2010 Fourth ACM/IEEE International Symposium on Networks-on-Chip

Traffic- and Thermal-Aware Run-Time Thermal Management Scheme for 3D NoC Systems

Chih-Hao Chao, Kai-Yuan Jheng, Hao-Yu Wang, Jia-Cheng Wu, and An-Yeu Wu
Graduate Institute of Electronics Engineering, National Taiwan University
Taipei 10617, Taiwan ROC

Abstract—**Three-dimensional network-on-chip (3D NoC), the combination of NoC and die-stacking 3D IC technology, is motivated to achieve lower latency, lower power consumption, and higher network bandwidth. However, the length of heat conduction path and power density per unit area increase as more dies stack vertically. Routers of NoC have comparable thermal impact as processors and contributes significant to overall chip temperature. High temperature increases the vulnerability of the system in performance, power, reliability, and cost. To ensure both thermal safety and less performance impact from temperature regulation, we propose a traffic- and thermal-aware run-time thermal management (RTM) scheme. The scheme is composed of a proactive downward routing and a reactive vertical throttling. Based on a validated traffic-thermal mutual-coupling co-simulator, our experiments show the proposed scheme is effective. The proposed RTM can be combined with thermal-aware mapping techniques to have potential for higher run-time thermal safety.**

Keywords - traffic-aware; thermal-aware; run-time thermal management; routing; throttling; 3D NoC; 3D IC

I. INTRODUCTION

As the complexity of the System-on-Chip (SoC) grows, on-chip interconnections gradually dominate the system performance. Network-on-Chip (NoC) has been proposed as a novel, practical and efficient communication infrastructure [1]. Recently, die-stacking three-dimensional (3D) IC technology is emerging for its capability to reduce wire delays by connecting with shorter vertical connections - Through Silicon Via (TSV) [2][3]. The combination of NoC and TSV, 3D NoC, is motivated to achieve lower transmission latency, lower network power consumption, higher device density, and higher platform bandwidth [4][5].

Thermal issues are significant challenges for developing 3D IC and also 3D NoC. As more dies stacked vertically, power density (in W/m^2) increases, and the length of heat conduction path increases. High temperature results in longer propagation delay and increases the leakage power. A chip operating above its thermal limit may generate incorrect output data and suffer from reliability degradation. Besides, heat also makes the cooling and packing cost increase. To solve the heat problem, various thermal management schemes have been proposed for 2D and 3D ICs, especially for high performance designs like chip multi-processor (CMP) [6] and NoC [7].

Keeping high performance under a certain thermal limit is the major goal of most thermal management schemes and thermal-aware design techniques. By characterizing thermal profile of the MIT Raw chip, [7] shows that NoC has comparable thermal impact as processors and contributes significant to overall chip temperature. Due to the high switching activity and the relative small area, [8] shows the average power density of a NoC router is even higher than the floating-point MAC and memory on Intel's 80-core processor. The thermal management scheme proposed in [7], ThermalHerd, regulates the network temperature and eliminates thermal emergencies of 2D NoC. With the distributed temperature-aware traffic throttling and the proactive/reactive thermal-correlation based routing, the performance degradation from regulation can be controlled. However, the vertical stacking makes the thermal design more difficult, and the traditional 2D techniques for NoC have their limitation. For the 3D case, herding the majority of the switching activities close to the heat sink is a specialized technique. For the 3D-integrated microprocessors, [9] proposes a micro-architecture design technique, Thermal Herding, to control hotspots. By partitioning the processor into multiple layers, Thermal Herding steers the majority of switching activity to the die that is closest to the heat sink. For 3D CMP that place cores on multiple layers, [6] identifies the critical concept for optimal thermal management, and derives guidelines for their near-optimal policy.

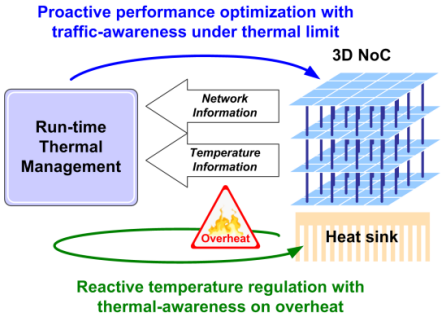

Figure 1. The proposed traffic- and thermal-aware run-time thermal management (RTM) for 3D NoC

To ensure the thermal safety and little performance degradation for 3D NoC, we propose a new run-time thermal management (RTM) scheme in this paper, as Fig. 1. We assume temperature distribution is available through the distributed thermal sensors in 3D NoC, and the scheme is composed of two techniques:

- *Traffic-aware downward routing*: an adaptive proactive technique to prevent heat accumulating inside 3D NoC and maximize achievable performance under network constraints. With monitoring the network status to prevent congestion, the workloads are migrated toward heat sink adaptively.
- *Thermal-aware vertical throttling*: an adaptive reactive technique to decrease temperature in emergency. We improve the traditional distributed traffic throttling with consideration of 3D characteristics, and provide a thermal-aware adaptation for network availability.

We develop and validate a traffic-thermal mutual-coupling co-simulation platform [15] for 3D NoC. The experimental result shows the proposed RTM scheme has better temperature controllability and less performance degradation on regulation. For temperature-limiting cases, the achievable throughput under the 80℃ thermal limit is improved around 7%. The average throttling time is reduced around 70%, and the average throttling ratio is reduced around 9-15%.

This work is supported in part by the National Science Council, Taiwan, ROC under Grant NSC98-2220-E-002-034.

978-0-7695-4053-5/10 $26.00 © 2010 IEEE
DOI 10.1109/NOCS.2010.32

The rest of paper is organized as the following. In Section II, we describe the problem and goal of RTM for 3D NoC. In Section III, the proposed traffic-aware and thermal-aware RTM scheme is described. In Section IV, the experiments are shown and discussed. The related work is introduced in Section V, and this paper is concluded in Section VI.

II. PROBLEM DEFINITION

A. Motivation

The major goal of the RTM for 3D NoC is jointly optimizing performance and temperature. The problem is complex due to the goal and the optimization constraint change as temperature changes. If temperature is lower than a given temperature limitation (thermal limit), the goal is to maximize performance (achievable throughput). If temperature is higher than thermal limit, the goal is to decrease the temperature of overheat routers with minimum performance impact.

