LMS-based Adaptive Temperature Prediction Scheme for Proactive Thermal-aware Three-Dimensional Network-on-Chip Systems

Kun-Chih Chen, Huai-Ting Li, and An-Yeu (Andy) Wu
Graduate Institute of Electronics Engineering, National Taiwan University, Taipei 10617, Taiwan

ABSTRACTION

The three-dimensional Network-on-Chip (3D NoC) has been proposed to solve the complex on-chip communication issues. Because of the die-stacking architecture, the thermal problem becomes more severe than in 2D NoC. To simultaneously consider the thermal safety and system performance, proactive thermal management (*PDTM*) has been proved as an efficient way to control the system temperature against overheat. Based on the information of predictive temperature, the *PDTM* can early control the system temperature. To predict the future temperature, adopting the Thermal Resistance and Capacitance (*Thermal RC*) model is a popular way to derive the thermal prediction scheme. However, the *Thermal RC* value is sensitive to temperature changes, which affect the accuracy of the future temperature estimation. Therefore, the current proactive thermal-aware NoC system still suffers from large performance impact because of imprecise future temperature estimation. In this paper, we propose an LMS-based adaptive thermal prediction (*LMS-ATP*) model, which can adaptively adjust the involved *Thermal RC* values for future temperature estimation. The experimental results show that the proposed *LMS-ATP* model can improve the precision of future temperature estimation by 72.96%. In addition, the system throughput can be enhanced by around 0.77% to 47.96%.

1. INTRODUCTION

As the complexity of System-on-Chip (SoC) grows with Moore's law, the three-dimensional Network-on-Chip (3D NoC) has been proposed to provide larger interconnection bandwidth to achieve higher performance with lower power consumption [1]. However, the thermal issues have emerged as the main challenges of 3D NoC due to die stacking [1][2]. The thermal issue results in large performance impact and increases the leakage power [3]. The increasing leakage power may further increase the temperature, which results in thermal runaway [4].

To keep the system operating under the safe temperature, the Dynamic Thermal Management (*DTM*) scheme is usually employed to prevent the NoC system from overheating and being damaged. Conventionally, for emergent cooling, the system is shut down (or fully throttled) as the temperature of the system reaches the alarming level, which results in large performance impact. Recently, the Proactive *DTM* (*PDTM*) has been proved as an efficient way to mitigate the performance impact caused by the reactive *DTM* [5]. By using the temperature prediction scheme, the future temperature can be estimated and the *PDTM* can early control the system temperature based on the information of the predictive temperature. To estimate the future temperature, employing the Thermal Resistance and Capacitance (*Thermal RC*) model is a popular way to derive the thermal prediction scheme [5][7].

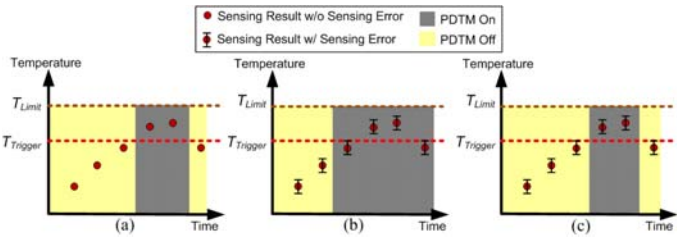

Fig. 1 (a) Ideal temperature estimation results ideal temperature control, (b) imprecise temperature estimation results in lower system availability, and (c) *LMS-ATP* can make temperature control scheme become more precise.

Although the *PDTM* can improve the system performance due to early temperature control, the accuracy of temperature estimation still affects the system performance and the reliability [6] because of temperature-varying *Thermal RC* values [8]. The *PDTM* control policy usually relies on the results of temperature prediction, and imprecise temperature estimation may result in late or early activation of the employed *PDTM*, as shown in Fig. 1(a)(b). Late activation of *PDTM* results in lower reliability because the temperature may exceed the hard thermal limit. On the other hand, early activation of *PDTM* would result in significant performance degradation. Therefore, accurate temperature estimation is necessary for guiding *PDTM* to make accurate and effective decisions.

To calibrate the results of future temperature estimation, many calibration schemes were proposed. In [9], the authors proposed to use Kalman filter to calibrate the results of imprecise temperature estimation. However, the disadvantage of the method employing Kalman filter is the large computational complexity and large area overhead. In [10], the authors separated the temperature estimation errors into Gaussian-typed ones and Non-Gaussian ones. Then, the authors proposed a statistical technique for the specific type of noise. Nevertheless, the accuracy of temperature estimation highly depends on the completeness of the offline analysis by using the statistic method. In [11], the authors employed the *Thermal RC* model to derive an error compensation term, which will be used to calibrate the results of temperature estimation. However, the authors assumed the error compensation term to be temperature independent, which may affect the efficiency of the calibration.

