat{D}_i = \widehat{D}(h_i)$, and $\widehat{P}_i$ are the local linear thermal matrices in (15) with T_i as the expansion point.

[6]We plot the end point of the red dashed line to be slightly off target, since predictive control does not guarantee perfect target hitting at the first control step.

[7]$T_r(t+h_n)$ is represented, but not computed, because P_d (the future power recommendation) is the actual unknown to be calculated.

Then, the reduced temperature states at the subsequent time points $t+h_{i+1}, t+h_{i+2}, \ldots, t+h_n$ are represented iteratively by using the compact local linear thermal models expanded at $T_{i+1}, T_{i+2}, \ldots, T_n$, respectively, as the following:

$$\begin{aligned}
T_r(t+h_{i+1}) =& \widehat{A}_{i+1}(U_{i+1})^\dagger U_i T_r(t+h_i) \\
&+ \widehat{D}_{i+1} P_d + \widehat{D}_{i+1}\widehat{P}_{i+1}, \\
T_r(t+h_{i+2}) =& \widehat{A}_{i+2}(U_{i+2})^\dagger U_{i+1} T_r(t+h_{i+1}) \\
&+ \widehat{D}_{i+2} P_d + \widehat{D}_{i+2}\widehat{P}_{i+2}, \\
&\vdots \\
T_r(t+h_n) =& \widehat{A}_n(U_n)^\dagger U_{n-1} T_r(t+h_{n-1}) \\
&+ \widehat{D}_n P_d + \widehat{D}_n \widehat{P}_n,
\end{aligned} \tag{17}$$

where $\widehat{A}_j = \widehat{A}(h_j - h_{j-1})$ and $\widehat{D}_j = \widehat{D}(h_j - h_{j-1})$ for $j = i+1, i+2, \ldots, n$. Please note that we have performed the reduced temperature state transformation in (17) using projection matrices at two adjacent expansion points to keep the boundary continuous at compact PWL local model switching, as shown in details in [15]. For example, $(U_{i+1})^\dagger U_i$ is multiplied to $T_r(t+h_i)$ in (17), where U_{i+1} and U_i are the projection matrices in (14) with T_i and T_{i+1} as the expansion points, respectively, and † denotes Moore-Penrose pseudoinverse.

Finally, the reduced temperature state at the end of the control step $(t+h_n)$ is expressed by combining the equations above as

$$T_r(t+h_n) = \widehat{\mathcal{A}} T_r(t) + \widehat{\mathcal{D}} P_d + \widehat{\mathcal{D}}_i \widehat{P}_i + \cdots + \widehat{\mathcal{D}}_n \widehat{P}_n, \tag{18}$$

where

$$\begin{aligned}
\widehat{\mathcal{A}} =& \widehat{A}_n(U_n)^\dagger U_{n-1} \widehat{A}_{n-1} \cdots \widehat{A}_{i+1}(U_{i+1})^\dagger U_i \widehat{A}_i, \\
\widehat{\mathcal{D}} =& \widehat{A}_n(U_n)^\dagger U_{n-1} \widehat{A}_{n-1} \cdots \widehat{A}_{i+1}(U_{i+1})^\dagger U_i \widehat{D}_i \\
&+ \widehat{A}_n(U_n)^\dagger U_{n-1} \widehat{A}_{n-1} \cdots \widehat{A}_{i+2}(U_{i+2})^\dagger U_{i+1} \widehat{D}_{i+1} \\
&+ \cdots + \widehat{D}_n, \\
\widehat{\mathcal{D}}_i =& \widehat{A}_n(U_n)^\dagger U_{n-1} \widehat{A}_{n-1} \cdots \widehat{A}_{i+1}(U_{i+1})^\dagger U_i \widehat{D}_i.
\end{aligned}$$

In order to be compatible with MPC, we rewrite (18) into the discrete form as

$$\begin{aligned}
T_r(k+1) &= \widehat{\mathcal{A}} T_r(k) + \widehat{\mathcal{D}} P_d(k) + \widehat{\mathcal{D}}_i \widehat{P}_i + \cdots + \widehat{\mathcal{D}}_n \widehat{P}_n, \\
Y(k+1) &= L_r T_r(k+1).
\end{aligned} \tag{19}$$

We call this newly formulated thermal model (19) as the *compact PWL thermal model*. The compact PWL thermal model matrices will be computed offline after Taylor expansion points selection.