The temperature distribution of a 3D NoC is correlated to power distribution, which depends on both application mapping and packet routing. The mapping determines the computation power, and the routing determines the communication power. For computation-intensive tasks, such as processors with iterative arithmetic operations, the power may be dominant by the local processor and memory. For communication-intensive tasks, such as a router that neighbor routers transmit many packets to it, the power consumption may be dominant by the traffic-hotspot router. We assume that mapping is fixed for RTM of NoC, and we focus on routing-based approach.

B. Problem Description

For 3D NoC, balanced traffic distribution does not result in balanced thermal distribution. The optimization is simultaneously constrained by network bandwidth and thermal limitation due to the mutual coupling effects. The major difference between 2D NoC and 3D NoC is the enlarged difference of thermal characteristic among routers. Especially the vertical aligned routers. Assume the same ambient temperature, fixed-sized heat sink, and constant air flow velocity i.e. the same cooling environment for cooling for simplicity. According to Fourier's law, the router which is on the layer closer to the heat sink has higher cooling efficiency. Inversely, the router that is farther to heat sink is has lower cooling efficiency and is thermal dominant. The thermal coupling of vertical aligned tiles is much larger than the coupling of horizontal aligned tiles [6]. As Fig. 2, we simplify the modeling of the heat flow from a 3D temperature profile to a vertical 1D profile for the discussion of the following two cases. Equation (1) shows the heat flow formula of 1D geometry derived from Fourier's law. $\Delta Q/\Delta t$ is the amount of heat transfer per unit time in Watt. k is the heat conductivity of the material in $W{\cdot}m^{-1}{\cdot}K^{-1}$, and A_{cross} is the cross surface area. ΔT is the temperature difference between the ends, and Δx is the distance between the ends. Equation (1) states that the rate of heat flow through a homogeneous solid is directly proportional to the area of the cross section of the direction of heat flow, the temperature difference of two terminals, and the conductivity.

$$\frac{\Delta Q}{\Delta t} = -kA_{cross}\frac{\Delta T}{\Delta x} \qquad (1)$$

We use the two following extreme cases to show the first two problems of joint optimization of performance and temperature. The third problem is caused by the non-ideality of proactive workload migration. For simplicity, we denote bottom layer (lowest) as the layer closest to the heat sink, and top layer (highest) as the farthest. Besides, we assume distributed constant power trace for the following discussion.

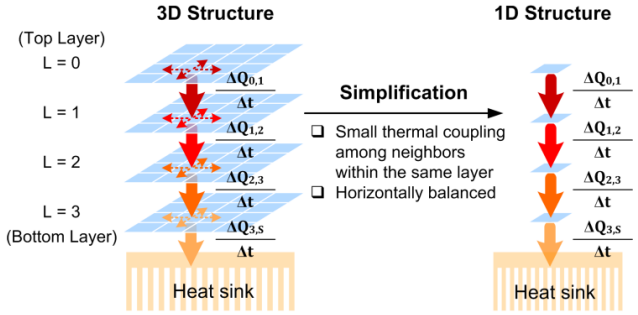

Figure 2. Thermal coupling and problem simplification

1) Temperature-limiting performance optimization

As mentioned above, maximize throughput needs to balance traffic to prevent channel loading above the bisection bandwidth. E.g., for uniformly distributed traffic on mesh or torus network, dimension-ordered routing such as XYZ routing has best performance, and is vertically balanced. The power consumption of each router is correlated to the traffic loading. For each layer, the injected heat from power equals to the conducted heat through the interface between layers. For the case that heat sink is only at one terminal of the 1D geometry, the temperature distribution forms a ladder. The heat flow is toward the terminal where heat sink is, i.e. heat flows gather from top to bottom and the difference of the temperature on each interface is getting larger. In this case, the top layer has highest temperature. If the temperature is low and no router is overheated, this scheme achieves best performance. When the power density is high, it is prone to overheat. This case is shown by Fig. 3(a).

2) Bandwidth-limiting thermal optimization

Assume we want to maximize heat conduction for a constant offered traffic. From (1) and Fig. 2, to maximize the heat transfer from bottom layer to heat sink, we assume the temperature of bottom layer is the highest. This assumption makes no heat conduction from top to bottom; only from bottom to top and sink. Since there is no other heat sink for inside layers, the steady state temperatures of all layers will be equal. Heat flow only exists between the interface of bottom layer and heat sink. Any workload migration from bottom layer to other layers makes heat generated in bottom layer decreases, and also decreases the heat flow through the interface between bottom layer and heat sink. In this case, all traffic is on the layer closest to heat sink. If the bottom layer is not saturated, this scheme has best heat conduction. Otherwise the network will suffer from congestion, which is shown by Fig. 3(b).

3) Non-ideality and assumption relaxation

The above two cases show the results of optimization for the steady state in the simplified ideal cases. For the transient state, several issues have to be considered. When the temperature is lower than the thermal limit, the optimization goal is to maximize performance with thermal consideration. Ideally this goal can be achieved by controlling the proactive work load migration. However, there are several assumptions too strong and have to be relaxed.

- The heat conduction between routers within a layer is not zero. Therefore the flow is not as simple as the 1D case. The router in higher layer may have lower temperature than the vertically aligned lower router in the transient state.
- The power trace over time is not constant and is not zero even without traffic. This relaxation makes the distribution of the temperature varies all the time.
- The granularity of control is not infinitely small: for implementation consideration, both the granularity of the amount of traffic migrated toward heat sink and the adjusting step of throttling ratio are quantized into several levels.

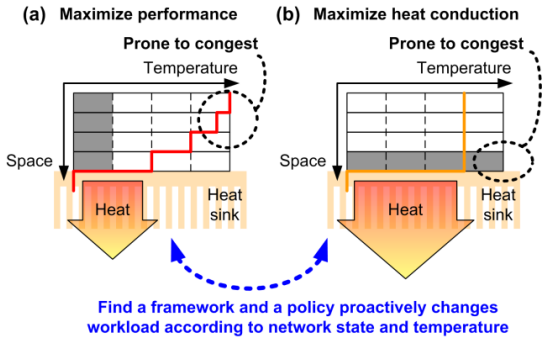

Figure 3. Two extreme optimization cases in the discussion of Section II.B.1) and Section II.B.2). For transient case, a framework and a policy are required.

