Immediate Neighborhood Temperature Adaptive Routing for Dynamically Throttled 3-D Networks-on-Chip

Sumeet S. Kumar, *Student Member, IEEE*, Amir Zjajo, *Member, IEEE*, and Rene van Leuken, *Member, IEEE*

***Abstract*—In this brief, we present the immediate neighborhood temperature (INT) routing algorithm, which balances thermal profiles across dynamically throttled 3-D networks-on-chip by adaptively routing interconnect traffic based on runtime temperature monitoring. INT avoids the overheads of system-wide temperature monitoring by relying on the heat transfer characteristics of 3-D integrated circuits that enable temperature information from routers in the immediate neighborhood to guide adaptive routing decisions. Experimental results indicate that INT yields balanced thermal profiles with up to 25% lower gradients than competing schemes and shortens communication latency by decreasing average network congestion by up to 50%, with negligible overheads.**

***Index Terms*—Adaptive routing, power management, thermal management, 3-D networks-on-chip (NoCs).**

I. Introduction

In large-scale multiprocessor systems-on-chip (MPSoCs), the utilization of processing elements (PEs) varies based on the nature of the workload under execution. Heavily utilized PEs dissipate a larger amount of power that in thermally constrained systems results in the formation of thermal hotspots. In 3-D stacked multiprocessors, due to the thermal coupling between stacked dies [1], high activity in one tier can result in the formation of thermal hotspots in the surrounding tiers. Sustained thermal gradients and high operating temperatures are detrimental to reliability and result in the accelerated degradation of devices [2]. Dynamic thermal managers (DTMs) control operating temperatures by reducing the switching activity and, thus, the power dissipation of components [3]. However, their invocation also results in decreased system performance [4].

The network-on-chip (NoC) interconnect in modern MPSoCs consumes a significant amount of power and can aggravate thermal imbalances in the system. For instance, interconnect traffic routed close to highly active PEs increases power density in the region, resulting in elevated operating temperatures, i.e., a thermal hotspot. If traffic were instead steered away from such critical regions by a temperature-aware routing strategy, the amount of interconnect power dissipated near high-activity nodes would reduce. This would yield more balanced operating temperatures, prevent invocation of the DTM, and minimize performance losses.

In this brief, we present the immediate neighborhood temperature (INT) adaptive routing algorithm, which balances operating temperatures across 3-D NoCs by incrementally routing packets along low-temperature minimal paths. INT eliminates the need for system-wide temperature awareness; instead, it utilizes only local temperature information from adjacent routers to drive output port selection for in-flight packets. INT outperforms state-of-the-art proposals [5], [6], yielding shorter communication latency, lower congestion, and balanced temperature profiles. Our work demonstrates the effectiveness of localized temperature information in driving adaptive routing in 3-D NoCs.

II. Background

The efficiency with which heat can be evacuated from a 3-D integrated circuit (IC) is effectively a function of the system's physical characteristics and the thermal efficiency of its package. For a tiled multiprocessor, this is shown with an equivalent of *Nagata's* equation [7], i.e.,

$$\frac{\alpha_t(N_tN_g)E_t}{t_{pd}} \leq g \cdot \Delta T_{\max} \quad \text{where } g = \kappa_{\text{eff}}\frac{A}{l_{x,y,z}} \tag{1}$$

where N_t represents the number of PE tiles, each with N_g gates, energy dissipation E_t, average activity rate α_t, and clock period t_{pd}. $\Delta T_{\max}$ represents the maximum temperature difference between the components on-chip and the ambience through heat transfer surfaces of area A, situated at a distance $l_{x,y,z}$ from the power dissipation site. κ_{eff} represents the effective thermal conductivity of the die stack and its through-silicon vias, whereas g is the effective thermal conductance between power-dissipating elements and the heatsink surface. Within MPSoC tiles, α_t can be broken down into its two constituent parts: activity due to functional operations (such as processing and memory loads/stores) and activity due to communication over the interconnect. The average switching activity per tile (α_t) is thus given as

$$\alpha_t = \alpha_p\frac{E_p}{E_t} + \alpha_r\frac{E_r}{E_t} \tag{2}$$

where E_p and E_r represent the average energy dissipation of the functional components within the tile and the interconnect router, respectively, with corresponding activity rates α_p and α_r. Adaptive routing strategies that aim to balance on-chip

Manuscript received April 30, 2015; revised July 25, 2015; accepted November 22, 2015. Date of publication November 24, 2015; date of current version June 23, 2017. This work was supported in part by the CATRENE program under the Computing Fabric for High Performance Applications (COBRA) project CA104. This brief was recommended by Associate Editor S. Hu.