Now, we have successfully approximated the original nonlinear temperature prediction using the compact PWL thermal model in (19). Next, we will demonstrate how to formulate the compact PWL thermal model based predictive control by replacing the original nonlinear thermal prediction (5) with the compact PWL model based thermal prediction (19).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2960414, IEEE Transactions on Industrial Informatics

D. Compact PWL model based predictive control

With the compact PWL thermal model (19), MPC should be able to calculate the power recommendation P_d to track a user defined output temperature as presented in this part.

By analyzing the MPC mechanism, we know the future temperature prediction trajectory can be described as the following. For the first control time step[8] into the future, the temperature prediction trajectory is similar to the red dashed line in Fig. 2, because the power recommendation will bring the temperature toward the target temperature. Assume the temperature prediction is close to the target temperature at time $k+1$, then at time $k+j$, where $j = 2, 3, \ldots, N_p$, all temperature prediction trajectories should look like the blue dot line in Fig. 2.

With the observation above, for N_p steps temperature prediction into the future (from k to $k+N_p$), we only need to use the temperature prediction with multiple Taylor expansion points at the first control step (from time k to $k+1$) expressed by the PWL thermal model (19).

For the rest of the control steps (from $k+1$ to $k+N_p$), only one segment of the PWL thermal model (15) is needed with target temperature Y_g (which equals to T_n) as the expansion point. The temperature predictions for these steps are written into discrete form as

$$\begin{aligned} T_r(k+j) &= \widehat{A}_t T_r(k+j-1) + \widehat{D}_t P_d(k+j-1) + \widehat{D}_t \widehat{P}_t, \\ Y(k+j) &= L_r T_r(k+j), \end{aligned} \tag{20}$$

where the matrices A_t, D_t, and P_t are obtained by setting the time discretization step as h_n and Taylor expansion point as the target temperature in (15), for $j = 2, 3, \ldots, N_p$.

Combining equations (19) and (20), we can get the predicted temperature trajectory $\mathcal{Y}$ as

$$\mathcal{Y} = F T_r(k) + V \mathcal{P}_d + \phi_1 \widehat{\mathcal{P}} + \phi_2 \widehat{\mathcal{P}}_t, \tag{21}$$

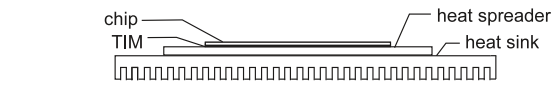
where $\widehat{\mathcal{P}} = [\widehat{P}_i^T, \widehat{P}_{i+1}^T, \ldots, \widehat{P}_n^T]^T$, $\widehat{\mathcal{P}}_t = [\mathbf{0}^T, \widehat{P}_t^T, \ldots, \widehat{P}_t^T]^T$,

$$F = \begin{bmatrix} L_r \widehat{\mathcal{A}} \\ L_r \widehat{A}_t \widehat{\mathcal{A}} \\ \vdots \\ L_r \widehat{A}_t^{N_p-1} \widehat{\mathcal{A}} \end{bmatrix},$$

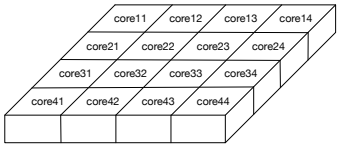
$$V = \begin{bmatrix} L_r \widehat{\mathcal{D}} & \mathbf{0} & \cdots & \mathbf{0} \\ L_r \widehat{A}_t \widehat{\mathcal{D}} & L_r \widehat{D}_t & \cdots & \mathbf{0} \\ L_r \widehat{A}_t^2 \widehat{\mathcal{D}} & L_r \widehat{A}_t \widehat{D}_t & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ L_r \widehat{A}_t^{N_p-1} \widehat{\mathcal{D}} & L_r \widehat{A}_t^{N_p-2} \widehat{D}_t & \cdots & L_r \widehat{A}_t^{N_p-N_c} \widehat{D}_t \end{bmatrix},$$

$$\phi_1 = \begin{bmatrix} L_r \widehat{\mathcal{D}}_i & L_r \widehat{\mathcal{D}}_{i+1} & \cdots & L_r \widehat{\mathcal{D}}_n \\ L_r \widehat{A}_t \widehat{\mathcal{D}}_i & L_r \widehat{A}_t \widehat{\mathcal{D}}_{i+1} & \cdots & L_r \widehat{A}_t \widehat{\mathcal{D}}_n \\ L_r \widehat{A}_t^2 \widehat{\mathcal{D}}_i & L_r \widehat{A}_t^2 \widehat{\mathcal{D}}_{i+1} & \cdots & L_r \widehat{A}_t^2 \widehat{\mathcal{D}}_n \\ \vdots & \vdots & \ddots & \vdots \\ L_r \widehat{A}_t^{N_p-1} \widehat{\mathcal{D}}_i & L_r \widehat{A}_t^{N_p-1} \widehat{\mathcal{D}}_{i+1} & \cdots & L_r \widehat{A}_t^{N_p-1} \widehat{\mathcal{D}}_n \end{bmatrix},$$