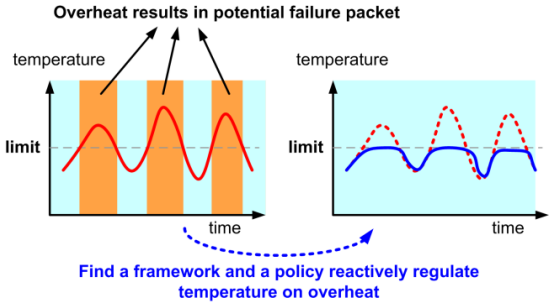

Figure 4. Temperature regulation in emergency

Therefore, the temperature varies over time. If the accumulated heat results in hotspot, overheat may occur. Hence a temperature monitoring infrastructure and a reactive mechanism are required for detection and cooling in emergency, which is shown in Fig.4. We define the thermal limit is a temperature should not be touched, and a threshold 1℃ below the limit is used for triggering the mechanism.

C. Design goal

Reference [10] shows that the change of temperature is much slower and smaller than the change of power. Therefore the required frequency of sensing the temperature and redistribute the information over the network is relatively small and negligible. With this assumption, the design goal of our RTM is as follows:

Given thermal limit, traffic distribution, network topology, router architecture, power model, and thermal model.

Find a framework and a policy for RTM.

Such that the achievable throughput is optimized with the constraint that the temperature never goes above the thermal limit, and the network has maximal availability.

III. Run-Time Thermal Management for 3D NoC

For RTM of 3D NoC, we start from the framework developed for 2D NoC [7], and extend it to the third dimension. Because the third dimension makes thermal management more difficult, the concepts and policies for thermal optimization of 3D CMP [6][9] are referred. Besides, for performance optimization of 3D NoC, benefits from the advantage of vertical links is necessary. Therefore we assume the crossbar-based network architecture is adopted [13][14]. Because the optimization goal varies on different temperature, an adaptive solution is preferred rather than a deterministic solution for achievable maximum performance and controlling the temperature under thermal limit.

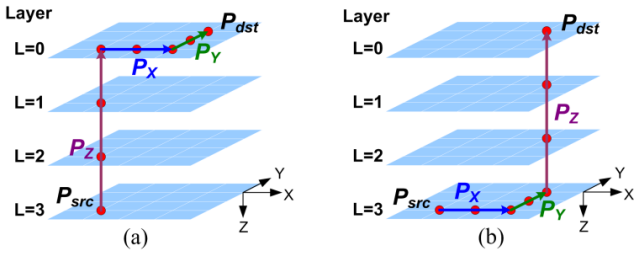

Figure 5. Power profile of routing paths in 3D NoC: (a) ZXY routing; (b) XYZ routing

A. Traffic-Aware Downward Routing

The two goals of the proposed traffic-aware downward routing are proactively migrating the power distribution from top to bottom and adaptively adjusting the amount of migration to prevent network saturation. Power distribution of a NoC is correlated to traffic distribution, and need to be characterized for workload migration. To prevent network congestion, the migration has to consider network status. We present the proposed technique in the following subsections.

1) Downward power migration

Since the mapping of task is predefined for RTM, the computation power cannot be migrated, and only the communication power can. Consider the power profile of the two routing paths shown in Fig. 5. The power P is composed of five parts, which is described by (2). P_{src} is the sum of the power consumed from traffic source to source router, including the buffering power inside source router. P_Z is the power consumed on the path of z direction, and similarly P_X and P_Y are the power consumed on the x and y directions. P_{dst} is the power consumed from destination router to traffic sink. Routing can only change the distribution of vertical routing power P_Z and horizontal routing power P_X and P_Y. For minimal path routing such as changing from XYZ to ZXY, the power migrates, and the overhead of migration is negligible. As we shown in Section II.B.2, the power should be concentrated on bottom layer for maximizing heat conduction. Minimal path routing cannot migrate power to bottom layer if none of source and destination routers are on the bottom layer.

$$P = P_{src} + P_Z + P_X + P_Y + P_{dst} \quad (2)$$

2) Downward routing

Downward routing is a non-minimal path routing that changes the lying layer of horizontal routing path toward heat sink. In each pillar, the number of downward level is determined according to the network status. Given a 3D NoC with *N* layers, the maximum downward level is *N-1*. Downward routing can adopt arbitrary traditional 2D routing algorithm for horizontal routing on XY plane.

Fig. 6 shows a four-layer example of downward routing where denotes downward by K layers. The horizontal routing algorithm is XY-routing. When downward level is set to zero (DW_level=0), the routing behavior is identical to XYZ routing. Fig. 6(b) shows the routing paths when downward level is set to one (DW_level=1). Because the destination routers are exactly one level below the source router, the routing behavior is identical to ZXY in this example. Fig. 6(c) and Fig. 6(d) show that the routing paths when downward level is set to two (DW_level=2) and three (DW_level=3). The routing behavior shows the non-minimal path property, and it is in the order of vertical-horizontal-vertical. If the destination is at the layer exactly lower by the number of downward level, the routing is reduced to vertical-horizontal e.g. ZXY routing. Fig. 6(c) and Fig. 6(d) also show the example that the number of assigned downward level is larger than the level distance between source and bottom. The horizontal routing path is lying on the bottom layer.

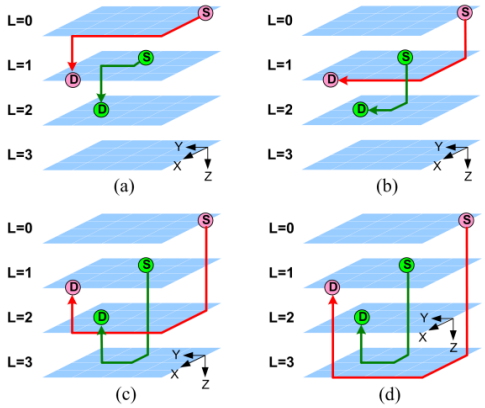

Figure 6. Examples of downward routing in a 4-layer 3D NoC : (a) without downward (downward level = 0); (b) downward level = 1; (c) downward level = 2; (d) downward level = 3.