The authors are with the Circuits and Systems Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: s.s.kumar@tudelft.nl; a.zjajo@tudelft.nl; t.g.r.m.vanleuken@tudelft.nl).

Color versions of one or more of the figures in this brief are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSII.2015.2503613

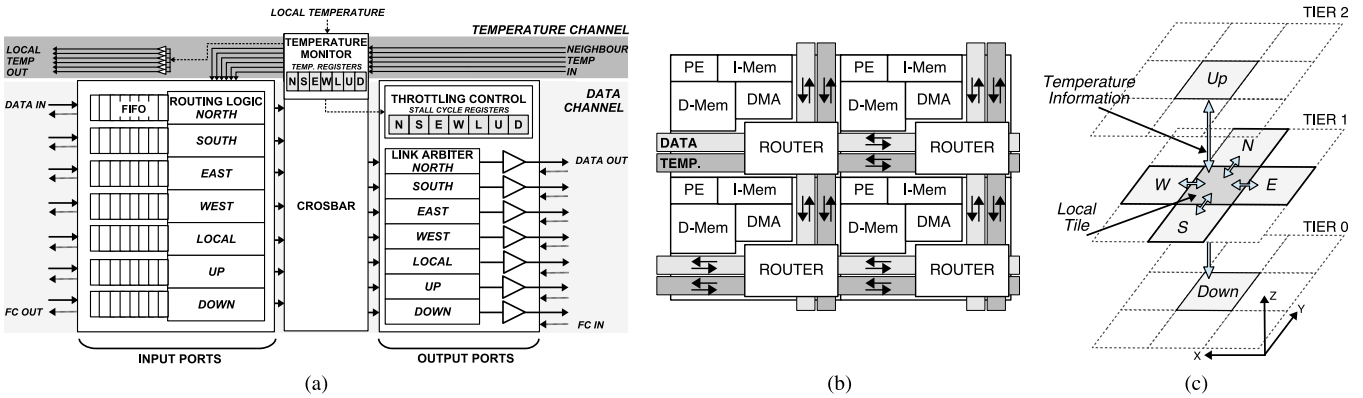

Fig. 1. Illustration of (a) INT router architecture. (b) MPSoC array with INT-based NoC with separate temperature and data channels. (c) Exchange of temperature information between immediate neighbor routers in a 3-D stack.

temperatures focus on controlling the ratio of α_r and other variables in (1) in a number of ways. Following Skadron's early work on temperature-aware microarchitectures [3], a number of proposals have addressed the issue of temperature management in SoCs. ThermalHerd [4], for instance, controls α_r in planar NoCs by routing interconnect traffic along paths with the least thermal correlation with a hotspot node. However, this approach relies on comprehensive design time analysis in terms of what the authors refer to as *heat spreading angle*. Given the complex nature of heat flow within die stacks, extending this notion of the heat spreading angle to 3-D NoCs is nontrivial in nature. Furthermore, ThermalHerd necessitates the storage of correlation factors for use at runtime. This is a significant overhead that is dependent on system size, physical dimensions, and magnitude of thermal coupling between nodes.

Traffic and thermal adaptive routing (TTAR) [5] implements thermal-aware dynamic throttling, which reduces the activity of routers (α_r) based on the available temperature margin $(\Delta T_{\max} - \Delta T)$. However, as temperature increases, routers are increasingly throttled, reducing not only power dissipation but also throughput. In highly active regions of the interconnect, the decreased throughput results in congestion. TTAR does not explicitly monitor temperatures. Instead, it uses network congestion as an indicator of high temperature. This is a significant drawback of the scheme since congestion can also be a consequence of actual interconnect traffic. Thus, a path of low congestion does not necessarily imply a path of low temperature. Furthermore, congestion information is aggregated and propagated over a dedicated regional congestion awareness (RCA) monitoring network [8], which poses an overhead in terms of both complexity and power. Proposals, such as *downward routing* [6], shift α_r toward tiers closer to the heatsink to decrease peak temperatures. However, the scheme aggravates congestion in higher tiers, thus increasing communication latency. Finally, other schemes [9] that rely on less complex runtime monitoring mechanisms often have a limited effect as they revert to conventional deterministic routing once thermal hotspots have formed in the system.

