

# Delay Analysis and Buffer Sizing for Priority-Aware Networks-on-Chip (NoC)

Baoliang Li, Zeljko Zilic, Wenhua Dou,

**Abstract**—Among all the implementation alternatives of Network-on-Chip (NoC), priority-aware wormhole-switched NoC promises both the worst-case and average-case performance requirements of on-chip communication. The worst-case end-to-end delay and buffer requirement analysis are especially important for the development of real-time applications on this platform. In this paper, we first build a Real-Time Calculus (RTC) based performance model for the priority-aware wormhole-switched NoC. Then, we propose an end-to-end delay analysis algorithm and a buffer sizing algorithm based on this model. The latency analysis algorithm can give much tighter delay bound than the deterministic network calculus based method, because it takes the maximum service capability and minimum arrival traffic into consideration. The buffer sizing algorithm tries to reduce the buffer space required for each flow without violating the deadline constraint, which improves the backlog bound obtained by link-level buffer-space analysis method. Our algorithms are topology-independent, taking the architecture parameters and the flow specifications as input, they can give the end-to-end delay bound and buffer requirement for each traffic flow. Our model enables the fast performance evaluation and buffer allocation of priority-aware wormhole-switched NoC, which can be used for application mapping, routing selection and power reduction, etc. Experimental results demonstrate the effectiveness and tightness of our model. In addition, a further comparison with other theoretical models also indicates that our method outperforms existing methods while the tightness of delay and backlog bounds are considered.

**Keywords**—Networks-on-Chip (NoC), priority-aware, real-time calculus, delay bound, buffer sizing

## I. INTRODUCTION

The conventional on-chip interconnection paradigms, e.g. bus, ring and point-to-point links, are not able to meet the strict and complex communication requirements of modern large-scale Chip-MultiProcessors (CMPs) and System-on-Chip (SoC). As an alternative, Networks-on-Chip (NoC) is proposed to provide better scalability and higher power efficiency. As a key component of CMPs and SoC, NoC must be well designed to meet the rigorous requirements on performance, power and implementation cost. Although various proposals have emerged, each focusing on improving different performance metrics of on-chip interconnects, e.g. end-to-end latency, throughput and power, most of the existing researches on NoC are focusing on the improvement of average performance, and simulation is the most widely used performance evaluation method. However, there are a variety of on-chip applications,

which are sensitive to the worst-case or real-time communication performance of NoC, e.g. cache coherent protocol [1] and multimedia application [2]. Designing the on-chip real-time communication infrastructure for these applications and analyzing its feasibility is a major challenge for researchers.

To meet the rigorous Quality-of-Service (QoS) requirement, various special hardware implementations have been proposed, e.g. Time-Division Multiplexing-Access (TDMA) [3], circuit-switch [4] and time-triggered switch [5]. Although providing strict real-time communication guarantees, the average performance and resource utilization of these proposals are very poor. In contrast, wormhole-switched NoC is widely used in on-chip network due to its simplicity and high-efficiency. Thus, providing real-time communication support on the conventional wormhole-switched NoC to meet both average-case and worst-case communication requirements is the most promising solution. To achieve this goal, a special scheduling policy (e.g. DifServ [6] or priority-aware implementation [7], [8], [9]) or flow control mechanism (e.g. [10], [11]) must be integrated into the conventional wormhole-switched NoC. For all these real-time communication proposals based on wormhole-switched NoC, a key step before their adoption as the platform of real-time communication is the analysis of the worst-case communication delay for all the real-time flows to guarantee satisfaction of the deadline constraint. An effective buffer analysis approach is also needed to optimize the buffer allocation under the real-time constraint, since the on-chip buffer usually contributes to a significant portion of the entire router's power and area [12], [13].

An accurate worst-case delay analysis is crucial for the application of wormhole-switched NoC in real-time communication, since an overoptimistic estimation will lead to the violation of the deadline, while an overly pessimistic estimation will make the utilization of on-chip resource very low. The conventional simulation-based method is not appropriate for this purpose, because the worst-case scenarios are hard to be captured by simulation. As an alternative, the analytical methods can establish the relationship between performance metrics and design parameters, and give the worst-case performance immediately. A lot of previous research [7], [14], [15], [16], [11], [8], [16] has focused on the analysis of worst-case delay bound for the priority-aware wormhole-switched networks. Among all these analytical methods, the Link-Level Analysis (LLA) [14] and Deterministic Network Calculus [15] based model outperform the others when the tightness of performance bound is considered. The LLA assumes that the traffic flows are periodic, and the buffer size of wormhole-switched NoC is sufficiently large to eliminate the influence

Baoliang Li and Wenhua Dou are with the College of Computer Science, National University of Defense Technology, Changsha 410073, P.R. China

Zeljko Zilic are with Department of Electrical & Computer Engineering, McGill University, Montreal H3A-2A7, Quebec, Canada

Manuscript received XX XX, 2014; revised XX XX, 2014.

of flow control on the delay bound. The DNC based method [15] overcomes these two limitations by utilizing the advanced operators and properties of the DNC theory. However, the delay bound obtained by the DNC method [15] can be further improved if we take the maximum service curve of routers and the minimum arrival curve of traffic into consideration. These two curves can be utilized to improve the output arrival curves of high-priority flows and the leftover service curves for the low-priority flows. The improved leftover service curves further lead to tighter delay bounds for the low-priority flows. We will further explain the reason and demonstrate the improvement in Section V.

Motivated by this observation, we construct a novel performance model for the priority-aware wormhole-switched NoC with credit-based flow control, and propose an delay analysis algorithm and a buffer sizing algorithm based on this model. The theoretical framework of our performance model is the Real-Time Calculus (RTC) theory [17], which is originally used for the real-time analysis of task scheduling. The main contribution of this paper is two-fold: (1) We propose an end-to-end delay analysis algorithm for the priority-aware wormhole-switched NoC based on our performance model. The delay bound obtained by our algorithm is much tighter than the DNC method [15]. The output of this algorithm can be used as the reference of IP core mapping, task mapping, routing selection, or NoC parameters configuration, etc. (2) Our buffer sizing algorithm considers the impact of flow control on the delay bound, and only allocates just enough buffer at each router for the real-time flows to meet their deadline. When applied to guide the buffer allocation of priority-aware wormhole-switched NoC, our algorithm significantly outperforms the previously proposed LLA-based buffer optimization method [18], because the buffer bound determined in [18] is the minimum buffer size required to prevent back-pressure caused by flow control. Our buffer sizing method can be used to minimize the power consumption and chip area for the application-specific NoC.

The rest of this paper is organized as follows: we present the existing real-time communication proposals and its related performance analysis methods in Section II. In Section III, the basic assumptions on priority-aware wormhole-switched NoC and a brief introduction to the RTC theory are presented. The detailed modeling process is presented in Section IV, where we also propose the end-to-end delay analysis algorithm and buffer sizing algorithm. We present the experimental results and comparison with other analytical methods in Section V. Finally, we summarize our paper in Section VI.

## II. RELATED WORK

Since introduced in 2001 [19], various NoC proposals have emerged to meet different on-chip communication requirements. The main requirements posed to NoC by on-chip applications are latency and bandwidth. To meet these demands, NoC is designed to be either best-effort or guaranteed-service, depending on the hardware cost and application requirements. Best-effort NoC can make better use of the on-chip shared resource, but it does not necessarily provide any performance

guarantee for the applications. To provide the guaranteed services for different applications, a simple and effective solution is classifying these applications into several service classes, each with different priorities, and the network provides services according to the priority of each class. Representative implementations of this idea include QNoC [20], fixed-priority NoC [7] and *Æthereal* [3] etc. The performance evaluation method for both best-effort and guaranteed service NoC include the average-case analysis and worst-case analysis. For the average-case analysis, simulation- and probability-based methods hold the dominant position for both of these two categories. However, for the worst-case analysis, simulation is not competent due to the difficulty in covering all the corner cases. The analytical worst-case analysis of these two categories is also slightly different.

Synchronous Data Flow (SDF) graph [21] and DNC [22] have been presented to model the worst-case performance bounds of best-effort NoC. The former method assumes the traffic flow to be periodical, and the latter one eliminates this constraint to allow the traffic to be arbitrary patterns. In [22], the authors build an analytical performance model with DNC taking the various contentions and flow control into consideration. This result is further extended in [23], where the traffic splitter is proposed to support the multi-path routing policies. Another method is presented in [24] to compute the worst-case delay for conventional wormhole-switched network, and a real-time Wormhole Channel Feasibility Checking (WCFC) algorithm is proposed. This research is further extended to calculate the bandwidth and delay bounds in [25], and used for topology synthesis of best-effort NoC in [26].

The worst-case delay bound for the priority-aware wormhole-switched networks has been extensively studied. In [16], the contention tree model was proposed to analyze the feasibility of real-time traffic delivered by priority-aware wormhole-switched NoC. It improves the previous results, e.g. lumped link model [11] and dependency graph model [8], by allowing the concurrent link usage. The Flow-Level Analysis (FLA) proposed in [7] improves the results obtained by contention tree model [16], lumped link model [11] and dependency graph model [8]. A comprehensive comparison between FLA and the other method can be found in [27], in which several defects of the previous method are illustrated and the advantages of FLA are highlighted. The LLA [14] improves the FLA by treating each link segment separately. Two buffer sizing methods based on FLA and LLA, i.e. Flow-Level Buffer-space Analysis (FLBA) and Link-Level Buffer-space Analysis (LLBA), are proposed in [18] to estimate the buffer size of priority-aware wormhole-switched NoC. Whereas, both FLA and LLA assume the traffic arrives periodically and the router has sufficiently large buffer size, which is a significant simplification to the realistic traffic pattern and router implementation. In addition, the FLBA and LLBA can only compute the minimum buffer size at each router which does not trigger the flow control. We can further reduce the buffer size as long as the deadline constraint is not being violated.

On the other hand, although the DNC based performance model for best-effort NoC proposed in [22] can also be

applied to the analysis of priority-aware wormhole-switched NoC, the obtained performance bounds are very conservative, especially for the high-priority flows. This is because it does not take the priority into consideration. To overcome this limitation, a revised DNC performance model was proposed to analyze the worst-case delay of priority-aware wormhole-switched NoC in [15]. But we found that the DNC method in [15] can be further improved if we take the maximum service curve of each router and minimum arrival curve of each flow into consideration. Motivated by this observation, we adopt the RTC theory [17] to build the worst-case performance model for the priority-aware wormhole-switched NoC. Real-time calculus extends the theory of DNC [28] by integrating the minimum arrival curve and maximum service curve to characterize more detailed information about the traffic and service processes. Due to its high accuracy, RTC theory has been widely used in the modeling and analysis of network processor [29], CAN [30], FlexRay [31] and DSP systems [32], etc. To ease the application of RTC, a real-time calculus toolbox has been implemented in [33] to support the numerical calculation.