In this paper, we assume there is one embedded thermal sensor at each NoC tile, and each thermal sensor can provide precise sensing results of current temperature. Based on the sensing results of the embedded thermal sensor, the predicted temperature can be obtained by using the *Thermal RC*-based Temperature Prediction (*RCTP*) scheme proposed in [5]. To calibrate the results of temperature prediction, we employ the Least-Mean-Square (LMS) adaptive filters to propose an LMS-based adaptive thermal prediction (*LMS-ATP*) model, which can help to improve the controlling efficiency of the involved *PDTM* scheme. The contributions of this paper are summarized as follows for clarification:

1) *Adaptive Thermal RC Value Fitting:* By applying the LMS adaptive filter, we can converge the involved *Thermal RC* value in the *RCTP* scheme based on the estimation error of temperature prediction.
2) *Online Temperature Calibration:* Because the involved *Thermal RC* value can be converged within a short response time, the proposed *LMS-ATP* model does not need to train the initial *Thermal RC* value during the offline phase. Therefore, we do not need to know the accurate *Thermal RC* value in early stage.

The experimental results show that the proposed *LMS-ATP* can help to reduce 72.96% Mean Absolute Error (*MAE*) of the temperature prediction by using the *RCTP* scheme proposed in [5]. With accurate temperature prediction results, the system throughput can be improved by 0.77% to 47.96%, as shown in Fig. 1(c).

The rest of this paper is organized as follows. In Section 2, we introduce some related calibration schemes. In Section 3, the proposed *LMS-ATP* scheme is described. In Section 4, the experiments are shown and discussed. Finally, we conclude this paper in Section 5.

2. Related Works

A. Noise Filtering by using Kalman Filter [9]

In [9], the authors proposed to employ the Kalman filter to calibrate the results of imprecise temperature estimation. The technique consists of offline parameter initialization and online procedure. In the offline phase, the authors first analyze the equivalent *Thermal RC* model of the system and create the corresponding Kalman filter to calibrate the initial parameter based on the assumed Gaussian noise. In the online phase, the authors use the Kalman filter to perform temperature estimation based on the parameters obtained during the offline phase. However, this Kalman filter-based method is not feasible for the 3D NoC system because the computational complexity of Kalman filter is very large, which results in long response time of the temperature estimation and large area overhead.

B. Calibration for Gaussian and Non-Gaussian Noise [10]

In [10], the authors separated the temperature estimation errors into Gaussian-typed and Non-Gaussian-typed ones. The authors proposed a statistical technique to solve the problem of estimating the temperature according to given noisy sensor readings. Obviously, the accuracy of temperature estimation highly depends on the completeness of the offline analysis of temperature estimation errors by using the statistic method. Hence, the statistic approach is not feasible to be applied to general-purpose 3D NoC system.

C. Error Compensation for Power Estimation [11]

In [11], the authors proposed a method to estimate and predict the temperature at runtime. The authors first derive the equivalent thermal circuit based on the ideal thermal model and accurate power estimation. Then, the error compensation term can be obtained through considering the inaccurate thermal model. However, the authors assume that the error compensation term is not temperature dependent, which may affect the efficiency of the calibration.

3. Proposed LMS-based Adaptive Temperature Prediction (LMS-ATP) Scheme

A. Baseline LMS-ATP Scheme

To predict the future temperature after $k\Delta t_s$ (*i.e.*, $T'(t+k\Delta t_s)$) in [5], the authors employed the *Thermal RC* model to propose the following *Thermal RC*-based Temperature Prediction (*RCTP*) model:

$$T'(t+k\Delta t_s) = T(t) + \Delta T(t) \cdot \frac{e^{-b\cdot\Delta t_s} \cdot (1-e^{-b\cdot k\Delta t_s})}{1-e^{-b\cdot\Delta t_s}} \quad , \tag{1}$$

where b is a technology-dependent constant that equals to the reciprocal of the *Thermal RC* value. In addition, $T(t)$ represents the current sensing results from the embedded thermal sensor; $\Delta T(t)$ means the temperature difference between the temperature at time t and time $(t-\Delta t_s)$; the Δt_s means the thermal sensing period. In this section, we first discuss the baseline *LMS-ATP* model (*i.e.*, the prediction distance is set to 1).