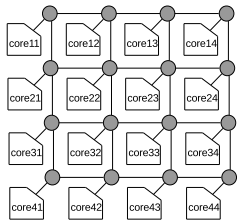
[8]Please note that the duration from t to $t+h_n$ in Fig. 2 equals to only one control step in MPC (for example, from k to $k+1$, or from $k+j-1$ to $k+j$).



(a) The package structure of the 16-core system. TIM stands for thermal interface material.



(b) The chip structure of the 16-core system.



(c) The architectural diagram of the 16-core system with cores connected by an NoC [30]. The circle denotes the router in the NoC.

Fig. 3: The configuration of the 16-core system used for the experiment.

$$\phi_2 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & L_r \widehat{D}_t & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & L_r \widehat{A}_t \widehat{D}_t & L_r \widehat{D}_t & \cdots & \mathbf{0} \\ \mathbf{0} & L_r \widehat{A}_t^2 \widehat{D}_t & L_r \widehat{A}_t \widehat{D}_t & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & L_r \widehat{A}_t^{N_p-2} \widehat{D}_t & L_r \widehat{A}_t^{N_p-3} \widehat{D}_t & \cdots & L_r \widehat{A}_t^{N_p-N_c} \widehat{D}_t \end{bmatrix},$$

with $\mathbf{0}$ as the zero matrix with suitable size.

Plugging (21) into (9), standard MPC optimization is performed to minimize (9) by making the first derivative of (9) (with respect to $\mathcal{P}_d$) equal to zero. The solution of $\mathcal{P}_d$ is

$$\mathcal{P}_d = (V^T V + R)^{-1} V^T (\mathcal{Y}_g - F T_r(k) - \phi_1 \widehat{\mathcal{P}} - \phi_2 \widehat{\mathcal{P}}_t). \tag{22}$$

At each MPC time k, only $P_d(k)$ (the first element of $\mathcal{P}_d$) will be outputted as the power recommendation for thermal management.

The proposed approach can be integrated into the thermal management of the multi-core system as a software implementation. It can be executed in one of the cores in the multi-core system or in a co-processor (if available), which distributes the results $P_d(k)$ to all the cores. Frequencies and task loads of the multi-core system will be adjusted according to $P_d(k)$. How to perform the management actions based on future power recommendation is presented in many DTM works such as [11], which will not be given here due to page limitation.

IV. Experimental results and discussions

In this section, we evaluate the performance of the compact PWL model based DTM method.

A. Basic experiment setup

The experiment is performed on a 16-core system plant with its package, chip structure, and architectural diagram shown in Fig. 3, where the cores are connected by a network-on-chip (NoC) as given in [30]. When the DTM computation is performed on one of the cores, the temperature information is gathered from all cores and the computed power adjustment information is sent to all cores through the NoC. The delay of information gathering and dispatching is around 100 cycles

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2960414, IEEE Transactions on Industrial Informatics

for the NoC [30]. The thermal models for the plant, the new method, and the traditional method share the same resolution of 20×20 for the chip, i.e., each core has 25 grids (thermal nodes). The plant provides the average temperature of each core for all DTM methods at the beginning of each control step. The ambient temperature is $20\,^{\circ}\mathrm{C}$, and the target temperature in DTM is $70\,^{\circ}\mathrm{C}$. We set the operating temperature range for the rising case as from around $30\,^{\circ}\mathrm{C}$ to $70\,^{\circ}\mathrm{C}$, and the range for the falling case as from around $110\,^{\circ}\mathrm{C}$ to $70\,^{\circ}\mathrm{C}$. All the experiments are performed on a PC with an Intel Core i7-8750H CPU and 16 GB memory.