III. INT Adaptive Routing

Temperature, unlike congestion information, does not need to be propagated across the network to support thermal-aware routing. The physical nature of heat transfer results in thermal hotspots influencing the temperature of tiles in their immediate vicinity. This observation is even more significant in 3-D ICs where the magnitude of thermal conduction between the thin stacked dies results in hotspots spread across multiple tiers [1]. Consequently, the temperature of candidate links in the direction of the thermal hotspot appears higher than others in the direction of cooler regions. This effectively removes the need for an aggregate–propagate-type monitoring network and enables temperature-aware adaptive routing to be implemented based on information available from routers in the immediate vicinity alone. In this brief, interconnect activity α_r is controlled in direct response to local temperature, using a thermal-aware throttling mechanism as well as an adaptive routing strategy that steers interconnect traffic away from regions of high temperature.

A. Temperature Monitoring

Temperature monitoring is performed locally through the use of one or more thermal sensors integrated within each tile. Sensors integrate an *analog-to-digital converter* to provide digital temperature measurements in every sampling interval. The digital temperature value is stored within the router's *temperature monitor*, shown in Fig. 1(a), and is used to control thermal-aware dynamic throttling in the local router as well as to provide a basis for adaptive routing at neighboring routers.

B. Temperature Channel Considerations

Unlike in RCA whereby congestion information must be aggregated and propagated across the network, INT relies only on temperature information available in the immediate neighborhood. This effectively reduces the amount of logic within the monitoring network. INT's temperature monitor consists of only a few registers to store the latest temperature measurements from neighboring routers as well as the local temperature. The local temperature is placed on the outgoing temperature channels to the router's immediate neighbors, as shown in Fig. 1(b) and (c). The area and power overheads imposed by this channel consist mainly of the link drivers. The magnitude of these overheads is determined by the width of the temperature vector, which depends on the resolution

of the temperature sensor and the maximum operating range. However, given the nonaggregating nature of this monitoring network, transitions on the temperature channel occur only when new temperature measurements are available from the sensor (typically every 50–100 μs [4]). The power overheads of the temperature channel are thus minimal as compared with the data channel and RCA-type networks.

C. Thermal-Aware Dynamic Throttling

Throttling influences the rate at which traffic flows through an interconnect router, thereby influencing activity (α_r) and power dissipation. The crossbar switch and output link drivers account for over 80% of the dynamic power dissipation of the complete router. Our throttling approach therefore focuses on controlling activity within these components and is implemented within the *link arbiters* of output ports, as shown in Fig. 1(a). When throttling is applied, a varying number of stall cycles are introduced between the polling of successive ports by the round-robin arbiter.

The amount of throttling invoked is dependent on temperature, and the range of available levels can be controlled by using a set of registers within each router. In addition to controlling α_r, throttling also affects the performance of PEs in dataflow architectures in which the execution of tasks is dependent on the presence of the necessary triggering data within the direct memory access (DMA) controller's message-passing buffers [10]. Since throttling decreases the throughput of routers, triggering of tasks is delayed, and PE activity (α_p) is thus decreased.

D. Temperature-Aware Adaptive Routing Algorithm

Temperature-aware path selection in INT consists of two steps. In the first step, an initial routing of the packet is performed by using the odd-even (OE) algorithm [11], identifying the candidate output ports that could be used to reach the destination. The algorithm returns a set of candidate output port options that correspond to the available minimal paths to the destination router. The generated set encapsulates the ports through which the waiting packet can be ejected, while respecting OE's turn restrictions that guarantee deadlock freedom within the network. INT's second step consists in identifying the candidate port with the least temperature from this set. This port is referred to as the preferred candidate, and it is used to eject the packet toward its destination. INT's output port selection is therefore thermal adaptive, steering interconnect traffic in response to temperature, within the set of available minimal paths. The INT routing algorithm is listed in Fig. 2. For a given quantum of traffic, INT influences the following terms of (1).