As mentioned before, the temperature-varying *Thermal RC* value results in imprecise results of temperature prediction. It is more important to have precise results of temperature prediction in cases of increasing temperature than in the situations of decreasing temperature because the utmost priority of the *DTM* is to prevent the temperature from achieving the alarming level. Hence, in this paper, we only consider the calibration of temperature prediction in cases where the temperature shows an exhaustive increasing trend. To detect the exhaustive temperature increasing trend, we employ a 4-step Markov Chain, as shown in Fig. 2. As the sensing temperature increases continuously during the three sensing periods, the proposed *LMS-ATP* scheme will be triggered to perform calibration of temperature prediction.

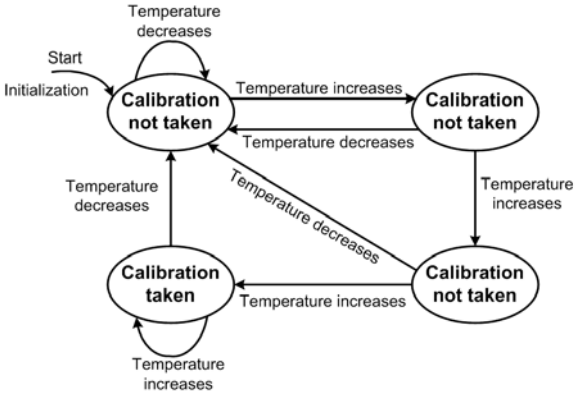

Fig. 2 The 4-step Markov Chain of calibration scheme.

According to the temperature prediction model of (1), which is used to predict the temperature with an increasing temperature trend, we can rewrite (1) as:

$$\begin{aligned} T'(t+\Delta t_s) &= T(t) + \Delta T(t) \cdot e^{-b\cdot\Delta t_s} \\ &= T(t) + (T(t) - T(t-\Delta t_s)) \cdot e^{-b\cdot\Delta t_s} \\ &= (1+e^{-b\cdot\Delta t_s}) \cdot T(t) - e^{-b\cdot\Delta t_s} \cdot T(t-\Delta t_s) \end{aligned} \quad , \tag{2}$$

where the prediction distance k equals to one. Note that $\Delta T(t)$ equals to $(T(t) - T(t-\Delta t_s))$ and Δt_s is the thermal sensing period. We use the parameters $w_0(t)$ and $w_1(t)$ to respectively substitute for $(1+e^{-b\Delta t_s})$ and $e^{-b\Delta t_s}$ in (2), and a two-tap filter-output form can be derived as:

$$T'(t+\Delta t_s) = w_0(t)\cdot T(t) - w_1(t)\cdot T(t-\Delta t_s). \quad (3)$$

Note that $\Delta T(t)$ and $\Delta T(t-\Delta t_s)$ are resulted from the embedded thermal sensor. According to (3), the block diagram of the two-tap filter is shown in Fig. 3. Obviously, $w_1(t)$ is equal to $(w_0(t)-1)$. Therefore, we can only adaptively adjust the $w_0(t)$ to perform calibration with lower area overhead. In this paper, we employ the LMS-based adaptive filter, which was introduce in [12], to adjust the $w_0(t+\Delta t_s)$ based on the estimation error $e(t)$. Therefore, $w_0(t+\Delta t_s)$ and $w_1(t+\Delta t_s)$ can be described as:

$$w_0(t+\Delta t_s) = w_0(t) + \mu e(t)T(t) \quad (4)$$

and

$$w_1(t+\Delta t_s) = w_0(t+\Delta t_s) - 1. \quad (5)$$

In this work, we set the step-size parameter μ of the involved LMS-based adaptive filter to the sensing period Δt_s.

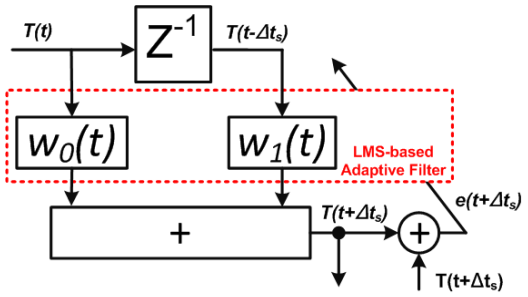

Fig. 3 Block diagram of the proposed *LMS-ATP* scheme.

