# Adaptive Machine Learning-based Temperature Prediction Scheme for Thermal-aware NoC System

Kun-Chih (Jimmy) Chen and Yuan-Hao Liao
Department of Computer Science and Engineering, National Sun Yat-sen University, Taiwan

***Abstract*—Because of the high-complex interconnection in the contemporary manycore system, the Network-on-Chip (NoC) technology is proven as an efficient way to solve the communication problem in multicore systems. However, the thermal problem becomes the main design challenge in the current NoC systems due to the high-diverse workload distribution and large power density. Therefore, the Proactive Dynamic Thermal Management (*PDTM*) is employed as an efficient way to control the system temperature in modern multicore systems. Based on the predicted temperature information, the *PDTM* can control the system temperature in advance, which helps to reduce the performance impact during the temperature control period. However, the conventional temperature prediction model is usually built based on several specific physical parameters, which are usually temperature-sensitive as well. As a result, the current temperature prediction models still suffer from large prediction errors, which reduces the benefit of the *PDTM*. To solve this problem, we combine the artificial neural network and LMS adaptive filter theory to propose an adaptive machine learning-based temperature prediction model. Because the proposed model can adapt to the hyperplane of the temperature behavior of NoC system during the runtime, the proposed approach can reduce average error by 37.2% to 62.3%, which helps to improve the system performance by 9.16% to 38.37% and can bring smaller area overhead than the related works by 18.59% to 22.11%.**

***Keywords—online learning; neural network; temperature prediction; thermal management; NoC; Network on Chip***

## I. INTRODUCTION

The interconnection complexity of the multiprocessor system grows with respect to the advance in the semiconductor technology. The Network-on-Chip (NoC) technology has been proposed as an efficient solution to solve the interconnection problem among each processing element [1]. However, because of the high power density and high-diverse workload distribution, the NoC system usually suffers from the severer thermal problem. The thermal issue leads to many negative impacts, such as lower system reliability and longer system latency. Consequently, in a practical way, the dynamic thermal management (*DTM*) is usually involved in the contemporary NoC systems to control the temperature of the system under the thermal limit [2].

To ensure thermal safety, many *DTM* schemes were proposed in recent years, which can be classified into reactive *DTM* (*RDTM*) and proactive *DTM* (*PDTM*). The conventional *RDTM* is triggered while the temperature of the NoC nodes reaches the triggering temperature [3]. Although the *RDTM* can cool down the thermal-emergent nodes quickly, it causes a large performance impact because the *RDTM* usually involves the fully throttling scheme to control the system temperature. On the other hand, the *PDTM* controls the system temperature in advance based on the information of temperature prediction. Because the *PDTM* usually employs the partially throttling strategy instead of fully throttling strategy in *RDTM*, the *PDTM* takes advantage of mitigating the performance impact during the temperature period [4]. Consequently, *PDTM* is proven as an efficient way to control the system temperature in recent years.

This work was supported by the Ministry of Science and Technology, TAIWAN, under Grant MOST 108-2218-E-110-010 and MOST 108-2218-E-110 -006

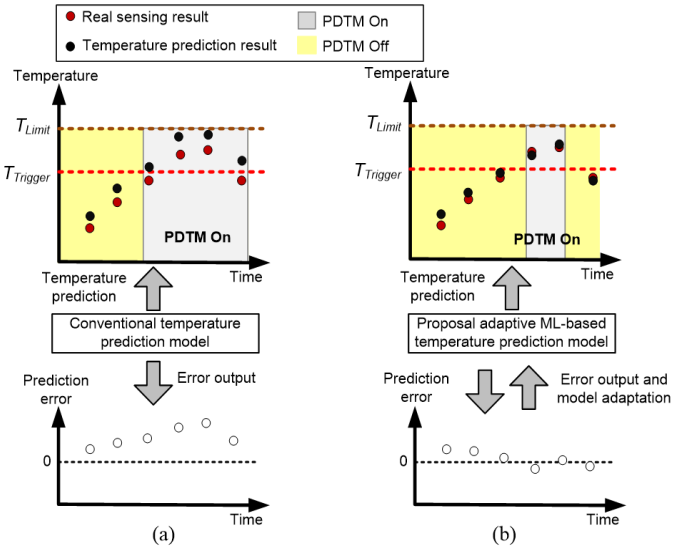


Fig. 1 (a) The conventional temperature prediction model cannot take the advantage of the *PDTM*; (b) the proposed method can adapt to the temperature behavior.

