Routing-Based Traffic Migration and Buffer Allocation Schemes for 3-D Network-on-Chip Systems With Thermal Limit

Chih-Hao Chao, Kun-Chih Chen, Student Member, IEEE, and An-Yeu (Andy) Wu, Senior Member, IEEE

Abstract—The 3-D network-on-chip (NoC) router is a major source of thermal hotspots, limiting the performance gain of 3-D integration. Due to the varying cooling efficiency of different silicon layers in 3-D NoC, the optimal criteria of traditional load balancing design (LBD) scheme and temperature balancing design (TBD) scheme may not be satisfied. To analyze the tradeoff between performance and temperature, we provide a new analytical model. The model shows that the LBD scheme and the TBD scheme can be considered as two corner cases in the design space, and design cases can be categorized by comparing the bandwidth bound and the thermal-limited bound. To find the optimal design criteria between the LBD and the TBD schemes in 3-D NoC, we propose a new routing-based traffic migration, vertical-downward lateral-adaptive proactive routing (VDLAPR), and buffer allocation methods, vertical buffer allocation (VBA). The VDLAPR algorithm enables to tradeoff between the LBD and the TBD schemes. The proposed VBA method mitigates the traffic congestion caused by traffic migration. To reach the optimal configuration, we propose a systematic design flow, which assists in finding the best design parameters in the expanded space between LBD and TBD. Based on the traffic-thermal co-simulation experiments, the achievable throughput can be improved from 2.7% to 45.2% using the proposed design scheme.

Index Terms—3-D IC, 3-D NoC, Buffer allocation, downward routing, network-on-chip (NoC).

I. INTRODUCTION

As the complexity of the system-on-chip (SoC) grows, on-chip data exchange gradually becomes a major challenge. To connect processors, memories, and other components in chip multiprocessors (CMPs) or in multiprocessor SoCs (MPSoCs), network-on-chip (NoC) has been proposed as a novel and practical solution for its regularity and scalability [1]. By combining with the emerging 3-D IC technology, 3-D NoC (3-D NoC) is able to achieve lower latency and higher performance [2], [3].

Thermal issues have been proven to be major factors that have been limiting improvement of system performance for decades. High temperature results in slower circuit switching, larger leakage power, and higher vulnerability of thermal runaway. In addition to processor cores, NoC routers were also proven to make significant impact to overall chip temperature [4]. The reason is that the power density of a high performance NoC router is usually higher than the average power density of the processor core [5]–[7]. As a result, NoC routers are one of the sources of thermal hotspots [4], [8]. As summarized in Fig. 1(a), the thermal problem is severer in 3-D NoC. The mean temperature of the routers is pushed higher by the longer heat conduction path and the larger chip cross-sectional power density. The temperature variance increases due to the varying cooling efficiency of different layers. As shown in Fig. 1(b), 3-D NoC tends to have more routers those producing unsafe temperature above the thermal limit. To handle the thermal hotspot generated by processors, thermal management techniques [9]–[11] had been proposed, including on-line temperature control and off-line thermal management schemes. For 3-D NoC systems with temperature sensors, a vertical throttling-based run-time thermal management scheme was proposed in [17]. However, for 3-D NoC systems without temperature sensors, there is no corresponding off-line design scheme aims at the performance-temperature problem caused by the vertical varying cooling efficiency. In this situation, to prevent 3-D NoC without temperature sensors from overheat, the worst-case design scenario has to be adopted, resulting in poor margin for high performance. Therefore, the throughput has to be very limited, resulting in less performance gain of 3-D integration.

Traditionally, improving the throughput of a NoC relies on balancing the loading of channels [13], which can be viewed as the load balancing design (LBD) scheme. Without thermal limits, LBD leads to the optimal performance because the stress of the heavily loaded channels can be relieved. However, the varying cooling efficiency of the different layers of 3-D NoC results in unbalanced temperature profile for the LBD scheme. The upper layers are thermal dominant and have higher temperature, as shown in Fig. 2(a). Therefore, the traffic
loading is limited by the temperature constraint. The reason is that the maximal temperature can easily become higher than the thermal limit. Unless thermal limit is extremely high, the traffic loading can be unbounded by the thermal-limit bound, until it reaches the bandwidth bound of the channel. Therefore, in typical cases, the traffic loading has to be small to prevent 3-D NoC from overheat, resulting in the problem that the throughput is very limited by temperature.

For systems in which performance is limited by temperature, eliminating hotspot could increase the margin for performance improvement. Temperature balancing design (TBD) scheme adopts this concept and tries to balance the network temperature. Shang et al. [4] proposed a proactive routing algorithm to shape the traffic loading for balancing the temperature. Addo-Quaye [14] proposed a mapping-based approach that adopting genetic algorithm with fitness function for reducing maximal temperature of the NoC system. Link and Vijaykrishnan [15] applied deterministic location transformations for balancing the temperature profile. Hung et al. [16] adopted genetic algorithm to solve the mapping problem of 2-D NoC, giving fitness functions for reducing maximal temperature. These works adopt the TBD scheme to eliminate thermal hotspots. In 2-D NoC, TBD can reduce the temperature variance for lower maximal temperature, allowing higher throughput without violating the temperature constraint. However, the varying cooling efficiency in the vertical direction of 3-D NoC reveals the unexpected shortage of the traditional TBD scheme, originally designed for 2-D systems. Adopting TBD scheme in 3-D NoC requires an extreme imbalance of vertical traffic loading, as shown in Fig. 2(b). When thermal limit is low, the throughput of the system is bounded by the thermal-limited bound, which is defined by the temperature constraint. However, in typical design cases, the traffic imbalance results in traffic congestion in the bottom silicon layer. Moreover, the traffic imbalance also brings about low channel utilization in the upper layers. The reason is that the bandwidth of the channels in the bottom silicon layer is insufficient for the heavy traffic loading there, and the bandwidth bound becomes tighter for channel loading. Therefore, the performance of a temperature-balanced 3-D NoC design is usually not the optimum.

Fig. 2 shows that the LBD scheme and the TBD scheme can be considered as two extreme cases in the design space. The LBD scheme has most balanced vertical traffic loading, and the problem results from the imbalanced temperature profile. Therefore, if the optimal criterion of LBD scheme is not satisfied, LBD cannot guarantee optimal performance. In contrast, the TBD scheme is one of the cases that has most imbalanced vertical traffic loading, and the problem comes from the consequent requirement of imbalanced vertical traffic loading. If the optimal criterion of TBD scheme is not satisfied, TBD cannot guarantee optimal performance. Our previous work [17] addressed these two factors that resisting the performance gain of 3-D stacking. However, there is still no analytical model presented for analyzing the problem of 3-D NoC, which is necessary to identify the impact of vertical varying cooling efficiency on performance-temperature trade-offs. Besides, there is no effective, controllable, and systematic approach for the case in which the optimal criteria of LBD and TBD schemes are not satisfied.

To maximize the achievable throughput of a 3-D NoC system under a certain thermal limit, the contributions of this paper are summarized as follows.

1) Propose an Analytical Model for Analyzing the Correlation Between Traffic and Thermal in 3-D NoC: Literally, we are the first group to identify the problem of the traditional LBD scheme and the problem of the traditional TBD scheme. Besides, we are also the first group that proposes the analytical model.
2) Propose a Routing-Based Traffic Migration Method for 3-D NoC: We propose the vertical-downward lateral-adaptive proactive routing (VDLAPR) algorithm. By adjusting one universal parameter, downward level, the vertical traffic loading can be redistributed between the LBD scheme and the TBD scheme.