B. *Enhanced LMS-ATP Scheme*

In cases where the prediction distance k (*i.e.*, *PD(k)*) is larger than one, one particular problem emerges when directly applying the method of (4) and (5) to update the $w_0(t)$ and $w_1(t)$. The reason is that we cannot obtain the actual temperature $T(t+k\Delta t_s)$ at time t, which results in an unknown $e(t+k\Delta t_s)$. Consequently, $w_0(t+k\Delta t_s)$ and $w_1(t+k\Delta t_s)$ cannot be adjusted due to the unknown error estimation $e(t+k\Delta t_s)$.

To make the proposed *LMS-ATP* scheme capable of calibrating the results of temperature prediction when the prediction distance k is larger than one, according to (2), we first name $w_1(t+\Delta t_s)$ to ρ, which is equal to $e^{-b\Delta t_s}$. Therefore, (1) can be rewritten as:

$$T(t+k\Delta t_s) = (1+\frac{e^{-b\cdot\Delta t_s}\cdot(1-e^{-b\cdot k\Delta t_s})}{1-e^{-b\cdot\Delta t_s}})\cdot T(t) - \frac{e^{-b\cdot\Delta t_s}\cdot(1-e^{-b\cdot k\Delta t_s})}{1-e^{-b\cdot\Delta t_s}}\cdot T(t-\Delta t_s) \quad (6)$$

$$= \underbrace{(1+\frac{\rho\cdot(1-\rho^k)}{1-\rho})}_{w_0(t+k\Delta t_s)}\cdot T(t) - \underbrace{\frac{\rho\cdot(1-\rho^k)}{1-\rho}}_{w_1(t+k\Delta t_s)}\cdot T(t-\Delta t_s)$$

where k is the prediction distance. Therefore, we can adaptively update $w_0(t+k\Delta t_s)$ and $w_1(t+k\Delta t_s)$ through updating ρ. In summary, three basic relations of the proposed *LMS-ATP* model can be described as follows:

1) *Temperature prediction:*

$$T'(t+k\Delta t_s) = w_0(t+k\Delta t_s)\cdot T(t) - (w_0(t+k\Delta t_s)-1)\cdot T(t-\Delta t_s). \quad (7)$$

2) *Estimation error:*

$$e(t+\Delta t_s) = T(t+\Delta t_s) - T(t). \quad (8)$$

3) *Tap-weight updating:*

- **Prediction distance *k*=1**

$$w_0(t+\Delta t_s) = w_0(t) + \mu e(t)T(t) \quad (9)$$

$$w_1(t+\Delta t_s) = w_0(t+\Delta t_s) - 1. \quad (10)$$

- **Prediction distance *k*>1**

$$w_0(t+k\Delta t_s) = 1 + \frac{\rho\cdot(1-\rho^k)}{1-\rho} \quad (11)$$

$$w_1(t+k\Delta t_s) = \frac{\rho\cdot(1-\rho^k)}{1-\rho}, \quad (12)$$

where ρ is equal to $w_1(t+\Delta t_s)$, which is equal to $e^{-b\Delta t_s}$.

4. Experiments and Discussions

To analyze the efficiency of the proposed *LMS-ATP* scheme, we implement the proposed method on the cycle-accurate traffic-thermal co-simulator [13]. The ambient temperature and the initial temperature are both set to 25°C. An 8x8x4 mesh-based 3D NoC is set as a design example in this work. For each router, the channel depth of the input buffer is 4 flits without virtual channel, and the wormhole flow control scheme is employed. In addition, the packet length is 8 flits. To simplify the problem, the XYZ routing algorithm is adopted.

The initial values of $w_0(t)$ and $w_1(t)$ (*i.e.*, $w_0(0)$ and $w_1(0)$) depends on the b value of (2). However, the value of b is difficult to obtain because it equals to the reciprocal of the *Thermal RC* value, which is temperature-varying. In this work, we set the initial b value to 100 as a design example. Therefore, $w_0(0)$ is equal to $(1+e^{-100\cdot 0.01})$ while the prediction distance is equal to one and the thermal sensing period is equal to 10*ms*.

A. *Accuracy Analysis of the Proposed LMS-ATP Scheme*

Fig. 4 shows the convergence of $w_0(t)$. The result of $w_1(t)$ convergence is the same as that of $w_0(t)$ because the value of $w_1(t)$ relies on that of $w_0(t)$. Compared to the previous work [5], the *LMS-ATP* can help to converge the value of $w_0(t)$ within 100ms based on the estimation error at each thermal sensing time (*i.e.*, the blue line). Obviously, the proposed *LMS-ATP* scheme provides a simple way to obtain the feasible b value, which is difficult to obtain. Therefore, we do not need to know the accurate b value in early stage.

