Ant Colony Optimization-Based Adaptive Network-on-Chip Routing Framework Using Network Information Region

Hsien-Kai Hsin, En-Jui Chang, Kuan-Yu Su, and An-Yeu (Andy) Wu, *Senior Member, IEEE*

Abstract—The network-on-chip (NoC) system can provide more scalable and flexible on-chip interconnection compared with system bus. The performance of on-chip adaptive routing algorithms greatly relies on the adopted network information. To the best our knowledge, previous routing algorithms utilize either spatial or temporal network information to improve performance. However, few works have established a framework on analyzing the network information nor showed how to integrate the spatial and temporal network information. In this paper, we define the network information region (NIR) framework for NoC systems. The NIR can indicate arbitrary combinations of network information and corresponding routing algorithms. We demonstrate how to apply NIR on analyzing the adaptive routing algorithms. To further demonstrate how NIR can help to integrate the spatial or temporal network information, we propose the ACO-based pheromone diffusion (ACO-PhD) adaptive routing framework based on the NIR. By diffusing the pheromone outward, spatial and temporal network information can be exchanged among adjacent routers. The range (i.e., size and shape) of the NIR is controllable by setting the parameters in the ACO-PhD algorithm. We show that we can reconfigure the ACO-PhD algorithm to each routing algorithm in its NIR subsets by adjusting the parameter settings. Finally, we implement and analyze the hardware design of corresponding router architecture. The results show an improvement of 4.86-16.93 percent on network performance and the highest area efficiency is achieved by the proposed algorithm.

Index Terms—Network-on-chip, adaptive routing, network information region, ant colony optimization

1 Introduction

Increased complexity and interconnection delay become the limiting factor of system-on-chip performance with the shrinking size of deep sub-micron technology [1], [2], [3]. To meet data transfer requirement and increase the interconnection efficiency in chip multiprocessor systems, the network-on-chip (NoC) systems have been developed in past decade and proven to be scalable and flexible solutions [4], [5], [6].

However, the NoC traffic load is highly unstable and unbalanced under various applications. The congested channels may shift from time to time and result in serious performance degradation. Furthermore, the increasing NoC size makes the congestion condition more severe and unpredictable [7]. Hence, many works have developed different adaptive routing algorithms to balance the traffic load [8], [9], [10], [11]. The adaptive routing algorithms have the advantage in selecting appropriate output channels based on the evaluation of network information. According to previous works, the precision of network information greatly affects network performance [12], [13].

Because the buffer space is more limited and the queueing dependency is more significant in NoC than in wide-area-network, the buffer utilization is considered as an essential network congestion information of the NoC systems [16], [17], [18], [19], [35], [36].

In order to study on network information acquirement, we visualize the network information, as shown in Figs. 1d, 1e, 1f and the *network information coordination* for current router at time T in Figs. 5a, 5b, 5c and 5d. We can locate the entire network information by this coordination. The vertical axis is the index of time slots. The time interval depends on information sampling rate. The top layer information is previous time slot, denote as $T-1$. The underneath layer is the historical information at two time slots ago $T-2$.

The horizontal layers are the spatial network information at different time slots. We can locate the NoC nodes with grid points on x- and y-axes. The origin of each layer is current router, and the intersection points around are information from corresponding neighboring routers.

The network information can be collected through propagation/aggregation process and/or storage process and are denoted as the *spatial* and *temporal* network information, respectively. Facing more complex load-balancing problems in the design of advanced NoC systems, we may have to consider integrating the spatial and temporal channel information in adaptive routing algorithm for better capturing of the traffic condition as follows:

- *Regional buffer utilization → spatial network information*. The propagating and aggregating buffer utilization information from adjacent routers are defined as spatial network information in our framework. As shown in Figs. 1a, 1b and 1c, output buffer level (OBL) [16] acquires the buffer utilization from

• *The authors are with the Graduate Institute of Electronics Engineering, National Taiwan University, Taipei 10617, Taiwan. E-mail: {ckcraig, enjui, Kuanyu}@access.ee.ntu.edu.tw, andywu@cc.ee.ntu.edu.tw.*

Manuscript received 25 Feb. 2014; revised 22 Sept. 2014; accepted 27 Oct. 2014. Date of publication 2 Nov. 2014; date of current version 10 July 2015. Recommended for acceptance by B. Ravindran.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TC.2014.2366768

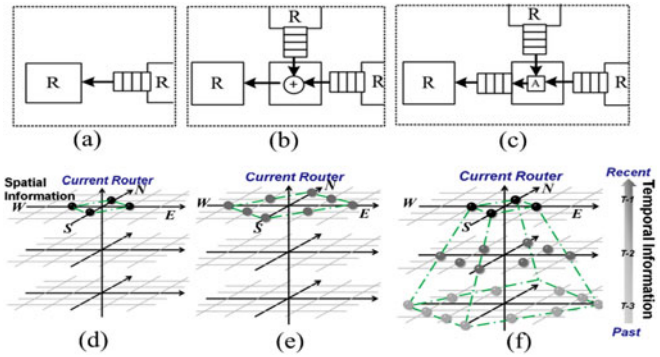


Fig. 1. Selection functions utilizing the spatial network information that come from adjacent routers. (a) OBL [16], (b) NoP [16], and (c) RCA [17], the block A is the aggregator of adjacent buffer status. (d)(e)(f) The NIR of OBL, NoP, and RCA, respectively.

adjacent routers. Neighbor-on-path (NoP) [16] acquires two-hop away information on path. Regional congestion awareness (RCA) [17] can acquire farther network information with an aggregation delay in each hop. Their network information region (NIR) are illustrated as in Figs. 1d, 1e and 1f, respectively.

- *Historical pheromone table → temporal network information.* In Fig. 2, ACO-based adaptive routing algorithm is a routing function coupled with ACO-based selection function [18], [19], [36]. The concept of ACO is shown in Fig. 3a, the ant colony optimization (ACO) [18] has ants diffusing pheromone on the paths. The later arrived ants are aware of the pheromone levels in the selection process. The ACO-based selection function utilizes an ACO routing table to store the pheromone and a state transition rule to update the table [21], [22], [23], [35]. We define the accumulating pheromone information as temporal network information in our framework. We can locate this NIR as in Fig. 3b. Since ACO-based routing can identify historically less-congested channels by utilizing the pheromone information, it has higher potential on balancing traffic load on NoC [18], [19], [20].

In this work, we focus on establishing a *network information region* framework on analyzing the network information, showing how to integrate the spatial and temporal network information with NIR, and investigating on how to use the proposed technique to develop advanced reconfigurable schemes. Our contribution are listed as follows:

1) We define the NIR in detail. We demonstrate how to apply NIR on analyzing four adaptive selection functions: output buffer level [16], neighbor-on-path [16], regional congestion awareness [17], and ant colony optimization [19], [36].
2) To demonstrate how NIR can help to integrate the spatial or temporal network information, we propose and implement the ant colony optimization-based pheromone diffusion (ACO-PhD) adaptive routing algorithm to provides the most attainable NIR for single-hop-propagation network as shown in Fig. 4a.
3) We show that we can reconfigure the ACO-PhD algorithm to each routing algorithm in its NIR subsets by adjusting the parameter settings. This concept can help in developing more advanced reconfigurable schemes for different types of traffic or design constraints.

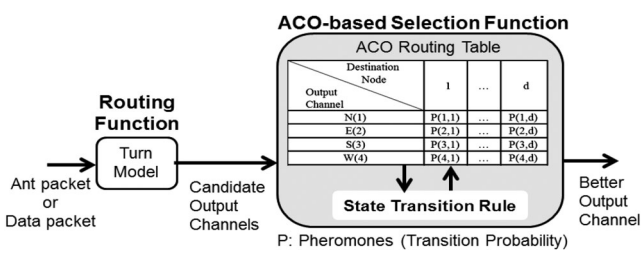


Fig. 2. ACO-based adaptive routing algorithm in NoC.

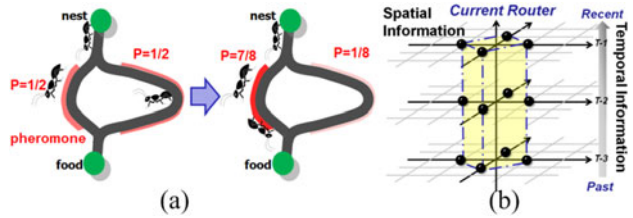


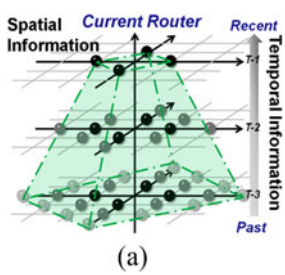
Fig. 3. The illustration of (a) ACO concept [20], and (b) the NIR of ACO.

We focus on three aspects, which are the information retrieving on spatial domain, temporal domain, and scalability. Fig. 4b shows the attributes of the related works and proposed algorithm. The proposed algorithm has higher ability in utilizing network information and balancing network traffic load in a scalable NoC system.

The rest of this paper is organized as follows: In Section 2, we define the NIR and apply it on analyzing the related works. In Section 3, we demonstrate how NIR can help to integrate the proposed ACO-PhD scheme. In Section 4, we evaluate the performance of the proposed framework. In Section 5, we implement the router architecture of ACO-PhD algorithm. Section 6 gives the conclusion and future work.