In order to forecast the future temperature of the NoC system, it is necessary to require a precise temperature prediction model. In a practical way, most of the temperature prediction models are formulated based on several physical parameters, such as capacitance, resistance, power, *etc.* [5][6]. However, these physical parameters are usually temperature-sensitive (*i.e.*, the value is different according to the temperature change). Besides, due to the varying workload distribution, the temperature behavior of each NoC node is different, which worsens the difficulty to catch a precise value of these physical parameters at runtime. Therefore, in a practical way, the designers usually see these physical parameters as constant values and use them to predict the system temperature with the involved temperature prediction model. Nevertheless, due to the characteristic of temperature-sensitive, the temperature prediction model with constant physical parameters leads to imprecise temperature prediction results. To prevent the system from overheating by using the imprecise information of the temperature prediction, the conventional *PDTMs* adopt a pessimistic temperature control strategy, as shown in Fig. 1(a), which still makes NoC systems suffer from large performance impact.

As mentioned before, due to the temperature-sensitive physical parameters, it is difficult to predict the future temperature of the NoC system precisely with the conventional temperature prediction model. In recent years, the approaches of artificial neural network (ANN) have shown the advantage in many fields, such as event prediction and classification [8]. Based on certain training methods, the artificial neural network can be used to satisfy the hyperplane of the system temperature behavior by adjusting the involved weights for ANN computing. Because the function of the trained ANN model is similar to the temperature behavior of the NoC system, it can be used to forecast the future temperature. However, the ANN is usually trained by a pre-defined dataset, which does not contain all kinds of

978-1-7281-3320-1/20/$31.00 ©2020 IEEE
Authorized licensed use limited to: National Institute of Technology Karnataka Surathkal. Downloaded on September 28,2024 at 04:51:41 UTC from IEEE Xplore. Restrictions apply.

temperature changes. The temperature prediction error will be larger if a certain situation of temperature change at runtime is not learned by the ANN in the offline training phase, which reduces the benefit, bringing from the *PDTM*.

To solve the aforementioned problem, we employ the Least Mean Square (LMS) adaptive filter theory [9] to make the involved ANN-based temperature prediction model adapt to the varying temperature behavior at runtime. Based on the information of temperature prediction error, the LMS adaptive filter can be used to adjust the involved weights for the ANN computing to dynamically fit the hyperplane of the temperature behavior, as shown in Fig. 1(b). With a precise adaptive temperature prediction model, the *PDTM* can control the system temperature efficiently and reduce the system performance impact significantly. The contributions of this paper are

1) An ANN-based temperature prediction model with low computational complexity is proposed to forecast the system temperature.
2) An LMS-based adaptive ANN model is further proposed to adapt to the varying temperature behavior dynamically.
3) The architecture of adaptive ANN-based temperature prediction is proposed to assist with the involved *PDTM*.

To evaluate the proposed temperature prediction scheme, we use a traffic-thermal co-simulator [10] to simulate a NoC system. We simulate difference synthetic traffic patterns to verify and analyze the prediction result. The experimental results show that the proposed approach can reduce average error by 37.2% to 62.3%, which helps to improve the system performance by 9.16% to 38.37% and can bring smaller area overhead than the related works by 18.59% to 22.11%

## II. Related Works

### A. RC-based temperature prediction [5]

In [5], Chen *et al.* employed the thermal RC model to propose an RC-based temperature prediction model. To predict the two kinds of temperature behavior (*i.e.*, increasing or decreasing trend), two distinguished temperature prediction models were proposed in this work, which reduces the temperature prediction error. However, because of the worst-case consideration, the authors in this paper assume that the temperature will increase after each temperature drop. Hence, the error of the temperature prediction is still very high, which reduces the benefit of the *PDTM*.

### B. Second derivative-based thermal predictive model [6]

To find the temperature increasing rate to predict the future temperature, Zhou *et al.* apply the second derivative to analyze a proposed thermal model, which integrates the thermal RC model and the information of power consumption. In this work, the authors assume that the system temperature is at the steady state. However, this model does not consider the time-varying system workload distribution, which makes the system temperature becomes non-stationary. Therefore, this approach still suffers from a large temperature prediction error.

