Overview of High-Efficiency Ant Colony Optimization (ACO)-based Adaptive Routings for Traffic Balancing in Network-on-Chip Systems

En-Jui Chang and An-Yeu (Andy) Wu, Fellow, IEEE
Graduate Institute of Electronics Engineering and the Department of Electrical Engineering,
National Taiwan University, Taipei 106, Taiwan
Email: enjui@access.ee.ntu.edu.tw; andyw@ntu.edu.tw

Abstract—The regular topology of mesh-based Network-on-Chip (NoC) provides flexible and scalable architecture for chip multiprocessor (CMP) systems. However, as the complexity of the network increases, the traffic congestion problems become major performance bottlenecks. An effective adaptive routing algorithm can help minimize path congestion through load balancing. However, conventional adaptive routing schemes only use current channel-based information to detect the congestion status. This information has only showing the real congestion status under time-variant traffic patterns. To predict temporal network congestion, Ant Colony Optimization (ACO) based routing was proposed to identify the near-future non-congested path to a desired target according to historical network information. To design a high-efficiency ACO-based routing for traffic balancing in resource-limited NoCs, two major design issues should be considered, including 1) the selection efficiency and 2) the cost efficiency. In this paper, the design challenges and concepts of ACO-based adaptive routings are introduced. Besides, the representative related works are reviewed and summarized. Finally, we conclude the paper and point out the future works of ACO-based routing algorithms.

1. Introduction

As semiconductor technology continues to advance, increasing complexity and interconnection delay are becoming limiting factors in system-on-chip (SoC) designs. To increase the efficiency of interconnections and meet data transfer requirements, network-on-chip (NoC) systems have proven to be a flexible, scalable, and reusable solution for chip multiprocessor (CMP) systems [1]. To achieve a high throughput rate, the packet-switched NoC multiplexes packets on channels and shares network resources among these packet flows. However, the packet congestion problem in channels results in unpredictable delays for each packet flow. As the system size increases, the network traffic load tends to become unbalanced with various applications, which not only causes network congestion but also dissipates additional energy. Therefore, to overcome the problem of traffic congestion, packet routing is a critical design challenge for high-performance NoCs [3].

To achieve balanced traffic of NoCs, a bio-inspired approach, Ant Colony Optimization (ACO) based

This research was supported in part by the Ministry of Science and Technology of Taiwan (MOST 105-2218-E-002-024, MOST 106-2633-E-002-001), National Taiwan University (NTU - 106R104045), and Intel Corporation.
2. Design Objectives for Traffic Congestion Problems

As mentioned, the adopted network information for evaluating the channel congestion status greatly influences the system performance [3]. Basically, the network congestion can be classified into two groups: 1) spatial network congestion, and 2) temporal network congestion.

2.1 Spatial Network Congestion

The spatial network congestion is mainly caused by the blocking condition in forwarding buffer. Because of the limited input buffer size, the channel soon runs out of buffer space. Input buffers cannot accommodate newly arrived packets. Notably, due to the backpressure effect of the link-level flow control. Therefore, path congestion can start to build and spread from a congested switch to source nodes, which grows into a congestion tree. It severely degrades the overall system performance, especially in real-time applications with strict latency requirements.

2.2 Temporal Network Congestion

On the other hand, on-chip network traffic has shown self-similar characteristics for many applications, which have been observed in the bursty traffic between on-chip modules, especially for multimedia traffic. Some data are considered to be self-similar in time, if the time series preserves its temporal properties with respect to scaling in time. That is, self-similarity can characterize burstiness and exhibit a long-range dependence (LRD) property of traffic in the temporal sense. Notably, bursty traffic can easily exhaust the bandwidth of channels, and create a congested node in the network.

2.3 Design Objectives

To achieve a good trade-off between performance and cost, the main two design goals of ACO-based adaptive routings include the following:

i) Selection efficiency: The conventional ACO-based selection only uses the historical traffic information. While additional temporal and spatial information provides better approximation of network status for global load balancing. Therefore, to improve the system performance, several spatial-temporal enhancements of ACO-based adaptive routings [4], [5] for traffic balancing were proposed to extend spatial range of congestion information and help to capture hidden-state dependencies of upcoming congestion status.

ii) Cost efficiency: If we directly apply ACO to NoC systems, the implementation cost of ACO is excessively high. To overcome this problem, the ACO-based adaptive routing must be reformulated while considering both router cost and NoC efficiency. In considering the NoC topology, router dependency, and pheromone characteristics, some cost-efficient ACO-based adaptive routings with a regional routing table was proposed [2], [6].

Then, we introduce the basic ACO-based routings firstly in the next section. Besides, in terms of these two design objectives, several performance-enhanced and low-cost ACO-based routings are reviewed.

3. Basic ACO-based Routing Algorithms

Generally, ACO-based adaptive routing can be defined as adaptive routing with an ACO-based selection function. This selection function has two working phases: the training phase and the normal phase. These two phases take turns by periodically sending two packet types—an ant packet and data packet.

3.1 The Training Phase and the Normal Phase

During the training phase, routing uses reactive ant packets to update the new pheromone to the routing table of currently used paths. Also, it tries to find new and uncongested paths. Besides, ACO-based routing uses the ant packet not only to discover paths but also to transfer payload. Notably, the ant updating frequency needs to faithfully keep track of the status of network, such as traffic and topology, changes. Therefore, to acquire instantaneous information, the generation rate of forward ants is proportional to the generation rate of data packets at source nodes. During the normal phase, routing uses a data packet to transfer payload according to the constructed routing table, which must be constructed to store pheromones in each router, as shown in Fig. 2.