- α_r/A: INT spreads the utilization of interconnect routers over a larger area by adaptively routing packets over multiple paths based on the available temperature margin $(\Delta T_{\max} - \Delta T)$.
- α_r/α_t: INT adaptively regulates interconnect traffic in regions of high activity (α_t), which typically exhibit higher operating temperatures.
- α_r/l_z: When temperatures permit, INT preferentially routes packets intended for tiers closer to the heatsink in the Z-dimension first.

```
1: if Address_local == Address_dest then
2:     Selected Port ← Local
3: else
4:     Candidate Ports ← Perform initial OE routing
5:     Fetch Temperatures of Candidates
6:     Preferred Candidates ← Candidate port with lowest temperature
7:     if Preferred Candidates >1 then
8:         if Up is a Preferred Candidate then
9:             Selected Port ← Up
10:        else
11:            Selected Port ← First Preferred Candidate
12:    else
13:        Selected Port ← First Preferred Candidate
14: Route packet onto Selected Port
```

Fig. 2. INT routing algorithm.

TABLE I
SYSTEM CONFIGURATION

MULTIPROCESSOR		NETWORK-ON-CHIP	
System Size	$4 \times 4 \times 4$ tiles	Flit Width	38-bit
D-/I-Mem	64KB/16KB	FIFO Depth	16 flit
DMA Buffers	8KB	Packet Size	4/32/64B
PE	32-bit RISC	Port Count	7
PHYSICAL			
Tech Node/Freq.	90nm/500MHz	Heat Transfer Co-eff.	100 W/(m^{2}K)
Tile Size	2mm×1.4mm	Trigger Temp.	332K
Temp. Range	300-370K	Temp. Sensors	64 (1 per tile)
Sensor Accuracy	0.5K	Sampling Interval	50μs

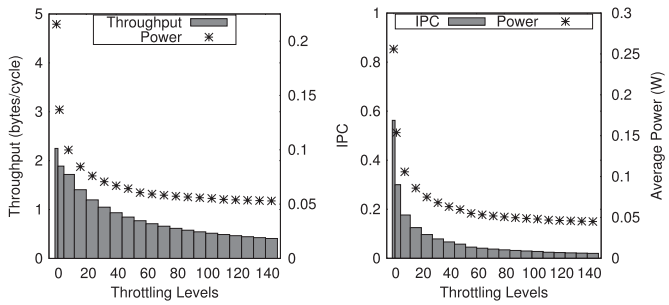

Fig. 3. Influence of throttling on router power and (a) throughput. (b) Effective *IPCs* of PE.

IV. EVALUATION

INT is evaluated using the *Ctherm* cycle-accurate thermal-functional cosimulation framework [12] with a *SystemC* model of the *NagaM* multiprocessor [10]. The evaluation consists of two parts: first, the characterization of the thermal-aware throttling mechanism and its effect on power and performance and, second, the evaluation of the INT routing algorithm under varying traffic conditions and number of thermal hotspots. The system configuration is listed in Table I. The physical model of the die stack used in this evaluation is based on data from a prototype 3-D chip, and its thermal behavior was previously modeled and characterized in our earlier work [1]. The test platform consists of 64 tiles arrayed over 4 stacked tiers (thus 16 tiles per tier). Tiles incorporate a processing element, memories, a DMA controller implementing the Pronto message-passing system [10], and a NoC router, as illustrated in Fig. 1(b).

A. Characterization of Throttling

The throttling mechanism is characterized using a single router with an emulated temperature input. Fig. 3(a) illustrates the relationship among average power dissipation, throughput, and throttling levels. Throttling is observed to steeply decrease