2 Network Information Region and Linking NIR with Existing Selection Functions

The routers under consideration are named *information nodes* (INs) and are represented with dots in the figures. We identify the router nodes as $N(x, y, t)$, where x, y denote the location of the node and t denotes the time slot.



Selection Function	Network Information		
	Spatial	Temporal	Scalability
OBL [16]	+		
NoP [16]	++		
RCA [17]	++		+
ACO [18]	+	+	
Proposed ACO-PhD	++	+	+

* ++ denotes relatively good, + denotes normal.

(b)

Fig. 4. (a) The NIR of the proposed ACO-PhD for demonstrating the integration of spatial and temporal network information. (b) The attribute of the NIR for different selection functions.

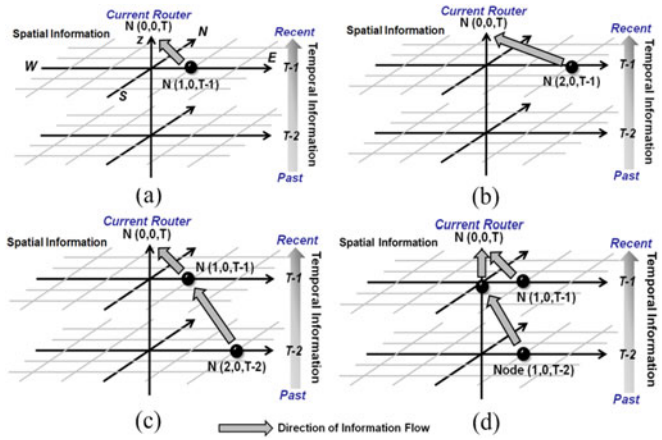


Fig. 5. Network information flow of the information nodes corresponding to (a) OBL, (b) NoP, (c) RCA, and (d) ACO.

Fig. 5 illustrates the *network information flow*, which is the information retrieving process from the INs.

- In Fig. 5a, corresponding to OBL of Fig. 1a, the information from the east port node $N(1, 0, T-1)$ is propagated to current router.
- In a similar case, Fig. 5b, corresponding to NoP of Fig. 1b, shows the information from two-hop away $N(2, 0, T-1)$ requires a two-hop wiring for propagation.
- In Fig. 5c, corresponding to RCA of Fig. 1c, the information from two-hop way node $N(2, 0, T-2)$ requires one cycle to propagate to node $N(1, 0, T-1)$, and finally to the current node.
- For the case in Fig. 5d, corresponding to ACO of Fig. 2, the information from node $N(1, 0, T-2)$ reaches current router at $N(0, 0, T-1)$. his historical information is then stored in current router.

We define the *network information region* as the network information set acquired by a routing algorithm for making routing decisions. The NIR contains multiple INs that are acquired by current network settings. We define these INs as the *available information nodes* (AINs) of the corresponding routing algorithm.

In the following, we introduce four related works, their NIR, and the corresponding wiring for information acquirement.

2.1 NIR of Output Buffer Level Selection Function [16]

OBL is a basic implementation of adaptive routing. It monitors the number of unoccupied buffer spaces (i.e., the *free slots*) of each output port. After the channel constraint function provides the candidate channels, OBL chooses the one with the most free slots. If there are more than one candidates having the most free slots, OBL selection randomly chooses one from them. The signal wiring and corresponding NIR of OBL are illustrated in Fig. 6.

As in Fig. 6a, each router spreads out its buffer information through $free\_slots\_out$ signals and receives the buffer information of neighboring routers from $free\_slots\_in$. By this method, each router can acquire information within distance of 1-hop, and the NIR of OBL is in Fig. 6b.

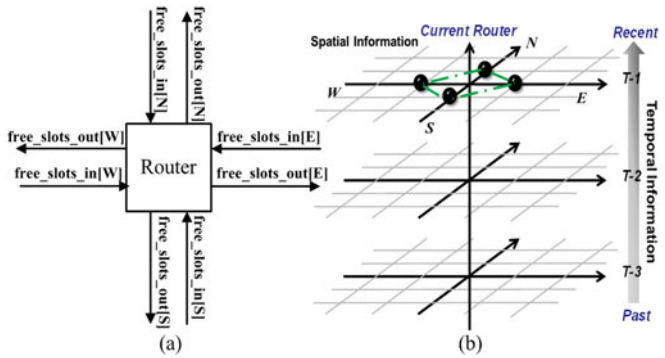


Fig. 6. (a) Signal wiring of OBL. (b) Network information region of OBL, with numerical form of $NIR\{1, T-1\}$.

The four dots are the AINs of the information acquired by the current router.

To describe NIR in numerical form, we define a form to represent the NIR as $NIR\{d, t\}$, where d denotes the routers that are d hops away from the current node and t denotes the cycle that the corresponding information originated from. For example, $NIR\{1, T-1\}$ represents nodes $N(1, 0, T-1)$, $N(0, 1, T-1)$, $N(-1, 0, T-1)$, and $N(0, -1, T-1)$. We can use the form $NIR\{1, T-1\}$ to describe the NIR of OBL selection function.

2.2 NIR of Neighbors-on-Path Selection Function [16]

The NoP selection function focuses on information with farther distance, which is two-hop-away buffer status, as shown in Fig. 7. However, in order to acquire the information of the second hop, more wiring cost is required. The signal wiring is in Fig. 7a, and the NIR is shown in Fig. 7b in the form of $NIR\{2, T-1\}$.

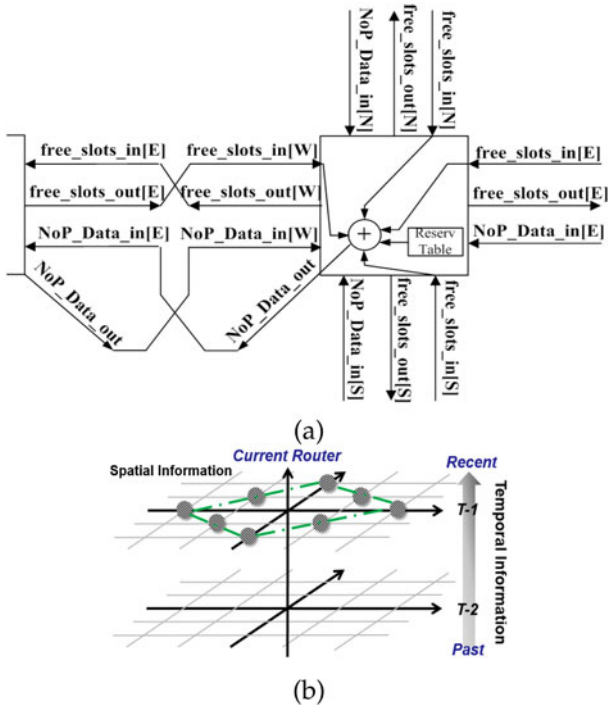


Fig. 7. (a) Signal wiring of NoP. (b) Network information region of NoP, with numerical form of $NIR\{2, T-1\}$.

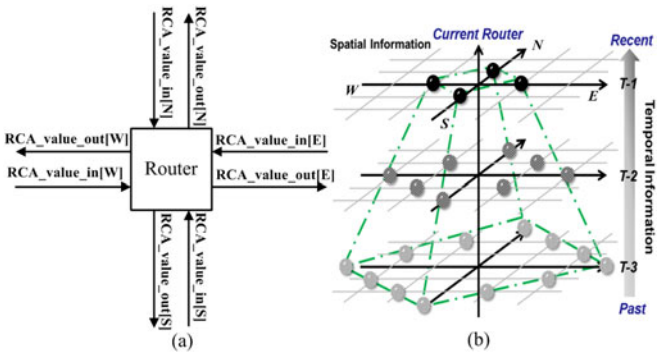


Fig. 8. (a) Signal wiring of RCA. (b) Network information region of RCA, with numerical form of $NIR\{n, T-n\}$.

The selection function of NoP works similarly as OBL but the criterion changes from $free\_slots\_in$ to the NoP_score. The channel constraint function offers the candidate channels to selection function. Selection function then eliminates the channels that are currently occupied. This is transmitted through NoP_Data_in. Finally, NoP_score is the accumulated number of free slots in the remaining channels. In other words, NoP considers the accessible channels by using two-step channel constraint function. At the cost of additional wiring for information in reservation table, the channels being occupied by other packets are also eliminated. With this consideration, NoP_score can represent more precise network information.

2.3 NIR of Regional Congestion Awareness Selection Function [17]

Unlike NoP which acquires long-range information with direct wiring, RCA gathers long-range information through aggregation and propagation. First, aggregation module receives the information from the downstream routers and adds it up with local buffer status. The propagation module then sums up information of different directions and sends it toward the upstream. With this approach, RCA can spread out the buffer status at the speed of one hop per cycle, and requires only adjacent router wiring.

The signal wiring of RCA is illustrated in Fig. 8a and is similar to OBL. Note that, the information contained in RCA_value_out and $free\_slots\_out$ are different. The aggregation module, RCA_value_out contains both buffer status of local router and adjacent routers. The NIR of RCA is shown in Fig. 8b. Since the propagation module spreads out information at the speed of one hop per clock cycle, the region thus contains the information of routers in n hop with a delay of n clock cycles. We may describe this NIR in the form of $NIR\{n, T-n\}$, where n is a positive integer.