![Fig. 2. The routing table of ACO-based routing.](image)

3.2 The Flow of Training Phase

Fig. 3 shows the flow of training phase. Two stages are used to train the table: 1) detection by forward ants; and, 2) notification by backward ants. First, the source node generates the forward ant packet to the network. This forward ant can be sent to the next router stochastically according to the per-destination pheromone of candidate channels. The pheromone, which is transition probability, is derived by the state transition rule, which is shown as

\[
P(j,d) = \frac{P(j,d) + \alpha L_j}{1 + \alpha |V_j| - 1}, \quad j \in \{\text{north}, \text{east}, \text{south}, \text{west}\}.
\]  
(1)
transition and local pheromone updating, can be processed repeatedly until the ant packet reaches its destination.

During the second stage, when forward ants are in the destination router, this router grades the path that was bypassed according to total queue latency, which is saved in forward ants along their paths. The destination router then generates a backward ant to the network. The backward ant retraces the paths bypassed by the forward ant to send routers congestion information. Therefore, the backward ant must update the pheromone to the routing table by

\[
P(j,d) = \begin{cases} 
P(j,d) + \gamma (1 - P(j,d)), & j \in \text{selected channel} \\
(P(j,d) - \gamma P(j,d)), & j \notin \text{selected channel}
\end{cases}
\]

(2)

Specifically, the pheromone of selected channel \( j \) in column \( d \) is increased and the other pheromones in column \( d \) are decreased. \( r \) is the reinforcement factor, a number in the range of 0–1, indicating congestion and underlying network dynamics. Therefore, non-local congestion information can be detected and propagated using this method. In summary, due to environmental exploration in the training phase, the ACO-based selection function has the high potential to estimate traffic-flow trends and then distribute traffic to uncongested routers.

### 4. Review of Congestion-Aware and High-Efficiency ACO-based Routing Algorithms

#### 4.1 Selection-Efficiency ACO-based Routing

**i) Temporal Enhancement: Multiple-Pheromone ACO-based (MP-ACO) selection** [4]

In Fig. 4(a), ant system with single pheromone can determine whether a channel is better than another. However, a traffic burst still reduces the system throughput. If pheromone is designed as in Fig. 4(b), multiple pheromones divide the system into more states and provide extra information to reduce traffic bursts. MP-ACO selection scheme was designed with adopting the Exponential Moving Average [4], and implement a system with two pheromones in each channel. The ants lay pheromones with different evaporation rates to capture the buffer state dependency and to be aware of upcoming congestion. With MP-ACO selection, we can better determine the channel congestion states.

**ii) Spatial Enhancement: Regional-Aware ACO-based (RA-ACO) selection** [4][5]

ACO-based selection synthesizes spatial and historical network congestion information. With the accumulation of local model information (i.e., output buffer length) in Fig. 5(a), the pheromone represents the historical buffer status. To further enhance system performance, the precise regional-aware congestion model Neighbor-on-Path (NoP) was integrated with the ACO historical model [5], as shown in Fig. 5(b). The proposed RA-ACO selection acquires information, including free slots, reservation status, and admissible

---

**TABLE 1 NOTATIONS OF ACO-BASED ADAPTIVE ROUTING**

<table>
<thead>
<tr>
<th>Parameters of ACO-based Adaptive Routing</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( j )</td>
<td>Channel index (North, East, South, and West)</td>
</tr>
<tr>
<td>( d )</td>
<td>Destination index</td>
</tr>
<tr>
<td>( k )</td>
<td>Router index</td>
</tr>
<tr>
<td>( P(j,d) )</td>
<td>Probability of selecting channel ( j ) for destination ( d )</td>
</tr>
<tr>
<td>( L_j )</td>
<td>Current spatial information of channel ( j )</td>
</tr>
<tr>
<td>( N_k )</td>
<td>The number of channels in router ( k )</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>Weighing coefficient for ( L_j ) and ( P(j,d) ), which is set to a number in the range of 0–infinity</td>
</tr>
<tr>
<td>( r )</td>
<td>Reinforcement factor defines the goodness of the path</td>
</tr>
</tbody>
</table>

---

![Fig. 3](image1.png)  
**Fig. 3.** The flowchart of ACO-based adaptive routing

![Fig. 4](image2.png)  
**Fig. 4.** Acquiring more detailed network information with multiple pheromones.

![Fig. 5](image3.png)  
**Fig. 5.** The information acquiring of (a) ACO with local model and (b) RA-ACO with NoP regional-aware model.
4.3 Development of ACO with More Advanced Features

The multiple-objective optimization is becoming more important because the on-chip traffic is subjected to the application mapping, channel bandwidth, scheduling, topology, and quality of service issues. The pheromone table is suitable for integrating multiple heuristic constraints. The possible researches can be: 1) Combined heuristic function: to build heuristic function considering bandwidth/latency/scheduling constraints, budgets and overall performance simultaneously and 2) ACO-based routing with multiple heuristic functions: The pheromone table records different heuristic constraints and weighting function. The routing algorithm then considers these factors before the selection process.

5. Conclusions

This work summarizes the ACO-based adaptive routing in NoCs. The proposed selection-efficient and cost-efficient ACO-based adaptive routings were appropriately transformed to fit NoC systems. These techniques not only reduce implementation cost of a routing table, but also explore the temporal and spatial NoC properties to balance network load. The main purposes of these works are to utilize the temporal variation of pheromone and spatial coverage of pheromone to sense and predict network congestion, which makes ACO-based routing more feasible for the implementation of NoCs.

References