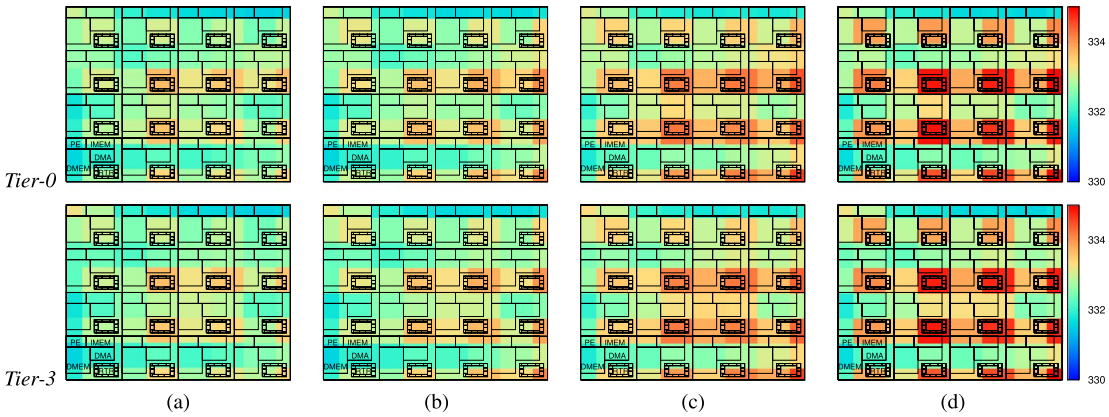

Fig. 4. Temperature maps from *uniform random* traffic with eight thermal hotspots using (a) INT (proposed). (b) Downward routing. (c) TTAR. (d) TTAR+. *Note*: Temperature sensors are considered to be located at the center of each tile.

power dissipation of the router and degrade throughput for each 0.5-K increment over the triggering temperature. To demonstrate the influence of α_r on α_p in dataflow and memory-bound systems, PEs in the characterization setup are configured to execute a fixed set of simple integer operations for every data word that arrives into the DMA's message-passing buffer. When the buffer is empty, execution is stalled. Fig. 3(b) reports the average instructions per cycle (IPCs) and power dissipation within a 50-K cycle window that corresponds to each router throttling level. This result indicates that in SoCs with significant intertile communication, the interconnect can be used to regulate α_p.

B. INT Evaluation

The evaluation of INT utilizes three synthetic traffic patterns: *hotspot* (10%), *uniform random*, and *bit transpose*. Packets are injected into the network by synthetic traffic generators that target destinations according to the probability density function of the traffic pattern. To test the ability of the various routing algorithms in balancing temperatures, the evaluation also uses a variable number of thermal hotspots placed at random locations within the system. These hotspots are obtained from the execution of a load–compute–store loop on tile PEs, resulting in a constant power dissipation that is interconnect independent. This behavior is characteristic of long-running compute-bound workloads that are unaffected by interconnect performance once their triggering data have arrived at the tile. Although the routing strategies are evaluated in the presence of four and eight such thermal hotspots inside the 3-D mesh, since the general performance trends of both are similar, this brief presents only the results from the evaluation with eight thermal hotspots due to space constraints. INT is compared with three other routing strategies: downward routing [6], *TTAR* [5], and TTAR *with temperature and congestion awareness* (TTAR+). The INT and TTAR cases use an 8-bit monitoring network for temperature and congestion, respectively. For TTAR+, temperature and congestion information are carried over independent 5-bit and 3-bit networks, respectively. To compensate for the loss in temperature resolution due to the decreased vector width, the monitoring temperature range was decreased to 330 K–345 K. The congestion resolution, although decreased, is still sufficient to discriminate between paths.

1) Thermal Performance: Fig. 4(a)–(d) illustrates temperature maps of tiers 0 and 3 of the die stack, with each routing algorithm following 750-K simulation cycles at the peak injection rate, with *uniform random* traffic and a total of eight thermal hotspots. INT is observed to provide the most balanced temperature profiles and the lowest peak temperatures among all the tested schemes. Downward routing provides a similar balance, however, with increased latency and congestion, as observed in Fig. 5(a). In the case of TTAR, the temperature imbalance occurs as a consequence of network traffic being routed through high-temperature regions with low congestion. The figures also illustrate the high degree of thermal coupling between stacked dies. The efficacy of the interconnect in balancing temperature differences in the system requires the presence of sufficient traffic. At low injection rates, the influence of network traffic on temperatures is relatively small. As a consequence, despite its temperature awareness, INT yields peak temperature differences identical to the competing schemes, as evidenced in Fig. 5(a). However, with increasing network traffic, the activity and power dissipation of the interconnect assumes increasing significance, and the effects of temperature-aware routing become evident.