Power estimator Wattch [31] is used to generate the dynamic power by running the standard SPEC benchmarks. The dynamic power traces from different SPEC benchmark applications are randomly assigned to different cores to create realistic thermal loads. The golden leakage power of the multi-core system plant is obtained by using the iteration based leakage-aware thermal simulation method with simulation step $0.01\,\mathrm{s}$. The power of the NoC is ignored in the experiment, because it is very small (less than $2\,\mathrm{mW}$ for each router including both dynamic power and leakage power) compared with the power of the core (several watts for each core) as analyzed in [32]. The control step of DTM is set as $1\,\mathrm{s}$. We assume the average temperature of each core is measurable at the beginning of each control step. Then the temperature state (T or T_r) is estimated by Kalman filter using the measured temperatures.

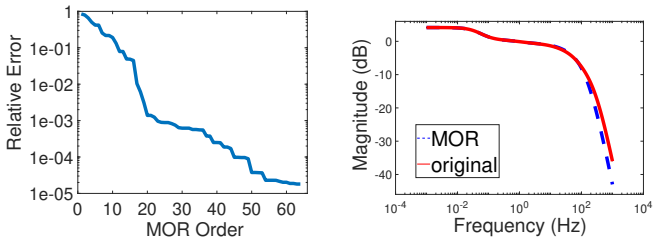
In order to show the advantage of the new DTM method with compact PWL model based predictive control (we call it *PWL DTM*), we compare it with two methods. The first one is the linear model based predictive control [11] (called the *traditional* DTM), which uses a least square regression based linear leakage model. We also compare the new method with the state-of-the-art leakage-aware DTM method [19] which is based on the echo state network (ESN) thermal model (called ESN DTM). The ESN thermal model has 200 neurons, which balances accuracy and speed. It is trained using 40000 input-output (dynamic power-temperature) samples generated by the plant covering the full temperature range (from $30\,^{\circ}\mathrm{C}$ to $140\,^{\circ}\mathrm{C}$) in DTM.

B. Accuracy verification of the compact local linear thermal model

First, we analyze and verify the accuracy of the compact local linear thermal model generated by the sampling based MOR.

We use three sampling points at $0\,\mathrm{Hz}$, $0.5\,\mathrm{Hz}$, and $1\,\mathrm{Hz}$ to reduce the 1612-order original model into compact models with different orders. Then, we apply a $60\,\mathrm{W}$ step power input to core11, and plot the average transient simulation errors (for the first $1\,\mathrm{s}$ simulation) of the reduced models with different orders in Fig. 4a. We see that the reduced model with order 32 already reaches high accuracy (with average relative error to be smaller than 0.001).

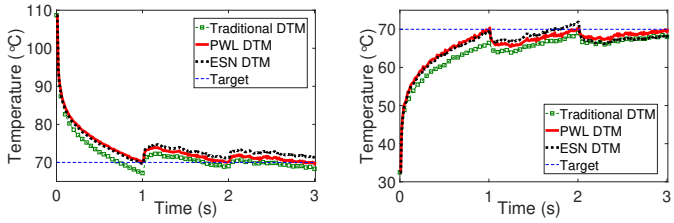
Then, we plot the frequency responses of the 32-order reduced model and the original model in Fig. 4b (with both input and output set at the grid at the upper left corner of the chip). From the figure, we can see that the reduced model, although much smaller in size, has a very similar frequency response compared with the original model, for all the important frequencies from DC to $100\,\mathrm{Hz}$. It only shows visible errors beyond $100\,\mathrm{Hz}$, but such errors do not affect the thermal estimation accuracy because the magnitudes at these frequencies are extremely small.



(a) Accuracy of the reduced model with different order.

(b) Frequency response comparison of the 32-order model and the 1612-order original model.

Fig. 4: Accuracy analysis of the compact local thermal model generated by the sampling based MOR.



(a) The falling case.

(b) The rising case.

Fig. 5: The transient temperatures of core22 using PWL DTM (with 7 expansion points), traditional DTM, and ESN DTM. All methods have $N_P = 1$ and $N_c = 1$. The temperature target is set as $70\,^{\circ}\mathrm{C}$. Power traces of different SPEC benchmark applications are randomly assigned to different cores.

C. Performance evaluation of leakage-aware DTM with compact PWL model based predictive control

We apply PWL DTM, traditional DTM, and ESN DTM to the 16-core system, and record the temperature control performance results in Table I. In order to see the accuracy of the PWL DTM with different configurations, we test it with different expansion point number, prediction horizon length N_p, and control horizon length N_c.