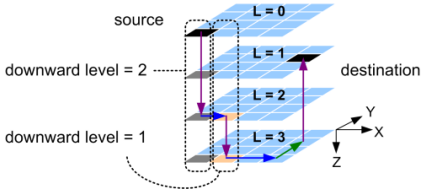

Figure 7. An example of traffic-aware downward routing

TABLE I. AGGREGATION OF TRAFFIC LOAD USING FIXED DOWNWARD LEVELS IN A 4-LAYER 3D NOC

DW_level	0	1	2	3
Layer 0 load	TL_0	0	0	0
Layer 1 load	TL_1	TL_0	0	0
Layer 2 load	TL_2	TL_1	TL_0	0
Layer 3 load	TL_3	TL_2+TL_3	TL_1+TL_2+TL_3	TL_0+TL_1+TL_2+TL_3

TL_k denotes original horizontal traffic load in layer k

Non-minimal path routing naturally increases the zero load latency and also has power overhead. However, due to the relatively short distance among layers, the latency and driving power of a vertical transfer is small. As [13] and [14], the crossbar switch-based 3D architecture is preferred due to its superior performance over symmetrical 3D structure, and can be implemented with cost-effective dimensionally-decomposed (DimDe) router. The latency overhead of non-minimal path routing on z direction is constrained to one cycle, no matter where the source and destination routers are.

3) *Traffic-aware level selection*

Downward routing provides a downward power migration mechanism by transporting horizontal traffic from upper layer to bottom layer. Fig.7 shows the traffic-aware downward routing. The downward level for each pillar is different. If the traffic load is small, more traffic can be migrated by a selecting a larger downward level. However, if the aggregated traffic load on bottom layer is larger than the bandwidth, the bottom layer will be congested. TABLE I. shows the aggregated traffic load of each layer with different fixed downward level. As the number of downward level increase, more traffic load is aggregated on bottom layer (layer 3). Fig. 8 shows the pseudo code of the proposed traffic-aware level selection. The selection of downward level depends on the traffic load estimation and prediction of each layer. To prevent network congestion, the selected downward level should not make aggregated traffic load over the limitation. The actual implementation of load estimation relies on counters, which is updated individually inside each router, and the summation of counter is taken only once on each interval.

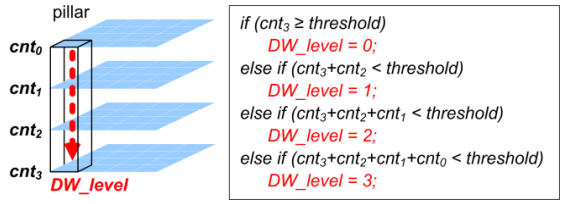

Figure 8. Traffic-aware level selection. All the routers inside the pillar adopt the same downward level *DW_level*. To prevent network congestion, the aggregated traffic load should be smaller than a given amount. cnt_i is the counter used for traffic load estimation and prediction.

B. *Thermal-Aware Vertical Throttling*

The goal of throttling is to effectively regulate the network temperature with minimal performance effects. The idea of throttling is creating a low power density region, and the heat generation rate of the region decreases. If a path for heat conduction exists, i.e. the neighbor's temperature is lower than the region, temperature decreases faster. Reference [7] shows the power impact of localized throttling. As throttling ratio increases, less traffic is allowed to pass through a router, and power consumed is reduced. Traffic throttling within a router also affects the traffic of the neighboring router that exchanges flits with the throttled router. This phenomenon helps cooling but affects performance. Throttling can be implemented with many different approaches, such as clock gating, dynamic voltage-frequency scaling, and can be combined with data isolation techniques. For simplicity and maximal cooling speed, global throttling (GT) is a commonly adopted approach. However, if there are only few routers overheated, GT inevitably throttles the remaining routers those are not overheated, and the system performance drops drastically. On the contrary, distributed traffic throttling (DTT) proposed in [7] only throttles the input traffic of the overheated router to the reduce workload. DTT has much less performance impact in comparison with GT and is suitable for 2D NoC.

1) *Vertical throttling*

Although the collaborative scheme adopted by DTT also works for 3D NoC, it is slow to form an effective heat conduction path. The throttling region gradually grows outward from the overheat router in DTT. Vertical throttling (VT) is a reactive mechanism that throttles routers in the direction of maximum temperature decreasing. For 3D NoC, VT throttles vertically aligned routers in the pillar. For an N-layer 3D NoC, VT throttles the routers on upper N-1 layers simultaneously and leaves the router in bottom layer. The routers in bottom layer are never throttled because they always have large heat conduction to the heat sink. Besides, downward routing aggregates traffic load on bottom layer. If the router in bottom layer is throttled, the performance impact will be large. For emergency cooling on overheat, VT has higher cooling speed than DTT and less performance impact than GT. However, if the temperature of the overheat router is not very high and the heat generation of the router is relatively slow, the cooling speed of DTT is enough and DTT has less performance impact. VT can be viewed as a specialized variation of collaborated DTT. The performance impact is larger due to the number of simultaneous throttled router is larger. To reduce the impact, we propose the thermal-aware vertical throttling (TAVT).

2) *Thermal-aware selection of throttling level*

Fig. 9 shows an example of TAVT in a 4-layer 3D NoC. There are two design parameters for TAVT: ratio set and level set. The ratio set determines how much a router is throttled. The level set determines all the combinations of the throttling ratio of vertically aligned routers. The number of throttling ratio can be determined as arbitrarily number. With consideration of the implementation cost of throttling, we choose a two throttling ratio for each router: full off (0%) and half off (50%). The number of throttling level does not required to be equal to the

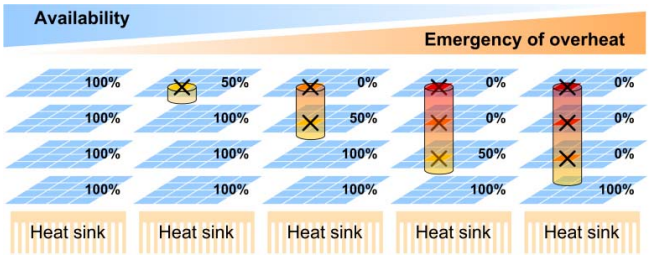

Figure 9. An example of thermal-aware vertical throttling with different throttling levels and different throttling ratios on each layer: (a) no throttling (normal mode); (b) throttling level =0 ; (c) throttling level =1; (d) throttling level =2; (e) throttling level =3