As OBL and NoP, the information RCA acquired is buffer status, indicating the remaining free slots of each direction. So while choosing from the candidate channels, RCA simply selects the one with the highest RCA_value_in.

2.4 NIR of ACO-Based Selection Function

2.4.1 ACO-Based Adaptive Routing in NoC [18]

As shown in Fig. 2, the new pheromone values are obtained using the *state transition rule*, which is derived from current and historical information of the network, as shown in (1). The pheromone $Ph(j,d)$ can be viewed as the probability of selecting channel index j, while $j \in \{North, South, East, West\}$ for packets heading to destination index d. L_j is normalized free slots of queue at channel j, and α is the weighting coefficient ranging from zero to one for current and historical information of the network

$$Ph'(j,d) = (1-\alpha) \times Ph(j,d) + \alpha \times L_j. \tag{1}$$

2.4.2 Regional ACO-Based Adaptive Routing [19]

Regional ACO-based adaptive routing (RACO) was proposed for improvements on both balancing traffic load and reducing the size of routing table. Instead of storing pheromone information of all the source-destination pairs in the network, RACO greatly reduces the table size from N^2 entries to several entries by solving the information sharing problem of the original routing table in ACO, where N is the dimension of the network. A more feasible on-chip implementation (e.g., memory cost, table access time, and power consumption) is thus provided. As in (2) and (3), the destination index d in (1) is replaced with region index R. N_E is the reduced number of entries with region setting k, which is a multiple of four according to previous research. By taking advantage of regional characteristic of NoC systems and the properties of ant colony, RACO greatly reduces implementation cost while maintaining similar performance

$$N_E(k) = \begin{cases} 1, & \text{if } k = 0, \\ 4k, & \text{if } k \in N, \end{cases} \tag{2}$$

$$Ph'(j,R) = (1-\alpha) \times Ph(j,R) + \alpha \times L_j. \tag{3}$$

2.4.3 Signal Wiring and NIR of ACO

The pheromone expression in (3) can be viewed as an exponential moving average (EMA) of the buffer status in the form (4). EMA_t is the moving average at a time slot t and can be substituted by the pheromone; D_t is the data observed at a time period t and can be substituted by buffer status; and α is the weighting between EMA_t and D_t:

$$EMA_{t+1} = (1-\alpha) \times EMA_t + \alpha \times D_t. \tag{4}$$

As a result, the meaning of pheromone in ACO can be interpreted as the buffer information. Therefore, a high pheromone value indicates a historically less-congested channel for packet transmission. Moreover, owing to the EMA characteristic, the pheromone in ACO can smooth out short-term fluctuations, highlight long-term trends of the buffer status, and provide more accurate network information. The signal wiring of ACO illustrated in Fig. 9a is the same as OBL, but the NIR of ACO contains the historical information, making it extend along the vertical temporal axis as in Fig. 9b. We may describe this NIR by the form of $NIR\{1, T-n\}$, where n is a positive integer.

3 Proposed ACO Framework with Pheromone Diffusion (ACO-PhD)

To demonstrate how NIR can help to integrate the spatial or temporal network information, we propose the ant colony optimization-based pheromone diffusion adaptive routing

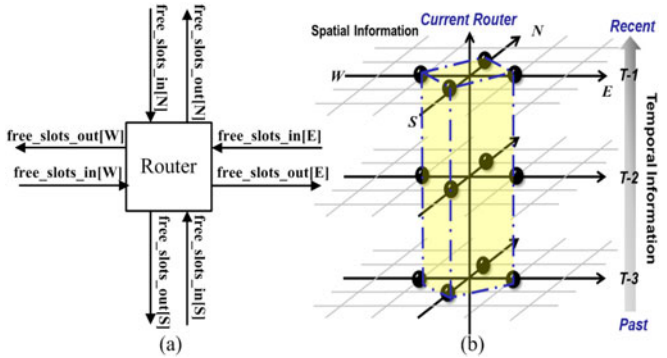


Fig. 9. (a) Signal wiring of ACO. (b) Network information region of ACO, with numerical form of $NIR\{1, T-n\}$.

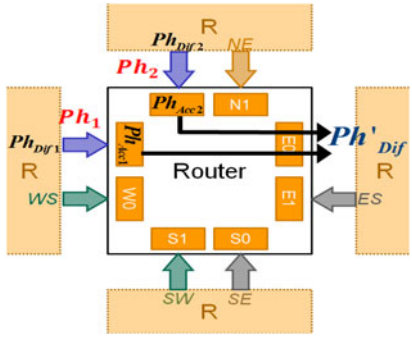


Fig. 10. Formation of Ph_{Dif} signals.

algorithm to provides the most attainable NIR for single-hop-propagation network.

The main concept of ACO-PhD is to utilize the pheromone information in ACO and further spreads it outward as an exchange of information. Since the pheromone is the exponential moving average of buffer status that represents more precise network information. It is beneficial to share this information with neighboring routers.

In this work, we define two kinds of pheromone as follows:

- *Diffusive pheromone* Ph_{Dif} is the propagated information from the different channels of the region {*NE*, *SE*,*SW*,*NW*} relative to current router. Ph_{Dif} contains the information of buffer utilization and historical pheromone in each level (i.e., distance of router) with certain weightings.
- *Accumulative pheromone* Ph_{Acc} is the accumulative Ph_{Dif} information in each router. Ph_{Acc} is stored in the pheromone table and will be propagated out by the diffusion mechanism.

3.1 Formulation of Diffusive Pheromone Ph_{Dif} to Link with Spatial Information

The pheromone accumulated in each router represents the observed historical information. By making use of this information and spreading it outward hop by hop, neighboring routers can acquire not only the present network status but also the status of the past. Therefore, the NIR of ACO-PhD can expand on both temporal domain and spatial domain.

Before introducing the formation of Ph_{Dif}, we want to explain the notation of the directions in our NoC system. For each router, the entire NoC is separated into four regions, which are NE, NW, SE, and SW. For example, for packets heading to NE region, selection function may choose *North* or *East* channel for transmission. To make a proper choice, network status of the two directions are required. Since there are four regions and two directions for each region, the network status is divided into a total number of eight directions. Note that *NE* and *EN* are both required for packets heading to NE region, while *NE* stands for the network information of *North* channel and *EN* stands for that of the *East* channel.

To receive the information from the eight directions, each router has to provide corresponding information for others. For ACO-PhD, it is the Ph_{Dif} signal of the eight directions. Take Fig. 10 for example, the purpose of Ph'_{Dif} is to inform the east neighbor router of the network status of sending packets through current router toward NW region. Thus, besides the information of local router, the information of *NW* region is also required, which is the information from west and north channel. Besides combining the information of the two directions, Ph_{Dif} further takes the temporal information Ph_{Acc} into consideration. Ph_{Acc} accumulates network information and is stored in the pheromone table. The eight orange blocks in Fig. 10 represent Ph_{Acc} of each direction. The detailed formula of Ph'_{Dif} is shown below:

$$\begin{aligned} Ph'_{Dif} &= \frac{1}{2}(Ph_1 + Ph_2) \\ &= \frac{1}{2}(\beta Ph_{Acc1} + (1-\beta)Ph_{Dif1}) \\ &\quad + \frac{1}{2}(\beta Ph_{Acc2} + (1-\beta)Ph_{Dif2}). \end{aligned} \tag{5}$$

Ph_1 and Ph_2 stand for information of the two directions, and each of them consists of two components, Ph_{Acc} and Ph_{Dif}. The ratio between Ph_{Acc} and Ph_{Dif} is determined by β, which is the weighting of pheromone diffusion ranging from zero to one. It is worth noticing that by combining Ph_{Acc} and Ph_{Dif}, the diffusing Ph'_{Dif} contains both temporal information, which is Ph_{Acc} of the local router, and spatial information, which is Ph_{Dif} diffused from neighboring routers. The Ph'_{Dif} then accumulates with local congestion status and becomes Ph_{Dif} at the next hop. The purpose of pheromone diffusion is thus achieved.

3.2 Modified State Transition Rule

To enable the information sharing of the pheromone information, the state transition rule is reformed as follows:

$$Ph'_{Acc}(j,R) = (1-\alpha)Ph_{Acc}(j,R) + \alpha Ph_{Dif}(j,R). \tag{6}$$

There are two types of pheromone in ACO-PhD, one is for information accumulating in the original ACO, and the other is for information spreading in pheromone diffusion. To distinguish, we name the pheromone in the original ACO as accumulative pheromone (Ph_{Acc}), and the new pheromone in ACO-PhD as diffusive pheromone (Ph_{Dif}). Comparing (6) with (3), the local model, which is the buffer information of neighboring routers and represented by L_j,

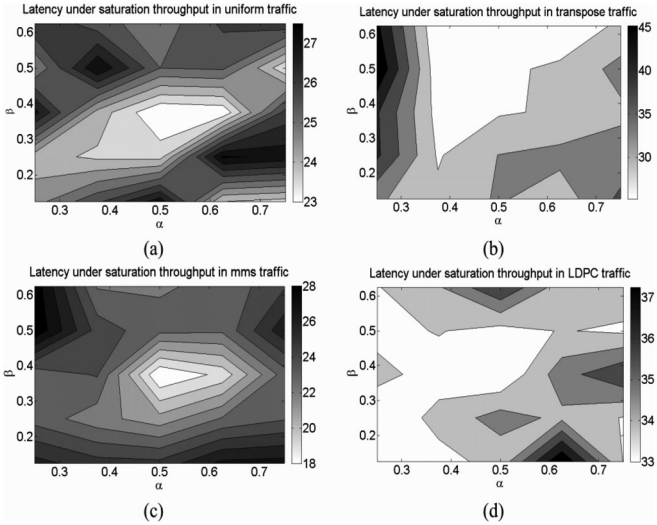


Fig. 11. The latency under saturation throughput for different settings of α and β in (a) uniform, (b) transpose, (c) MMS [30], and (d) LDPC traffic [36].

is replaced by Ph_{Dif}. Unlike L_j which contains spatial information only, Ph_{Dif} contains both the spatial and temporal network information.