3) Propose a Vertical Buffer Allocation Method for 3-D NoC: The proposed vertical buffer allocation method supports the VDLAPR algorithm to solve the shortages resulting from traffic imbalance.

4) Propose a Systematic Design Scheme for Maximizing the Thermal-Limited Throughput of 3-D NoC: The presented design flow of the systematic design scheme can reach the optimum in the design space of traffic loading and buffer allocation for any given thermal limit.

The design scenario of this paper is focused on off-line thermal management for systems without temperature sensors, where traffic patterns are assumed known and stationary in design stage.

An $8 \times 8 \times 4$ 3-D NoC system with 256 tiles is modeled for our experiments. Based on our traffic-thermal co-simulator [8], the experiments show that the proposed design methodology can effectively improve the thermal limited performance of 3-D NoC. For typical thermal limits from 100 °C to 150 °C, the throughput can be averagely improved by 13.3% compared to the traditional LBD scheme; in comparison with the traditional TBD scheme, the throughput can be averagely improved by 13.7%.

The rest of this paper is organized as follows. In Section II, we review the related works. In Section III, we describe the thermal-limited throughput problem of 3-D NoC. In Section IV, we describe the proposed routing-based traffic migration method and the corresponding buffer allocation method. In Section V, we use experiments to evaluate the proposed methods. In Section VI, we present our systematic design flow and design examples. We conclude this paper in Section VII.

II. REVIEW OF RELATED WORKS

It is infeasible to eliminate hotspot solely by enhancing the cooling device. Therefore, many technology-based general approaches and design-based architecture-specific thermal-management approaches were proposed. In this section, we briefly review these related works.

A. Technology-Based Approaches for Eliminating Hotspots

To eliminate thermal unsafe thermal hotspots, technology-based general approaches focus on reducing heat generation and increasing the speed of heat removal. For reducing heat generation, the dynamic voltage frequency scaling (DVFS) technique is applied in many high performance processors. Because the power consumption is reduced by DVFS, the heat generation is also reduced. However, even though power and temperature are correlated, they are still fundamentally different in nature, and design techniques focused on power cannot address NoC thermal issues very efficiently [4].

To increase the speed of heat removal of 3-D IC, the major two proposed technologies are: 1) microchannel fluid cooling (MFC, also named for microfluidic channel cooling) [18], [19], and 2) thermal-TSV (TTSV) [20], [21]. MFC works very well for removal of internal hotspot [22], but it requires extra device for stress and drain water, bringing reliability issues and extra cost. TTSV works well if its size is large and the number of TTSV is big [21], [23]. However, large size and large number of TTSVs may generate routing difficulties among within-die metal wire, TTSV, signal TSV, power/ground TSV [20], [24]–[26]. The same issues exist for MFC [27], [28]. The routing problem leads to complex performance-temperature tradeoffs. Besides, the additional cost of these technology-based approaches is an issue.

B. Design-Based Architecture-Specific Approaches for Eliminating Hotspots

Design-based architecture-specific approaches focus on balancing the temperature profile for eliminating unsafe thermal hotspots. Therefore, the following works can be categorized to the TBD scheme. For 2-D NoC systems, Shang et al. [4] aimed at the heat of the network and proposed ThermalHerd, a routing-based approach for thermal problem. The concept of proactive routing was proposed to shape the traffic distributions for reducing potential hotspot routers. Link and Vijaykrishnan [15] applied deterministic location transformations for mapping workload and evaluated the scheme with an NoC-based low density parity check decoder. Hung et al. [16] adopted genetic algorithm to solve the mapping problem of 2-D NoC, giving fitness functions for reducing maximal temperature. For 3-D NoC systems, Addo-Quaye [14] proposed a mapping-based approach that adopting genetic algorithm with fitness function for reducing maximal temperature. Our previous work [17] identified the divergence of performance optimization and thermal optimization in off-line thermal management.

III. PROPOSED ANALYTICAL MODEL AND ANALYSIS OF TRADITIONAL DESIGN SCHEMES

A. Definition of the Scope for the Correlation Analysis Between Traffic and Thermal in 3-D NoC

To derive the analytical model, the involved notations in the following analysis are shown in Table I. The problem is to maximize the achievable throughput, the maximal throughput that keeps maximal temperature of the routers below the thermal limit, described by

$$\theta_a \triangleq \max \theta, \text{ where } P(T > T_L) = 0$$

where the $P(T > T_L)$ is the probability of thermal emergency (i.e., the current temperature $T$ is larger than the thermal limit $T_L$). The temperature profile of 3-D NoC is correlated to the power profile and the chip floorplan. By following the concept of compact thermal modeling [29], the architectural-level thermal model is adopted for analysis of heat transfer and thermal behavior in early-stage design. In this paper, we choose mesh-based NoC as the design example, because it
TABLE I
PARAMETER NOTATION FOR PROBLEM DESCRIPTION

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>superscript</td>
<td>( R ): router; ( M ): memory; ( P ): processor; ( BW ): bandwidth; ( LBD ): load balancing design; TBD: traffic balanced design</td>
</tr>
<tr>
<td>subscript</td>
<td>( x ), ( y ), ( z ): location variable; ( X ), ( Y ), ( Z ): upper bounds of ( x ), ( y ), ( z );</td>
</tr>
<tr>
<td>( T ), ( T )</td>
<td>Matrix/vector and scalar of temperature; ( T^{R}_{R,x,y,z} ): temperature of router at ((x, y, z));</td>
</tr>
<tr>
<td>( P ), ( P )</td>
<td>Matrix/vector and scalar of power; ( P^{R}_{R,x,y,z} ): power of router at ((x, y, z));</td>
</tr>
<tr>
<td>( L ), ( L )</td>
<td>Matrix/vector and scalar of traffic loading (rate); ( L_{x,y,z} ): traffic loading of router at ((x, y, z));</td>
</tr>
<tr>
<td>( g_{x,y,z} )</td>
<td>Thermal conductance between router ((x, y, z)) and ((x, y, z + 1)); ( g_{x,y,z} ) is the effective thermal conductance between bottom layer and ambient;</td>
</tr>
<tr>
<td>( f )</td>
<td>Function describes traffic and power, router power = ( f ) (channel loading);</td>
</tr>
<tr>
<td>( \theta, \theta )</td>
<td>Normalized throughput and achievable throughput; traffic rate per network I/O channel;</td>
</tr>
<tr>
<td>( T^L, T^A )</td>
<td>Thermal limit and ambient temperature;</td>
</tr>
<tr>
<td>( P(T &gt; T^L) )</td>
<td>The probability ratio of thermal unsafe routers.</td>
</tr>
</tbody>
</table>

is the popular implementation in the related works and the research field of NoC. In Fig. 3(a), the temperature of router, memory, and processor at node \((x, y, z)\) are \( T^{R}_{R,x,y,z} \), \( T^{M}_{x,y,z} \), and \( T^{P}_{x,y,z} \), and the corresponding power are \( P^{R}_{R,x,y,z} \), \( P^{M}_{x,y,z} \), and \( P^{P}_{x,y,z} \), respectively. As shown in Fig. 3(b), the \( X \times Y \times Z \) 3-D NoC is composed of identical single tiles. The temperature profile \( T^{R} \) of routers and the power profile \( P^{R} \) of routers can be represented as follows:

\[
T^{R} = \begin{bmatrix} T^{R}_{1,1} & \cdots & T^{R}_{1,X} \\ \vdots & \ddots & \vdots \\ T^{R}_{Y,1} & \cdots & T^{R}_{Y,X} \end{bmatrix}, \quad P^{R}_{x,y} = \begin{bmatrix} P^{R}_{x,y,1} \\ \vdots \\ P^{R}_{x,y,Z} \end{bmatrix}
\]

\( T^{R}_{x,y,z} \) is the 1-D vertical temperature profile of routers at \((x, y)\), and \( T^{R} \) is the entire 3-D temperature profile of routers. \( P^{R}_{x,y,z} \) is the 1-D vertical power profile of routers laterally at \((x, y)\), and \( P^{R} \) is the entire 3-D power profile of routers. Similarly, the temperature profile and power profile of memory part are represented by \( T^{M} \) and \( P^{M} \), and the temperature profile and power profile of processor part are represented by \( T^{P} \) and \( P^{P} \).