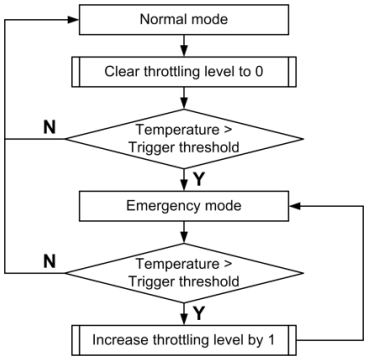

Figure 10. Thermal-aware selection of throttling level

number of downward level or the fully combination of the ratio set of all routers in a pillar (e.g. 3^4=81 combinations for a 4-level 3D NoC with 3 ratios in each router: full throttled, half throttled, and no throttled in normal mode). For simplification of control, we set four levels of throttling in emergency as shown in Fig. 9. When the temperature is lower than the trigger threshold $T_{trigger}$, the 3D NoC runs normal mode without throttling, which is shown as Fig. 9(a). If some router is detected overheated i.e. temperature over trigger threshold, it enters emergency mode. Since top layer is usually hotter than the layer below it, TAVT throttles the top router first. As Fig. 9(b) shows, the input bandwidth is limited to half of the full bandwidth. If the temperature decreases slow or does not decrease, the overheat router needs higher cooling speed. , The degree of emergency can be detected by continuously comparing the temperature with the trigger threshold. If the router keeps in emergency mode, it needs higher throttling level. The vertical aligned routers gradually throttle from top toward bottom to provide a faster heat conduction channel. The flow chart of selection of throttling level is shown in Fig. 10.

3) *Reactive routing on throttling*

If a router is throttled, the latency of packets routed through the throttled router increases drastically. If the NoC adopts deterministic routing, the packets will be blocked by the throttled router until the temperature is lower than trigger threshold. Reactive routing is an adaptive routing algorithm that prevents packets route through throttled routers. The bypassing path is not limited to the path through the neighboring routers of the overheated router. We use the never throttled bottom router and downward routing for reactive routing.

C. *Proof of Deadlock Freedom for the Proposed RTM*

Any deadlock from traffic migration will cause system stall. Therefore the routing has to be deadlock-free. Deadlock occurs when the four following necessary conditions are all true: mutual exclusion, partial allocation, no preemption, and circular waiting. Using virtual channel to allow preemption or adopting turn-model to prevent circular waiting are both effective approaches. However, the cost of virtual channel is high for NoC, therefore many turn-model based routing algorithms are proposed. Besides, virtual channel requires more buffer, which makes extra power consumed and heat generated. Reference [13] derives a theorem to guarantees a routing algorithm to be deadlock-free in 3D NoC. The theorem and proof is shown in Fig. 11. We follow the restrictions of this theorem to prove the deadlock-freedom of the proposed routing algorithm in both proactive and reactive states

Theorem *The routing designed with the following procedures guarantees deadlock-freedom.*

Proof *The routing guarantees deadlock-freedom, because no dependency occurs as the following restrictions all true:*

1. *No cyclic dependency is formed in each tier, because a packet must follow the restrictions for deadlock-freedom, as long as the packet is transferred on the single tier.*
   Downward routing adopts deadlock-free routing (e.g. XY routing in our experiment) for routing in a layer.
2. *No cyclic dependency is formed across tiers, because a packet is passed between tiers only in the descending order.*
   Downward routing never routes a packet above the source router. The horizontal route is always below or within the same layer of the source router.
3. *No cyclic dependency is formed within a pillar, because a pillar router is a crossbar switch.*
   The vertical transfer is through the crossbar switch in downward routing. The downward transfer can be arbitrary level in each pillar, which is controlled by the level selection mechanism. The upward transfer only occurs at the bottom of destination router and is directly to the destination.

Figure 11. Theorem and proof for deadlock-free routing in 3D NoC [13]

IV. EXPERIMENT AND DISCUSSION

A. *Simulation environment*

The simulation environment for RTM of 3D NoC couples the network model, power model and thermal model. We integrate Noxim [11] and Hotspot [10] as our simulator, and validate with CFD-RC [12] for the accuracy of the vertical temperature distribution. We adopt the tile geometry and power model of Intel's 80-core processor [8]. We add the model of basic 3D router and the DimDe router, and we extend NoC simulator to generate a 4x4x4 3D architectures of NoC. During network traffic simulation, the power trace is generated based on the power model. The power trace and the physical floorplan are used as inputs of the thermal simulation. For each router, the depth of the buffering channel is 4 flits and the no virtual channel is used.

B. *Evaluation of proposed traffic-aware downward routing*

The first experiment shows the effectiveness of temperature control and performance optimization of the proposed proactive traffic-aware downward routing. First we show the steady state maximum temperature of each layer in Fig. 12. The two extreme cases are Fig. 12(a) and Fig. 12(d). With uniform traffic offered, XYZ is the optimal routing algorithm that achieves maximum throughput and distributes traffic evenly. This is the case of optimizing performance without thermal consideration. Although the network is not saturated for all packet injection rate (PIR) between 0.001 and 0.029, the maximum temperature of the 3D NoC is over 140 ℃, which is above