3.3 Design Parameters of ACO-PhD

There are two design parameters in ACO-PhD. One is the *ACO* weighting α. The other is the pheromone diffusion *PhD* weighting β. We further describe the physical meaning and the proper value of each weighting.

3.3.1 ACO Weighting—α

The ACO weighting α determines the composition of Ph_{Acc}. In (6), α determines the ratio between the historical information stored in the local router and the incoming information sent from the neighboring routers. A higher α leads to a higher ratio of Ph_{Dif}, indicating that the routing algorithm relies more on the incoming information over the historical information stored. A lower α, on the other hand, indicates more trust on the historical information that the local router observed.

3.3.2 PhD Weighting—β

In (5), Ph'_{Dif} is the pheromone value that propagates the network information outward. The ratio between Ph_{Acc} and Ph_{Dif} is controlled by PhD weighting β. Higher β indicates higher proportion of Ph_{Acc} in Ph_{Dif}, which increases the significance of the temporal information. Lower β, on the other hand, indicates higher proportion of Ph_{Dif} in Ph'_{Dif}, which increases the significance of the spatial information.

3.3.3 ACO-PhD Weighting Settings

The value of α and β ranges from zero to one. We show the latency of different settings of α and β under saturation throughput in Fig. 11. For network performance and hardware friendliness, we choose the optimized value of (α, β) in uniform traffic to be $(0.5, 0.375)$ in our simulations in Sections 4.2 and 4.3. In addition, we find this setting is also a suitable choice in different kinds of synthetic and real traffic patterns as shown in Fig. 11b, and Figs. 11c, 11d, respectively.

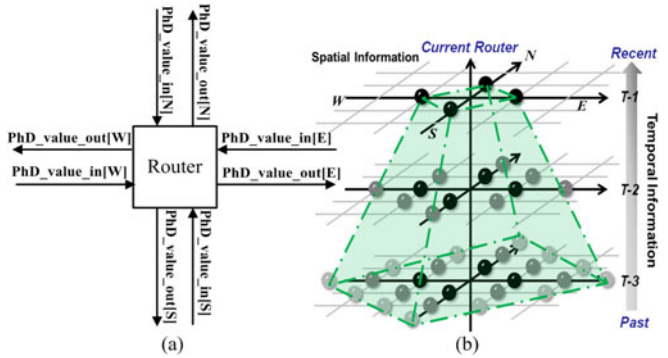


Fig. 12. (a) Signal wiring of ACO-PhD. (b) Network information region of ACO-PhD, with numerical form of $NIR\{\langle 1, n\rangle, T-n\}$.

3.4 NIR and Operations of ACO-PhD

The signal wiring and NIR of ACO-PhD are illustrated in Fig. 12. We may describe this NIR in the form of $NIR\{\langle 1, n\rangle, T-n\}$, where n is a positive integer and $\langle 1, n\rangle$ is an integer ranging from 1 to n. This form means, for each $T-n$ time slot, the NIR includes the AINs with hop counts within range of $\langle 1, n\rangle$. This NIR provides the most attainable network information with information propagation speed equals to one hop per cycle. It has the shape of a pyramid, which is the bound of network information.

Let us now analyze how an ACO-PhD algorithm works. Fig. 13 shows the pseudocode of the ACO-PhD selection executed at node n_c with destination node n_d. The input parameter AOC stands for the set of admissible output channels, which is computed by the routing function RF. (i.e. $AOC = RF(n_c, n_d)$). The output parameter is the selected channel $sc \in AOC$:

1) *Region computation.* For each channel in AOC (line 2), we first compute the region for reaching certain columns in the pheromone tables (line 3).
2) *Pheromone accumulation.* We use (6) to update the accumulation pheromone Ph_{Acc} (line 4).
3) *Path selection.* We select the route from the highest value of pheromone and compute the corresponding channel from region (lines 6 and 7).

```
1  ACO-PhD Selection (in : AOC, out: sc) {
2    for ch ∈ AOC {
3      R←ComputeRegion (n_c, n_d, ch)
4      Ph_Acc[R]←(1-α)Ph_Acc[R]+ α Ph_Dif[R]      //Eq. (6)
5    }
6      sr ←R s.t. Ph_Acc[R]= max(Ph_Acc[])
7      sc ←ComputeChannel(sr)
8  }
9  ACO-PhD Diffusion () {
10   for d ∈ Direction {
11     Ph_1[d] ←β Ph_Acc1[d]+ (1-β) Ph_Dif1[d]
12     Ph_2[d] ←β Ph_Acc2[d]+ (1-β) Ph_Dif2[d]
13     Ph'_Dif[d]← ½ Ph_1[d]+ ½ Ph_2[d]      //Eq. (5)
14   }
15 }
```

Fig. 13. Pseudo code of ACO-PhD selection function and diffusion process.

TABLE 1
Different Parameter Settings of ACO-PhD

Selection Function	Weighting of ACO - α	Weighting of PhD - β	Norm. W_{Acc}	Norm. W_{Dif}
ACO-PhD$_{OBL}$	1.0	–	–	1
ACO-PhD$_{ACO}$	0.5	–	N	1
ACO-PhD$_{RCA}$	1.0	0.0	–	M
ACO-PhD	0.5	0.375	N	M

4) *Pheromone diffusion*. For the eight directions (line 9), the propagation process is conducted every cycle (lines 11-13), collecting information from corresponding part of the network.

3.5 Reconfigurable NIR Framework of ACO-PhD

The coverage of NIR of advanced routing algorithms may need to be adaptively reconfigured for different types of traffic or design constraints. For example, network information can be simple under simple traffic patterns to save energy and can be more precise under complexed traffic patterns to improve performance.

According to our observation, the NIR of OBL, RCA, and ACO are the subsets of the ACO-PhD NIR. Therefore, we can adjust the parameters to reconfigure the coverage (i.e. size and shape) of ACO-PhD NIR. By reconfiguring, ACO-PhD can be tuned to have the same behavior with the corresponding algorithm. That is to say, ACO-PhD can be seen as a general case of these selection schemes, while OBL, RCA, and ACO are the special cases of ACO-PhD.

For demonstration, the settings are shown in Table 1 and Fig. 14. ACO-PhD$_{OBL}$, ACO-PhD$_{RCA}$, and ACO-PhD$_{ACO}$ stand for ACO-PhD with the same NIR as OBL, RCA, and ACO, respectively.

Note that the wordlength W_{Acc} and W_{Dif} of Ph_{Acc} and Ph_{Dif} are normalized according to the exact buffer length. The physical meaning of W_{Acc} and W_{Dif} are the temporal and spatial distance that each router can observe, respectively.

3.5.1 Settings of ACO-PhD$_{OBL}$

ACO-PhD$_{OBL}$ does not requires historical information. Therefore, the ACO weighting α is set to 1, making Ph_{Acc} contains present information only. In addition, ACO-PhD$_{OBL}$ requires the spatial information of the adjacent routers. Therefore, the W_{Dif} is set to 1.

3.5.2 Settings of ACO-PhD$_{ACO}$

For ACO-PhD$_{ACO}$, the ACO weighting α is set to a specific value that determines the composing of Ph_{Acc}. In addition, ACO-PhD$_{ACO}$ requires historical information. The larger observation window provides more precise network status. W_{Acc} is set to N, which represents a observation window of N cycles. The influence of the size N on the network performance is simulated by fixed-point analysis and set to 7. The W_{Dif} set to 1 because ACO-PhD$_{ACO}$ also requires the adjacent routers information.

3.5.3 Settings of ACO-PhD$_{RCA}$

Since ACO-PhD$_{RCA}$ does not require historical information, the ACO weighting α is set to 1. In addition, because ACO-PhD$_{RCA}$ propagate information outward, the *PhD* weighting β is set to 0 and W_{Dif} is set to M, which indicates the acquirement of spatial information within M-hops. The influence of the size M is simulated by fixed-point analysis and set to 4.

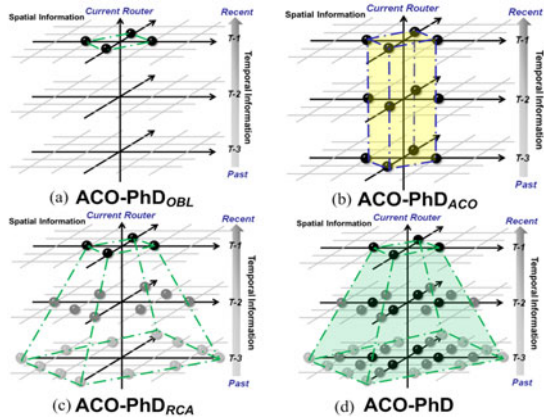


Fig. 14. NIR settings to different selection functions: (a) ACO-PhD$_{OBL}$, (b) ACO-PhD$_{ACO}$, (c) ACO-PhD$_{RCA}$, (d) ACO-PhD.