### III. PRELIMINARIES

#### A. Basic Assumptions

In this paper, the entire priority-aware NoC is represented as a directional network topology graph  $G : V \times E$ , where  $V$  and  $E$  represent the set of routers and links respectively. Each link  $e_{i,j} \in E$  corresponds to a physical channel connecting the two routers  $R_i$  and  $R_j$ . A flow is a sequence of packets with the same transmission path, source address and destination address. Packet of different flows generated by a Intellectual Property (IP) core are buffered at different queues within the corresponding Network Interface (NI). Each packet is comprised of one head flit, one tail flit and several body flits. The path of a flow  $f_i$  traversed is defined as a router chain starting from the injection router (denoted as  $start_i$ ) and ending at the ejection router (denoted as  $end_i$ ). The set of all the flows in the network is denoted as  $\mathcal{F}$ , and each flow  $f_i \in \mathcal{F}$  has a fixed-priority  $P_i$  and deadline  $D_i$ . The set of routers along the path of  $f_i$  is denoted as  $\mathcal{R}_i$ , and the set of links a flow  $f_i$  traversed is denoted as  $\Gamma_i$ . There exists contention between flow  $f_i$  and  $f_j$ , if and only if  $\Gamma_i \cap \Gamma_j \neq \emptyset$ . For all the router  $R_j$  along the path of flow  $f_i$ , denote the set of contending flows at  $R_j$  sharing the same priority with  $f_i$  as  $\Theta_{R_j, f_i}$ , the set of contending flows at  $R_j$  with lower priorities as  $\Omega_{R_j, f_i}$ , and the buffer size reserved at  $R_j$  for  $f_i$  as  $B_{R_j, f_i}$ .

The router we considered is the priority-aware wormhole-switched router proposed in [7] and further discussed in [9][11][14]. Each router has the same number of input and output port, and each input port has sufficient number of FIFO buffer, i.e. Virtual Channel (VC), to accommodate all the incoming packets of different priority levels. The allocation of VC is determined by the VC allocator. The buffer depth of each VC is finite, and the credit-based flow control [34] is adopted between adjacent routers to prevent buffer overflow. To ensure the predictable transmission delay, we assume that, a deterministic routine computation module is used to determine

the output port of each packet. The crossbar is utilized to switch traffic from input ports to the output ports, and the switch operation is determined by the switch allocator. The switch allocator is priority-aware, if multiple flits from different input ports or different VCs of the same input port contend for the same output port, it will only grant the flit with highest priority. Flits from a lower priority can transmit a flit, if and only if there are no flits from higher priority in the input buffer or the flits with higher priority are self-blocked due to the insufficiency of VC buffer at the downstream router.

The micro-architecture of the priority-aware router considered in this paper has standard pipeline stages, i.e. Buffer-Write (BW), Route Computation (RC), VC Allocation (VA), Switch Allocation (SA), Switch Traversal (ST) and Link Traversal (LT), as shown in Fig. 1b. Each head flit should go through all these stages to determine the path and reserve a VC for the following non-head flits. Non-head flits skip the RC and VA stages since the routine and VC have been determined by head flit. Router resource and control information reserved for a packet will be released only after the tail flit of this packet has been departed from the router. An additional priority field in the head flit is required for the routers to schedule multiple contending flows according to their priority. For the detailed description about the implementation and functionality of these pipeline stages, please refer to [35]. Although we focus on the standard router, our method can be easily modified to support other router micro-architecture, e.g. single-cycle router [9][7][11][14] and speculation-based router [35]. We will demonstrate the adoption of our model in a single-cycle router in subsection V-A. To simplify our analysis, we also assume that the entire chip is synchronous, with clock frequency  $f$  and period  $T$ . Our method can also be applied to analyze Global Asynchronous Local Synchronous (GALS) NoC with little modification, because the routers located in different voltage-frequency islands can be synchronized with a half cycle synchronizer [36], corresponding to a fixed-latency element in DNC theory [28].

Our performance model is topology-independent, but to demonstrate the basic idea, we take the mesh topology shown in Fig. 1a as an example throughout this paper. Routers in the mesh topology have at most five input/output ports, corresponding to the four cardinal directions (West, East, North and South) and the Local IP core. There are four traffic flows in Fig. 1a, i.e.  $f_1$ ,  $f_2$ ,  $f_3$  and  $f_4$ . We must emphasize that, although there are only four flows in the network, it is sufficient to demonstrate the idea of our method, and our method can handle more traffic flows efficiently. Our method extends the existing methods in [14][15] to allow multiple flows to share the same priority. Flits of different flows sharing the same priority are served in round-robin order when they are designated the same output port. Since the minimum transmission unit in the priority-aware wormhole-switched NoC is flit and a high-priority flit can preempt the transmission of a low-priority flit, the NoC architecture considered in this paper is flit-level preemptive [24].

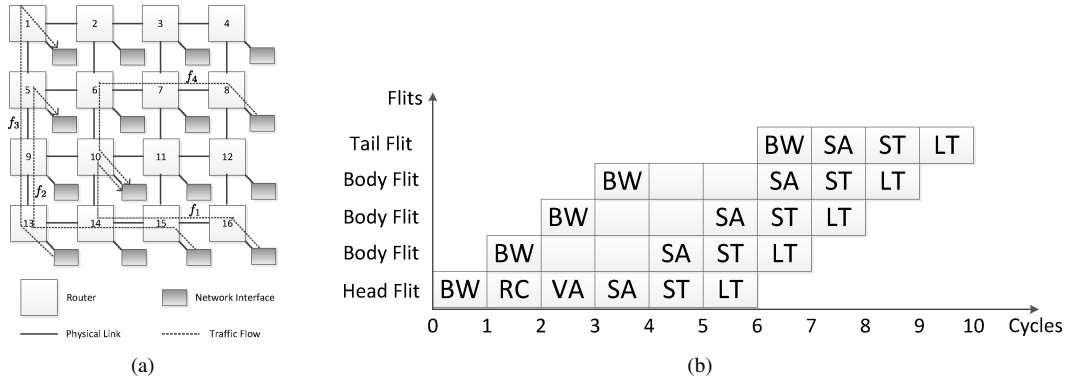


Fig. 1: Priority-aware wormhole-switched NoC. (a) Mesh topology with four real-time traffic flows. (b) Time-line graph of a packet going through the standard router pipeline. The delayed tail flit enters the ST stage immediately after it is written into the dedicated buffer.

### B. Introduction to Real-Time Calculus

Real-time calculus [17] is the theoretic extension of the DNC theory [28], by adding the upper service curve and lower arrival curve to describe the maximum service capability of a system and the minimum arrival rate of an event stream. It is the mathematical basis of the Modular Performance Analysis (MPA) [37] technique used for real-time task scheduling. Due to the space limitation, we only present the definitions of the RTC arrival curve and service curve in this subsection. For more details about this theory, please refer to [17].

**Definition 1 (Real-Time Arrival Curve [17]):** Let  $R[s, t]$  denote the number of events that arrived within the time interval  $[s, t]$ . The lower and upper bounds on  $R[s, t]$  are called the lower arrival curve  $\alpha^l$  and upper arrival curve  $\alpha^u$  which satisfy

$$\alpha^l(t - s) \leq R[s, t] \leq \alpha^u(t - s), \forall s < t$$

where  $\alpha^l(0) = \alpha^u(0) = 0$ . The RTC arrival curve for an event stream is denoted as  $\langle \alpha^l, \alpha^u \rangle$  for short.

**Definition 2 (Real-Time Service Curve [17]):** Let  $S[s, t]$  denote the total number of events that can be processed by the system in the time interval  $[s, t]$ . The lower and upper bounds on  $S[s, t]$  are called the lower service curve  $\beta^l$  and upper service curve  $\beta^u$  which satisfy

$$\beta^l(t - s) \leq S[s, t] \leq \beta^u(t - s), \forall s < t$$

where  $\beta^l(0) = \beta^u(0) = 0$ . The RTC service curve for a system is denoted as  $\langle \beta^l, \beta^u \rangle$  for short.

From these two definitions, we find that the upper arrival curve and lower service curve correspond to the arrival curve and service curve of DNC theory [28]. Similarly, the upper service curve is identical to the maximum service curve of DNC theory. Thus, the two concatenation theorems for service curve (see Theorem 1.46 in [28]) and maximum service curve (see Theorem 1.6.1 in [28]) together form the concatenation theorem for the RTC service curve. Assume an event stream traverses two systems  $S_1$  and  $S_2$  in sequence, and  $S_i$  offers an RTC service curve  $\langle \beta_i^l, \beta_i^u \rangle$  ( $i = 1, 2$ ) to this event stream. The concatenation theorem gives the equivalent RTC service

curve offered by these two systems to this event stream, which is  $\langle \beta_1^l \otimes \beta_2^l, \beta_1^u \otimes \beta_2^u \rangle$ .

In this paper, we will utilize the discrete time RTC arrival curve and service curve to characterize the arrived traffic and service capability of the wormhole-switched NoC, since the minimum time unit of this system is the clock period  $T$ . Events in the definitions of arrival curve and service curve refer to the arrival and service of flits, respectively. If we obtain the arrival curve  $\langle \alpha^l, \alpha^u \rangle$  of a specific flow at specific router and the service curve  $\langle \beta^l, \beta^u \rangle$  provided by this router, we can get the output arrival curve  $\langle \alpha^{l'}, \alpha^{u'} \rangle$  of this flow and leftover service curve  $\langle \beta^{l'}, \beta^{u'} \rangle$  of this router with the following equations [17]:

$$\alpha^{l'} = \min\{(\alpha^l \otimes \beta^u) \otimes \beta^l, \beta^l\} \quad (1)$$

$$\alpha^{u'} = \min\{(\alpha^u \otimes \beta^u) \otimes \beta^l, \beta^u\} \quad (2)$$

$$\beta^{l'} = (\beta^l - \alpha^u) \bar{\otimes} 0 \quad (3)$$

$$\beta^{u'} = \max\{(\beta^u - \alpha^l) \bar{\otimes} 0, 0\} \quad (4)$$

where  $\otimes$ ,  $\bar{\otimes}$ ,  $\bar{\otimes}$ ,  $\bar{\otimes}$  correspond to the min-plus convolution, min-plus de-convolution, max-plus convolution and max-plus de-convolution [28].

After we obtain the arrival curve  $\langle \alpha_f^l, \alpha_f^u \rangle$  of flow  $f$  and the equivalent service curve  $\langle \beta_f^l, \beta_f^u \rangle$  offered by the system to flow  $f$ , we can get the end-to-end delay bound by the following equation [28]

$$\text{Delay}(f) = H(\alpha_f^u, \beta_f^l) \quad (5)$$

where operator  $H(\cdot, \cdot)$  means the maximal horizontal deviation between the two operands.

### IV. DELAY ANALYSIS AND BUFFER SIZING

In this section, we first build an RTC based performance model for the priority-aware wormhole-switched NoC. Based on the constructed performance model, we then propose an end-to-end delay analysis algorithm and a buffer sizing algorithm.

The performance model comprises two parts, i.e. traffic model and service model. The traffic model utilizes the RTC

arrival curve to describe the arrival process of each flow. We will introduce two methods to obtain the arrival curve in subsection IV-A. The service model characterizes the services offered by the priority-aware NoC to each flow. The construction of service model is much more complicated than the traffic model. While constructing the service model, the following three issues should be considered: (1) Only the head flit needs to traverse the RC and VA stages, because the non-head flits of a packet follow the data-path built by the head flit. To simplify our RTC model, we need a special mechanism to characterize the service offered to head and non-head flits in a unified way. (2) Our model extends the existing approach [14][15] by allowing priority sharing among flows. Thus, the leftover service curve provided to the lower-priority flows can only be derived when all the service curves of high-priority flows have been computed. (3) We should first break the cyclic-dependence between the adjacent routers caused by flow control before analyzing the end-to-end delay bound with Eq.(5). We will discuss the first two issues in subsection IV-B, and the last issue is discussed in subsection IV-C. Finally, we present the delay analysis algorithm and buffer sizing algorithm in subsection IV-D and subsection IV-E, respectively.