2) Latency and Congestion: Fig. 5(a) also illustrates the average packet latency and network congestion resulting from the use of each routing strategy. In a network with temperature-dependent throttling, packet latency is influenced by operating temperatures and network congestion. Congestion-aware routing strategies such as TTAR offer significantly low latency on account of their avoiding high-traffic regions. However, this often results in the routing of packets close to interconnect-independent thermal hotspots. Consequently, operating temperatures rise, and as throttling is invoked, packet latency increases. Downward routing, on the other hand, prevents the formation of high-temperature regions by routing traffic toward higher tiers. This has the effect of increasing network congestion on the cooler tiers and drastically increasing packet delivery latency. INT's routing of packets based on operating temperatures results in a decreased chance of encountering a heavily throttled router. This ensures that packets remain on higher throughput paths, reducing the congestion caused due to slow-moving traffic in the network. The latency benefits obtained as a result of INT's temperature-adaptive routing are observed in Fig. 5(b), which reports the latency distribution across network nodes at the peak injection rate. In the case of TTAR+, the power overheads incurred due to the propagation of monitoring information across the system are so considerable

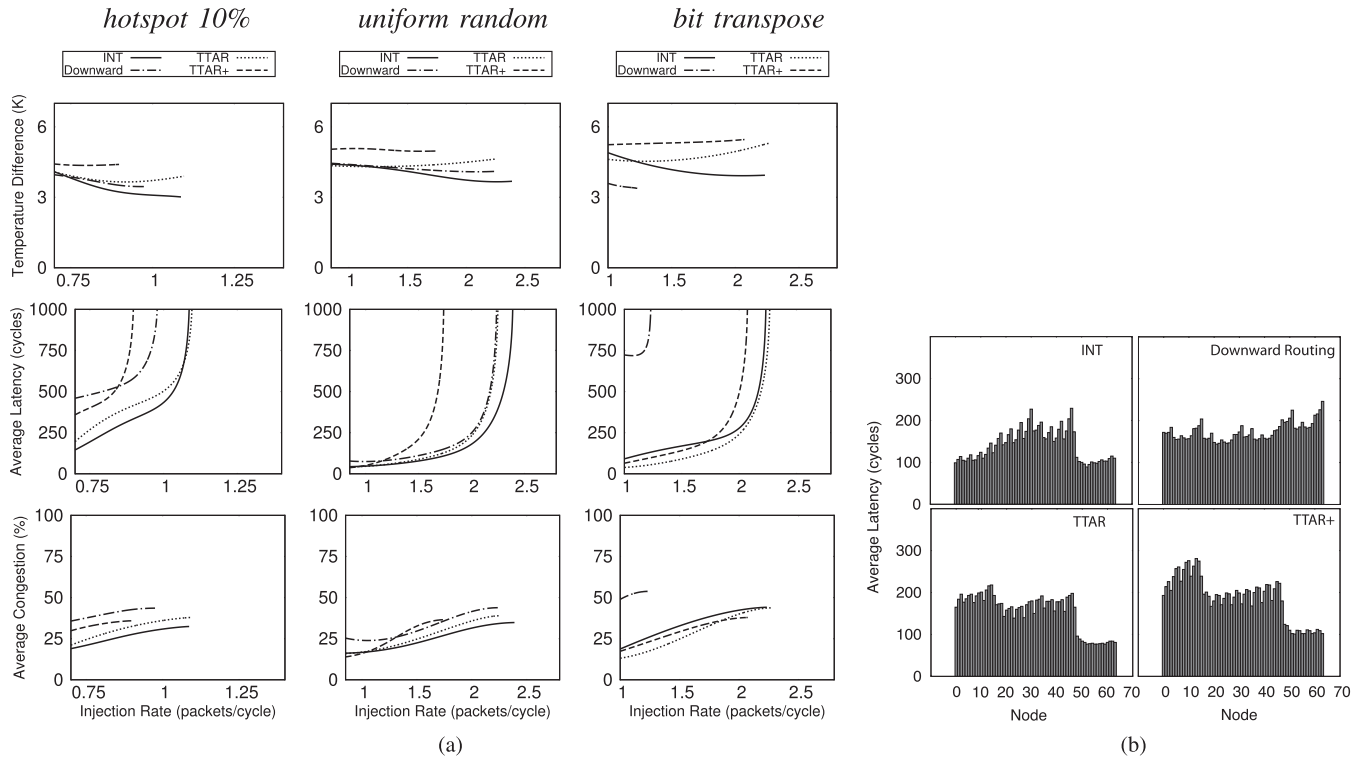

Fig. 5. (a) Peak temperature difference, average packet delay, and average congestion with eight thermal hotspots for different traffic patterns. In the figures, the data lines end at different injection rates, corresponding with the saturation throughput of the network. (b) Latency distribution among network nodes for *uniform random* traffic at a peak injection rate. *Note*: Since each tier in the stack contains 16 tiles, the distributions in (b) also indicate the variation of latency across tiers.