3.5.4 Settings of ACO-PhD

For ACO-PhD, the value of ACO weighting α and Ph_{Acc} wordlength W_{Acc} are the same as ACO-PhD$_{ACO}$. The PhD weighting β is set to 0.375 and W_{Dif} is set to same with ACO-PhD$_{RCA}$ for diffusing the pheromone outward.

4 Performance Evaluation

4.1 Environment Settings

The simulation results are evaluated by Noxim [24], which is a flit- and cycle-accurate SystemC simulator for Network-on-Chip systems. An 8-ary 2-mesh network topology is constructed under wormhole switching mechanism and round-robin arbitration [8]. Each channel has an input queuing buffer with the size of four flits and each packet is 8-flit-long. Our setting is consistent with work [16].

The simulation time is 20,000 cycles and the first 2,000 cycles is the warm-up time for the NoC system, in order to measure the steady-state performance of network. The average latency under different packet injection rate (*pir*) is used as the performance index of the simulations. We also adopt the *saturation throughput* [25], which is the throughput where average latency equals to twice of the zero-load latency, as the evaluation metric.

4.2 Experimental Results on Synthetic Traffic

For each traffic scenario and algorithm, we compare the average packet delay with various *pir* values. The selection schemes compared are OBL, NoP, RCA-Quadrant, ACO, and ACO-PhD, all using an Odd-Even routing algorithm.

4.2.1 Uniform Traffic

According to [8], the *uniform* traffic, in which each source is equally likely to send to each destination, is the most commonly used traffic pattern in network evaluation. This pattern is very benign because uniform traffic making the traffic uniformly distributed for the routing algorithms. The results are illustrated in Fig. 15a. As it shows, the performance of

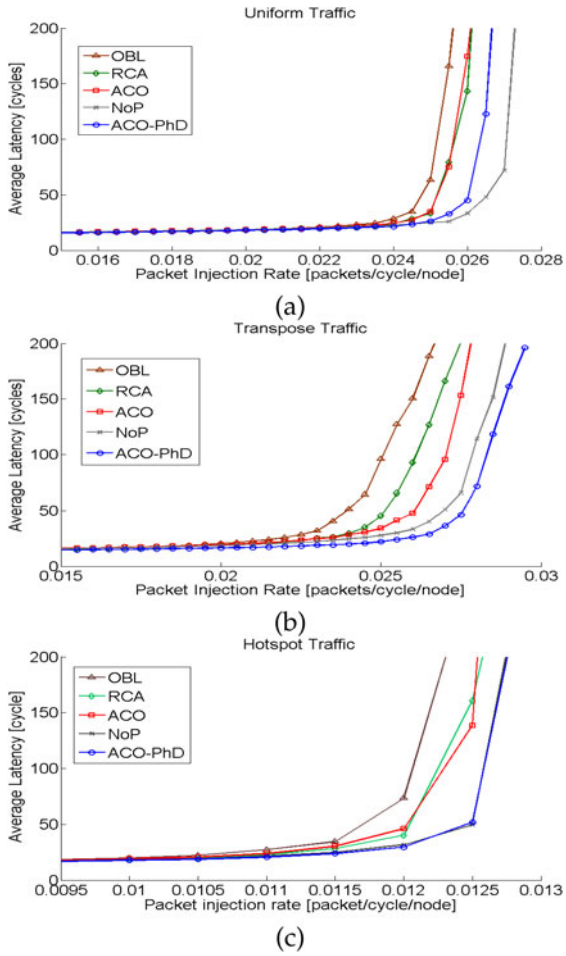


Fig. 15. Performances evaluation under (a) uniform (b) transpose, and (c) hotspot traffic.

OBL is the least among all, owing to its smallest network information region. RCA and ACO acquire better performance over OBL by extending their NIR along the spatial and temporal domains, respectively. The performance of ACO-PhD further improves by considering both spatial and temporal domain. The additional information provided by the enlarged NIR of ACO-PhD gives ACO-PhD a higher chance of making better decisions for traffic load-balancing. However, as we can see, the performance of NoP outperforms ACO-PhD under this traffic pattern. This is owing to the fact that NoP employs additional routing functions in order to eliminate the useless network information.

4.2.2 Transpose Traffic

Also as mentioned in [8], to stress the routing algorithm, transpose traffic is induced from matrix transpose or corner-turn operations. Because it concentrates load on individual source-destination pairs, transpose traffic stress the load balance of a routing algorithm. The results are shown in Fig. 15b. ACO outperforms RCA in this case. This phenomenon is caused by the traffic distribution, which is highly unbalanced under transpose traffic. With the pheromone indicating the historically less congested channels, ACO can avoid sending packets toward the center region and attain better performance. The performance of NoP still exceeds OBL, RCA, and ACO, but it is worth noticing that ACO-PhD

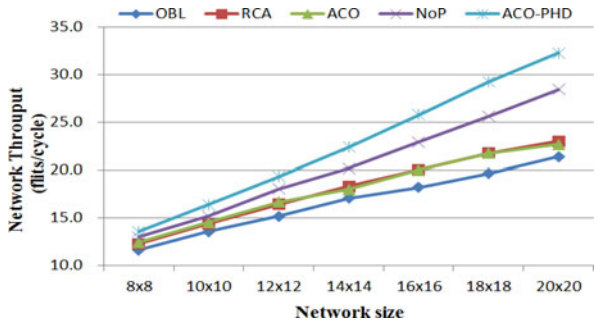


Fig. 16. The network throughput of different network sizes.

achieves the best performance among all under transpose traffic. This is because the pheromone can provide more advantage under highly unbalanced and regular traffic.

4.2.3 Hotspot Centred Traffic

The *hotspot* pattern [16] simulates that the presence of concentrated traffic on several tiles. The hotspot can be a caching process or frequent access on certain IPs. This pattern also stresses the load balancing ability of the routing algorithms. In Fig. 15c, four hotspot nodes are located at the centre of the mesh, that is nodes [(3,3),(3,4),(4,3),(4,4)] with 10 percent hotspot traffic. The results are close to each other in hotspot traffic. The performance of ACO-PhD is slightly better than NoP and is the highest among all routing algorithms.

4.3 Analysis on Performance Scalability

The performance scalability of routing algorithm is important in determining the performance of large-scale NoC systems. To investigate the effect of network scaling on the performance, we analyze the network throughput (NT) of different routing algorithms under different NoC sizes [36]. The experimental process changes the network sizes (i.e., these sizes are 8×8, 10×10, 12×12, 14×14, 16×16, 18×18, and 20×20) and records each NT, which is defined as

$$NT = \text{total received flits/total cycles.} \tag{7}$$

The experiments use transpose traffic to evaluate NT We show NT for OBL, RCA, ACO, NoP, and ACO-PhD in different NoC sizes. in Fig. 16, ACO-PhD has higher improvement especially for large network sizes. For instance, in a 20×20 mesh, improvements of ACO-PhD to its NIR subset OBL, RCA, ACO are 50.07, 40.28, and 42.25 percent in NT, respectively. ACO-PhD also outperforms NoP in NT by 13.5 percent. Notably, the NT of OBL, RCA and ACO gradually become saturated as the network size increases. The NT of ACO-PhD, in contrast, grows steadily with the increase of network size. This implies ACO-PhD can improve the scalability of performance. Namely, it has higher potential to estimate traffic-flow trends in a large NoC by properly integrating the spatial and temporal congestion information.

4.4 Experimental Results on Real-Traffic Data

To simulate a realistic traffic load in NoC system, we perform a complete network analysis, including the low-density parity-check (LDPC) codes and the MMS [30].

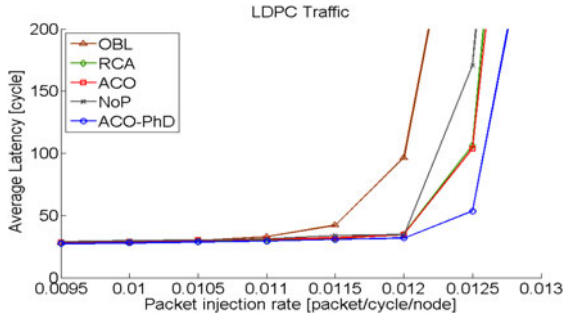


Fig. 17. Performances evaluation under LDPC traffic.

4.4.1 LDPC Traffic

LDPC code is one kind of linear block codes, first proposed in [32]. Due to the excellent error-correcting performance, the LDPC codes have been widely adopted by many advanced wire-line and wireless communication systems.

To increase the hardware implementation efficiency, a parallel LDPC decoding was proposed on an NoC-based multiprocessor system because of its regularity and scalability [33]. Given the LDPC design in [32] and the evaluation method in [34], we evaluate different routing schemes with the data flow of the (1,944, 972) LDPC code, which is defined in *IEEE 802.11n*. The mapping of bit node unit (BNU) and check node unit (CNU) onto processing elements can be done by

$$N_{map} = (BNU/CNU\ number) mod(N_{NoC}), \tag{8}$$

where N_{map} denotes the mapped node number and N_{NoC} denotes the total number of NoC nodes. We map the (1,944, 972) LDPC codes to a 16×16 mesh-based NoC system.