We mainly focus on two DTM performances in the comparison. The first is the average temperature tracking difference between the actual plant temperature and the target temperature for the first three control steps, which indicates the effectiveness and accuracy of the DTM. The second is the overhead (computing time and memory cost) of the DTM for each control step.

For traditional DTM, the difference between the actual temperature and the target temperature is large for all cases as shown in Table I. Even for the best case, the average difference is over $1.5\,^{\circ}\mathrm{C}$, because the linear model cannot approximate the nonlinearity in leakage power accurately.

For ESN DTM, the temperature tracking difference is smaller than the traditional DTM, because ESN, being a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2960414, IEEE Transactions on Industrial Informatics

TABLE I: Computing time (time), storage memory (mem), and tracking difference (difference) comparison results of the PWL DTM with traditional DTM and ESN DTM. Computing time is recorded as the average computing time for each thermal management action (every 1 s), including the state estimation time by Kalman filter. The tracking difference is in °C.

Case	Methods	$N_c = 1, N_p = 1$				$N_c = 3, N_p = 4$				$N_c = 5, N_p = 7$			
		time (s)	mem (KB)	difference (°C)	speed up	time (s)	mem (KB)	difference (°C)	speed up	time (s)	mem (KB)	difference (°C)	speed up
rise	Traditional	0.21	18488	1.77	NA	0.23	18488	1.74	NA	0.24	18488	1.76	NA
	ESN	0.0035	113	1.29	NA	0.087	113	1.29	NA	0.22	113	1.60	NA
	PWL (3 points)	0.0021	33	0.438	102×	0.0025	33	0.448	91×	0.0025	33	0.439	95×
	PWL (5 points)	0.0023	55	0.435	93×	0.0026	55	0.444	87×	0.0026	55	0.436	92×
	PWL (7 points)	0.0024	77	0.433	89×	0.0026	77	0.441	87×	0.0028	77	0.434	85×
	PWL (9 points)	0.0025	99	0.432	86×	0.0026	99	0.440	87×	0.0028	99	0.433	85×
	PWL (11 points)	0.0026	121	0.431	82×	0.0027	121	0.439	84×	0.0029	121	0.432	82×
fall	Traditional	0.22	18488	1.58	NA	0.23	18488	1.60	NA	0.24	18488	1.58	NA
	ESN	0.0043	113	0.93	NA	0.090	113	0.82	NA	0.23	113	0.69	NA
	PWL (3 points)	0.0021	33	0.225	105×	0.0025	33	0.216	92×	0.0025	33	0.224	96×
	PWL (5 points)	0.0022	55	0.221	100×	0.0025	55	0.213	92×	0.0027	55	0.219	89×
	PWL (7 points)	0.0024	77	0.218	92×	0.0026	77	0.211	87×	0.0028	77	0.216	86×
	PWL (9 points)	0.0024	99	0.216	92×	0.0027	99	0.210	85×	0.0029	99	0.215	83×
	PWL (11 points)	0.0025	121	0.215	88×	0.0028	121	0.209	82×	0.0029	121	0.214	83×

nonlinear model, is able to model the nonlinearity of leakage power. However, ESN DTM still shows some temperature tracking difference of around 0.6 °C to 1.6 °C as shown in Table I. The main reason is that the ESN thermal model is a black-box model trained directly from the input-output data samples, which lacks the detailed structural information of the packaged multi-core system.

On the contrary, for the PWL DTM, the temperature tracking difference is smaller than the traditional DTM for all cases. This tracking accuracy improvement is achieved by approximating the nonlinearity accurately using the PWL thermal model. Especially, the average tracking difference is only 0.209 °C when the number of expansion points is 11 for the falling case with $N_c = 3$ and $N_p = 4$. The PWL DTM also shows higher temperature tracking accuracy than the ESN DTM. The reason is that the PWL thermal model is a white-box model directly built from the detailed system structural information using physical laws, which has an accuracy advantage to the data training based black-box model in ESN DTM.

On the runtime side, we observe from Table I that the computing time of the PWL DTM is much smaller than the traditional method, with up to 105× speed up. This is because PWL DTM is based on the compact PWL thermal model generated by sampling based MOR (composed of 32-order local linear thermal models), which is much smaller than the original thermal model (with order 1612) used in traditional DTM. With much faster computing speed, PWL DTM still achieves high temperature tracking accuracy as discussed previously. PWL DTM also shows advantage to ESN DTM in computing time, because ESN DTM has to perform iterations to solve the nonlinear optimization problem in DTM as its thermal model is nonlinear.