### C. Linear Regression-based temperature prediction [7]

Because the future temperature depends on the current temperature and the power consumption within a period, Eduardo *et al.* proposed to predict the future temperature based on the information of the two previous successive sensing temperature and the power consumption by using the linear regression approach. However, the model cannot adapt to the high-diverse temperature behavior of the system. Besides, it is hard to obtain the power consumption data of the system during the runtime operation.

## III. Proposed Adaptive Machine Learning-based Temperature Prediction Model

### A. ANN-based Temperature Prediction Model

In the traditional temperature prediction model, it is usually built based on physical knowledge. Zhang *et al*. proposed a temperature prediction model in [11], which is formulated as

$$T(k+N \mid k) = B^N T(k \mid k) + \sum_{i=k}^{q} B^{q-i} D u(i), \qquad (1)$$

where $B$ and $D$ are both coefficients depending on the parameters of the thermal conductance and capacitance circuit; $u(i)$ is the power consumption model. Besides, the $T(k \mid k)$ is the temperature at time $k$ and the $T(k + N \mid k)$ represents the future temperature after $N$ time periods. Note that the $q$ is the time after $N - 1$ time periods at time $k$ (*i.e.*, $q = k + N - 1$). Obviously, it is necessary to understand the detail of the circuit to obtain these physical parameter values. However, it is difficult to find these detailed physical characteristics as the scale of the system becomes large.

To solve the aforementioned problem, a method to find the hyperplane of the system temperature changing behavior during the runtime is needed. Because the model in (1) is composed of a series of multiply-accumulate operations, the model can be seen as an operation of a neuron in ANN. The ANN is usually used to transform the input to the proper expected target, which can be used to approximate a hyperplane of temperature behavior. Thus, we can rewrite the (1) as

$$T(k+N \mid k) = \sum_{n=1}^{k-q+2} w_n \cdot X_n, \qquad (2)$$

where $w_n$ is an integrated factor combing $B$ and $D$ in (1). $X_n$ in (2) is the input data $T(k \mid k)$ and $u(i)$. Through the ANN approach, we can calculate a similar computation in (1). Besides, based on the weight training approach, we can obtain the weight (*i.e.*, $w_n$ in (2)) to build a model instead of certain physical coefficients.

By following (2), a single-neuron ANN-based temperature prediction unit can be embedded in each NoC node, as shown in Fig. 2. Obviously, the result of ANN computation depends on the input features significantly; thereby, it is necessary to find the proper factors as the input of the proposed ANN-based temperature prediction model. Due to the spatial correlation of the temperature distribution, the temperature of each node depends on the temperature and the throttling activities of the local node and surrounding nodes. Therefore, in this work, we define a temperature-influence window, which includes one target node (*i.e.*, $N_{11}$ in Fig. 2) and eight nodes (*i.e.*, $N_6$, $N_7$, $N_8$, $N_{10}$, $N_{12}$, $N_{14}$, $N_{15}$, and $N_{16}$ in Fig. 2) surrounding the target node. According to the input of the ANN-based temperature predictor, the temperature and the throttling activities of the target node and surrounding nodes will be used for the further calculation of the temperature prediction. In this work, we adopt the ReLU as the activation function because the

Authorized licensed use limited to: National Institute of Technology Karnataka Surathkal. Downloaded on September 28,2024 at 04:51:41 UTC from IEEE Xplore. Restrictions apply.

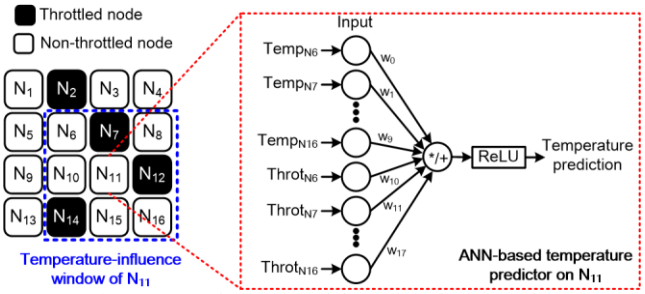


Fig. 3 Proposed ANN-based thermal prediction model

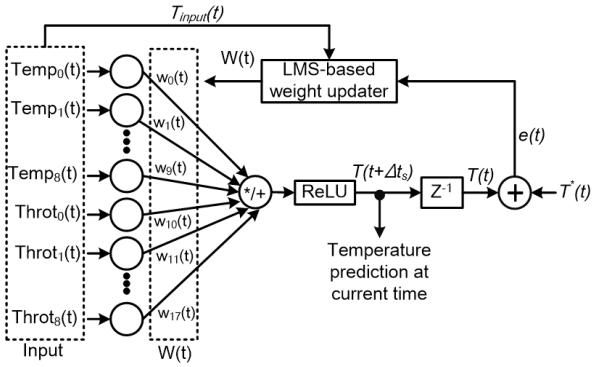


Fig. 4 Block diagram of proposed LMS-based thermal prediction model.

temperature of the chip is positive and the temperature change is linear within a short time. After the calculation of the proposed ANN-based temperature prediction model, the information of the predicted temperature can be obtained.