The routing-based design scheme to redistribute the power profile \( P^{R} \) for eliminating hotspots in \( T^{R} \) by adjusting the traffic loading \( L \), which is defined as

\[
L = \begin{bmatrix} L_{1,1} & \cdots & L_{1,X} \\ \vdots & \ddots & \vdots \\ L_{Y,1} & \cdots & L_{Y,X} \end{bmatrix}, \quad L_{x,y,z} = \begin{bmatrix} L_{x,y,1} \\ \vdots \\ L_{x,y,Z} \end{bmatrix}
\]

\( L \) represents the entire traffic loading of the 3-D NoC, and \( L_{x,y,z} \) is the average channel loading of the router at \((x, y, z)\). Because the source-destination pairs are not changed by the routing-based design scheme, the processor power profile \( P^{R}_{x,y} \) is not changed. Similarly, the memory power profile \( P^{M} \) is not changed.

B. Analysis of Traditional Routing-Based Design Schemes

1) LBD Scheme [13]: Conventionally, throughput maximization relies on balancing channel loading of the network [13]. Consider the performance optimum for a homogeneous 3-D NoC, the channel loading matrix \( L \) is balanced and equal to for all routers as

\[
L_{x,y} = \begin{bmatrix} L_{x,y,1} \\ \vdots \\ L_{x,y,Z} \end{bmatrix} = \begin{bmatrix} L^{LBD} \\ \vdots \\ L^{LBD} \end{bmatrix}.
\]

The channel bandwidth \( L^{BW} \) is an upper bound that limits the transfer rate. The channel loading is bounded as

\[
L^{LBD} \leq L^{BW} - L^{LBD}.
\]

Assume that the power of a router is an increasing function of the channel loading, where \( P^{LBD} = f_{L\rightarrow P}(L^{LBD}) \). Then, \( P^{R} \) is balanced as \( L \) is balanced, as

\[
P^{R}_{x,y} = \begin{bmatrix} P^{R}_{x,y,1} \\ \vdots \\ P^{R}_{x,y,Z} \end{bmatrix} = \begin{bmatrix} f_{L\rightarrow P}(L^{LBD}) \\ \vdots \\ f_{L\rightarrow P}(L^{LBD}) \end{bmatrix} = \begin{bmatrix} P^{LBD} \\ \vdots \\ P^{LBD} \end{bmatrix}.
\]

According to the Appendix, the vertical temperature profile \( T^{R}_{x,y} \) of each router at \((x, y)\) is mainly determined by the vertical power profile \( P^{R}_{x,y} \). In LBD, the vertical temperature profile results in the thermal differences as

\[
T^{R}_{x,y} = \begin{bmatrix} T^{R}_{x,y,1} \\ \vdots \\ T^{R}_{x,y,Z-1} \\ T^{R}_{x,y,Z} + P^{LBD} \cdot g_{x,y,Z}^{-1} \end{bmatrix} = \begin{bmatrix} T^{R}_{x,y,2} + P^{LBD} \cdot g_{x,y,1}^{-1} \\ \vdots \\ T^{R}_{x,y,Z-1} + P^{LBD} \cdot g_{x,y,Z-1}^{-1} \\ T^{A} + P^{LBD} \cdot g_{x,y,Z}^{-1} \end{bmatrix}.
\]

Because \( P^{LBD} \cdot g_{x,y,z}^{-1} \) is never negative, the following thermal gradient is always true:

\[
T^{A} \leq T^{R}_{x,y,Z} \leq T^{R}_{x,y,Z-1} \leq \cdots \leq T^{R}_{x,y,2} \leq T^{R}_{x,y,1} \leq T^{L}.
\]

Since we must keep the temperatures of all routers not above thermal limit, the channel loading \( L^{LBD} \) is bounded.
By combining (8) and (9), we can derive the thermal-limited bound of channel loading, as follows:

\[ L_{\text{LBD}}^{\text{LBD}} = \int_{L_{\text{L}}^{\text{L}} - \text{p}}^{1} \left( p_{\text{LBD}} \right) \leq \int_{L_{\text{L}}^{\text{L}} - \text{p}}^{1} \left( \frac{(T_{\text{L}} - T_{\text{A}})}{\sum_{z=1}^{1} g_{x,y,z}} \right) \]

\[ = L_{\text{TL-LBD}}^{\text{LBD}}. \]  

(10)

The optimal criterion of LBD scheme is that the bandwidth bound \( L_{\text{BW-LBD}}^{\text{LBD}} \) is a tighter bound than the thermal-limited bound \( L_{\text{TL-LBD}}^{\text{LBD}} \). Because \( T_{\text{L}} \) is not extremely large in most cases, \( L_{\text{TL-LBD}}^{\text{LBD}} \) is usually smaller than \( L_{\text{BW-LBD}}^{\text{LBD}} \). Therefore, the achievable throughput is not guaranteed maximum by using LBD scheme.

2) TBD Scheme [4]: The goal of TBD scheme is to balance the temperature profile for eliminating hotspots. Assume there is only one heat sink at bottom of the chip. The optimum point of temperature balancing scheme is that the every temperature difference between vertical neighbor routers is zero. Such result leads to the extreme unbalanced vertical power profile for each \( P_{x,y}^{R} \), as follows:

\[ T_{x,y}^{R} |_{\text{w}_{x,y}} = \left[ \begin{array}{c} T_{x,y,1}^{R} \\ \vdots \\ T_{x,y,Z}^{R} \end{array} \right] = \left[ \begin{array}{c} T_{\text{TBD}}^{\text{TBD}} \\ \vdots \\ T_{\text{TBD}}^{\text{TBD}} \end{array} \right]. \]  

(11)

By following (11), the heat transfer toward ambient is maximum, if the temperature difference between bottom layer and ambience is \( T_{\text{L}} - T_{\text{A}} \), and the temperature difference between vertical neighbor routers is zero. Such result leads to the extreme unbalanced vertical power profile for each \( P_{x,y}^{R} \), as follows:

\[ P_{x,y}^{R} = \left[ \begin{array}{c} p_{x,y,1}^{R} \\ \vdots \\ p_{x,y,Z}^{R} \end{array} \right] = \left[ \begin{array}{c} (T_{x,y,1}^{R} - T_{x,y,2}^{R}) \cdot g_{x,y,1} \\ \vdots \\ (T_{x,y,Z-1}^{R} - T_{x,y,Z}^{R}) \cdot g_{x,y,Z-1} \end{array} \right] \]

\[ = \left[ \begin{array}{c} (T_{x,y,Z}^{R} - T_{\text{A}}) \cdot g_{x,y,Z} \\ 0 \cdot g_{x,y,1} \\ \vdots \\ 0 \cdot g_{x,y,Z-1} \end{array} \right] \]

\[ = \left[ \begin{array}{c} (T_{\text{L}} - T_{\text{A}}) \cdot g_{x,y,Z} \\ 0 \cdot g_{x,y,1} \\ \vdots \\ 0 \cdot g_{x,y,Z-1} \end{array} \right] \]

\[ = \left[ \begin{array}{c} 0 \\ \vdots \\ 0 \end{array} \right]. \]  

(12)

To fit such a power profile, all the traffic must concentrate on the bottom silicon layer and lead to huge traffic congestion.