#### A. Traffic Model

The communication in a priority-aware wormhole-switched NoC is realized by transmitting packets, and the packet is further divided into flits, which is the minimum transmission unit in wormhole-switched NoC. Denote by  $\langle \alpha^l(\Delta), \alpha^u(\Delta) \rangle$  the flit arrival curve of a flow, namely, the minimum and maximum number of flits can be seen within any time window of length  $\Delta$ . We can extract the flit arrival curve from the synthetic traffic or communication trace with the sliding window method [17]. For each window length  $\Delta$ , this method tries to find the maximal and minimal number of arrived flits (corresponding to  $\alpha^l(\Delta)$  and  $\alpha^u(\Delta)$ ) by analyzing the time series of flits. For the synthetic traffic or communication trace, the obtained arrival curve might not be periodic. But, it does not hinder the application of RTC theory, because the RTC arrival curve can characterize arbitrary traffic patterns.

However, the obtained flit arrival curve can only be applied to compute the worst-case performance bound at the flit level. To obtain the packet level delay bound, this arrival curve must be  $L$ -packetized [28]. Denote by  $L(n)$  the cumulative packet length<sup>1</sup> of the first  $n$  packets in a flow,  $R(t)$  the cumulative arrived flits by time  $t$ . Then, the  $L$ -packetizer operator  $\mathcal{P}^L(\cdot)$  is defined as  $\mathcal{P}^L(R(t)) = \sup_{n \in \mathcal{N}} \{L(n)1_{L(n) \leq R(t)}\}$ <sup>2</sup>. Intuitively,  $\mathcal{P}^L(\cdot)$  can be interpreted as the largest cumulative packet length contained in  $R(t)$ , as shown in Fig. 2a. For any flit arrival curve  $\langle \alpha^l(\Delta), \alpha^u(\Delta) \rangle$ , the  $L$ -packetized arrival curve can be obtained by applying the following theorem.

**Theorem 1 ( $L$ -packetized arrival curve):** If a flow has a flit arrival curve  $\langle \alpha^l(\Delta), \alpha^u(\Delta) \rangle$ , the flow also has a  $L$ -

packetized arrival curve  $\langle \alpha^l(\Delta) - l_{max}1_{\{\Delta > 0\}}, \alpha^u(\Delta) + l_{max}1_{\{\Delta > 0\}} \rangle$ , where  $l_{max}$  is the maximum packet length (in flits) of a flow.

*Proof:* According to the basic properties of  $L$ -packetizer [28], we have

$$R(t) - l_{max} \leq \mathcal{P}^L(R(t)) < R(t)$$

and

$$R(s) - l_{max} \leq \mathcal{P}^L(R(s)) < R(s).$$

The inequalities above indicate

$$R(t) - R(s) - l_{max} \leq \mathcal{P}^L(R(t)) - \mathcal{P}^L(R(s))$$

and

$$\mathcal{P}^L(R(t)) - \mathcal{P}^L(R(s)) \leq R(t) - R(s) - l_{max}.$$

Based on the definition 1, the  $L$ -packetized flow has a packet arrival curve  $\langle \alpha^l(\Delta) - l_{max}1_{\{\Delta > 0\}}, \alpha^u(\Delta) + l_{max}1_{\{\Delta > 0\}} \rangle$ , which ends the proof. ■

We can also directly obtain the  $L$ -packetized arrival curve instead of transformation from flit arrival curve for some special cases. For example, suppose all the packets in a flow have the same length  $F$  and arrived periodically with period  $I$ . By applying the sliding window method [17], we can obtain the flit arrival curve of this flow, which is a pair of staircase functions<sup>3</sup>  $\langle F \cdot u_{I,0}, F \cdot u_{I,I} \rangle$ . As shown in Fig. 2b, the obtained flit arrival curve which is equal to the  $L$ -packetized arrival curve  $\mathcal{P}^L(\alpha)$ , since  $\mathcal{P}^L(R(t)) = R(t)$ ,  $\mathcal{P}^L(\alpha^l(t)) = \alpha^l(t)$  and  $\mathcal{P}^L(\alpha^u(t)) = \alpha^u(t)$ .

#### B. Basic Feed-forward Service Model

The service model characterizes the service obtained by each flow at its source NI and entire path, which will be discussed as follows.

1) *Service Curve at Source NI:* If a IP core generates more than one flows simultaneously, the source NI will schedule these flows to go through the output link connecting the source NI and injection router according to their priorities. Figure ?? illustrates the internal structure of this priority-aware NI, where the IP core generates four flows simultaneously. Messages of different flows are encapsulated and stored in the dedicated buffer for that flow. The priority-aware scheduler selects one flit with the highest-priority at a time for transmission, and imposes an additional latency  $T$  to all the flits which traverse it. Thus, the RTC service curve of the source NI (denoted as  $\langle \beta_{NI}^l, \beta_{NI}^u \rangle$ ) can also be obtained by applying the sliding window method [17], as shown in Fig. 2c, which is a pair of staircase functions  $\langle u_{T,0}, u_{T,T} \rangle$ . Given the set of flow specifications, Algorithm 1 can derive the service curve obtained by each flow at the source NI. The flow specification of  $f_i$  is a quadruple  $\langle \alpha^l, \alpha^u, \mathcal{R}_i, D_i, P_i \rangle$ , which specifies the arrival curve, routine, deadline and priority of a flow.

<sup>1</sup>Let  $l_i$  be the length (in flits) of  $i$ -th packet, the cumulative packet length  $L(n)$  is defined as  $L(n) = \sum_{i=1}^n l_i$ .

<sup>2</sup> $\mathcal{N}$  is the set of natural numbers and  $1_{\{val\}}$  is the indicator function,  $1_{\{val\}} = 1$  if and only if  $val$  is true.

<sup>3</sup>A staircase function  $u_{T,\tau} = \lceil \frac{t+\tau}{T} \rceil$  for  $t > 0$  and 0 otherwise, where  $0 \leq \tau \leq T$  and  $\lceil \cdot \rceil$  is the ceiling operator.

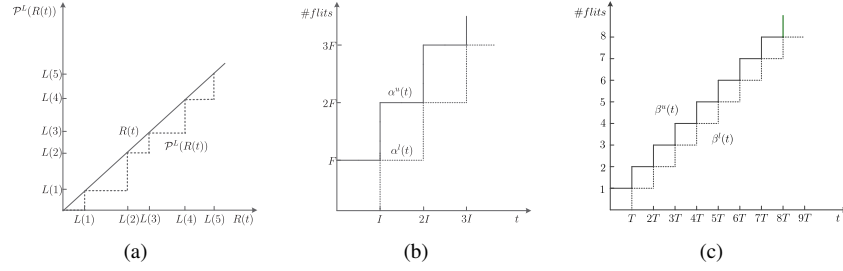


Fig. 2: Traffic model and service model. (a) Definition of  $\mathcal{P}^L(R(t))$ . Cumulative arrival function  $R(t)$  and the  $L$ -packetized cumulative arrival function  $\mathcal{P}^L(R(t))$  are represented by the dotted line and solid line, respectively. (b) Real-time calculus arrival curve for periodically arrived traffic with period  $I$  and packet length  $F$ . The solid line and dotted line represent the upper arrival curve and lower arrival curve. (c) Service model for each pipeline stage. The solid lines and dotted lines represent the upper service curves and lower service curves, respectively.

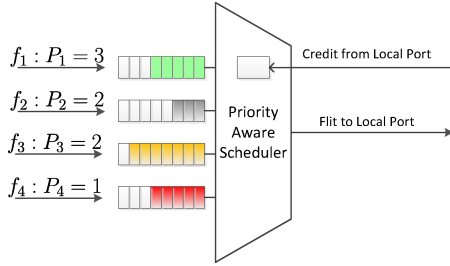


Fig. 3: Priority-aware network interface. Each flow has its dedicated buffer, and the scheduler selects flits with highest-priority for transmission at each cycle.

---

**Algorithm 1** Service curve at the source NI

---

**Require:** The set of flow specifications

**Ensure:** The service curve obtained by each flow

```

1: Group the flows with priority  $P_i$  into subset  $\mathcal{F}_i$ .
2:  $\beta_{NI}^l = \lfloor \frac{t}{T} \rfloor$ ;  $\beta_{NI}^u = \lceil \frac{t}{T} \rceil$ .
3: for each  $\mathcal{F}_i$  from highest priority to lowest priority do
4:   for each flow  $f_j \in \mathcal{F}_i$  do
5:      $\beta_{NI,f_j}^l = \lfloor \frac{\beta_{NI}^l}{|\mathcal{F}_i|} \rfloor$ ;  $||\mathcal{F}_i||$  denotes the cardinality of  $\mathcal{F}_i$ 
6:      $\beta_{NI,f_j}^u = \lceil \frac{\beta_{NI}^u}{|\mathcal{F}_i|} \rceil$ ;
7:   end for
8:    $\alpha^l = \sum_{f_j \in \mathcal{F}_i} \alpha_{f_j}^l$ ;  $\alpha^u = \sum_{f_j \in \mathcal{F}_i} \alpha_{f_j}^u$ ;
9:    $\beta_{NI}^l = (\beta_{NI}^l - \alpha^u) \otimes 0$ ;  $\beta_{NI}^u = \max\{(\beta_{NI}^u - \alpha^l) \otimes 0, 0\}$ ;
10: end for
```

---

2) *Service Curve of a Router:* While modeling the service capability of routers with RTC, we can analyze the data-path of a flow in a router stage-by-stage. On obtaining the service curves offered by each stage, the service curve provided by the router to a flow can be obtained by concatenating all the service curves of these stages. This is significantly different from the existing DNC based model [22], [15], where they treat the entire router as a whole and designate a Latency-Rate (LR) service curve [28] to simplify the performance derivation. Whereas, our model uses the staircase functions to characterize the detailed behavior of this discrete time system. The advantage of our method is that, it can be easily modified

to characterize the non-standard router micro-architectures, by simply letting the service curve of non-existed stages to be a burst delay function  $\delta_0(t)$ <sup>4</sup>. Next, we try to derive the service curves of all these stages:

(1) BW stage, SA stage and LT stage: all the flits within a traffic flow will go through these three stages, and experience a fixed delay  $T$  at each stage. The service curves provided by these stages, i.e.  $\langle \beta_{BW}^l, \beta_{BW}^u \rangle$ ,  $\langle \beta_{SA}^l, \beta_{SA}^u \rangle$  and  $\langle \beta_{LT}^l, \beta_{LT}^u \rangle$ , can be derived by applying the sliding window method [17], which are the same as the source NI, as shown in Fig. 2c.

(2) RC stage and VA stage: the latency of head flit experienced at these two stages is  $T$ . Although the non-head flits do not go through these two stages, they have to wait for two cycles before entering the SA stage at the worst-case, e.g. the three body flits of the same packet shown in Fig. 1b. Thus, a sophisticated solution to construct a unified lower service curve for head flit and non-head flits at these two stages comes from viewing each of these two stages impose an additional delay  $T$  for all the flits. Thus, the equivalent lower service curve of these two stages, i.e.  $\beta_{RC}^l$  and  $\beta_{VA}^l$ , can be easily obtained by the sliding window method [17], as shown in Fig. 2c. To derive the upper service curve of these two stages, let us consider the most ‘lucky’ flits of a flow, e.g. the tail flit shown in Fig. 1b. This flit can enter the SA stage immediately after it was written into the dedicated VC buffer. For this case, the RC and VA stages impose a zero latency to it. Thus, we can utilize the burst delay function  $\delta_0(t)$  to represent the upper service curve of these two stages.