### B. *LMS-based Adaptive ANN Model*

As mentioned before, due to the varying workload on the NoC system, the traditional temperature model cannot adapt to the behavior of system temperature. Although the ANN-based temperature prediction model can fit the hyperplane of the temperature behavior of a system through the training phase, it still suffers from the large temperature prediction error due to the fixed weight for the ANN computing. The reason is that the trained weights highly depend on the pre-defined training dataset, which cannot cover every situation of temperature change. Hence, it is necessary to adjust the weights during the runtime based on the different temperature situation.

To adjust the weight during the runtime operation, we adopt the LMS-based adaptive filter theory, which is widely used to calibrate the filter coefficient in signal processing. As the example in Fig. 2, the predicted temperature of $N_{11}$ at time $(t+\Delta t_s)$ by using the proposed ANN-based temperature prediction model can be formulated as

$$\begin{aligned} T(t+\Delta t_s) &= w_0(t)Temp_{N_6}(t)+w_1(t)Temp_{N_7}(t)+\ldots+w_8(t)Temp_{N_{16}}(t)+ \\ &\quad w_9(t)Throt_{N_6}(t)+w_{10}(t)Throt_{N_7}(t)+\ldots+w_{17}(t)Throt_{N_{16}}(t) \\ &= \sum_{n=0}^{17} w_n(t)T_{input}(t) \end{aligned} \quad (3)$$

where the $w_n(t)$ is the trained weight in the model and $T_{input}(t)$ is the input of the proposed ANN-based temperature prediction model (*i.e.*, the information of the temperature and the throttling state of the nodes in the temperature-influence window.) Besides, the $\Delta t_s$ is the thermal sensing period. By following the LMS-based adaptive filter theory, the weight can be adjusted based on the information of the estimation error (*i.e.*, the temperature prediction error). Fig.3 shows the block diagram of the LMS-based adaptive ANN

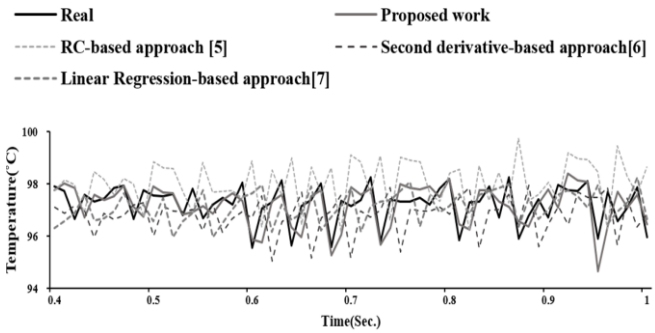


Fig. 2 Comparison of predictive temperature under Uniform Random traffic.

model. According to the estimation error at current thermal sensing time, we can formulate it to

$$e(t) = T^*(t) - T(t), \quad (4)$$

where the $T^*(t)$ is the real temperature resulted from the embedded thermal sensor. We assume that each node on the NoC has an embedded thermal sensor. Therefore, the $w_n(t)$, used to calculate the predicted temperature at next thermal sensing time, can be formulated as

$$w_n(t_s) = w_n(t-\Delta t_s) + \mu e(t) T_{input}(t) \quad (5)$$

by following the LMS-based adaptive filter theory, Note that the $\mu$ is the step-size parameter, which affects the convergence speed of the adopted weight adjustment method. In this work, we set the $\mu$ to the sensing period $\Delta t_s$. Besides, the $w_n(t-\Delta t_s)$ is the current weight before weight adjustment. According to (5), the involved weights can be updated adaptively with low computational complexity by using the input data and the estimation error of predicted temperature. Therefore, the proposed ANN-based temperature prediction model can adapt to different temperature behavior at runtime.