The channel loading is subjected to the thermal-limited bound in TBD scheme, which can be derived as

\[ L_{\text{TBD}}^{\text{TBD}} \leq \int_{L_{\text{L}}^{\text{L}} - \text{p}}^{1} \left( \left( T_{\text{L}} - T_{\text{A}} \right) \cdot g_{x,y,z} \right) = L_{\text{TL-TBD}}^{\text{TBD}}. \]  

(14)

Similarly, the channel loading in TBD scheme is subjected to channel bandwidth, as follows:

\[ L_{\text{TBD}}^{\text{TBD}} \leq L_{\text{BW-TBD}}^{\text{TBD}}. \]  

(15)

The optimal criterion of TBD scheme is that the thermal limited bound \( L_{\text{TL-TBD}}^{\text{TBD}} \) is a tighter bound than the bandwidth bound \( L_{\text{BW-TBD}}^{\text{TBD}} \). By observing (14), if \( T_{\text{L}} \) is not very small, \( L_{\text{TL-TBD}}^{\text{TBD}} \) will be relaxed, resulting in the possibility that \( L_{\text{TL-TBD}}^{\text{TBD}} \) becomes larger than \( L_{\text{BW-TBD}}^{\text{TBD}} \). Therefore, the TBD scheme cannot guarantee to achieve the optimal throughput.

3) Remarks for LBD Scheme and TBD Scheme: To discuss whether LBD or TBD can achieve maximal throughput for 3-D NoC, we have to compare the bandwidth bound and the thermal-limited bound of LBD and TBD, correspondingly. Table II shows all the four cases. In Case 1, both criteria are

<table>
<thead>
<tr>
<th>Case 1: ( L_{\text{BW-LBD}}^{\text{LBD}} \leq L_{\text{TL-LBD}}^{\text{LBD}} )</th>
<th>Case 2: ( L_{\text{BW-LBD}}^{\text{LBD}} &gt; L_{\text{TL-LBD}}^{\text{LBD}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBD is optimal, TBD is optimal.</td>
<td>Case 2: LBD is nonoptimal, TBD is optimal.</td>
</tr>
</tbody>
</table>

Therefore, all the channel bandwidth in the routers in the upper \( Z-1 \) silicon layers is not utilized, which is shown as follows:

\[ L_{x,y} = \left[ \begin{array}{c} L_{x,y,1} \\ \vdots \\ L_{x,y,Z-1} \\ L_{x,y,Z} \end{array} \right] = \left[ \begin{array}{c} f_{L_{\text{L}} - \text{p}}^{-1}(P_{x,y,1}^{R}) \\ \vdots \\ f_{L_{\text{L}} - \text{p}}^{-1}(P_{x,y,Z-1}^{R}) \\ f_{L_{\text{L}} - \text{p}}^{-1}(P_{x,y,Z}^{R}) \end{array} \right] = \left[ \begin{array}{c} 0 \\ \vdots \\ 0 \end{array} \right]. \]  

(13)
satisfied, so LBD and TBD are both optimal. However, by comparing (10) and (14), Case 1 only happens when \( Z = 1 \), which is a 2-D NoC case with very high thermal limit. Case 2 shows the situation that thermal limit is low, leading to LBD nonoptimal and TBD optimal. Case 3 shows the situation that thermal limit is high, resulting in LBD optimal and TBD nonoptimal. Case 4 shows the situation that a middle thermal limit is given. In this case, none of LBD and TBD can guarantee maximal throughput, because their optimal criteria are not satisfied.

By comparing (5) and (13), we discover that these two schemes are extremes for traffic load distribution \( L \) and the corresponding buffer allocation \( K \). LBD has the most balanced vertical traffic distribution; TBD has the most unbalanced vertical traffic distribution. Because these two extremes cannot achieve the maximal throughput for Case 4, we need a new design scheme to amend the shortage of the LBD scheme and the TBD scheme, which will be introduced latter.

IV. PROPOSED ROUTING-BASED TRAFFIC MIGRATION AND BUFFER ALLOCATION

In this section, we present the proposed routing-based traffic migration and buffer allocation method to find the feasible design criteria between TBD and LBD for maximal throughput. First, we describe the proposed routing-based traffic migration method in Section IV-A. Then, we present the proposed vertical buffer allocation in Section IV-B for supporting the routing-based traffic migration method.

A. Proposed Routing-Based Traffic Migration Method

1) Desired Traffic Loading of Different Thermal Limits: The proposed routing-based traffic migration method is a proactive routing that changes the traffic loading \( L \) for improving the thermal-limited throughput. Because the key difference between 2-D NoC and 3-D NoC is the varying cooling efficiency of routers in different silicon layers, 3-D NoC has the lateral-homogeneous vertical-heterogeneous thermal characteristics. For each silicon layer (i.e., XY-plane), the traditional LBD scheme still works for eliminating hotspots. However, the vertical-heterogeneous thermal characteristics of 3-D NoC causes the LBD scheme and the TBD scheme fail to reach maximal throughput. To solve this problem, the routing path is decomposed into the lateral paths on XY plane and the vertical paths along Z direction. Then, we partition the proactive routing into two parts: the lateral routing and the vertical routing. The design goal of the lateral routing is to balance the lateral traffic loading on each XY plane; the design goal of the vertical routing is to accommodate the various vertical traffic loadings for different thermal limits.

For different thermal limit, the desired traffic loading is different. If the thermal limit is high, the temperature bounded traffic load \( L^{LDB} \) can be high. The throughput maximization problem is similar to traditional NoC performance optimization problem. Therefore, the LBD scheme is preferred, and the target vertical traffic distribution is (5), as shown in Fig. 4(a). In contrast, if the thermal limit is low, the throughput is more limited by temperature. Hence, eliminating overheated hotspots is primary objective for proactive routing. Therefore, TBD scheme is preferred, and the target vertical distribution is (13), as shown in Fig. 4(c). If the thermal limit is in the middle, both temperature and channel loading may limit the throughput. Therefore, a vertical traffic distribution between (5) and (13) is preferred to achieve maximum throughput, as shown in Fig. 4(b).

2) VDLAPR Algorithm: To accommodate these three traffic loadings shown in Fig. 4, we propose VDLAPR algorithm. Fig. 5 shows the routing scheme of traffic migration. Traditional dimension-ordered routing algorithms, such as XYZ routing in Fig. 5(a), cannot accommodate all the vertical traffic requirements of different thermal limits shown in Fig. 4. We relieve the minimal routing constraint in vertical direction and define downward level \( D \), a scalar control parameter representing the vertical misroute distance for gradually redistributing the traffic. Assume the source address of a packet is \((X, Y, Z)\), and the destination address is \((X, Y, Z)\). With given downward level, the address of target lateral routing layer of the packet, defined as \(Z^{TLRL}\), can be found by the following equation:

\[
Z^{TLRL} = Z^S + D.
\]

Because the routing layer is within the 3-D mesh, the actual lateral routing layer \(Z^{TLRL}\) is bounded by \([1, Z]\), as

\[
Z^{LRL} = \begin{cases} 
1, & \text{if } Z^{TLRL} < 1 \\
Z, & \text{if } Z^{TLRL} > Z \\
Z^{TLRL}, & \text{otherwise}.
\end{cases}
\]

To deliver a packet from source to destination, it is first injected to the source router, which is at \((X^S, Y^S, Z^S)\). The source router routes the packet to the intermediate router at \((X^S, Y^S, Z^{LRL})\), and then the intermediate router routes the packet to \((X^D, Y^D, Z^{LRL})\). When the packets reach the router that has the same XY address as the destination, they are directly routed to the destination \((X^D, Y^D, Z^D)\). Fig. 5(b)
illustrates an example. \( D = 1 \) represents the case that the lateral routing path is migrated by one layer.