Fig. 17 shows the result of the LDPC traffic pattern. We observe that ACO-PhD outperforms the other routing algorithms. Then RCA and ACO have similar performance, and outperform OBL and NoP.

4.4.2 MMS Traffic

The multi-media system (MMS) pattern adopts the model from [18] and was discussed in [30]. It includes an H. 263 video codec and an MP3 audio codec. The application is partitioned into 40 distinct tasks and assigned on 25 IPs. The mapping of IPs into nodes of a 5×5 mesh-based NoC architecture has been obtained by using the method presented in [31]. As we observe from Fig. 18, the result shows that ACO-PhD outperforms other algorithms. Then RCA and NoP have similar performance, and outperform OBL and ACO.

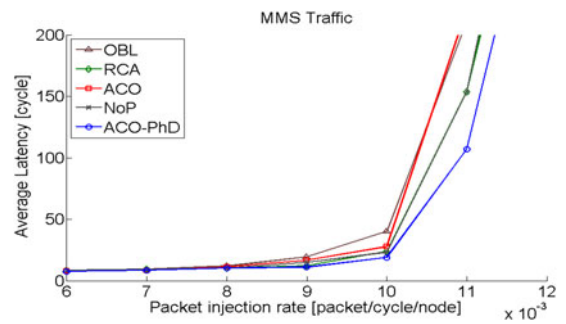


Fig. 18. Performances evaluation under MMS traffic.

TABLE 2
Saturation Throughput of Different Routing Algorithms (Flits/Cycle)

Selection Function	OBL	NoP	RCA	ACO	ACO-PhD
Uniform	12.35	13.20	12.62	12.63	12.95
Transpose	11.64	13.06	12.31	12.47	13.61
Hotspot	5.73	6.08	5.93	5.88	6.15
LDPC	23.76	24.58	24.68	24.68	25.19
MMS	1.70	1.76	1.84	1.80	1.90
Normalized Improvement (in percent)	0.00	6.43	4.71	4.36	9.42

In order to give a quantitative analysis on the network performance, the saturation throughput under various traffic patterns and the normalized improvement of each routing algorithms are tabulated in Table 2. As we can see, ACO-PhD acquires the highest the highest normalized improvement.

4.5 Performances of Different Parameter Settings

The statement in Section 3.5 can be verified by the following simulation results. Fig. 19 shows the performances of OBL, RCA, ACO, and ACO-PhD with different parameter settings under transpose traffic. The performances of OBL, RCA, ACO, and ACO-PhD are represented by the solid lines, while the dotted lines are the performances of ACO-PhD under different settings.

As we can see the ACO-PhD$_{OBL}$ has the same performance as OBL. The reason is that by having the same NIR, the selection function will receive the same network information. Then the selection function makes the same decision and thus provide the same network performance. The slight difference between the two lines is due to the randomly generated traffic, and it is verified that under exactly the same traffic pattern, two selection functions make exactly the same decision as declared. The same results also stand on the cases in RCA and ACO, making each pair of lines overlap with each other.

5 Hardware Implementation of ACO-PhD Router

There are several fundamental components in the basic router architecture shown in Fig. 20a. First of all, the packets

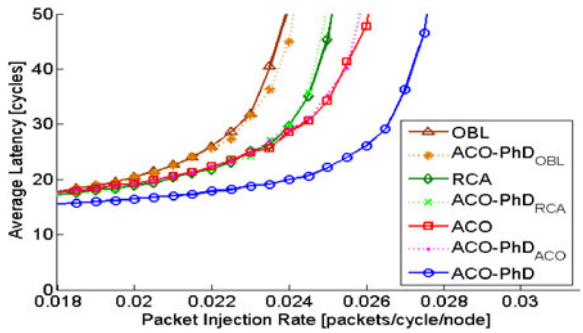


Fig. 19. Performance of OBL, RCA, ACO, and ACO-PhD with different parameter settings.

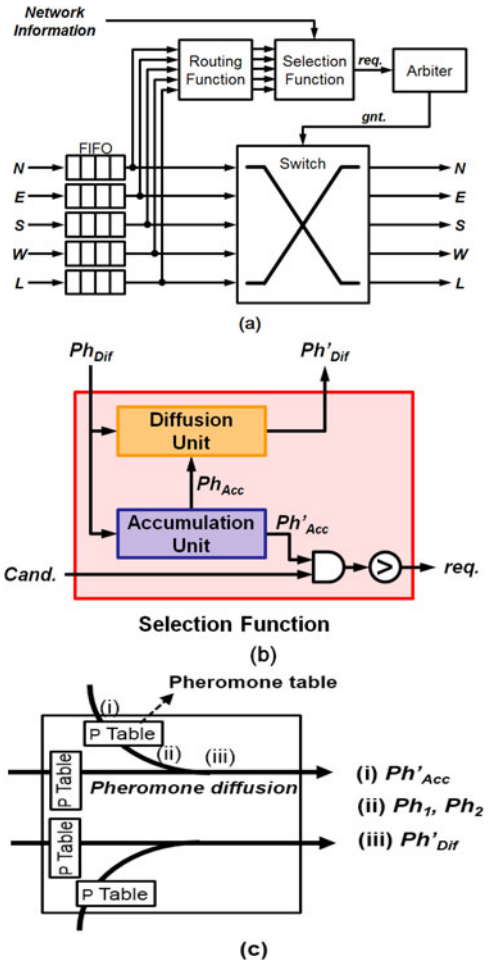


Fig. 20. (a) The router architecture. (b) The accumulation unit and diffusion unit for storing and propagating pheromone information in the selection function. (c) The quadrant-based propagation of the diffusive pheromone and the accumulation of the historical pheromone.

received from the five directions (*North*, *East*, *South*, *West*, and *Local*) are stored in the FIFOs. The routing function unit then evaluates the candidate channels and sends it to selection function unit. Base on the network information received or stored, selection function unit chooses a proper output channel and generate the request signal (abbreviated as *req.* in Fig. 20a). According to Round-Robin arbitration, the arbiter then generates the grant signal (abbreviated as *gnt.*), controlling the switch to transmit the packets through the corresponding channels. Fig. 20b is the framework located in the selection function block in Fig. 20a. The critical path of selection scheme is on the Accumulation unit and thus is similar to that of RACO.

Fig. 20c shows the process of quadrant-based pheromone diffusion. The downstream information are collected and propagated to the upstream routers. The extra wiring cost of ACO-PhD is of the same level with RCA Quadrant [17]. The wiring costs of Ph_{Dif} in each channel (i.e., North, East, West, and South) are computed as: two quadrants in a channel multiplied by 4-bits of Ph_{Dif} wordlength. Therefore, the sizes of *PhD_value_in* and *PhD_value_out* are both 8-bits. We use the input signal for computing the wiring overhead. We consider the settings of 128-bits flit size as discussed in [40], 2-bits parity check,

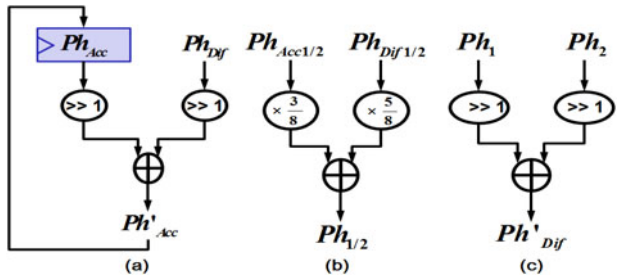


Fig. 21. Hardware implementation of (a) Ph'_{Acc}, (b) $Ph_{1/2}$, and (c) Ph'_{Dif}.

and 3-bits of free-slot information. The overhead is computed as: 8-bits divided by (128 + 2 + 3) = 133-bits, which is less than 7 percent of the total wiring cost.

The ant packets, as discussed in [36], only affect the update of pheromone table in the Accumulation Unit. The data packets and ant packets use the same data network. In addition, ant packet carries the same payload as a data packet. The only difference is an identifier in the *header flit*: When an ant packet header-flit is being routed, the selection process triggers the pheromone table update. In this work, we set all the packets to be ant packets for demonstrating the proposed network information region framework.

The hardware implementation of ACO-PhD can be separated into two parts, which are the implementation of Ph'_{Acc} and Ph'_{Dif}. The formula of Ph'_{Acc} is shown in (6) and the implementation of Ph'_{Acc} is shown in Fig. 21a. As we can see, Ph'_{Acc} simply adds up the previous Ph_{Acc} value and the incoming Ph_{Dif} value with a specific weighting α. The value of α in ACO-PhD is discussed in Section 3.3 and is set to 0.5, making the multiplication of α can be implemented by shifters. After evaluating Ph'_{Acc}, it is stored into the register at every clock cycle.