(3) ST stage: each output port of the wormhole-switched NoC has a switch allocator to schedule the switch traversal among all the contending flows at each clock cycle. The notation  $\langle \beta_{ST,R_i^p}^l, \beta_{ST,R_i^p}^u \rangle$  is used to identify the service curve obtained by all the contending flows injected into switch port  $p$ . For the mesh topology, the port indicator  $p$  can be concreted with  $W$  (West port),  $E$  (East port),  $S$  (South port) and  $N$  (North port) or  $L$  (Local port). Thus, Following the same procedure as BW stage, we can get the service curves  $\langle \beta_{ST,R_i^p}^l, \beta_{ST,R_i^p}^u \rangle$ , as shown in Fig. 2c.

Alert readers have noticed that, the contention of different

<sup>4</sup> $\delta_{val}(t) = +\infty$  if  $t > val$ , and 0 otherwise.

flows within a router only occurs at ST stage. For the fixed-priority scheduling policy, switch allocators schedule the flow with the highest priority first, flows with the same priority will be served with Round-Robin order. All the unscheduled flows will be imposed an additional latency  $T$  due to the failure of switch arbitration.

Denote by  $\langle \beta_{ST,R_i^p}^l, \beta_{ST,R_i^p}^u \rangle$  the total service curve provided by the ST stage,  $\langle \beta_{ST,R_i^p,f_j}^l, \beta_{ST,R_i^p,f_j}^u \rangle$  the service curve provided to flow  $f_j$  by SA stage of router  $R_i$ , and  $\langle \beta_{ST,R_i^p}^l, \beta_{ST,R_i^p}^u \rangle$  the leftover service curve after serving the flows with higher priority than  $f_j$ . In order to obtain the service curve  $\langle \beta_{ST,R_i^p,f_j}^l, \beta_{ST,R_i^p,f_j}^u \rangle$ , we should consider the following two cases:

(a) All the flows contending with  $f_j$  at  $R_i$  have lower priorities. For the synchronized router architecture, flow  $f_j$  gets the total leftover service curve  $\langle \beta_{ST,R_i^p}^l, \beta_{ST,R_i^p}^u \rangle$ .

(b) There exists some contention flows with the same priority as  $f_j$ . Denote by  $\Theta_{R_i,f_j}$  the set of contention flows at router  $R_i$  with the same priority as  $f_j$ , and let  $N_{R_i,f_j}$  be the number of flows in  $\Theta_{R_i,f_j}$ . Since all the flows in  $\Theta_{R_i,f_j}$  got serviced in Round-Robin order, the service curve provided to  $f_j$  is  $\langle \lfloor \beta_{ST,R_i^p}^l / (|\Theta_{R_i,f_j}| + 1) \rfloor, \lceil \beta_{ST,R_i^p}^u / (|\Theta_{R_i,f_j}| + 1) \rceil \rangle$ . After serving all the flows in  $\Theta_{R_i,f_j}$ , the leftover service curve for low-priority flows can be derived by applying Eq.(3) and Eq.(4).

After we obtained the service curve provided by ST stage to flow  $f_j$ , we can obtain the service curve of the router directly. The equivalent service curve of router  $R_i$  provided to  $f_j$ , i.e.  $\langle \beta_{R_i,f_j}^l, \beta_{R_i,f_j}^u \rangle$ , can be obtained by concatenating the service curves of all these stages together:

$$\begin{aligned} \beta_{R_i,f_j}^l &= \beta_{BW}^l \otimes \beta_{RC}^l \otimes \beta_{VA}^l \otimes \beta_{SA}^l \otimes \beta_{ST,R_i^p,f_j}^l, \\ \beta_{R_i,f_j}^u &= \beta_{BW}^u \otimes \beta_{RC}^u \otimes \beta_{VA}^u \otimes \beta_{SA}^u \otimes \beta_{ST,R_i^p,f_j}^u. \end{aligned}$$

### C. Service Model with Feedback Flow Control

To this end, we have construct the traffic model and the feed-forward service model. Whereas, the credit-based flow control introduces cyclic-dependence between the adjacent routers, and leads to self-blocking within a flow due to the insufficiency of buffer space at the downstream router. The cyclic-dependence between the adjacent routers prevents us from deriving the performance bound directly even after we have obtained the service curve reserved at each router for the target flow. In existing literature, this cyclic-dependence is addressed by fixed-point iteration [38] or transformation from marked dataflow graph [39].

In this subsection, we will try to tackle the flow control problem with another solution motivated by [22], where the authors abstract the flow control as a network element (called flow controller) providing a service curve (corresponding to the lower service curve of RTC theory). Then, this service curve is obtained by applying some basic properties of DNC theory [28]. To make the discussion concrete, we take flow  $f_2$  in Fig. 1a as an example and utilize the scheduling network model [17] in RTC theory to visualize the credit-based flow control and complex relationship among  $f_2$  and the other flows, as shown in Fig. 4. We ignore flow  $f_4$  and the flow

control of the other flows for brevity and clarity. We also assume that, all the destination IP cores can consume the arrived flits immediately, thus there is no flow control between the ejection router and destination NI. However, to prevent the buffer overflow, the flow control between source NI and injection router is necessary.

In a realistic wormhole router with credit-based flow control, a flit will be locked if the all the credits have been used up. We can treat the blocking caused by credit-insufficiency as traversing a virtual pipeline stage, called Flow Control (FC) stage, as shown in Fig. 4. Then, we follow the similar procedure as [22] to derive the service curve for the stage, as summarized in Theorem 2. The obtained service curve enable us to break the cyclic-dependence caused by flow control and build a comprehensive performance model with RTC.

**Theorem 2:** Suppose the router provides an RTC service curve  $\langle \beta^l, \beta^u \rangle$ , the buffer size and credit feedback delay are denoted as  $B$  and  $\sigma$ , respectively. Then, the flow controller provides an equivalent RTC service curve  $\langle \beta^l \otimes \beta_{LT}^l \otimes \delta_\sigma + B, \beta^u \otimes \beta_{LT}^u \otimes \delta_\sigma + \bar{B} \rangle$ , where  $\bar{f}$  is the sub-additive closure of  $f$  [28].

*Proof:* The lower service curve has been derived in [22]. In the rest of this proof, we will take the flow control between  $R_9$  and  $R_5$  in Fig. 4 as an example to derive the upper service curve. Denote the amount of injected and departed flits at  $R_5$  by time  $t$  as  $I(t)$  and  $D(t)$ , and the amount of flits served by  $R_9$  by time  $t$  as  $A(t)$ . The feedback link can be represented as a network element providing upper service curve  $\delta_\sigma(t)$ .

For the flow control between router  $R_9$  and  $R_5$ , we have  $I(t) \leq A(t)$  for causality and  $I(t) \leq D'(t) + B$  due to the effect of flow control, where  $D'(t) \leq D \otimes \delta_\sigma(t)$ . Thus,

$$I(t) \leq \min\{A(t), D'(t) + B\}.$$

Based on the equivalent definition of upper service curve<sup>5</sup>, we have

$$D(t) \leq I \otimes \beta_{R_5,f_2}^u \otimes \beta_{LT}^u(t).$$

Bring  $I(t)$  and  $D'(t)$  into this equality, we get

$$\begin{aligned} D(t) &\leq I \otimes \beta_{R_5,f_2}^u \otimes \beta_{LT}^u(t) \\ &\leq \min\{A \otimes \beta_{R_5,f_2}^u \otimes \beta_{LT}^u(t), D \otimes \delta_\sigma \otimes \beta_{R_5,f_2}^u \otimes \beta_{LT}^u(t) + B\}. \end{aligned}$$

By applying Theorem 4.31 in [28], we have

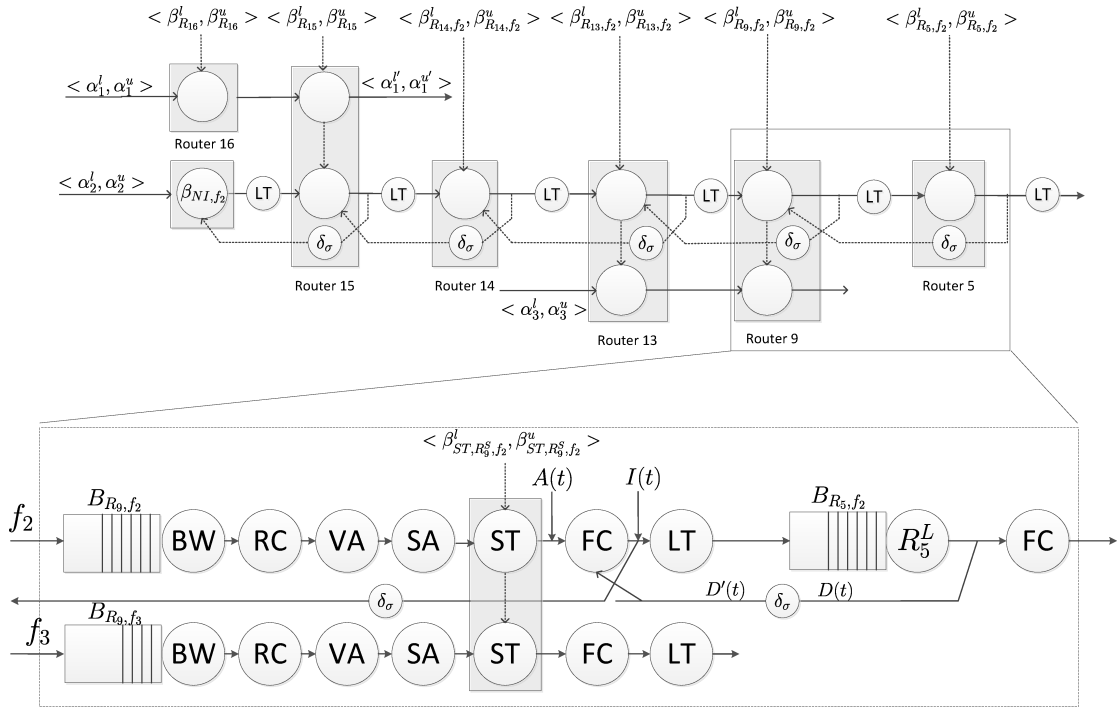
$$D \leq A \otimes \beta_{R_5,f_2}^u \otimes \beta_{LT}^u \otimes \overline{\beta_{R_5,f_2}^u \otimes \beta_{LT}^u \otimes \delta_\sigma + B}.$$

Thus,

$$\begin{aligned} I &\leq \min\{A, D' + B\} \\ &\leq \min\{A, D \otimes \delta_\sigma + B\} \\ &\leq \min\{A, A \otimes \beta_{R_5,f_2}^u \otimes \beta_{LT}^u \otimes \overline{\beta_{R_5,f_2}^u \otimes \beta_{LT}^u \otimes \delta_\sigma + B} \otimes \delta_\sigma + B\} \\ &= \min\{A \otimes \delta_\sigma, A \otimes \delta_\sigma \otimes \beta_{R_5,f_2}^u \otimes \beta_{LT}^u + B\} \\ &= A \otimes \min\{\delta_\sigma, \overline{\beta_{R_5,f_2}^u \otimes \beta_{LT}^u \otimes \delta_\sigma + B}\} \\ &= A \otimes \overline{\beta_{R_5,f_2}^u \otimes \beta_{LT}^u \otimes \delta_\sigma + B} \end{aligned}$$

where the steps from the third line to the fifth line holds due to the general properties of  $\otimes$  operator (see Rule 6 and Rule

<sup>5</sup>please refer to definition 1.6.1 in [28] for more details.