For the lateral routing, in our previous work [17], \( XY \) routing is adopted for simplicity. However, for nonuniform random traffic patterns, \( XY \) routing is not able to balance the loading of the network. Many 2-D adaptive routing algorithms had been presented for balancing the traffic loading on \( XY \) plane. These adaptive routing algorithms adaptively select the routing paths from the set of output channels according to the channel status. To determine the set of output channels, odd–even turn model [30] is well known and widely accepted. In comparison with the other turn models, such as west-first, north-last, and negative-first. The odd–even turn model can provide more even adaptiveness to nonuniform traffic patterns. The selection strategy we adopted in the proposed lateral routing path is migrated by one layer.

The examples of path diversity are shown in Fig. 5(c). For per other turn models, such as west-first, north-last, and negative-first. The odd–even turn model can provide more even adaptiveness to nonuniform traffic patterns. The selection strategy we adopted in the proposed lateral routing algorithm is neighbor-on-path (NOP) [31], which outperforms than traditional output-buffer-level selection strategy.

The examples of path diversity are shown in Fig. 5(c). For the eastward source-destination pairs, such as \( S_3 - D_3 \) and \( S_5 - D_6 \), the east-north (EN) turn and east-south (ES) turn are only allowed at even columns (even \( X \) addresses). For the westward pairs, such as \( S_4 - D_4 \) and \( S_5 - D_5 \), the west-north (WN) turn and west-south (WS) turn are only allowed at odd columns (odd \( X \) addresses).

If thermal limit is high, setting \( D = 0 \) can keep the lateral routing layer unchanged. If thermal limit is low, setting a larger can form a traffic distribution more like Fig. 4(c). From (16) and (17), setting \( D \geq Z \) leads to the result that all the lateral routing paths lie on the bottom layer, which is the extreme case that makes the vertical traffic loading of the lateral routing closest to (13). The power profile of the 3-D NoC is closest to the TBD scheme, approaching (12). For medium thermal limits, setting suitable \( D \) is essential to improve the thermal-limited throughput. Because the vertical traffic loading is redistributed by the proposed routing-based traffic migration method, the channel loading will be unbalanced. In Section IV-B, we will propose the vertical buffer allocation method to solve this problem, and we introduce our systematic design flow to find the most suitable downward level in Section VI.

Fig. 6 shows an example of VDLAPR in a \( 4 \times 4 \times 4 \) 3-D NoC. Assume the routers in the purple are traffic hotspots. As mentioned before, the path from source to destination can be decoupled to vertical paths and lateral paths. In vertical direction, by setting different downward levels, downward routing changes the \( Z \) location of the lateral path. The lateral path selection is determined by the lateral adaptive routing. Because the channels of traffic hotspot routers are more congested, the NOP selection scheme will not select these channels as output channels.

3) Proof of Deadlock Freedom in Terms of Turn Model: The proposed routing algorithm adopts turn model-based approach to prevent deadlock. In 3-D, we need to prove that circular waiting will not occur on \{\( XY, YZ, XZ \}\) planes and also the combination of \( \{XY – YZ, XY – XZ, YZ – XY, YZ – ZX, XZ – XY, XZ – YZ\}\) planes. The deadlock-freedom of the turns on \( XY\)-plane is achievable through the odd–even turn model. The circular waiting condition is never true on \( XZ\)-plane and \( YZ\)-plane since we remove the turns of Up-then-East (UE), Up-then-West (UW), Up-then-North (UN), and Up-then-South (US), as shown in Fig. 7. Therefore, we can guarantee the routing algorithm is deadlock-free on the combination of \( \{YZ – XY, YZ – XZ, XZ – XY, XZ – YZ\}\) planes and also deadlock-free on the combination of \( \{XY – YZ, XY – XZ\}\) planes.

B. Proposed Vertical Buffer Allocation Method

To relieve the channels with heavy traffic loading in the lower silicon layers, adjusting the depth of buffer is an effective method. Ogras et al. [32] formalized the buffer sizing problem, and Hu et al. [33] proposed the design flow and the required analytical model for the buffer allocation of 2-D NoC. In this paper, we follow the same design concept to derive the model for 3-D NoC, and marginalize the model for the VDLAPR algorithm.

1) Derivation and Marginalization for Vertical Queuing Model of 3-D NoC: Here, we extend the queuing model proposed in [34] and [33], which adopts M/M/1/K queues. Several assumptions are made for simplification of the derivation: Poisson packet arrival time and exponential service time for each channel are required. Besides, each packet is viewed as an atomic data in this derivation. We use the north input channel of the router at \((x, y, z)\) as an example. Assume the network is not overloaded (i.e., \( \lambda_{x,y,z,dir} < \mu_{x,y,z,dir} \)) in steady state, the main target of buffer allocation is to calculate the full probability for each channel. For channel \( C_{x,y,z,N} \), the full probability \( \rho_{x,y,z,N} \) is

\[
\rho_{x,y,z,N} = \frac{1 - \rho x,y,z,N}{1 + \rho x,y,z,N} \times \rho_{x,y,z,N}
\]

where

\[
\rho_{x,y,z,N} = \frac{\lambda_{x,y,z,N}}{\mu_{x,y,z,N}}
\]

(18)
The full probability depends on two parameters: arrival rate \( \lambda_{x,y,z,N} \) and service rate \( \mu_{x,y,z,N} \), as shown by (18). One problem is to obtain the arrival rate for each channel. In the original works [34], [33], the arrival rate is computable if the routing path is deterministic. However, the path selection in this paper depends on the channel status, which is not predictable. Therefore, the arrival rate has to be obtained through simulation. The other problem is that the service rate of a cascaded M/M/1/K queue depends on the full probability of all its downstream channels. Consider some of the downstream packets are delivered toward east, which is to say \( C_{x+1,y,z,W} \). The full probability \( b_{x+1,y,z,W} \) will affect the effective service rate of \( C_{x,y,N} \). The effective service rate of \( C_{x+1,y,z,W} \) can be approximated by \( 1/b_{x+1,y,z,W} \). When \( C_{x+1,y,z,W} \) is full, the reciprocal of the decreasing rate of the occupation in queue equals to the average waiting time for entering the queue. Hence, the average waiting time for entering the queue of \( C_{x+1,y,z,W} \) can be approximated by

\[
W_{x+1,y,z,W} = \left( \frac{1}{b_{x+1,y,z,W}} - \lambda_{x+1,y,z,W} \right)^{-1}. \tag{19}
\]

At steady state, by applying Little’s formula, the average waiting time for entering the queue of \( C_{x+1,y,z,W} \) can be written as

\[
W_{x+1,y,z,W} = \left( \frac{1}{b_{x+1,y,z,W}} - \lambda_{x+1,y,z,W} \right)^{-1} \cdot \bar{\lambda}_{x,y,z,N} \times \bar{\mu}_{x,y,z,N} \tag{20}
\]