Ph'_{Dif} is the outwarding pheromone indicating the network status. Owing to the fact that the formula of Ph'_{Dif} is quite similar to the formula of Ph'_{Acc}, as shown in Eq. (5), the hardware implementation of Ph'_{Dif} is also similar to the one of Ph'_{Acc}. There are two differences between Ph'_{Acc} and Ph'_{Dif}. The first one is that Ph'_{Dif} contains two components, which are the pheromone from two directions, Ph_1 and Ph_2. The second one is that the PhD weighting β, which is discussed in Section 3.3 and is set to 0.375. Fig. 21b shows the formation of Ph_1 and Ph_2. Since the value of β is a constant and the denominator 8 is a power of 2, the multiplication of β and $(1-\beta)$ can be implemented by adders and shifters. After evaluating Ph_1 and Ph_2, Ph'_{Dif} simply adds up Ph_1 and Ph_2 with two shifters, as illustrated in Fig. 21c, and accumulates with local congestion information later for propagation.

5.1 Analysis of Hardware Overhead

Based on the router architecture in Fig. 20, routers with different routing algorithms are designed and implemented. Furthermore, each router is synthesized with *TSMC 90nm* technology. The synthesis results are tabulated in Table 3 .

As we can see, the router of OBL consumes the lowest area since it requires the least network information and

TABLE 3
Area Cost of Different Routing Algorithms

Synthesis Results (TSMC 90nm @ 360MHz)					
Selection Function	OBL	NoP	RCA	ACO	ACO-PhD
Area (k μm^2)	54.4	62.7	56.0	56.5	57.5
Overhead (in percent)	0.00	15.35	3.01	3.82	5.71

thus the least computation. In order to propagate the network information hop by hop, RCA needs to receive the incoming information and generate the outgoing information, and the hardware cost is thus increased. ACO requires an ant table to store the pheromone information accumulated over time, which becomes the main hardware overhead. As previously described, the routing resource in NoP is three times larger, and the cost on signal wiring is also the highest among all selection schemes. As a result, the hardware cost of NoP is comparatively high comparing to others. The hardware requirement of ACO-PhD includes the storage of Ph_{Acc} and the computation of Ph_{Dif}. Although the area cost of ACO-PhD is the second highest among all, the following area efficiency analysis shows that ACO-PhD can acquire the highest efficiency comparing to other related works.

5.2 Evaluation of Area Efficiency

In order to evaluate the effectiveness of each routing algorithm, the area efficiency [36] is calculated by dividing the saturation throughput by the corresponding total area of each routing algorithm (the throughput under uniform, transpose, and LDPC patterns are chosen for evaluation). The area efficiency is tabulated in Table 4. As it shows, RCA and ACO both acquire area efficiency improvements, implying that the ratio of improvement gained over overhead consumed is worthwhile. In contrast, although NoP attains the second high performance improvement, its area efficiency is much smaller than others. This is due to the high hardware cost caused by the tripled routing resource required. Due to the highest network improvement acquired and the moderate hardware cost required, ACO-PhD achieves the highest area efficiency among all routing algorithms.

5.3 Energy Consumption with Real-Traffic Data

To predict the realistic energy consumption of an NoC system, we perform a complete network analysis. We consider the congestion effects in draining a fixed workload (10 Mbytes) with two real applications, including the LDPC codes and the MMS [30], as discussed in Section 4.4.

TABLE 4
Area Efficiency of Different Routing Algorithms

Selection Function	OBL	NoP	RCA	ACO	ACO-PhD
Sat. Throughput ($Gbps/node$)	8.08	8.75	8.40	8.44	8.87
Total Area (k μm^2)	54.4	62.8	56.0	56.5	57.5
Area Efficiency ($Gbps/node/mm^2$)	148.6	139.3	149.9	149.4	154.3

TABLE 5
Total Energy Consumption to Drain 10M Bytes of Data for Parallel LDPC System and Multimedia System

Real Traffic Data	*pir*	Energy Consumption (mJ)				
		OBL	NoP	RCA	ACO	ACO-PhD
(1944, 972) LDPC codes	0.0095	6.15	6.21	6.20	6.21	6.15
	0.011	6.20	6.21	6.19	6.24	6.19
	0.0115	6.35	6.31	6.26	6.26	6.20
	0.0125	7.39	6.66	6.56	6.54	6.32
	0.0135	8.07	8.46	7.97	8.23	8.09
MMS	0.007	1.65	1.65	1.67	1.64	1.66
	0.008	1.64	1.65	1.64	1.66	1.65
	0.009	1.64	1.66	1.63	1.65	1.66
	0.010	1.67	1.65	1.67	1.65	1.64
	0.011	1.68	1.68	1.66	1.66	1.66

Table 5 shows the total energy consumption of five representative pir under the LDPC codes and MMS traffic. For example, at the lower pir of LDPC (i.e., 0.0095-0.0115), we find that the energy used by ACO-PhD is comparable to OBL, RCA and ACO. Because there is less congestion at these pir values, all of selection schemes can rapidly forward the packets to the destination nodes. Therefore, the amount of energy consumption is directly related to the complexity of routing computation.

On the other hand, at higher pir values of LDPC (*i.e.*, 0.0125-0.0135), the energy consumption of all selection schemes increases significantly. Due to the high probability of contention under heavy traffic workloads, many packets could be blocked. The routing computation is re-executed which would consumes additional energy. Meanwhile, the buffer to store these packets also requires extra energy.

Notably, we find out that a routing algorithm with better performance can balance its power dissipation overhead, which is consistent with the observation of [16]. The total energy consumption of ACO-PhD is in the same level as other routing algorithms. For example, at the saturation throughput of 0.0125 in LDPC, ACO-PhD scheme has an energy saving of 14.5 percent compared to OBL. Similarly, at the saturation throughput pir 0.01 of MMS, ACO-PhD scheme has an energy saving of 1.4 percent compared to OBL.

6 Conclusions and Future Work

We bring out the concept of *network information region*, which indicates the network information utilized by each routing algorithms. We use the NIR to analyze each related work and propose the *ACO-based adaptive routing with pheromone diffusion* (ACO-PhD) algorithm. We show that we can reconfigure the ACO-PhD algorithm to each routing algorithm in its NIR subsets by adjusting the parameter settings. This concept can help in developing more advanced reconfigurable schemes for different types of traffic or design constraints. The overall performance is evaluated and compared with other related works, showing that ACO-PhD can achieve the highest performance among all schemes on both saturation throughput and area efficiency.

The propagation-based mechanism has been widely discussed and analyzed. Ramanujam and Lin [15] and Ma et al. [27] focus on further improving the accuracy of RCA model. CATRA [28] is an extension of NoP by adopting more precise evaluation metric. GCA [37] proposed a light-weight adaptive routing algorithm by propagating precise congestion information through a separate side-band network. CARS [38] can effectively mitigate the congestion by prioritizing requests based on the congestion level information. DeBAR [39] proposed a series of design optimization on minimally buffered NoC routers. HARAQ algorithm [29] is an advanced method that considers the routing and selection function simultaneously. PCAR [41] proposed a selection strategy that simultaneously considers switch congestion and channel congestion. The switch-based information design of PCAR can be integrated with different kinds of channel congestion models. In network information aspect, future work can be directed towards the design of more advanced adaptive routing on NoC systems where the NIR integration can be applied to the propagation-based channel congestion and switch-congestion aware mechanism to improve performance.

ACKNOWLEDGMENTS

This work was supported by the National Science Council of Taiwan under Grants NSC 100-2221-E-002-091-MY3 and NSC 102-2220-E-002-001.

REFERENCES

[1] D. Hodges, H. Jackson, and R. Saleh, *Analysis and Design of Digital Integrated Circuits: In Deep Submicron Technology*. New York, NY, USA: McGraw-Hill, 2004.

[2] P. Magarshack and P. Paulin, "System-on-chip beyond the nanometer wall," *Proc. Des. Autom. Conf.*, Jun. 2003, pp. 419–424.

[3] J. Howard, S. Dighe, S. R. Vangal, G. Ruhl, N. Borkar, S. Jain, V. Erraguntla, M. Konow, M. Riepen, M. Gries, G. Droege, T. Lund-Larsen, S. Steibl, S. Borkar, V. K. De, and R. Van Der Wijngaart, "A 48-core IA-32 processor in 45 nm CMOS using on-die message-passing and DVFS for performance and power scaling," *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 173–183, Jan. 2011.

[4] W. J. Dally and B. Towles, "Route packets, not wires: On-chip interconnection networks," *Proc. ACM/IEEE Des. Autom. Conf.*, 2001, pp. 684–689.

[5] L. Benini and G. D. Micheli, "Network on chip: A new SoC paradigm for systems on chip design," *Proc. IEEE Des., Autom. Test Conf.*, 2002, pp. 418–419.

[6] Cloud Computing Intel Labs, (2009, Dec.). Single Chip Cloud Computer: Project. [Online]. Available: http://www.intel.com/content/www/us/en/research/intel-labs-single-chip-cloud-computer.html

[7] R. Marculescu, U. Y. Ogras, L.-S. Peh, N. E. Jerger, and Y. Hoskote, "Outstanding research problems in NoC design: System, microarchitecture," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 28, no. 1, pp. 3–21 Jan. 2009.

[8] W. J. Dally and B. Towles, *Principles and Practices of Interconnection Networks*. San Mateo, CA, USA: Morgan Kaufmann, 2004, pp. 159–244.

[9] T. M. Pinkston and S. Warnakulasuriya, "On deadlocks in interconnection networks," in *Proc. Int. Symp. Comput. Archit.*, Jun. 1997, pp. 38–49.