Fig. 4: Scheduling network model for flow  $f_2$ 

7 of Theorem 3.1.5 in [28] for more details), and the last step follows from the definition of sub-additive closure.

The inequality  $I \leq A \otimes \beta_{R_5, f_2}^u \otimes \beta_{LT}^u \otimes \delta_\sigma + B$  implies that the flow controller between  $R_9$  and  $R_5$  provides an equivalent upper service curve  $\beta_{R_5, f_2}^u \otimes \beta_{LT}^u \otimes \delta_\sigma + B$ . Thus, we can conclude that for any router providing upper service curve  $\beta^u$ , the corresponding flow controller has an equivalent upper service curve  $\beta_\tau^u(t) = \beta^u \otimes \beta_{LT}^u \otimes \delta_\sigma(t) + B$ . ■

On obtaining the equivalent service curve of flow controller (denoted as  $\langle \beta_{FC, f_i}^l, \beta_{FC, f_i}^u \rangle$ ), we get the equivalent service curve of each router after breaking the cyclic-dependence loop, which is

$$\begin{aligned} \beta_{R_i, f_j}^l &= \beta_{BW}^l \otimes \beta_{RC}^l \otimes \beta_{VA}^l \otimes \beta_{SA}^l \otimes \beta_{ST, R_i^p, f_j}^l \otimes \beta_{FC, f_j}^l, \\ \beta_{R_i, f_j}^u &= \beta_{BW}^u \otimes \beta_{RC}^u \otimes \beta_{VA}^u \otimes \beta_{SA}^u \otimes \beta_{ST, R_i^p, f_j}^u \otimes \beta_{FC, f_j}^u. \end{aligned}$$

Theorem 2 derives the RTC service curve of a single flow controller, and we can get the service curves of all the flow controllers along the router chain of any flow by applying Theorem 2 iteratively. As shown in Fig. 4, the service curve of a flow controller is determined by the service curves of the downstream flow controllers and routers. Hence, for each flow, we should compute the service curves of flow controllers from the ejection router to the injection router. Take flow  $f_2$  as an example, we should first compute the service curve of flow controller between  $R_5$  and  $R_9$  (denoted as  $\langle \beta_{\tau_9, f_2}^l, \beta_{\tau_9, f_2}^u \rangle$ ), which is  $\langle \beta_{R_5, f_2}^l \otimes \beta_{LT}^l \otimes \delta_\sigma + B, \beta_{R_5, f_2}^u \otimes \beta_{LT}^u \otimes \delta_\sigma + B \rangle$ . Then, By applying the concatenation theorem, we can obtain the equivalent service curve provided to  $f_2$  by router  $R_9$ , i.e.  $\langle \beta_{R_9, f_2}^l \otimes \beta_{\tau_9, f_2}^l, \beta_{R_9, f_2}^u \otimes \beta_{\tau_9, f_2}^u \rangle$ , which can be utilized to derive  $\langle \beta_{\tau_{13}, f_2}^l, \beta_{\tau_{13}, f_2}^u \rangle$  further. Follow the same

procedure,  $\langle \beta_{\tau_{14}, f_2}^l, \beta_{\tau_{14}, f_2}^u \rangle$ ,  $\langle \beta_{\tau_{15}, f_2}^l, \beta_{\tau_{15}, f_2}^u \rangle$  and  $\langle \beta_{\tau_{NI}, f_2}^l, \beta_{\tau_{NI}, f_2}^u \rangle$ <sup>6</sup> can be derived iteratively.

#### D. End-to-End Delay Analysis

In this subsection, we present the delay analysis algorithm, as shown in Algorithm 2. This algorithm takes the architecture parameters and flow specifications as input, and gives the worst-case end-to-end delay for all the flows. The architecture parameters specify the network topology graph, buffer size of each VC and the service curve of each pipeline stage. The flow specifications specifies the arrival curve, routine, deadline and priority of each flow.

In this algorithm, the arrival curve of flow  $f_i$  at the source NI and router  $R_j$  are denoted as  $\langle \alpha_{f_i}^l, \alpha_{f_i}^u \rangle$  and  $\langle \alpha_{R_j, f_i}^l, \alpha_{R_j, f_i}^u \rangle$ , respectively. The leftover service curve of ST stage at output port  $p$  is represented as  $\langle \beta_{ST, R_j^p}^l, \beta_{ST, R_j^p}^u \rangle$  (Initially, let  $\beta_{ST, R_j^p}^l = \beta_{ST, R_j^p}^l$  and  $\beta_{ST, R_j^p}^u = \beta_{ST, R_j^p}^u$ ). In the fixed-priority flit-level preemptive NoC, only the leftover service curve can be used by the low-priority flows. Thus, our algorithm compute the leftover service curve and delay bound from high-priority flows to low-priority flows. For each iteration, it performs the following four steps in sequence: (1) Calculating the service curves provided by the routers (lines 4-5) and flow controllers (lines 6-8) along the path. The computed service curve for flow controller can only be applicable for the specified flow. (2) Computing the worst-case end-to-end delay of the flow (lines 10-12). The service curve provided by the source NI of  $f_i$ , i.e.  $\beta_{NI, f_i}$  has been computed with Algorithm 1. (3) A strong point of

<sup>6</sup>  $\langle \beta_{\tau_{NI}, f_2}^l, \beta_{\tau_{NI}, f_2}^u \rangle$  denotes the service curve of flow controller between source NI and injection router, as shown in Fig.4.



performance model when compared with the LLA method [14] and DNC method [15] is that our algorithm supports the priority-sharing. Thus, the leftover service curve at each router for low-priority flows can only be calculated when all the flows sharing the same priority have been calculated. To calculate the leftover service curve at ST stage, we have to first derive the partial equivalent service curve (lines 16-17) and arrival curve (lines 18-19). Then, derive the leftover service curve with the aggregate arrival curve of the same priority-level (lines 21-24). The overall algorithm has two-level embedded loops, and the computation complexity for this algorithm is  $O(HN)$ , where  $N$  and  $H$  is the number of flows and the hop count of each flow. This algorithm can be easily integrated into the RTC toolbox [33] to compute the end-to-end delay bound automatically. Since our algorithm takes the maximum service capability and minimum service rate into consideration, our algorithm can give much tighter delay bound than the DNC-based delay analysis algorithm proposed in [15].

---

**Algorithm 2** End-to-End Delay Analysis Algorithm

---

**Require:** Architecture parameters and flow specifications

**Ensure:** Worst-case end-to-end delay for all the flows

```

1: for each flow  $f_i \in \mathcal{F}$  with priority order do
2:    $\beta_\tau^l = \delta_0(t)$ ;  $\beta_\tau^u = \delta_0(t)$ ;
3:   for each router  $R_j \in \mathcal{R}_i$  from  $end_i$  to  $start_i$  do
4:      $\beta_{R_j, f_i}^l = \beta_{BW}^l \otimes \beta_{RC}^l \otimes \beta_{VA}^l \otimes \beta_{SA}^l \otimes \lfloor \frac{\beta_{ST, R_j}^{l'}}{|\Theta_{R_j, f_i}|+1} \rfloor \otimes \beta_\tau^l$ ;
5:      $\beta_{R_j, f_i}^u = \beta_{BW}^u \otimes \beta_{RC}^u \otimes \beta_{VA}^u \otimes \beta_{SA}^u \otimes \lceil \frac{\beta_{ST, R_j}^{u'}}{|\Theta_{R_j, f_i}|+1} \rceil \otimes \beta_\tau^u$ ;
6:      $\beta_{\tau_j, f_i}^l(t) = \beta_\tau^l$ ;  $\beta_{\tau_j, f_i}^u(t) = \beta_\tau^u$ ;
7:      $\beta_\tau^l = \frac{\beta_{R_j, f_i}^l \otimes \beta_\tau^l \otimes \beta_{LT}^l \otimes \delta_\sigma(t) + B_{R_j, f_i}}{\beta_{R_j, f_i}^l \otimes \beta_\tau^l \otimes \beta_{LT}^l \otimes \delta_\sigma(t) + B_{R_j, f_i}}$ ;
8:      $\beta_\tau^u = \frac{\beta_{R_j, f_i}^u \otimes \beta_\tau^u \otimes \beta_{LT}^u \otimes \delta_\sigma(t) + B_{R_j, f_i}}{\beta_{R_j, f_i}^u \otimes \beta_\tau^u \otimes \beta_{LT}^u \otimes \delta_\sigma(t) + B_{R_j, f_i}}$ ;
9:   end for
10:   $\beta_{\tau_{NI}, f_i}^l = \beta_\tau^l$ ;  $\beta_{\tau_{NI}, f_i}^u = \beta_\tau^u$ ;
11:   $\beta_{f_i}^l = \beta_{NI, f_i}^l \otimes \beta_{\tau_{NI}, f_i}^l \otimes (\bigotimes_{R_k \in \mathcal{R}_{f_i}} (\beta_{R_k, f_i}^l \otimes \beta_{LT}^l))$ ;
12:   $Delay(f_i) = H(\alpha_{f_i}^u, \beta_{f_i}^l)$ ;
13:   $\beta_{f_i}^l = \beta_{NI, f_i}^l \otimes \beta_{\tau_{NI}, f_i}^l$ ;  $\beta_{f_i}^u = \beta_{NI, f_i}^u \otimes \beta_{\tau_{NI}, f_i}^u$ ;
14:  for  $\forall R_j \in \mathcal{R}_{f_i}$  from  $start_i$  to  $end_i$  do
15:    if  $\Omega_{R_j, f_i} \neq \emptyset$  then
16:       $\beta^l = \beta_{f_i}^l \otimes \beta_{BW}^l \otimes \beta_{RC}^l \otimes \beta_{VA}^l \otimes \beta_{SA}^l$ ;
17:       $\beta^u = \beta_{f_i}^u \otimes \beta_{BW}^u \otimes \beta_{RC}^u \otimes \beta_{VA}^u \otimes \beta_{SA}^u$ ;
18:       $\alpha_{R_j, f_i}^l = \min\{(\alpha_{f_i}^l \otimes \beta^u) \otimes \beta^l, \beta^l\}$ ;
19:       $\alpha_{R_j, f_i}^u = \min\{(\alpha_{f_i}^u \otimes \beta^u) \otimes \beta^l, \beta^u\}$ ;
20:      if  $\forall f_k \in \Theta_{R_j, f_i}$  have been calculated then
21:         $\alpha_{R_j, f_i}^l = \alpha_{R_j, f_i}^l + \sum_{f_k \in \Theta_{R_j, f_i}} \alpha_{R_j, f_k}^l$ ;
22:         $\alpha_{R_j, f_i}^u = \alpha_{R_j, f_i}^u + \sum_{f_k \in \Theta_{R_j, f_i}} \alpha_{R_j, f_k}^u$ ;
23:         $\beta_{ST, R_j}^{l'} = (\beta_{ST, R_j}^{l'} - \alpha_{R_j, f_i}^l) \bar{\otimes} 0$ ;
24:         $\beta_{ST, R_j}^{u'} = \max\{(\beta_{ST, R_j}^{u'} - \alpha_{R_j, f_i}^u) \bar{\otimes} 0, 0\}$ ;
25:      end if
26:    end if
27:     $\beta_{f_i}^l = \beta_{f_i}^l \otimes \beta_{R_j, f_i}^l$ ;  $\beta_{f_i}^u = \beta_{f_i}^u \otimes \beta_{R_j, f_i}^u$ ;
28:  end for
29: end for

```

---

### E. Buffer Sizing

The priority-aware wormhole-switched NoC [7], [8], [9] requires the same amount of VCs as the priorities to prevent priority inversion, which refers to the blocking of high-priority flows when the low priority flows occupy all the VCs [11]. To reduce the buffer area and power consumption of priority-aware wormhole-switched NoC, priority sharing [40] and buffer optimization [18] techniques have been proposed. However, the backlog bound derived in [18] is the minimum buffer size that does not trigger the flow control. Although reducing this buffer size will cause the back-pressure between adjacent routers and leads to a larger end-to-end delay, we can further reduce the buffer reserved at each router as long as the deadline constraint is not being violated. The amount of cycles that a packet can be stalled in the network without violating the deadline is defined as ‘slack’. Our buffer sizing algorithm can reduce the buffer size iteratively as long as slack is greater than zero and the buffer size is greater than one.