Substituting \( W_{x+1,y,z,W} \) in (19) and (20), we get the effective service rate toward east of \( C_{x,y,N} \) as follows:

\[
\bar{\mu}_{x,y,z,N} = \frac{1}{b_{x+1,y,z,W}} - \lambda_{x+1,y,z,W} + \bar{\mu}_{x,y,z,N} \times \bar{\lambda}_{x,y,z,N} \tag{21}
\]

and

\[
\bar{\mu}_{x,y,z,N} = \sum_{\text{dir}} \bar{\mu}_{x,y,z,N} \times \bar{\mu}_{x,y,z,N}. \tag{22}
\]

By merging the cascaded the M/M/1/K queues, we can get a simplified model for average queue length of \( C_{x,y,z,N} \), which can be approximated by

\[
q_{x,y,z,N} = \frac{\bar{\lambda}_{x,y,z,N}}{\mu_{x,y,z,N} - \bar{\lambda}_{x,y,z,N}}. \tag{23}
\]

By viewing the two queues as two independent queues, the average queueing length is equal to the sum of the length of the two queues, which can be described as

\[
q_{x,y,z,N} = \frac{\bar{\lambda}_{x,y,z,N}}{S^{-1} - \bar{\lambda}_{x,y,z,N}} + \frac{\bar{\lambda}_{x,y,z,N}}{\bar{\mu}_{x,y,z,N} - \bar{\lambda}_{x,y,z,N}}. \tag{24}
\]

Substituting \( q_{x,y,z,N} \) in (23) and (24), we can get the channel service rate \( \mu_{x,y,z,N} \)

\[
\mu_{x,y,z,N} = \bar{\lambda}_{x,y,z,N} \left( S^{-1} - \bar{\lambda}_{x,y,z,N} + \frac{1}{\bar{\mu}_{x,y,z,N} - \bar{\lambda}_{x,y,z,N}} \right)^{-1}. \tag{25}
\]

Then we can compute the full probability \( b_{x,y,z,N} \) for \( C_{x,y,z,N} \). Similar computation is applied to all the channels in the network. The marginalized service rate \( \mu_{z,N} \) for the north input channel in layer \( z \) can be described as

\[
\mu_{z,N} = \bar{\lambda}_{z,N} + \left( \frac{1}{S - 1 - \bar{\lambda}_{z,N}} + \frac{1}{\bar{\mu}_{z,N} - \bar{\lambda}_{z,N}} \right)^{-1} \tag{26}
\]

which is dependent on the marginalized arrival rate \( \bar{\lambda}_{z,N} \) and the marginalized effective service rate \( \bar{\mu}_{z,N} \). Here, we simplify the computation of by directly taking average of \( \bar{\lambda}_{x,y,z,N} \) in each layer

\[
\bar{\lambda}_{z,N} = \frac{1}{XY} \sum_{x,y} \bar{\lambda}_{x,y,z,N}. \tag{27}
\]

Similarly, \( \bar{\mu}_{z,N} \) is calculated by taking average of \( \bar{\mu}_{x,y,z,N} \). Therefore, we can get \( \bar{\mu}_{z,N} \) the for calculating the full probability \( b_{z,N} \) for layer \( z \) as

\[
b_{z,N} = \frac{1 - \rho_{z,N}}{1 - \rho_{z,N}} \times \rho_{z,N} \times \frac{k_{z,N}}{\mu_{z,N}}. \tag{28}
\]

2) Design Flow for Vertical Buffer Allocation: As mentioned before, we derive the queuing model for 3-D NoC. Buffer allocation for wormhole-based router is still an open problem as mentioned in [32]. When packet size is small, the behavior of fit arrival is similar to Poisson process. Because the traffic pattern is assumed known in design stage for offline thermal management of 3-D NoC, we can run statistics for estimating the arrival rate. We use Poisson arrivals see time averages [35] to obtain the packet arrival rate and the effective fit arrival rate for each layer. For implementation in our simulator, we use a counter for each input channel of every router. After the network is warmed up, the packet arrival rate and the effective fit arrival rate can be estimated by the tick number of the counter. The proposed design flow for vertical buffer allocation of 3-D NoC is shown in Fig. 8, in which we use a four-layer 3-D NoC as an example. The design flow is applied to each direction of channels, including north, south, east, west, up, and down. We assumed the base router is an isotropic six-way structure for connecting up, down,
east, west, north, and south directions, and each channel has a minimum depth for a buffered mesh network. The initialization procedure is based on this assumption, so the queueing depth is initialized to unity. The total buffer budget count in flit is \( N_B \) for each direction. First, we run statistics for downward level \( L \) to obtain the arrival rate \( \lambda_{x,y,z,N} \) of each channel. Then, we initialize the depth of each queue to one flit. For each downward level \( D \), the statistic of arrival rate is obtained by counting the number of arrived flits for each channel after the network warms up, to the end of simulation. The full probability \( b_Z \) of each layer is calculated, and the buffer depth of the layer that has maximum \( b_Z \) is increased. Finally, the iteration runs until all the total budgets are consumed, and the buffer allocation is done.

V. PERFORMANCE EVALUATION OF PROPOSED METHODS

Here, we start from model construction and simulation setting, and then we use two experiments to support our claim. We model the 2-D NoC system implemented in [5] and stack it to multiple layers. Fig. 9 shows the employed manually constructed physical model. For temperature simulation, Hotspot [4] is integrated into our simulator. To obtain more precise results of thermal simulation, the grid-level simulation of Hotspot is applied. For the granularity alignment of thermal model between the proposed analytical model and the simulation results of Hotspot, the architecture-level temperature of a node is translated by averaging the grids within the floorplan region of the node. We use the default parameters of Hotspot for the thickness and the thermal conductance of the four silicon layers and the four isolator layers. For traffic simulation, Noxim [36] is adopted and modified for using vertical buses to connect routers in different silicon layers. Wormhole flow control and random arbitration for switch allocation are adopted in all NoC routers. The geometry and the power model of each single tile are based on Intel’s 80-core processor [5]. The power scaling factor is set to one-fourth of the 80-core processor in all the following simulations.

For the following experiments, if there is no special description, we use the following settings, which are defaults in Hotspot [4]. The thermal conductance of heat sink is set to 400 Wm\(^{-1}\)K\(^{-1}\), and the ambient temperature is set to 25 °C. To reduce the impact of HoL blocking, the packet size equals to two flits, because HoL blocking is not the focus of this paper. Without buffer allocation, the default queue depth of each input channel is four flits. Besides, the network size is \( 8 \times 8 \times 4 \).

A. Reducing Maximum Temperature by Proposed Routing-Based Traffic Migration Method

We use the temperature profile of the modeled \( 8 \times 8 \times 4 \) 3-D NoC system to show the effectiveness of the proposed VDLAPR algorithm. The normalized flit injection rate is set to 0.08 to make maximum temperature high enough without saturating the network. Fig. 10(a) shows the temperature profile with uniform random traffic offered. The mean of temperature, the standard deviation of temperature, and the maximum temperature are all decreased as the downward level increases. Fig. 10(b) shows the temperature profile with transpose traffic offered. Similar temperature trend can be found, except for the overall temperature standard deviation, which is affected by the natural lateral traffic imbalance of transpose traffic. In both traffic patterns, as downward level increases, the vertical temperature difference is reduced. These experiments show VDLAPR can be used to move the traffic loading profile from TBD scheme toward TBD scheme, resulting in lower peak temperature. However, as the traffic loading profile approaches TBD scheme, the network also suffers from traffic congestion. In Fig. 10(b), the temperature decreases in the \( D = 2 \) and \( D = 3 \) cases, because the bandwidth bound is reached. In these cases, the bottom layer is already saturated. Hence, the traffic in the upper layer becomes less. The total throughputs become 0.073 and 0.069, resulting in less power consumption.