The memory cost of PWL DTM is also much smaller than the traditional DTM thanks to its compact PWL model generated by sampling based MOR. The memory cost of PWL DTM increases linearly with the expansion point number as shown in Table I. This is because more matrices computed offline need to be stored, such as the PWL thermal model matrices F and ϕ_1, which is a trade-off between accuracy and overhead in PWL DTM. The memory cost of PWL DTM exceeds the 200-neuron ESN DTM when its expansion point number is 11, but the accuracy of PWL DTM is much higher than the 200-neuron ESN DTM with this expansion point number.

Finally, we plot the transient plant temperature comparison results for both falling case and rising case in Fig. 5 by assigning power traces of different SPEC benchmark applications randomly to different cores and activating all DTMs at 0 s. We only plot the results of core22 due to page limitation. It is observed that the temperature controlled by traditional DTM shows large tracking difference especially when the current temperature is far from target (the first control step from 0 s to 1 s). ESN DTM, although being more accurate than the traditional DTM, still shows visible tracking differences for all three control steps. On the other hand, the temperature controlled by PWL DTM tracks the target accurately even for the first control step. This clearly demonstrates the advantage of PWL DTM in temperature control quality.

In summary, experimental results show that PWL DTM outperforms both the traditional DTM and ESN DTM in temperature control quality because the PWL thermal model approximates the nonlinear leakage power effects accurately. It even achieves lower computing overhead thanks to the compact thermal model generated by the sampling based MOR.

V. Conclusion and Future Work

In this paper, we have proposed a new leakage-aware DTM method for multi-core systems using compact PWL model based predictive control. We built a compact PWL thermal model by combining multiple compact local linear thermal models which are expanded at several Taylor expansion points. These compact local linear thermal models are obtained by sampling based MOR with Taylor expansion points selected by a systematic scheme which exploits the thermal behavior property of the multi-core chips. Based on the compact PWL thermal model, predictive control is used to find the opti-

mal future power recommendations for thermal management. Experimental results show the new method outperforms the linear model based MPC method and the echo state network based predictive thermal management method in temperature management quality with lower computing overhead.

With PWL approximation of the nonlinearity between temperature and leakage power, the new method is accurate with DTM actions including dynamic frequency scaling and task migration. In order to work with dynamic voltage scaling, the nonlinearity between supply voltage and leakage power should also be considered. The future research direction is to extend the 1D PWL model based DTM to 2D PWL model based DTM which also approximates the nonlinearity between supply voltage and leakage power.

REFERENCES

[1] G. Quan and V. Chaturvedi, "Feasibility analysis for temperature-constraint hard real-time periodic tasks," *IEEE Trans. on Industrial Informatics*, vol. 3, no. 6, pp. 329–339, Aug. 2010.

[2] E. Rotem, A. Naveh, D. Rajwan, A. Ananthakrishnan, and E. Weissmann, "Power-management architecture of the intel microarchitecture code-named sandy bridge," *IEEE MICRO*, vol. 32, no. 2, pp. 20–27, March–April 2012.

[3] H. Wang, D. Tang, M. Zhang, S. X.-D. Tan, C. Zhang, H. Tang, and Y. Yuan, "GDP: A greedy based dynamic power budgeting method for multi/many-core systems in dark silicon," *IEEE Trans. on Computers*, vol. 68, no. 4, pp. 526–541, Apr. 2019.

[4] A. Schranzhofer, J.-J. Chen, and L. Thiele, "Dynamic power-aware mapping of applications onto heterogeneous MPSoC platforms," *IEEE Trans. on Industrial Informatics*, vol. 6, no. 4, pp. 692–707, Nov. 2010.

[5] X. Wang, X. Fu, X. Liu, and Z. Gu, "PAUC: Power-aware utilization control in distributed real-time systems," *IEEE Trans. on Industrial Informatics*, vol. 6, no. 3, pp. 302–315, Aug. 2010.

[6] B. Zhao, H. Aydin, and D. Zhu, "On maximizing reliability of real-time embedded applications under hard energy constraint," *IEEE Trans. on Industrial Informatics*, vol. 3, no. 6, pp. 316–328, Aug. 2010.

[7] J. Donald and M. Martonosi, "Techniques for multicore thermal management: Classification and new exploration," in *Proc. Int. Symp. on Computer Architecture (ISCA)*, pp. 78–88, Jun. 2006.