Suppose the applications have been mapped onto the NoC, and each flow  $f_i$  has been assigned to their corresponding priority  $P_i$  and deadline  $D_i$ . Following the same notation as Algorithm 1, we propose the buffer sizing algorithm to further reduce the buffer size, as shown in Algorithm 3. It tries to reduce the buffer size for each flow from high-priority to low-priority gradually. For each iteration, it performs the following four steps: (1) Calculating the service curves provided by the routers (lines 3-5). Initially, we set the service curve of flow controller to be  $< \delta_0(t), \delta_0(t) >$ , because the initial buffer is large enough to avoid flow control. (2) Calculate the minimum buffer size that can avoid flow control for each router (lines 6-8). (3) Reduce the initial buffer size gradually as long as the constraint of deadline is not being violated (lines 10-26). (4) Calculating the leftover service curve at each router for low-priority flows (lines 27-42). This algorithm can be implemented in RTC toolbox [33] to optimize the buffer size automatically.

Our buffer sizing algorithm can be used to reduce the router area and power consumption. However, it is significantly different from the DNC based slack optimization in [41]. In [41], the energy optimization is achieved by adjusting the voltage, frequency and link bandwidth of on-chip routers for the fixed configuration and deadline. In contrast, our method tries to optimize the buffer size under the deadline constraint, and the buffer reduction directly leads to the area and power saving. In addition, our algorithm can be used in conjunction with the priority sharing techniques [40] to optimize the buffer size of priority-aware wormhole-switched NoC.

## V. EXPERIMENTS

In this section, we validate the correctness and tightness of our performance model by comparison with other analytical methods and simulation. Several analytical methods exist for the delay analysis of priority-aware NoC, examples include contention tree model [16], lumped link model [11], dependency graph model [8], FLA [7], LLA [14] and DNC [15], etc. There are also extensive research on the buffer sizing of the priority-aware NoC, representative methods include shaping

**Algorithm 3** Buffer Sizing Algorithm**Require:** Architecture parameters and flow specifications**Ensure:** Optimized buffer size

```

1: for each flow  $f_i \in \mathcal{F}$  with priority order do
2:   for each router  $R_j \in \mathcal{R}_i$  do
3:      $\beta_{\tau_j, f_i}^l = \delta_0(t)$ ;  $\beta_{\tau_j, f_i}^u = \delta_0(t)$ ;
4:      $\beta_{R_j, f_i}^l = \beta_{BW}^l \otimes \beta_{RC}^l \otimes \beta_{VA}^l \otimes \beta_{SA}^l \otimes \lfloor \frac{\beta_{ST, R_j}^{l'}}{|\Theta_{R_j, f_i}|+1} \rfloor \otimes \beta_{\tau_j, f_i}^l$ ;
5:      $\beta_{R_j, f_i}^u = \beta_{BW}^u \otimes \beta_{RC}^u \otimes \beta_{VA}^u \otimes \beta_{ST}^u \otimes \lceil \frac{\beta_{ST, R_j}^{u'}}{|\Theta_{R_j, f_i}|+1} \rceil \otimes \beta_{\tau_j, f_i}^u$ ;
6:      $B_{R_j, f_i}^l = \inf\{B | \beta_{R_j, f_i}^l \otimes \beta_{LT}^l \otimes \frac{\beta_{R_j, f_i}^l \otimes \beta_{LT}^l \otimes \delta_\sigma(t) + B}{\beta_{R_j, f_i}^l \otimes \beta_{LT}^l} \geq \beta_{R_j, f_i}^l \otimes \beta_{LT}^l\}$ ;
7:      $B_{R_j, f_i}^u = \inf\{B | \beta_{R_j, f_i}^u \otimes \beta_{LT}^u \otimes \frac{\beta_{R_j, f_i}^u \otimes \beta_{LT}^u \otimes \delta_\sigma(t) + B}{\beta_{R_j, f_i}^u \otimes \beta_{LT}^u} \geq \beta_{R_j, f_i}^u \otimes \beta_{LT}^u\}$ ;
8:      $B_{R_j, f_i} = \max\{B^l, B^u\}$ ;
9:   end for
10:   $\beta_{\tau_{NI}, f_i}^l = \delta_0(t)$ ;  $\beta_{\tau_{NI}, f_i}^u = \delta_0(t)$ ;
11:  for each router  $R_j \in \mathcal{R}_i$  from  $end_i$  to  $start_i$  do
12:     $\beta_{f_i}^l = \beta_{NI, f_i}^l \otimes \beta_{\tau_{NI}, f_i}^l \otimes \bigotimes_{R_k \in \mathcal{R}_{f_i}} (\beta_{R_k, f_i}^l \otimes \beta_{LT}^l)$ ;
13:    while  $H(\alpha_{f_i}^u, \beta_{f_i}^l) \leq D_i$  and  $B_{R_j, f_i} > 1$  do
14:       $B_{R_j, f_i} = B_{R_j, f_i} - 1$ ;
15:      for all  $R_k$  from  $R_j$  to  $start_i$  do
16:         $\beta_{\tau_k, f_i}^l = \frac{\beta_{R_j, f_i}^l \otimes \beta_{LT}^l \otimes \beta_\sigma + B}{\beta_{R_j, f_i}^l \otimes \beta_{LT}^l}$ ;
17:         $\beta_{\tau_k, f_i}^u = \frac{\beta_{R_j, f_i}^u \otimes \beta_{LT}^u \otimes \beta_\sigma + B}{\beta_{R_j, f_i}^u \otimes \beta_{LT}^u}$ ;
18:      end for
19:       $\beta_{f_i} = \beta_{NI, f_i}^l \otimes \beta_{\tau_{NI}, f_i}^l \otimes \bigotimes_{R_k \in \mathcal{R}_{f_i}} (\beta_{R_k, f_i}^l \otimes \beta_{LT}^l)$ ;
20:    end while
21:    if  $H(\alpha_{f_i}^u, \beta_{f_i}^l) > D_i$  then
22:       $B_{R_j, f_i} = B_{R_j, f_i} + 1$ ;
23:       $\beta_{\tau_j, f_i}^l = \frac{\beta_{R_j, f_i}^l \otimes \beta_{LT}^l \otimes \beta_\sigma + B}{\beta_{R_j, f_i}^l \otimes \beta_{LT}^l}$ ;
24:       $\beta_{\tau_j, f_i}^u = \frac{\beta_{R_j, f_i}^u \otimes \beta_{LT}^u \otimes \beta_\sigma + B}{\beta_{R_j, f_i}^u \otimes \beta_{LT}^u}$ ;
25:    end if
26:  end for
27:   $\beta_{f_i}^l = \beta_{\tau_{NI}, f_i}^l$ ;  $\beta_{f_i}^u = \beta_{\tau_{NI}, f_i}^u$ ;
28:  for  $\forall R_j \in \mathcal{R}_{f_i}$  from  $start_i$  to  $end_i$  do
29:    if  $\Omega_{R_j, f_i} \neq \emptyset$  then
30:       $\beta^l = \beta_{f_i}^l \otimes \beta_{BW}^l \otimes \beta_{RC}^l \otimes \beta_{VA}^l \otimes \beta_{SA}^l$ ;
31:       $\beta^u = \beta_{f_i}^u \otimes \beta_{BW}^u \otimes \beta_{RC}^u \otimes \beta_{VA}^u \otimes \beta_{SA}^u$ ;
32:       $\alpha_{R_j, f_i}^l = \min\{(\alpha_{f_i}^l \otimes \beta^u) \otimes \beta^l, \beta^l\}$ ;
33:       $\alpha_{R_j, f_i}^u = \min\{(\alpha_{f_i}^u \otimes \beta^u) \otimes \beta^l, \beta^u\}$ ;
34:      if  $\forall f_k \in \Theta_{R_j, f_i}$  have been calculated then
35:         $\alpha_{R_j, f_i}^l = \alpha_{R_j, f_i}^l + \sum_{f_k \in \Theta_{R_j, f_i}} \alpha_{R_j, f_k}^l$ ;
36:         $\alpha_{R_j, f_i}^u = \alpha_{R_j, f_i}^u + \sum_{f_k \in \Theta_{R_j, f_i}} \alpha_{R_j, f_k}^u$ ;
37:         $\beta_{ST, R_j}^{l'} = (\beta_{ST, R_j}^{l'} - \alpha_{R_j, f_i}^l) \bar{\otimes} 0$ ;
38:         $\beta_{ST, R_j}^{u'} = \max\{(\beta_{ST, R_j}^{u'} - \alpha_{R_j, f_i}^u) \bar{\otimes} 0, 0\}$ ;
39:      end if
40:    end if
41:     $\beta_{f_i}^l = \beta_{f_i}^l \otimes \beta_{R_j, f_i}^l$ ;  $\beta_{f_i}^u = \beta_{f_i}^u \otimes \beta_{R_j, f_i}^u$ ;
42:  end for
43: end for

```

delay analysis [42] and LLA [18]. Among all these analytical methods, LLA [14][18] and DNC [15] based model outperform the others when the tightness of delay and backlog bound are considered. Thus, we will only perform the comparison with LLA and DNC to demonstrate the improvement of our results on the delay bound and buffer sizing, as presented in subsection V-A and subsection V-B respectively. We also present the simulation results to validate the correctness of our method in subsection V-C.