As mentioned before, the unfeasible b value will result in large estimation error of temperature prediction. Fig. 5 shows the comparison of Mean Absolute Error (*MAE*) as applying the *RCTP* model, which was proposed in [5], and using the proposed *LMS-ATP*. Compared to the previous work in [5], the *LMS-ATP* can improve the precision of future temperature estimation by 72.96%.

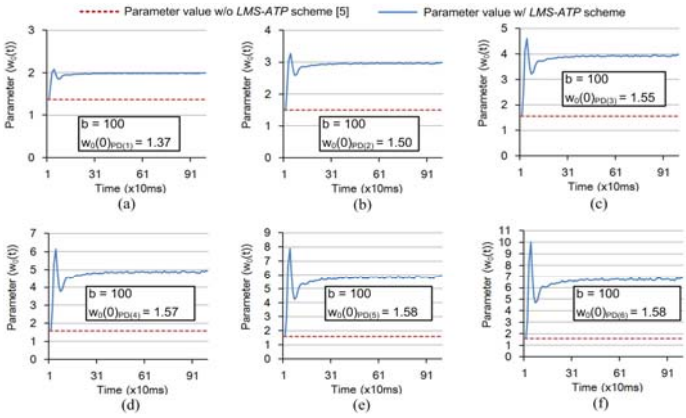

Fig. 4 Parameter analysis as b=100 under (a) *PD(1)*, (b) *PD(2)*, (c) *PD(3)*, (d) *PD(4)*, (e) *PD(5)*, and (f) *PD(6)*.

Table 1 Comparison of the *PDTM-VT* with different temperature prediction scheme under different prediction distance.

	Temperature prediction by using *RCTP* [5]		Temperature prediction by using the proposed *LMS-ATP*	
	Numbers of thermal-emergent node	Throughput (flits/cycle)	Numbers of thermal-emergent node	Throughput (flits/cycle)
PD(1)	18,436	3.91	15,688 (-14.91%)	3.94 (+0.77%)
PD(2)	17,988	3.92	14,464 (-19.59%)	5.71 (+45.66%)
PD(3)	17,988	3.92	16,108 (-10.45%)	4.52 (+15.31%)
PD(4)	17,988	3.92	14,036 (-21.97%)	5.80 (+47.96%)
PD(5)	18,436	3.91	15,396 (-16.49%)	4.66 (+19.18%)

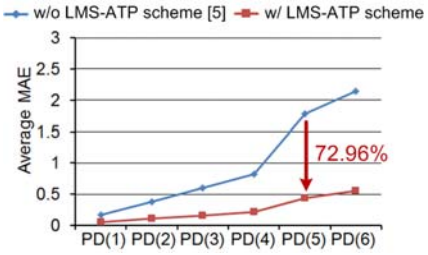

Fig. 5 The *LMS-ATP* can improve the precision of temperature prediction.

To verify the proposed *LMS-ATP* scheme on the physical platform, Vertex-6 FPGA ML605 is employed in this work. There is one embedded temperature monitor in each Vertex-6 FPGA family. In addition to the proposed *LMS-ATP* unit, we also used a 3-bit adder as a heater, which is used for heating the FPGA. The maximum measurement error of the embedded sensor is ±4°C, and the precision of the LSB of the embedded sensor is 0.5°C. Besides, the thermal sensor is built around a 10-bit, 200-kSPS (kilo-samples per second) Analog-to-Digital Convertor (ADC) with digital averaging. Fig. 6 shows the result of temperature prediction while the prediction distance is equal to one (*PD(1)*). The proposed *LMS-ATP* can predict the future temperature before FPGA reaches that. Obviously, the proposed *LMS-ATP* can accurately predict the future temperature due to the adaptive value adjustment of $w_0(t)$ and $w_1(t)$.

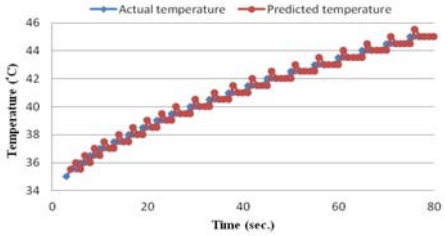

Fig. 6 Results of temperature prediction via FPGA verification.