[10] C. J. Glass and L. M. Ni, "The turn model for adaptive routing," *J. ACM*, vol. 40, no. 1, pp. 874–902, Sep. 1994.

[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] J. C. Martnez, F. Silla, P. Lopez, and J. Duato, "On the influence of the selection function on the performance of networks of workstations," in *Proc. IEEE High Performance Comput. Conf.*, 2000, pp. 292–299.

[13] L. Schwiebert and R. Bell, "Performance tuning of adaptive wormhole routing through selection function choice," *IEEE Trans. Parallel Distrib. Comput.*, vol. 13, no. 7, pp. 1121–1141, Jul. 2002.

[14] U. Y. Ogras, J. Hu, and R. Marculescu, "Key research problems in NoC design: A holistic perspective," in *Proc. IEEE/ACM/IFIP Int. Conf. Hardware/Softw. Codesign Syst. Synth.*, 2005, pp. 69–74.

[15] R. S. Ramanujam and B. Lin, "Destination-based adaptive routing on 2D mesh networks," in *Proc. ACM/IEEE Symp. Comput. Archit.*, 2010, pp. 1–12.

[16] G. Ascia, V. Catania, M. Palesi, and D. Patti, "Implementation and analysis of a new selection strategy for adaptive routing in networks-on-chip," *IEEE Trans. Comput.*, vol. 57, no. 6, pp. 809–820, Jun. 2008.

[17] P. Gratz, B. Grot, S. W. Keckler, "Regional congestion awareness for load balance in networks-on-chip," in *Proc. IEEE Int. Symp. High Performance Comput. Archit.*, 2008, pp. 203–214.

[18] M. Daneshtalab and A. Sobhani, "NoC hot spot minimization using AntNet dynamic routing algorithm," in *Proc. IEEE App. Specific Syst., Archit. Processors Conf.*, 2006, pp. 33–38.

[19] H.-K. Hsin, E.-J. Chang, C.-H. Chao, and A.-Y. Wu, "Regional ACO-based routing for load-balancing in NoC systems," *Proc. IEEE Nature Biol. Inspired Comput. Conf.*, Dec. 2010, pp. 370–376.

[20] M. Dorigo , M. Birattari, and T. Sttzle, "Ant colony optimization," *IEEE Comput. Intell. Mag.*, vol. 1, no. 4, pp. 28–39, Nov. 2006.

[21] M. Dorigo, V. Maniezzo, and A. Colorni, "The ant system: Optimization by a colony of cooperating agents," *IEEE Trans. System, Man, Cybern.*, vol. 26, no. 2, pp. 29–41, Feb. 1996.

[22] M. Dorigo and L. M. Gambardella, "Ant colony system: A cooperative learning approach to the travelling salesman problem," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 53–66, Apr. 1997.

[23] K. M. Sim and W. H. Sun, "Ant colony optimization for routing and load-balancing: Survey and new directions," *IEEE Trans. Syst., Man Cybern.*, vol. 33, no. 5, pp. 560–572, Sep. 2003.

[24] (2008). Noxim: Network-on-Chip Simulator. [Online]. Available: http://sourceforge.net/projects/noxim

[25] L. Shang, L.-S. Peh, and N. K. Jha, "Powerherd: A distributed scheme for dynamically satisfying peak-power constraints in interconnection networks," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 25, no. 1, pp. 92–110, Jan. 2006.

[26] E. Shin, V. Mooney, and G. Riley, "Round-robin arbiter design and generation, " *Proc. IEEE Int. Symp. Syst. Synth.*, Oct. 2002, pp. 243–248.

[27] S. Ma, N. E. Jerger, and Z.-Y. Wang, "DBAR: An efficient routing algorithm to support multiple concurrent applications in networks-on-chip," in *Proc. Int. Symp. Comput. Archit*, 2011, pp. 413–424.

[28] M. Ebrahimi, M. Daneshtalab, P. Liljeberg, J. Plosila, and H. Tenhunen, "CATRA—Congestion aware trapezoid-based routing algorithm for on-chip networks, " in *Proc. Des., Autom. Test in Eur. Conf. Exhib.*, 2012, pp. 320–325.

[29] M. Ebrahimi, M. Daneshtalab, F. Farahnakian, J. Plosila, P. Liljeberg, M. Palesi, and H. Tenhunen, "HARAQ: Congestion-aware learning model for highly adaptive routing algorithm in on-chip networks," in *Proc. IEEE/ACM Int. Symp. Netw. Chip*, 2012, pp. 19–26.

[30] J. Hu and R. Marculescu, "Energy- and performance-aware mapping for regular NoC architectures," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 24, no. 4, pp. 551–562, Apr. 2005.

[31] G. Ascia, V. Catania, and M. Palesi, "Multi-objective mapping for mesh-based NoC architectures," in *Proc. Second IEEE/ACM/IFIP Intl Conf. Hardware/Softw. Codes. Syst. Synth.*, Sep. 2004, pp. 182–187.

[32] R. Gallager, "Low-density parity-check codes," *IRE Trans. Inf. Theory*, vol. 7, pp. 21–28, Jan. 1962.

[33] W. H. Hu, J. H. Bahn, and N. Bagherzadeh, "Parallel LDPC decoding on a network-on-chip based multiprocessor platform," in *Proc. Int.l Symp. Compt. Archit. High Performance Comput.*, Oct. 2009, pp. 35–40.

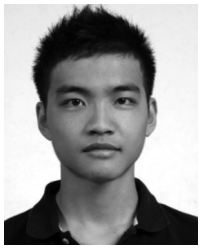
[34] K.-C. Chen, S.-Y. Lin, H.-S. Hung, A.-Y. Wu, "Topology-aware adaptive routing for non-stationary irregular mesh in throttled 3D NoC systems," *IEEE Trans. Parallel and Distrib. Syst.*, vol. 24, no. 10, pp. 2109-2120, Oct. 2013.

[35] G. Di Caro and M. Dorigo, "AntNet: Distributed stigmergetic control for communications networks," *J. Artif. Intell. Res.*, vol. 9, pp. 317–365, 1998.

[36] E.-J. Chang, H.-K. Hsin, C.-H. Chao, S.-Y. Lin, and A.-Y. Wu, "Regional ACO-based cascaded adaptive routing for load balancing in mesh-based network-on-chip systems," *IEEE Trans. Comput.*, 2014, doi 10.1109/TC.2013.2296032.
[37] M. Ramakrishna, P. V. Gratz, and A. Sprintson, "GCA: Global congestion awareness for load balance in networks-on-chip," in *Proc. Int. Symp. Netw. Chip*, 2013, pp. 1–8.
[38] M. Daneshtalab, N. Ebrahimi, J. Plosila, and H. Tenhunen, "CARS: Congestion-aware request scheduler for network interfaces in NoC-based manycore systems," in *Proc. Des., Autom. Test Eur. Conf. Exhib.*, 2013, pp. 1048–1051.
[39] J. Jose, B. Nayak, K. Kumar, and M. Mutyam, " DeBAR: Deflection based adaptive router with minimal buffering," in *Proc. Des., Autom. Test Eur. Conf. Exhib.*, 2013, pp. 1583–1588.
[40] J. Lee, C. Nicopoulos, S. J. Park, M. Swaminathan, and J. Kim, "Do we need wide flits in networks-on-chip?" in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI*, 2013, pp. 2–7.
[41] E.-J. Chang, H.-K. Hsin, S.-Y. Lin, and A.-Y. (Andy) Wu, "Path-congestion-aware adaptive routing with a contention prediction scheme for network-on-chip systems," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 33, no. 1, pp. 113–126, Jan. 2014.

Hsien-Kai Hsin received the BS degree in electronic engineering from the National Taiwan University, Taiwan,in 2009, and the PhD degree from the Graduate Institution of Electronic Engineering, National Taiwan University, 2014. His current research interests include biologically inspired systems and reliable interconnects. In particular, his research focuses on swarm intelligence design methodologies for multicore SoCs, with special interest on network-on-chip communication architecture.

En-Jui Chang received the BS degree in electrical engineering from National Central University, Zhongli, Taiwan, in 2008, and the PhD degree from the Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan, 2014. His research interests include network-on-chip algorithms/architectures, bioinspired algorithms/architectures, fault-tolerance algorithms/architectures, and VLSI architectures for DSP in communication systems.



Kuan-Yu Su received the BS degree in electronic engineering from the National Taiwan University, Taipei, Taiwan, in 2010, and the MS degree from the Graduate Institute of Electronics Engineering, National Taiwan University in 2012. His research interests include network-on-chip algorithms/architectures, bioinspired algorithms/architectures, and adaptive routing algorithms/architectures in VLSI systems.

An-Yeu (Andy) Wu (S'91-M'96-SM'12) received the BS degree from National Taiwan University in 1987, and the MS and PhD degrees from the University of Maryland, College Park, in 1992 and 1995, respectively, all in electrical engineering. In 2000, he joined the faculty of the Department of Electrical Engineering and the Graduate Institute of Electronics Engineering, National Taiwan University, where he is currently a Professor. His research interests include low-power/ high-performance VLSI architectures for DSP and communication applications, adaptive/multi-rate signal processing, reconfigurable broadband access systems and architectures, and system-on-chip/network-on-chip platform for software/hardware co-design.