**A. Comparison with Link Level Analysis**

The network topology and flows are shown in Fig. 1a. There are four flows (i.e.  $f_1, f_2, f_3$  and  $f_4$ ) in the network, with different priority  $P_4 > P_1 > P_2 > P_3$ . Suppose the packet length of flow  $f_i$  is  $F_i$  (in flits), and the injection period is  $I_i$ . We perform the comparison on a set of periodical traffic due to the restriction of LLA method [14][18], and the traffic jitter for all the flows are set to be zero. To ease the analysis, LLA supposes the number of bits in a flit is the same as the physical channel width, and the latency of a router is one cycle. To compare with LLA, our model for the standard wormhole-switched router should be specialized to a single-cycle router, this is achieved by letting the service curve of BW, RC, VA, SA and LT stage be a burst delay function  $\delta_0(t)$ . Under this condition, the service curve of the entire router is the same as the service curve provided by the SA stage, which is  $< \beta_{ST, R_j}^{l'}, \beta_{ST, R_j}^{u'} >$ . In addition, the LLA method does not consider the self-blocking caused by flow control. Thus, we set the credit feedback delay  $\sigma = 0$  cycle in our model for a fair comparison. Next, we compare the end-to-end delay and buffer requirement computed with LLA and our model.

1) *End-to-End Delay*: For this scenario, we suppose the VC buffer is large enough so that we can neglect the back-pressure caused by flow control and directly apply LLA for the delay analysis. Suppose all the flows have the same injection period  $I_i$  (in cycles) and packet length  $F_i$  (in flits), and the release jitter of all the flows are set to zero. We examine the end-to-end delay of the four flows in Fig. 1a under different packet length  $F_i$  and injection period  $I_i$ , the calculated result is shown in Table I. Each quaternion in the table corresponds to the worst-case delay of  $f_1, f_2, f_3$  and  $f_4$  (in cycles) under given configuration. The blank items corresponding to LLA columns indicate that the worst-case delay of a flow is greater than its injection period, which cannot be analyzed with LLA [14][18], and the blank items corresponding to RTC columns indicate that the network is unstable because the injection rate exceeds the service capability of the network. As shown in Table I, the RTC method is applicable to these scenarios which can not be analyzed by LLA. In addition, we also observed from the table that the RTC result is as tight as that of LLA except for the scenarios that the worst-case delay of a flow is close to the injection period, e.g.  $F_i = 1$  flit and  $I_i = 6$  cycles in Table I. The root cause of these exceptions is that RTC theory we used in this paper is a counter-based algebra approach, which ignores some state information of the entire network. Although the counter-based approach, e.g. timed automata [43] and event count automata [44], can be applied to improve our

results, they will lead to higher computation complexity and state space explosion.

TABLE I: Delay comparison with link level analysis

$I_i$	$F_i = 1$		$F_i = 2$		$F_i = 4$	
	RTC	LLA	RTC	LLA	RTC	LLA
3	7,7,8,4	—	—	—	—	—
4	6,6,6,4	—	—	—	—	—
5	6,6,6,4	—	9,10,12,5	—	—	—
6	5,6,6,4	5,6,5,4	9,8,9,5	—	—	—
7	5,6,5,4	5,6,5,4	9,8,9,5	—	—	—
8	5,6,5,4	5,6,5,4	7,8,9,5	7,8,7,5	—	—
9	5,6,5,4	5,6,5,4	7,8,9,5	7,8,7,5	15,16,19,7	—
10	5,6,5,4	5,6,5,4	7,8,7,5	7,8,7,5	15,12,15,7	—
11	5,6,5,4	5,6,5,4	7,8,7,5	7,8,7,5	15,12,15,7	—
12	5,6,5,4	5,6,5,4	7,8,7,5	7,8,7,5	11,12,15,7	—
13	5,6,5,4	5,6,5,4	7,8,7,5	7,8,7,5	11,12,15,7	—
14	5,6,5,4	5,6,5,4	7,8,7,5	7,8,7,5	11,12,15,7	—
15	5,6,5,4	5,6,5,4	7,8,7,5	7,8,7,5	11,12,15,7	11,12,11,7
16	5,6,5,4	5,6,5,4	7,8,7,5	7,8,7,5	11,12,11,7	11,12,11,7
17	5,6,5,4	5,6,5,4	7,8,7,5	7,8,7,5	11,12,11,7	11,12,11,7

2) *Buffer Sizing*: The LLBA method [18] can only give the required buffer size at each VC to prevent flow control. By taking the flow control into consideration, our buffer sizing algorithm can be utilized to reduce the buffer size calculated with LLBA method further as long as the deadline constraint is not violated. We take a simple example to demonstrate the benefit of our method. Suppose all the flows have the same packet injection period  $I_i = 20$  cycles and packet length  $F_i = 4$  flits ( $i = 1, 2, 3, 4$ ), and the release jitter of all the flows are set to zero. We can also get the buffer size reserved at each router for the four flows with LLBA method, which are  $(1, 1, 1, 4)$ ,  $(4, 1, 1, 1, 1)$ ,  $(4, 4, 1, 1)$  and  $(1, 1, 1, 1)$ , respectively. The total buffer size required by the four flows can be obtained by summing up these buffer size, which is 29 flits. For the same configuration, if we change the deadline constraint from 15 cycles to 27 cycles, the buffer required for each flow can be further reduced with our buffer sizing algorithm, as shown in Fig. 5. Take the deadline  $D_i = 27$  cycles as an example, the buffer size calculated by our buffer sizing algorithm is  $17/29 \approx 58.62\%$  smaller than the total buffer size calculated with the LLBA method.

### B. Comparison with Network Calculus

In this subsection, we present the numerical results to demonstrate the improvement of our method over DNC method proposed in [15]. We use the periodical traffic as an example to make the comparison. Denote by  $I_i$  the injection period,  $F_i$  the packet length, the arrival curve can be easily obtained according to the method introduced in subsection IV-A. Suppose the priority of each flow in Fig. 1a satisfies  $P_4 > P_1 > P_2 > P_3$ , the buffer depth at each router  $B = 15$  flits, credit feedback delay  $\sigma = 0$  cycle. Define injection rate  $V_i = F_i/I_i$ , we change  $V_i$  ( $i = 1, 2, 3, 4$ ) from  $1/3$  to  $1/6$  flits/cycle and packet length from 1 to 8 flits, the end-to-end delay of flow  $f_3$  calculated with the DNC-based and our method are plotted in Fig. 6. By comparison, we find that our method can derive a much tighter delay bound than the DNC method proposed in [15]. The root cause for this improvement lies in the fact that our method utilizes the upper service curve

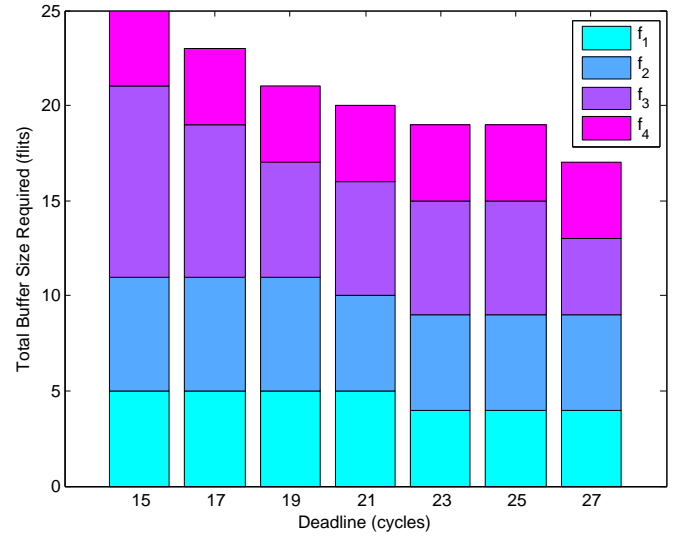


Fig. 5: Buffer requirement computed with RTC model

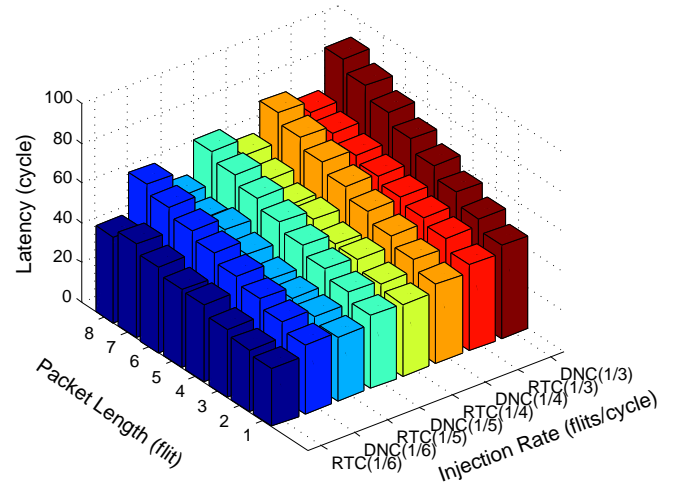


Fig. 6: Comparison with network calculus

to limit the output upper arrival curve, and further leads to a tighter lower service curve for the low-priority flows. To demonstrate this, let  $F_i = 8$  flits and  $V_i = 1/6$  flits/cycle, we plot the service curve for flow  $f_3$  calculated with the DNC and RTC in Fig. 7. The service curve is calculated with RTC toolbox. From Fig. 7, we find that the calculated service curve with DNC is indeed looser than the service curve calculated with our method. Since the delay bound is the maximal horizontal deviation between upper arrival curve and lower service curve in this figure, this looser service curve will finally lead to a looser end-to-end delay bound.

The comparison results in Fig. 6 also exhibits two special data points calculated with our method need to be explained. For the given injection rate  $V_i = 1/6$  ( $1/5$ ) flits/cycle, when we increase the packet length  $F_i$  from 7 flits to 8 flits, the worst-case delay calculated with our method decreases from 47 to 43 (49 to 47) cycles. To explain this phenomenon, for the given injection rate  $V_i = 1/6$  flits/cycle, we plot the upper arrival curve and lower service curve of flow  $f_3$  in Fig. 8 for

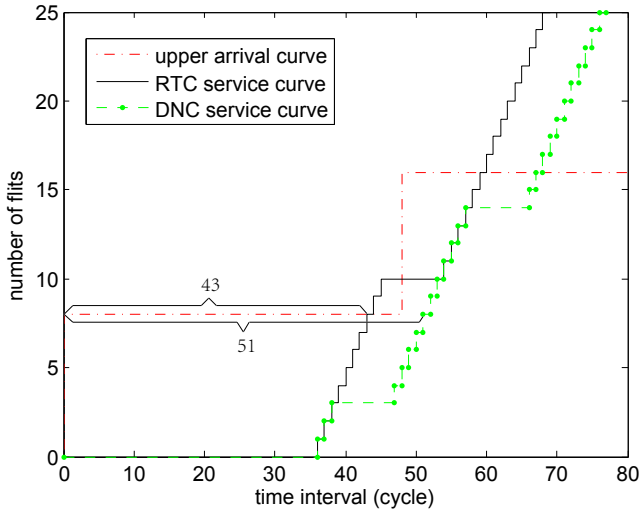
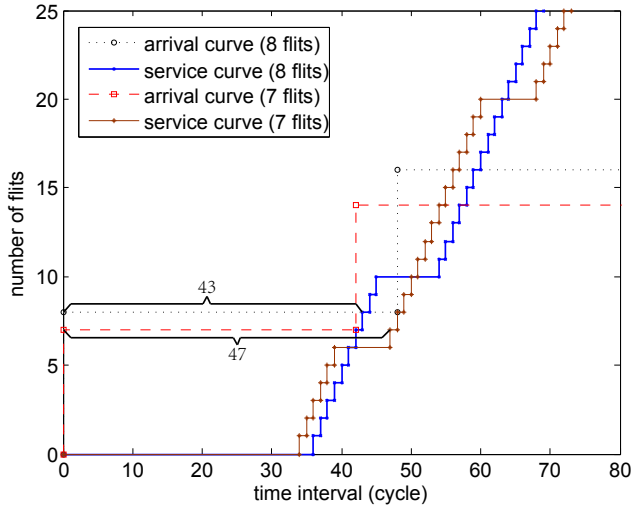


Fig. 7: Service curve derived from RTC and DNC

Fig. 8: Delay Comparison between  $F_i = 7$  and  $F_i = 8$ 

both  $F_i = 7$  flits and  $F_i = 8$  flits scenarios. As indicated in this figure, when  $F_i = 8$  flits, the lower service curve for  $f_3$  is greater than that of  $F_i = 7$  flits between the time interval  $[41, 51]$ , which leads to a smaller delay bound than  $F_i = 7$ .