B. Performance Analysis

To analyze the system performance after applying the proposed *LMS-ATP*, we combine the proactive dynamic thermal management with the vertical throttling (*PDTM-VT*) scheme, which was proposed in [2]. Table 1 shows the comparison between the system performance and the transient numbers of thermal-emergent nodes when adopting the *RCTP* [5] and the proposed *LMS-ATP* respectively as the temperature prediction scheme. Because of the temperature-varying *Thermal RC* value, the estimation error of temperature prediction will be large, which results in large performance impact. Compared with the previous work [5], the proposed *LMS-ATP* can help to reduce the numbers of thermal-emergent nodes by around 10.45% to 21.97%. Besides, the system performance can be improved by around 0.77% to 47.96%.

5. Conclusions

In this paper, we address the problem of future temperature estimation with given temperature-varying *Thermal RC* values. To calibrate the results of future temperature estimation, we propose an LMS-based adaptive thermal prediction (*LMS-ATP*) scheme by applying the Least-Mean-Square (LMS) adaptive filter. The *LMS-ATP* scheme can calibrate the results of predicted temperature even if the *Thermal RC* value is unknown. Compared to the previous work, the proposed *LMS-ATP* can reduce the Mean Absolute Error (*MAE*) by 72.96%. In addition, due to the precise future temperature estimation results, the proposed *LMS-ATP* can improve the system throughput by 0.77% to 47.96%.

Acknowledgement

This work was supported by the National Science Council under NSC 100-2220-E-002-013 and NSC 100-2220-E-002-016.

References

[1] B.S. Feero and P.P. Pande, "Networks-On-Chip in a Three Dimensional Environment: A Performance Evaluation," *IEEE Trans. Comput.*, vol.58, no.1, pp.32-45, Jan. 2009.

[2] I. Yeo *et al.*, "Predictive Dynamic Thermal Management for Multicore Systems," *ACM/IEEE Design Automation Conference (DAC)*, pp.734-739, Jun. 2008.

[3] C.H. Chao *et al.*, "Traffic- and Thermal-Aware Run-Time Thermal Management Scheme for 3D NoC System" IEEE Intl. Symp. Network-on-Chip (NOCS), pp.223-230, May 2010.

[4] I. Yeo *et al.*, "Predictive Dynamic Thermal Management for Multicore Systems," ACM/IEEE Design Automation Conference (DAC), pp.734-739, Jun. 2008.

[5] K.C. Chen *et al.*, "Design of Thermal Management Unit with Vertical Throttling Scheme for Proactive Thermal-aware 3D NoC Systems," *IEEE Int'l Symp. VLSI Design, Automation, and Test (VLSI-DAT-2013)*, pp.118-121, Hsinchu, Taiwan, Apr. 2013.

[6] K. Skadron, M.R. Stan, K.Skadron, *et al.*, "Temperature aware microarchitecture," *Proc. Int'l Symp. Comput. Architect.*, pp.2-13, Jun. 2003.

[7] R. Ayoub *et al.*, "Energy Efficient Proactive Thermal Management in Memory Subsystem," *ACM/IEEE Int'l Symp. Low-Power Electronics and Design (ISLPED)*, pp.195-200, Aug. 2010.

[8] Cengel and Boles, *Thermodynamics: An engineering approach*, 7th edition, McGraw Hill, 2011.

[9] S. Sharifi and T.S. Rosing, "Accurate Direct and Indirect On-Chip Temperature Sensing for Efficient Dynamic Thermal Management," *IEEE Trans. Computer-Aided Design Integr. Circuits*, vol.29, no.10, pp.1586-1599, Oct. 2010.

[10] Y. Zhang and A. Srivastava, "Accurate Temperature Estimation Using Noisy Thermal Sensor for Gaussian and Non-Gaussian Cases," *IEEE Trans. Very Large Scale Integr. Syst.*, vol.19, no.9, pp.1617-1626, Sept. 2011.

[11] H. Wang, S.X.D. Tan, G. Liao, *et al.*, "Full-Chip Runtime Error-Tolerant Thermal Estimation and Prediction for Practical Thermal Management," *IEEE/ACM International Conf. Computer-Aided Design (ICCAD)*, pp.716-723, Nov. 2011.

[12] S. Haykin, Adaptive Filter Theory (5th Edition), Prentice-Hall, 2013.

[13] K.Y. Jheng, C.H. Chao, H.Y. Wang, and A.Y. Wu, "Traffic-Thermal Mutual-Coupling Co-Simulation Platform for Three-Dimensional Network-on-Chip," in *Proc. IEEE Intl. Symp. on VLSI Design, Automation, and Test* (VLSI-DAT'10), pp.135-138, Apr. 2010.