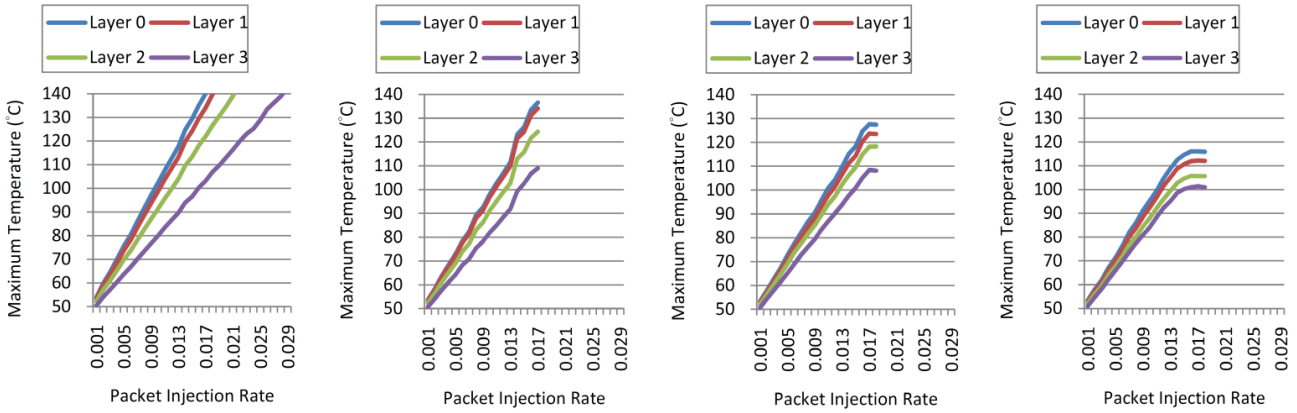

Figure 12. Maximum temperature of each layer with different fixed downward level, 1packet = 6 flits: (a) fixed downward level = 0 (XYZ routing); (b) fixed downward level = 1; (c) fixed downward level = 2; (d) fixed downward level = 3

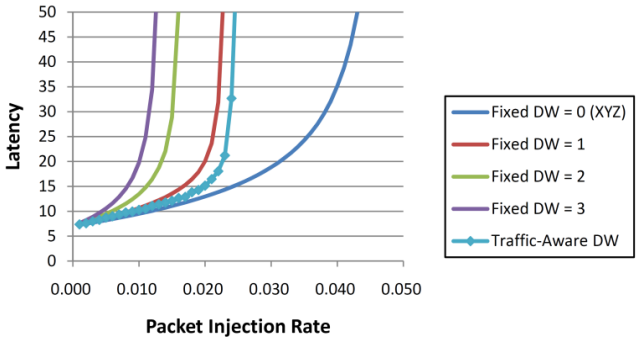

Figure 13. The latency verses injection rate with uniform traffic offered, 1 packet = 6 flits, (without temperature limitation)

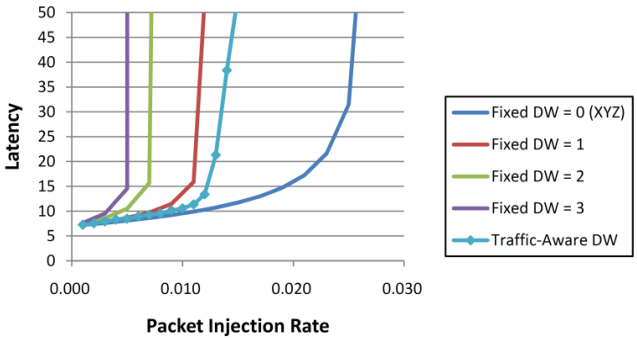

Figure 14. The latency verses injection rate with transpose traffic offered, 1 packet = 6 flits, (without temperature limitation)

TABLE II. ACHIEVABLE THROUGHPUT WITH UNIFORM TRAFFIC

Thermal limit	60℃	80℃	100℃	120℃	Infinite
XYZ (DW_level = 0)	0.0020	0.0059	0.0096	0.0133	0.0230
Fixed DW_level = 1	0.0023	0.0064	0.0104	0.0137	0.0160
Fixed DW_level = 2	0.0024	0.0065	0.0108	0.0112	0.0112
Fixed DW_level = 3	0.0025	0.0067	0.0083	0.0083	0.0083
Traffic-aware	0.0023	0.0063	0.0101	0.0140	0.0195
Improvement	15.00%	6.78%	5.21%	5.26%	-15.22%

TABLE III. ACHIEVABLE THROUGHPUT WITH TRANSPOSE TRAFFIC

Thermal limit	60℃	80℃	100℃	120℃	Infinite
XYZ (DW_level = 0)	0.0021	0.0058	0.0094	0.0131	0.0186
Fixed DW_level =1	0.0024	0.0064	0.0103	0.0104	0.0104
Fixed DW_level =2	0.0024	0.0066	0.0067	0.0067	0.0067
Fixed DW_level =3	0.0025	0.0050	0.0050	0.0050	0.0050
Traffic-aware	0.0023	0.0062	0.0096	0.0121	0.0121
Improvement	9.52%	6.90%	2.13%	-7.63%	-34.95%

normal thermal limit. Therefore Fig. 12(a) is a thermal-limited case for performance optimization. In our experiments, all the fixed downward level cases except level = 0 are network limited. Therefore the maximum temperatures of all fixed downward cases are just plotted slightly over the PIR that reaches saturation throughput. Fig. 12(d) is the case that maximizing heat conduction by concentrating all horizontal routing power on the bottom layer. As we expect, it has the lowest steady state temperature. Fig. 12 also shows that maximum temperature changes as workload migration, and for uniform traffic, the difference of temperature between layers is smaller when more traffic is migrated downward. Fig. 13 and Fig. 14 show the latency versus PIR of the proposed traffic-aware downward routing without thermal consideration. Because the behavior of traffic-aware downward routing is a combination of all the others, the latency is among the two extreme cases. However, with thermal limit, the achievable throughput is not necessarily the saturation throughput defined by the PIR that corresponding to the twice zero-load latency.

We use TABLE II. and TABLE III. to show the thermal-limited achievable throughput of the downward routing algorithms. The rightmost column shows the saturation throughput calculated from the PIR where the latency is double of zero-load latency, which can be viewed as the thermal limit is at positive infinite. As thermal limit becomes lower, the network is more prone to suffer from the problem of temperature-limiting performance. Proactive workload migration reduces the maximum temperature, and therefore eases the problem of temperature-limiting. However, too much workload migration makes the packets on bottom layer congested. The proposed traffic-aware level selection adjusts the downward level according to the congestion degree, so it prevents saturation on bottom layer. TABLE II. and TABLE III. show that the proposed proactive technique improves the achievable throughput for most temperature-limited cases in uniform and transpose traffic. When thermal limit is 80℃, the improvement is around 7%.