B. Improving Performance by Proposed Vertical Buffer Allocation Method

In this section, we show the performance improvement with the proposed vertical buffer allocation. The configuration is the traditional LBD scheme as \( D = 0 \). Therefore, the 3-D NoC has balanced traffic loading, and the buffer depth are all the same for each channel in the four layers. In this case, vertical buffer allocation does not bring performance improvement. However, for all the other three cases, vertical buffer allocation improves the saturation throughput, as shown in Fig. 11.
For $D = 1$, the saturation throughput is improved from 0.14 to 0.16 flits/node/cycle. For $D = 2$, the saturation throughput is improved from 0.09 to 0.10 flits/node/cycle. For $D = 3$, which migrates most lateral routing traffic to the bottom layer and most like the traditional TBD scheme, the saturation throughput is improved from 0.086 to 0.095 flits/node/cycle. The average performance improvement is around 11%.

VI. PROPOSED SYSTEMATIC DESIGN SCHEME AND DESIGN EXAMPLES

In this section, we present the proposed design scheme for maximizing the thermal-limited performance of 3-D NoC. First, we present the design flow of the scheme in Section VI-A. Then, we apply the design flow for typical thermal limits in Section VI-B as design examples.

A. Proposed Design Flow of the Systematic Design Scheme

Fig. 12 shows the proposed design flow for 3-D NoC to combine the proposed routing-based traffic migration method and the proposed vertical buffer allocation method. The flow is composed of two loops.

1) $D$-loop: It moves in the expanded design space and determines the buffer depth of each layer.

2) $\mu$-loop: It pushes the network to limit by increasing the injection rate $\mu$.

The traffic-thermal model has to be created before entering the design flow. $D$-loop searches $D$ from zero, where the network is configured as the traditional LBD scheme. For each downward level, the optimal buffer depth is set as $K$, which is determined by the proposed buffer allocation method.

Then, the traffic-thermal co-simulator simulates for obtaining the distributions for a certain injection rate $\mu_{IR}$, and we check whether the temperature constraint and maximum average latency constraint are violated. The injection rate is gradually increased to find the achievable throughput, which is the maximal $\mu_{IR}$ that satisfies both constraints. If the injection rate increased in $\mu$-Loop violates latency constraint before temperature constraint, this configuration is channel loading limited. Otherwise, if the network is thermal-limited, the temperature constraint will be violated earlier. If downward level does not reach downward limit $D_{Limit}$, we keep searching in the expanded design space by increasing $D$ for further power migration. Finally, the configuration $D$ with maximal achievable throughput will be chosen. Because $D$-loop exhaustively searches the expanded design space for the configuration with maximal achievable throughput, the final result is guaranteed optimum.

B. Design Examples of Proposed Systematic Design Scheme

In this section, we demonstrate the design examples by applying the proposed design scheme. With consideration of the latency in real application is not infinite, the maximum average latency is set to 500 cycles for estimating the saturation throughput. The thermal limit is set from 100 °C to 150 °C to cover Cases 2 and 4 of Table II, and a 200 °C case is used to demonstrate the Case 3 of Table II. The buffer constraint $N_B$ is set to 16 flits for VBA. Uniform random traffic and hotspot traffic are offered as an example. With known traffic pattern and given thermal limit, the $D$-loop traverses downward level from 0 to 3, and VBA produces the optimal buffer depth for each downward level, as in Table V. Then, the achievable throughput is obtained by simulation with applying each configuration with increasing injection rate.

We use the Table III to show the achievable throughput comparison for uniform random traffic. First, we observe the achievable throughput of the traditional LBD scheme. The achievable throughput increases as the thermal limit becomes higher, but LBD is not the scheme achieves maximal throughput. The reason is that the thermal-limited bound of LBD is tighter than the bandwidth bound of LBD, so the optimal criterion of LBD is not satisfied. Therefore, LBD is optimal only for the 200 °C design case, in which the optimal criterion of LBD scheme is satisfied. Second, we observe the achievable throughput of the traditional TBD scheme. The achievable throughput increases as the thermal limit becomes higher and
TABLE III
RANDOM TRAFFIC DESIGN EXAMPLES, ACHIEVABLE THROUGHPUT (FLIT/NODE/CYCLE)

<table>
<thead>
<tr>
<th>Design Case</th>
<th>TBD is optimal</th>
<th>TBD is nonoptimal, and LBD is nonoptimal</th>
<th>LBD is optimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermal limit</td>
<td>100 °C</td>
<td>110 °C</td>
<td>120 °C</td>
</tr>
<tr>
<td>LBD scheme</td>
<td>0.0554</td>
<td>0.0660</td>
<td>0.0767</td>
</tr>
<tr>
<td>TBD scheme</td>
<td>0.0642</td>
<td>0.0764</td>
<td>0.0850</td>
</tr>
<tr>
<td>Proposed scheme</td>
<td>0.0642</td>
<td>0.0764</td>
<td>0.0885</td>
</tr>
<tr>
<td>Downward level</td>
<td>D = 3</td>
<td>D = 3</td>
<td>D = 3</td>
</tr>
<tr>
<td>Improvement over LBD</td>
<td>15.8%</td>
<td>15.6%</td>
<td>15.5%</td>
</tr>
<tr>
<td>Improvement over TBD</td>
<td>0.0%</td>
<td>0.0%</td>
<td>4.2%</td>
</tr>
</tbody>
</table>

TABLE IV
HOTSPOT TRAFFIC DESIGN EXAMPLES, ACHIEVABLE THROUGHPUT (FLIT/NODE/CYCLE)

<table>
<thead>
<tr>
<th>Design Case</th>
<th>TBD is optimal</th>
<th>TBD is nonoptimal, and LBD is nonoptimal</th>
<th>LBD is optimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermal limit</td>
<td>100 °C</td>
<td>110 °C</td>
<td>120 °C</td>
</tr>
<tr>
<td>LBD scheme</td>
<td>0.0599</td>
<td>0.0664</td>
<td>0.0769</td>
</tr>
<tr>
<td>TBD scheme</td>
<td>0.0699</td>
<td>0.0761</td>
<td>0.0760</td>
</tr>
<tr>
<td>Proposed scheme</td>
<td>0.0699</td>
<td>0.0761</td>
<td>0.0824</td>
</tr>
<tr>
<td>Downward level</td>
<td>D = 3</td>
<td>D = 3</td>
<td>D = 3</td>
</tr>
<tr>
<td>Improvement over LBD</td>
<td>7.2%</td>
<td>7.1%</td>
<td>7.1%</td>
</tr>
<tr>
<td>Improvement over TBD</td>
<td>0.0%</td>
<td>0.0%</td>
<td>8.4%</td>
</tr>
</tbody>
</table>

TABLE V
CHANNEL BUFFER DEPTH FROM RESULTS OF VERTICAL BUFFER ALLOCATION

<table>
<thead>
<tr>
<th>Layer</th>
<th>Z = 0</th>
<th>Z = 1</th>
<th>Z = 2</th>
<th>Z = 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Traffic</td>
<td>D = 1</td>
<td>D = 2</td>
<td>D = 3</td>
<td></td>
</tr>
<tr>
<td>Lateral: E, S, W, N</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Up</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Down</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Hotspot Traffic</td>
<td>D = 1</td>
<td>D = 2</td>
<td>D = 3</td>
<td></td>
</tr>
<tr>
<td>Lateral: E, S, W, N</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Up</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Down</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Then saturates at 0.086 flit/node/cycle. For the 100 °C and 110 °C cases, TBD achieves maximal throughput because the optimal criterion of TBD scheme is satisfied. However, as the thermal limit becomes higher, the thermal-limited bound of TBD is relaxed. Then, the optimal criterion of TBD scheme is not satisfied. The bandwidth bound of TBD scheme is tighter.