[8] D. Shin, S. W. Chung, E.-Y. Chung, and N. Chang, "Energy-optimal dynamic thermal management: Computation and cooling power co-optimization," *IEEE Trans. on Industrial Informatics*, vol. 3, no. 6, pp. 340–351, Aug. 2010.

[9] Z. Liu, S. X.-D. Tan, X. Huang, and H. Wang, "Task migrations for distributed thermal management considering transient effects," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 23, no. 2, pp. 397–401, Feb. 2015.

[10] C. Tan, T. Muthukaruppan, T. Mitra, and L. Ju, "Approximation-aware scheduling on heterogeneous multi-core architectures," in *Proc. Asia South Pacific Design Automation Conf. (ASP-DAC)*, pp. 618–623, 2015.

[11] H. Wang, J. Ma, S. X.-D. Tan, C. Zhang, H. Tang, K. Huang, and Z. Zhang, "Hierarchical dynamic thermal management method for high-performance many-core microprocessors," *ACM Trans. on Design Automation of Electronic Systems*, vol. 22, no. 1, pp. 1:1–1:21, Jul. 2016.

[12] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach, Fifth Edition*. Elsevier, 2012.

[13] A. Rowe, K. Lakshmanan, H. Zhu, and R. Rajkumar, "Rate-harmonized scheduling and its applicability to energy management," *IEEE Trans. on Industrial Informatics*, vol. 6, no. 3, pp. 265–275, Aug. 2010.

[14] Q. Xie, X. Lin, Y. Wang, S. Chen, M. J. Dousti, and M. Pedram, "Performance comparisons between 7-nm FinFET and conventional bulk CMOS standard cell libraries," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 62, no. 8, pp. 761–765, 2015.

[15] H. Wang, J. Wan, S. X.-D. Tan, C. Zhang, H. Tang, Y. Yuan, K. Huang, and Z. Zhang, "A fast leakage-aware full-chip transient thermal estimation method," *IEEE Trans. on Computers*, vol. 67, no. 5, pp. 617–630, May. 2018.

[16] V. Chaturvedi, H. Huang, and G. Quan, "Leakage aware scheduling on maximum temperature minimization for periodic hard real-time systems," in *Proc. International Conference on Computer and Information Technology*, pp. 1802–1809, 2010.

[17] V. Hanumaiah, S. Vrudhula, and K. Chatha, "Performance optimal online DVFS and task migration techniques for thermally constrained multi-core processors," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 30, no. 11, pp. 1677–1690, Nov. 2011.

[18] B. Shi and A. Srivastava, "Dynamic thermal management considering accurate temperature-leakage interdependency," in *ENCYCLOPEDIA OF THERMAL PACKAGING: Thermal Packaging Tools*, pp. 39–60. World Scientific, 2015.

[19] H. Wang, X. Guo, S. X.-D. Tan, C. Zhang, H. Tang, and Y. Yuan, "Leakage-aware predictive thermal management for multi-core systems using echo state network," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 2019.

[20] H. Wang, S. X.-D. Tan, X.-X. Liu, and A. Gupta, "Runtime power estimator calibration for high-performance microprocessors," in *Proc. European Design and Test Conf. (DATE)*, pp. 352–357, Mar. 2012.

[21] Y. Liu, R. Dick, L. Shang, and H. Yang, "Accurate temperature-dependent integrated circuit leakage power estimation is easy," in *Proc. European Design and Test Conf. (DATE)*, pp. 1–6, 2007.

[22] R. Shen, S. X.-D. Tan, H. Wang, and J. Xiong, "Fast statistical full-chip leakage analysis for nanometer VLSI systems," *ACM Trans. on Design Automation of Electronic Systems*, vol. 17, no. 4, pp. 51:1–51:19, Oct. 2012.

[23] Predictive technology model. [Online]. Available: http://ptm.asu.edu

[24] E. Rotem, A. Naveh, D. Rajwan, A. Ananthakrishnan, and E. Weissmann, "TILTS: A fast architectural-level transient thermal simulation method," *Journal of Low Power Electronics*, vol. 3, no. 1, pp. 13–21, Apr. 2007.

[25] B. C. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," *IEEE Trans. on Automatic Control*, vol. 26, no. 1, pp. 17–32, Feb. 1981.

[26] S. X.-D. Tan and L. He, *Advanced Model Order Reduction Techniques in VLSI Design*. Cambridge University Press, 2007.