TABLE II
OVERHEADS

	WIDTH	AREA	POWER
INT	8-bit	0.9%	$< 0.1\%$
TTAR	8-bit	10.7%	25%
TTAR+	8-bit	10.7%	61%
DOWNWARD	X	0%	0%

that they result in elevated operating temperatures, yielding higher latency.

3) Overheads: Table II lists the area and power overheads imposed by each routing strategy. Note that the area overheads listed for RCA derivatives, such as TTAR and TTAR+, do not include the computational resources required to perform the weighted summation. Here, these are considered to be integrated within the area of the crossbar switch. The aggregate–propagate network for these cases is approximated as an 8-bit point-to-point network with a single-entry input buffer per port.

V. CONCLUSION

In this brief, we have presented the INT adaptive routing algorithm for dynamically throttled NoCs. INT bases adaptive routing decisions solely on temperature information from neighboring routers, thus minimizing monitoring overheads. Interconnect traffic is steered away from regions of high temperature, yielding balanced thermal profiles, with up to 25% lower thermal gradients. Furthermore, as a consequence of lower operating temperatures and adaptive routing, communication latency is improved, and network congestion decreased by up to 50% even in the presence of system thermal hotspots.

REFERENCES

[1] S. S. Kumar, A. Zjajo, and R. van Leuken, "Physical characterization of steady-state temperature profiles in three-dimensional integrated circuits," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2015, pp. 1969–1972.

[2] K. Ramakrishnan *et al.*, "Variation impact on SER of combinational circuits," in *Proc. Int. Symp. Qual. Electron. Des.*, Mar. 2007, pp. 911–916.

[3] K. Skadron *et al.*, "Temperature-aware microarchitecture: Modeling and implementation," *ACM Trans. Architect. Code Optim.*, vol. 1, no. 1, pp. 94–125, Mar. 2004.

[4] K. Puttaswamy and G. H. Loh, "Thermal herding: Microarchitecture techniques for controlling hotspots in high-performance 3D-integrated processors," in *Proc. IEEE Int. Symp. High Perform. Comput. Architect.*, 2007, pp. 193–204.

[5] S.-Y. Lin, T.-C. Yin, H.-Y. Wang, and A.-Y. Wu, "Traffic-and thermal-aware routing for throttled three-dimensional network-on-chip systems," in *Proc. Int. Symp. VLSI Des., Autom. Test*, Apr. 2011, pp. 1–4.

[6] C.-H. Chao *et al.*, "Traffic- and thermal-aware run-time thermal management scheme for 3D NoC systems," in *Proc. ACM/IEEE Int. Symp. Netw.-on-Chip*, May 2010, pp. 223–230.

[7] M. Nagata, "Limitations, innovations, and challenges of circuits and devices into a half micrometer and beyond," *IEEE J. Solid-State Circuits*, vol. 27, no. 4, pp. 465–472, Apr. 1992.

[8] P. Gratz, B. Grot, and S. Keckler, "Regional congestion awareness for load balance in networks-on-chip," in *Proc. IEEE Int. Symp. High Perform. Comput. Architect.*, Feb. 2008, pp. 203–214.

[9] F. Liu, H. Gu, and Y. Yang, "DTBR: A dynamic thermal-balance routing algorithm for network-on-chip," *Comput. Elect. Eng.*, vol. 38, no. 2, pp. 270–281, Mar. 2012.

[10] S. S. Kumar, M. T. A. Djie, and R. van Leuken, "Low overhead message passing for high performance many-core processors," in *Proc. Int. Symp. Comput. Netw.*, Dec. 2013, pp. 345–351.

[11] G.-M. Chiu, "The odd-even turn model for adaptive routing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 11, no. 7, pp. 729–738, Jul. 2000.

[12] S. S. Kumar, A. Zjajo, and R. van Leuken, "Ctherm: An integrated framework for thermal-functional co-simulation of systems-on-chip," in *Proc. Euromicro Int. Conf. Parallel, Distrib. Netw.-Based Process.*, Mar. 2015, pp. 674–681.