Comparing to LBD and TBD scheme, the proposed design scheme always achieves maximal throughput because it covers both LBD and TBD. For the cases with 100 °C and 110 °C thermal limits, the proposed scheme is identical to TBD; for the case with 200 °C thermal limit, the proposed scheme is identical to LBD. Besides, for the cases in which the optimal criteria of LBD and TBD are not satisfied, the proposed scheme can reach the optimal configuration in the design space expanded between LBD and TBD. For these TBD-nonoptimal LBD-nonoptimal cases, the proposed scheme improves the achievable throughput from 2.7% to 45.2%. Similarly, the downward level and the depth of channel buffers in each layer produced by the proposed design scheme for hotspot traffic can be found in the upper part in Table V.

Table IV shows the achievable throughput comparison for design examples with hotspot traffic offered. Eight hotspot are set in the network, and the traffic ratio is 2% of total amount for each router. Comparing to other routers, in which the traffic ratio is 0.4% of total amount, the hotspot locations are about 5 times burdened. Similar to the results in Table III, for the cases in which the optimal criteria of LBD and TBD are not satisfied, the proposed scheme can reach the optimal configuration in the design space expanded between LBD and TBD. For these TBD-nonoptimal LBD-nonoptimal cases, the proposed scheme improves the achievable throughput from 2.7% to 45.2%. Similarly, the downward level and the depth of channel buffers in each layer produced by the proposed design scheme for hotspot traffic can be found in the lower part in Table V.

VII. Conclusion

In this paper, we demonstrated the performance-temperature problem encountered for off-line thermal management of a 3-D NoC system. Due to the different cooling efficiency on vertical direction, the optimal criteria of traditional LBD and TBD schemes may not be satisfied. We proposed an analytical model to analyze the correlation between performance...
Note that the first term represents the temperature of the current router; the second term represents the temperature of the memory and the processor in the current tile; the third term represents the thermal coupling effect of the lateral neighbor memory and processor; and the fourth term represents the thermal coupling effect of the vertical aligned routers.

According to Fourier’s heat conduction law, heat flow rate is in proportion to the temperature difference $\Delta T$, cross-sectional surface area $A$, conductivity $k$, and the reciprocal of distance $(\Delta d)^{-1}$. Therefore, the Fourier’s heat conduction law can be rewritten as

$$\frac{\Delta Q}{\Delta t} = -k \cdot \left(\frac{A_{\text{interface}}}{\Delta d}\right) \cdot \Delta T = -k \cdot \beta \cdot \Delta T.$$  (31)

Assume there is only one heat sink attached to the chip in the bottom, as shown in Fig. 13(a). In steady state, the main heat flow in the system is in toward heat sink for heat dissipation, as shown in Fig. 13(b). Besides, due to the 3-D IC physical characteristics, $\beta$ of vertical neighbor routers are much larger than that of the lateral neighbor routers. The reason is that the cross-sectional surface area between the source router and its vertical neighbor router is much larger than the cross-sectional surface areas between its lateral neighbors. Literally the thermal impact coefficient between a source node and its vertical neighbor router is typically 16 times larger than the thermal impact coefficient between the source node and its lateral neighbor [10], [37]. For 3-D NoC routers, the lateral coupling is even smaller, due to the lateral neighbor routers are much larger than the cross-sectional surface areas of the lateral neighbor routers. Literally the thermal impact coefficient between a source node and its lateral neighbor router is typically 8 times larger than the thermal impact coefficient between the source node and its lateral neighbor. Literally the thermal impact coefficient between a source node and its vertical neighbor node is typically 16 times larger than the thermal impact coefficient between the source node and its lateral neighbor [10], [37]. For 3-D NoC routers, the lateral coupling is even smaller, due to the lateral neighbor routers are not directly placed side-by-side to the source router. Therefore $T_{s,y,z}^R$ can be further reduced to

$$t_{s,y,z}^R \approx p_{s,x'}^R \cdot \xi_{s,y',z',x,y,z}^R + \sum_{i \neq z} p_{s,x',y',z',x,y,z}^R \cdot \xi_{s,y',z',x,y,z}^R \quad (32)$$

which means the temperature of a router is mainly determined by its power consumption and the power consumption of its vertical neighbor routers. Assume there is only one heat sink attached to the chip in the bottom, as shown in Fig. 13(a). As show in [38], heat sink takes about 90% of the total heat dissipation. Therefore, in steady state, the main heat flow in the system is toward the heat sink, and the top silicon layer of a typical 3-D NoC is thermal-dominant. As shown in Fig. 13(b), the coupling form can be approximated by

$$T_{s,y,z}^R \approx \sum_{i \neq z} p_{s,x',y',z',x,y,z}^R \cdot g_{s,x',y',z'} \quad (33)$$

which is equivalent to the compact thermal model of Fig. 13(c). Namely, $g_{s,x,y,z}$ represents the thermal conductance between router $(x, y, i)$ and router $(x, y, i + 1)$; $g_{s,x,y,z}$ is the thermal conductance between router $(x, y, Z)$ and ambience. According to this model, the temperature of a router is mainly determined by the power consumption of the current router and the power consumption of the vertical aligned routers.

**ACKNOWLEDGMENT**

The authors would like to thank the reviewers and the editors for their helpful comments. The authors would also like to thank S.-Y. Lin, E.-J. Chang, H.-K. Hsin, and the members of NTU Access Lab.
REFERENCES


An-Yeu (Andy) Wu (S’91–M’96–SM’12) received the B.S. degree from National Taiwan University, Taipei City, Taiwan, in 1987, and the M.S. and Ph.D. degrees from the University of Maryland, College Park, in 1992 and 1995, respectively, all in electrical engineering.

He joined the Department of Electrical Engineering, and the Graduate Institute of Electronics Engineering, National Taiwan University (NTU), in 2000, where he is currently a Professor, and, since 2012, he has been the Deputy Director, Graduate Institute of Electronics Engineering. From 2007 to 2009, he was on leave from NTU and was the Deputy General Director of the SoC Technology Center, Industrial Technology Research Institute, Hsinchu, Taiwan, where he supervised the Parallel Core Architecture VLIW DSP Processor and Multi-core/Android SoC Platform projects. His current research interests include low-power/high-performance VLSI architectures for digital signal processing and communication applications, adaptive and multirate signal processing, reconfigurable broadband access systems and architectures, and system-on-chip/network-on-chip platform for software and hardware co-design.

Dr. Wu was a recipient of the Outstanding EE Professor Award from The Chinese Institute of Electrical Engineering, Taiwan, in 2010. He was an Associate Editor of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I: REGULAR PAPERS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: EXPRESS BRIEFS, and the IEEE TRANSACTIONS ON SIGNAL PROCESSING. He is currently an Associate Editor of the Journal of Signal Processing Systems (JSPS), and was the Lead Guest Editor of the special issue of the 2010 IEEE Workshop on Signal Processing Systems (SiPS) in JSPS, in November 2011. He was on the technical program committees of many major IEEE International Conferences, such as SiPS, AP-ASIC, ISCAS, ISPACS, ICME, SOCC, and A-SSCC. He is currently the Chair of the VLSI Systems and Architectures (VSA) Technical Committee in the IEEE Circuits and Systems (CAS) Society.