## IV. Experimental Results

To evaluate the proposed temperature prediction model, we use a traffic-thermal co-simulator [10] to simulate an 8x8 NoC system. We set the initial temperature to 80˚C. In addition, the buffer depth of each router is 8 flits without virtual channels. To simplify the routing problem, the XY routing algorithm is adopted. We verify the proposed temperature prediction model under three synthetic traffic patterns, *Uniform Random, Transpose-1,* and *Hotspot*. To compared with the related works, we implement three different approaches: 1) RC-based temperature prediction model [5], 2) Second derivative-based thermal predictive model [6] and 3) linear regression-based temperature prediction model [7].

### A. *Precision Analysis and the Advantage for the System Performance*

Fig. 4 shows a comparison between the actual temperature and predicted temperature by using different approaches. Compared with the conventional approaches, the proposed approach can achieve more precise temperature prediction results. The reason is that the conventional approaches predict the temperature based on certain physical parameters, which are usually temperature-sensitive. Therefore, the conventional approaches cannot adapt to the temperature behavior efficiently at runtime. On the other hand, the proposed thermal prediction model can dynamically adapt to the temperature behavior of the system, which helps to reduce the temperature prediction error. As

Authorized licensed use limited to: National Institute of Technology Karnataka Surathkal. Downloaded on September 28,2024 at 04:51:41 UTC from IEEE Xplore. Restrictions apply.

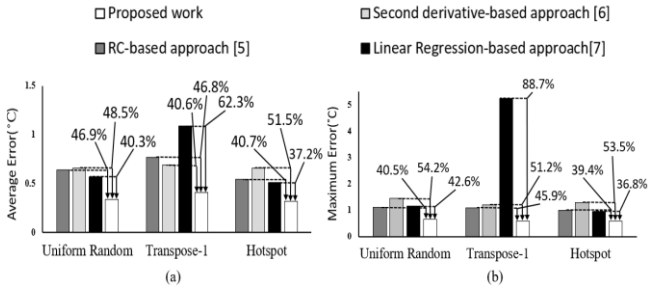


Fig. 5 The comparison of (a) the average and (b) maximum temperature prediction error under three traffic patterns.

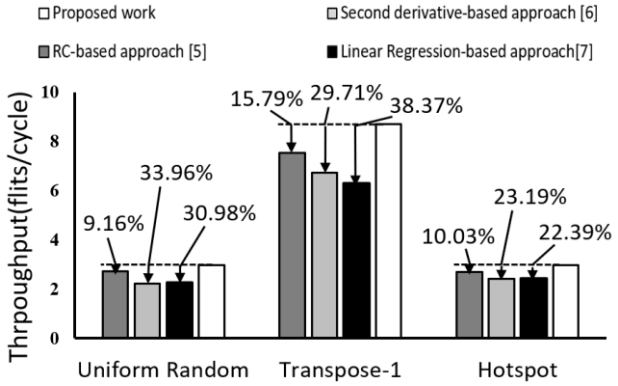


Fig. 6 Comparison of throughput with other tradictional approaches under difference traffic pattern.

shown in Fig. 5, compared with the related works, the proposed approach can reduce 37.2% to 62.3% average error and 36.8% to 88.7% maximum error.

To evaluate the system performance with our proposed temperature prediction model, similar to [5], the distributed *PDTM* is adopted in this work. Based on the information of the temperature prediction given by different approaches, we evaluate the performance of the *PDTM* and compare the throughput of the system under three synthetic traffic patterns. As shown in Fig. 6, compared with the conventional approaches, the proposed approach can make the involved *PDTM* control the system temperature precisely and improve the system throughput by 9.16% to 38.37%.

### *B. Architecture Analysis of the Proposed Model*

In this work, we assume that there is one embedded thermal sensor and one temperature prediction unit (*TPU*) on each NoC tile. The *TPU* is composed of a one-layer ANN and an LMS-based adaptive weight adjustment computation unit, as shown in Fig. 7. After the weight adjustment, the updated weight will be stored in the weight bank. Furthermore, the ANN computation unit is composed of a multiplication array and an accumulator. The multiplication array is to multiply the input data (*i.e.*, the temperature and the throttling activities of the local node and surrounding nodes) and the corresponding weights; the accumulator is to accumulate the result of the multiplication. According to the LMS computation block, it calculates the new weight by using multiplication and subtraction based on (4) and (5). To analyze the area overhead of the proposed LMS-based *TPU*, we implement it and three other related works with TSMC 90*nm* technology process. As shown in TABLE I, the TPU area of the proposed approach is smaller than the related works by 18.59% to 22.11%. Therefore, the proposed LMS-based *TPU* can bring the benefit of smaller area overhead than the related works.