[27] K. Skadron, M. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, "Temperature-aware microarchitecture," in *Proc. Int. Symp. on Computer Architecture (ISCA)*, pp. 2–13, Jun. 2003.

[28] H. Wang, S. X.-D. Tan, G. Liao, R. Quintanilla, and A. Gupta, "Full-chip runtime error-tolerant thermal estimation and prediction for practical thermal management," in *Proc. Int. Conf. on Computer Aided Design (ICCAD)*, pp. 716–723, Nov. 2011.

[29] H. Wang, S. X.-D. Tan, and R. Rakib, "Compact modeling of interconnect circuits over wide frequency band by adaptive complex-valued sampling method," *ACM Trans. on Design Automation of Electronic Systems*, vol. 17, no. 1, pp. 5:1–5:22, Jan. 2012.

[30] L. Huang, S. Chen, Q. Wu, M. Ebrahimi, J. Wang, S. Jiang, and Q. Li, "A lifetime-aware mapping algorithm to extend MTTF of networks-on-chip," in *Proc. Asia South Pacific Design Automation Conf. (ASP-DAC)*, Jan. 2018.

[31] D. Brooks, V. Tiwari, and M. Martonosi, "Wattch: A framework for architectural-level power analysis and optimizations," in *Proc. Int. Symp. on Computer Architecture (ISCA)*, pp. 83–94, Jun. 2000.

[32] L. Huang, J. Wang, M. Ebrahimi, M. Daneshtalab, X. Zhang, G. Li, and A. Jantsch, "Non-blocking testing for network-on-chip," *IEEE Trans. on Computers*, vol. 65, no. 3, pp. 679–692, Mar. 2016.

Hai Wang (M'19) received the B.S. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2007, and the M.S. and Ph.D. degrees from the University of California at Riverside, Riverside, CA, USA, in 2008 and 2012, respectively.

He is currently an associate professor with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include modeling, optimization, and artificial intelligence assisted design automation of VLSI circuits and systems.

Dr. Wang was a recipient of the Best Paper Award nomination from ASP-DAC in 2019. He has served as a Technical Program Committee Member of several international conferences, including DATE, ASP-DAC and ISQED, and also served as a Reviewer of many journals including the IEEE Transactions on Computers, the IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, the IEEE Transactions on Parallel and Distributed Systems, and ACM Transactions on Design Automation of Electronic Systems.

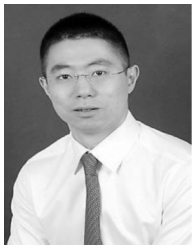
Liwen Hu received the bachelor's degree from Xi'an University of Posts & Telecommunications in 2017. Currently, he is a master student at the University of Electronic Science and Technology of China.

His current research interests include deep learning, thermal management of integrated circuit and thermal map recovery.

Xingxing Guo received the bachelor's degree from Anhui University in 2016 and the master's degree from the University of Electronic Science and Technology of China in 2019.

Her current research interests include deep learning, thermal analysis, power analysis, and thermal management of integrated circuit and building systems. She was a recipient of the Best Paper Award nomination from ASP-DAC in 2019.

Yang Nie received the M.E. degree from Beijing University of Posts and Telecommunications, China in 2010, and also got a double M.E. degree from Tokushima University, Japan in 2010. He is currently a Chief Technical Officer of Sichuan Haoxunda Technology. His research mainly focuses on natural language processing, computer vision, machine learning, and their industrial applications. He is a member of Chinese Association for Artificial Intelligence, and he got a patent in the field of natural language processing (CN102081598B).



He Tang (M'09) received the BSEE degree from the University of Electronic Science and Technology of China, Chengdu, China, the MS degree in electrical and computer engineering from the Illinois Institute of Technology, Chicago, and the PhD degree in electrical engineering from University of California, Riverside, in 2005, 2007, and 2010. From 2010 to 2012, he was with OmniVision Technologies Inc., in Santa Clara, California, as an Analog IC Designer, where he worked on high-speed I/O interface. Since 2012, he has been an associate professor and subsequently a professor with the University of Electronic Science and Technology of China, Chengdu, China. He has authored or coauthored more than 40 papers. His research interests focus on data converters and analog/mixed-signal IC designs. His past work includes high-speed high-resolution pipelined ADCs with digital calibration and high-performance ultra-low-power SAR ADCs. He has served on IEEE CAS Analog Signal Processing Technical Committee (ASPTC) since 2013. He is a member of the IEEE.