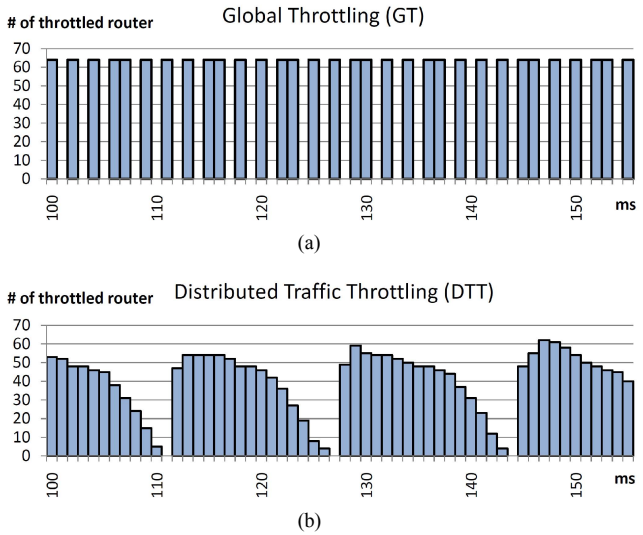

Figure 15. Number of the throuttled router over time for:
(a) global throttling (GT); (b) distributed traffic throttling (DTT)

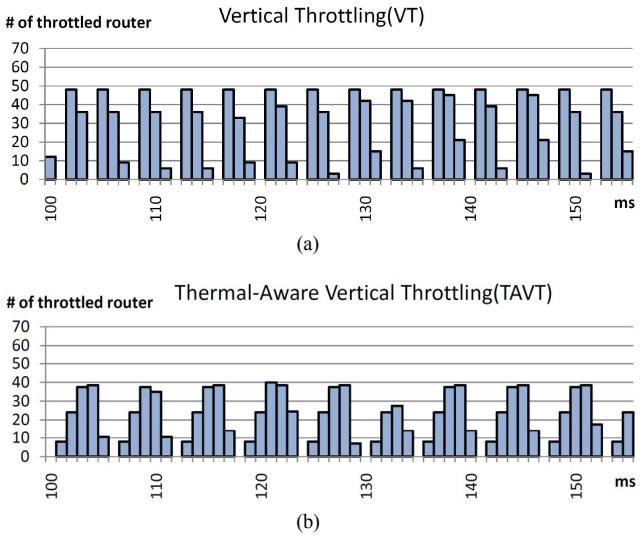

Figure 16. Number of the throttoled router over time for the proposed:
(a) vertical throttling (VT); (d) thermal-aware vertical throttling (TAVT).

C. Evaluation of proposed thermal-aware vertical throttling

The second experiment shows the availability impact of throttling. GT, DTT, VT, and TAVT are all implemented and simulated individually with the same PIR. The throttling ratio will keep the setting of the selected throttling level. In this experiment, GT, DTT, and VT have only one level of throttling, and TAVT has four as Fig. 9. The throttling ratio of GT and DTT is 1.0, and the throttling ratio of VT is combined by three 1.0 and one 0.5. Once the temperature of a router is above the trigger threshold, GT throttles all 64 routers. DTT only throttles the overheated router, and VT throttles the corresponding pillar of 4 routers. TAVT throttles according the selected level.

Fig. 15 and Fig. 16 show the effective number of throttled router in simulation. The number is calculated by accumulating the throttling ratio of all 64 routers on each updating of throttling mechanism. In this experiment, the interval of updating is 10^{-3} second. From Fig. 15 (a) and Fig. 15 (b), the two extremes of throttling are shown. GT has highest cooling speed for 3D NoC, therefore it usually throttle only 1 ms. DTT throttles only the overheated routers. Therefore overheat region grows if the heat conduction is not enough. The throttling region reforms horizontally first and then vertically toward heat sink When the cooling speed increases, the number of required throttled router may decrease. Fig. 16 shows the proposed VT and TAVT. Since VT directly throttles the vertical aligned routers to create a most effective heat conduction path, the number of throttled router decreases much faster than DTT. The throttling behavior of TAVT is a combination of DTT and VT. TAVT adaptive changes the level, and trades off between cooling speed and performance impact.

To analyze the cooling speed of different throttling approaches, we record the throttling time of each throttled router. TABLE IV. shows the statistics of the time. Because GT creates the largest low power density region over the NoC, the cooling of GT is the fastest. DTT has longer throttling time due to the reforming region is not guided. Since the temperature of the router beneath the overheated router is usually lower, the region tends to grows horizontally, not vertically. The region reforms toward bottom layer unless the accumulated heat makes the router in the beneath layer overheated. However, the router close to heat sink has higher cooling efficiency. Therefore DTT takes longer time to form a effective cooling region.

TABLE IV. STATISTICS OF THE THROTTLING TIME (UNIT: 1MS)

	GT	DTT	VT	TAVT	Reduction (VT vs. DTT)	Reduction (TAVT vs. DTT)
Mean	1.147	14.333	2.481	4.353	82.69%	69.63%
Var.	0.125	2.222	0.324	0.581	85.42%	73.85%

TABLE V. AVERAGE NETWORK THROTTLING RATIO

Injection Rate (flits/node/cycle)	GT	DTT	VT	TAVT	Reduction (TAVT vs. DTT)
0.016	0.0160	0.0085	0.0105	0.0077	9.12%
0.022	0.0256	0.0126	0.0155	0.0108	14.68%
0.030	0.0320	0.0169	0.0175	0.0143	15.28%

TABLE VI. AVERAGE NETWORK AVAILABILITY

Injection Rate (flits/node/cycle)	GT	DTT	VT	TAVT	Improvement (TAVT vs. DTT)
0.016	0.9840	0.9915	0.9895	0.9923	1.0008×
0.022	0.9744	0.9874	0.9846	0.9893	1.0019×
0.030	0.9680	0.9832	0.9825	0.9857	1.0026×

When the router in the lower layer is cooled and not overheated, its temperature may be still close to the thermal limit. The conceptual difference between DTT and VT/TAVT is similar to the conceptual difference between breadth-first search and depth-first search. The variation and standard deviation also shows that overheat region reforms but not directly toward heat sink. VT has much less average throttling time than DTT. In this experiment, the reduction is 82.7%. Because TAVT is adaptive and starts from the lower level, the average throttling time is longer than VT. From TABLE IV. , the reduction of TAVT in comparison with DTT is 69.6%. Both VT and TAVT have smaller variation on throttling time.