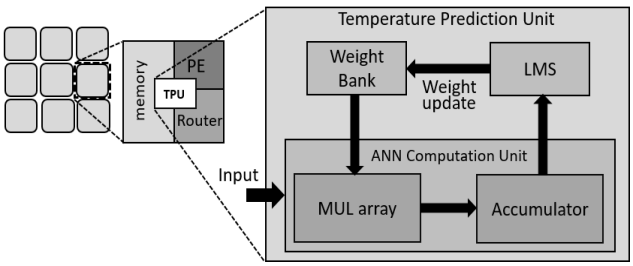


Fig. 7 Hardware design of proposed temperature prediction model

TABLE I. The comparison results of TPU area.

| ($\mu m^2$) | [5] | [6] | [7] | Proposed work |
|---|---|---|---|---|
| **TPU** | 193,908 | 207,540 | 191,703 | 161,640 |
| **Area comparison** | +19.96% | +22.11% | +18.59% | — |

## V. CONCLUSION

In this work, an adaptive machine learning-based temperature prediction model is proposed to forecast the system temperature precisely. By using the LMS-based weight adjustment method, the proposed temperature prediction model can adapt to the temperature behavior dynamically. Compared with the related works, the proposed temperature prediction model can reduce 37.2% to 62.3% average error and 36.8% to 88.7% maximum error. With the precise information of the temperature prediction, the involved *PDTM* can control the system temperature properly and helps to improve the system throughput by 9.16% to 38.37% and bring smaller area overhead than the related works by 18.59% to 22.11%.

## REFERENCES


[1] Y. Hoskote, *et al*., "A 5-GHz mesh interconnect for a teraflops processor," *IEEE Micro,* vol.27, no. 5, pp. 51-61, Sep. 2007.

[2] A.K. Coskun *et al*., "Dynamic thermal management in 3D multicore architectures," *Proceedings -Design Automation and Test in Europe DATE*, Apr. 2009, pp. 1410-1415.

[3] C. Ciordas *et al.*, "An Event-based Monitoring Service for Networkson Chip", *ACM Trans. Desg. Automation of Electric Systems (TOADES)*, vol. 10, no. 4, pp. 702-723, Oct. 2005.

[4] T. Wegner *et al.*, "Impact of proactive temperature management on performance of networks-on-chip," *Proc. Int. Symp. System on Chip*, Oct. 2011, pp. 116–121.

[5] K.-C. Chen *et al.*, "RC-based temperature prediction scheme for proactive dynamic thermal management in throttle-based 3D NoCs," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 1, pp. 206-218, Jan. 2015.

[6] Z. Lei *et al*., "A Predictive Thermal Model for Multiprocessor System-on-Chip," *Proc. of the World Congress on Engineering and Computer Science*, Oct. 2015, pp. 27-32.

[7] E. Weber *et al.*, "Predictive Thermal Management for Energy-Efficient Execution of Concurrent Applications on Heterogeneous Multicores," *IEEE Trans. Very Large Scale Integration (VLSI) Syst.*, vol. 27, no. 6, pp. 1404-1415, Jun. 2019.

[8] A. K. Jain *et al.*, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3, pp. 31-44, Mar 1996.

[9] P. L. Feintuch, "An adaptive recursive LMS filter," *Proc. of the IEEE* vol. 64, no.11, pp. 1622-1624, Nov. 1976.

[10] K.-Y. Jheng *et al.*, "Traffic-thermal mutual-coupling co-simulation platform for three-dimensional network-on-chip," *Proc. Int. Symp. VLSI-DAT*, Apr. 2010, pp. 135-138.

[11] Y. Zhang *et al.*, "Accurate temperature estimation using noisy thermal sensors for Gaussian and Non-Gaussian cases," *IEEE Trans. Very Large Scale Integration (VLSI) Syst.*, vol. 19, no. 9, pp. 1617-1626, Sep. 2011.

[12] L. Shang *et al.*, "Thermal modeling, characterization and management of on-chip network," *IEEE Micro*, Dec. 2004, pp. 67-68.


Authorized licensed use limited to: National Institute of Technology Karnataka Surathkal. Downloaded on September 28,2024 at 04:51:41 UTC from IEEE Xplore. Restrictions apply.