Finally we show and discuss performance impact from the view of the average network throttling ratio and the average network availability. TABLE V. shows the average network throttling ratio and TABLE VI. shows the average network availability. Although the throttling time of GT is shortest, too many routers are throttled; therefore the average throttling ratio of GT is the highest, and the

availability is the lowest. In comparison with VT, DTT has higher network availability. TAVT has the highest availability among all four approaches. It should be noted that the average network throttling ratio is a joint view of the throttling time and the number of throttled routers, and the average network availability is equal to one subtracting the average throttling ratio. From the analysis result of 3D NoC, faster cooling speed means more routers are throttled. Fewer router throttled means longer throttling time. The trade off is dependent on application, and should consider the data dependency among packets.

V. RELATED WORK

For 2D NoC, [7] provides a framework for RTM: ThermalHerd. ThermalHerd is a distributed scheme where routers collaboratively regulate the network temperature and work towards averting thermal emergencies while minimizing performance impact. The network simulation environment involves NoC model, power model, thermal model, and timing driven simulators. Thermal profile of the MIT Raw chip is characterized for constructing the thermal model of NoC. With traffic monitoring and prediction, routers change their behavior based on the temperature monitoring. To minimize the performance loss in emergency, distributed traffic throttling (DTT) is adopted, which is a reactive technique for thermal management. Over the temperature trigger threshold, DTT decreases the available bandwidth of the router at the hotspot to reduce the switching activity. The power consumption of the throttled region is reduced, thus cooling the hotspot. To reduce the performance penalty of thermal management, [7] provides a proactive routing protocol and a reactive routing protocol. The proactive protocol adjusts traffic to balance the network temperature profile when the maximum temperature is below the emergency limit (normal mode). The reactive protocol replaces the proactive protocol in emergency mode and steers packets away from throttled region to minimize the performance penalty due to throttling.

VI. CONCLUSION

To ensure thermal safety and gracefully degradation for 3D NoC, we define the problem as performance optimization with temperature consideration, and propose a traffic-aware and thermal-aware run-time thermal management scheme. The proposed techniques utilize the larger thermal coupling among the vertically stacked routers of the 3D NoC. For optimizing performance with the varying temperature, a proactive technique is used. The traffic-aware downward routing migrates horizontal routing power toward bottom layer and adjusts the amount of traffic to prevent packet congestion. The experimental results show the proactive approach is effective. The proposed proactive technique improves the achievable throughput around 7% for a 3D NoC with 80℃ thermal limit. For overheat routers, the proposed vertical throttling creates a heat conduction path to heat sink, and has higher cooling speed than the traditional distributed traffic throttling. Besides, we propose the thermal-aware vertical throttling to throttle overheat routers with smaller performance impact. In comparison with distributed traffic throttling, the experimental result shows the average throttling time of the proposed vertical throttling is reduced by 82.7%, and the throttling time of the proposed thermal-aware throttling is reduced by 69.6%. The average network throttling ratio reduced by the proposed thermal-aware vertical throttling is 9.1% to 15.3%. The improvement of average network availability over distributed traffic throttling is from 1.0008× to 1.0026×.

REFERENCE

[1] L. Benini and G. De Micheli, "Networks on chip: a new paradigm for systems on chip design," *Proc. Proc. IEEE Conference on Design, Automation, and Test in Europe (DATE'02)*, Mar. 2002, pp. 418–419.

[2] A. W. Topol *et al.*, "Three-dimensional integrated circuits," *IBM J. Research Development*, pp. 491–506, Jul. 2006.

[3] B. Black *et al.*, "Die Stacking (3D) Microarchitecture," in *Proc. IEEE/ACM International Symposium on Microarchitecture (Micro'06)*, Dec. 2006.

[4] V. Pavlidis and E. Friedman, "3-D Topologies for Networks-on-Chip," *IEEE Trans. Very Large Scale Integration Systems*, vol. 15, no. 10, pp. 1081-1090, Oct. 2007.

[5] D. Park *et al.*, "MIRA: A Multi-Layered On-Chip Interconnect Router Architecture," in *Proc. Intl. Symp. Computer Architecture (ISCA'08)*, pp.251-261, June 2008.

[6] C. Zhu, Z. Gu, L. Shang, R. P. Dick, and R. Joseph, "Three-Dimensional Chip-Multiprocessor Run-Time Thermal Management," *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, vol. 27, no. 8, pp. 1479-1492, Aug. 2008.

[7] L. Shang, L. Peh, A. Kumar, and N. K. Jha, "Thermal modeling, characterization and management of on-chip networks," in *Proc. IEEE/ACM International Symposium on Microarchitecture (Micro'04)*, Dec. 2004, pp. 67–78.

[8] Y. Hoskote, S. Vangal, A. Singh, N. Borkar, and S. Borkar, " A 5-GHz Mesh Interconnect for A Teraflops Processor", *IEEE MICRO*, vol. 27, pp. 51-61, 2007.

[9] K. Puttaswamy and G. H. Loh, "Thermal herding: microarchitecture techniques for controlling hotspots in high-performance 3D-integrated processors," in *Proc. IEEE High Performance Computer Architecture (HPCA),* Feb. 2007, pp. 193–204.

[10] W. Huang *et al.*, "HotSpot : A compact thermal modeling methodology for early-stage VLSI design," *IEEE Trans. Very Large Scale Integration Systems*, vol. 14, no. 5, pp.501-513, May 2006.

[11] Noxim: network-on-chip simulator [Online]. Available: http://sourceforge.net/projects/noxim/

[12] CFD-RC [Online]. Available: http://www.cfdrc.com/

[13] H. Matsutani, M. Koibuchi, and H. Amano, "Tightly-Coupled Multi-Layer Topologies for 3D NoCs," in *International Conference on Parallel Processing (ICPP)*, pp. 75-85, 2007

[14] J. Kim *et al.*, "A Novel Dimensionally-Decomposed Router for On-Chip Communication in 3D Architectures," in *Proc. Intl. Symp. Computer Architecture (ISCA'07)*, June 2007, pp.138-149.

[15] 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)*, Apr. 2010.