### C. Comparison with Simulation

The correctness of our delay analysis algorithm and buffer sizing algorithm is verified by simulation. We modified the Booksim 2.0 simulator [45] to support the specified traffic pattern and injection process. Examples investigated in this section include the traffic pattern shown in Fig. 1a and a realistic example provided by Ericsson Radio Systems and discussed in [46][47]. We adopt the optimized lookahead pipeline router [35] to construct the mesh network presented in Fig. 1a, which remove the RC stage from the pipeline. Our service model is customized to fit this optimization by letting the service curve of RC stage be  $\delta_0(t)$ . Other architecture and simulation parameters are listed in the Table II.

For the traffic pattern shown in Fig. 1a, we set the injection

TABLE II: Architecture parameters used in the simulation

network topology	$4 \times 4$ mesh	routing algorithm	X-Y routing
clock cycle	1ns	channel width	128
buffer size	16	switch allocator	priority-aware
VC allocator	reserved	credit delay	1 cycle
sampling period	1e5 cycles	warmup period	3e5 cycles

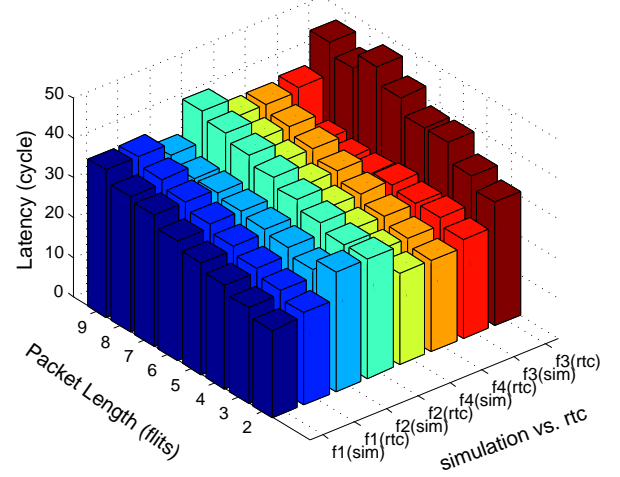


Fig. 9: Comparison with simulation

rate  $V_i$  to  $1/6$  flits/cycle, and change the packet length from two flits to nine flits. The collected maximum end-to-end delay of the four flows obtained by simulation and our RTC method is shown in Fig. 9. To prevent the results of flow  $f_2$  from shading the results of  $f_3$ , we exchange the order of  $f_3$  and  $f_4$  in this figure. As indicated in the comparison, for the given configuration, delay calculated with our method is indeed an upper bound of the simulation results, which verifies the correctness of our methods. In addition, we also found that, the delay bound of high-priority flows (e.g.  $f_1$  and  $f_4$ ) is tighter than that of low-priority flows.

We also take the real application discussed in [46][47] as an example to demonstrate the accuracy and ability to analyze multi-flows. This application is comprised of 16 IP cores. The 26 communication flows among these 16 IPs are classified into nine groups, and each group has their bandwidth requirement. When mapped to a  $4 \times 4$  mesh network, the traffic pattern of this application is demonstrated in Fig. 10(a). We assume the flows in a group have the same injection period and priority. The flow specification of each group is listed in Table. 10(b). We set the packet size to 128 bits, and collect the end-to-end delay of each flow obtained by simulation and our method, the related results is shown in Fig. 11.

## VI. CONCLUSION

The priority-aware wormhole-switched NoC is a promising platform for the on-chip real-time communication if the worst-case performance can be accurately analyzed and guaranteed. Simulation is not well suited for this purpose because it is difficult to cover all the corner cases. In this paper, we propose an RTC based performance model to achieve this goal. We first build the traffic model and service model for this NoC,

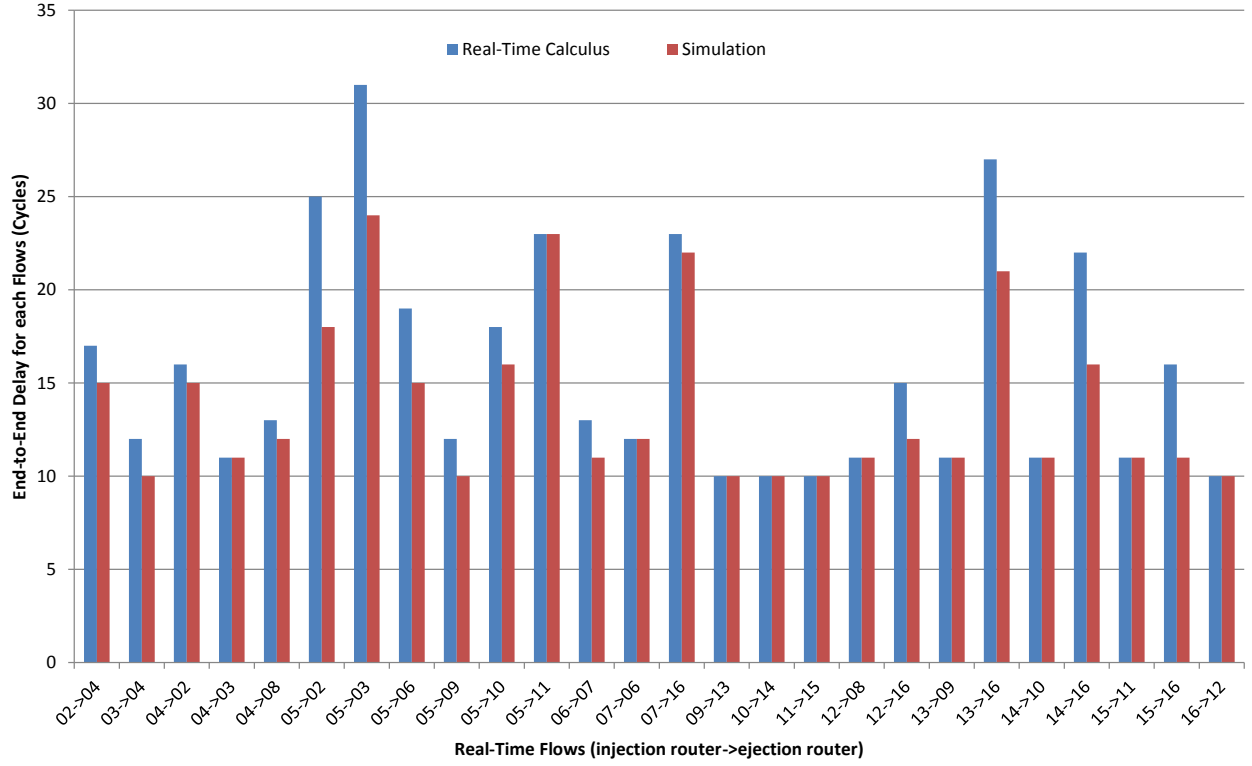


Fig. 11: Comparison with simulation

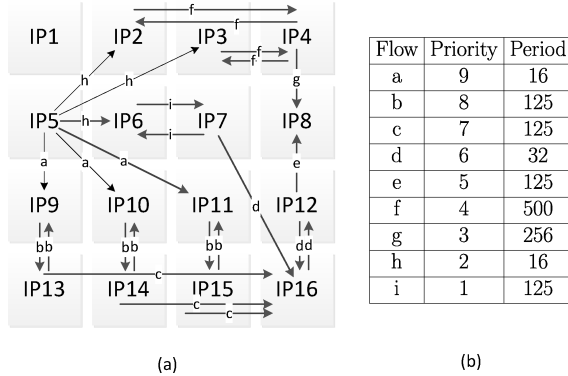


Fig. 10: Traffic pattern of ericsson radio system application

and propose a novel method to derive the upper service curve of credit-based flow control. Compared with the FLA and LLA methods which assume the router to be single cycle and free of flow control, our performance model is more general and comprehensive. Based on the proposed RTC model, we then proposed an end-to-end delay analysis algorithm and a buffer sizing algorithm. The delay analysis algorithm can be implemented to compute the end-to-end delay for each flow automatically, and verify whether all these flows meet their deadline under this configuration. To improve the calculated delay bound, we also proposed the concept of collapsible sub-path. The proposed buffer sizing algorithm can optimize the buffer size from high-priority flows to low-priority flows. It can also be implemented to perform the buffer reduction automatically under the constraint of deadline. Compared with

the DNC based performance model, our model can give tighter performance bound, because the RTC-based model takes the upper service curve and lower arrival curve into consideration. Experimental results also illustrate that our method indeed outperforms the conventional analytical methods, e.g. LLA and DNC, when the tightness of performance bound are considered. Our results can be applied to the mapping, routing and power reduction of NoC.

#### ACKNOWLEDGEMENT

The authors thank the reviewers for their suggestions and comments, and all the experiments are carried out at the Integrated Microsystem Lab (IML) of McGill University. The first author also thank Ari Ramdial at McGill University for his helpful comments. This research is supported by High Technology Research and Development Program of China (Grant No. 2012AA012201, 2012AA011902).

#### REFERENCES

- [1] E. Bolotin, Z. Guz, I. Cidon, R. Ginosar, and A. Kolodny. The power of priority: Noc based distributed cache coherency. In *Networks-on-Chip (NoCs), 2007. First International Symposium on*, pages 117–126, May 2007.
- [2] Jörn Ostermann, Jan Bormans, Peter List, Detlev Marpe, Matthias Narroschke, Fernando Pereira, Thomas Stockhammer, and Thomas Wedi. Video coding with h. 264/avc: tools, performance, and complexity. *Circuits and Systems magazine, IEEE*, 4(1):7–28, 2004.
- [3] Kees Goossens, John Dielissen, and Andrei Rădulescu. The Æthereal network on chip: Concepts, architectures, and implementations. *IEEE Design & Test of Computers*, 22(5):414–421, Sep.-Oct. 2005.
- [4] Shaoteng Liu, A. Jantsch, and Zhonghai Lu. Analysis and evaluation of circuit switched noc and packet switched noc. In *Digital System Design (DSD), 2013 Euromicro Conference on*, pages 21–28, Sept 2013.

- [5] C. Paukovits and H. Kopetz. Concepts of switching in the time-triggered network-on-chip. In *Embedded and Real-Time Computing Systems and Applications, 2008. RTCSA '08. 14th IEEE International Conference on*, pages 120–129, 2008.
- [6] M.D. Harmanci, N.P. Escudero, Y. Leblebici, and P. Ienne. Providing qos to connection-less packet-switched noc by implementing diffserv functionalities. In *System-on-Chip, 2004. International Symposium on*, pages 37–40, Nov 2004.
- [7] Zheng Shi and Alan Burns. Real-time communication analysis for on-chip networks with wormhole switching. In *Proceedings of the Second ACM/IEEE International Symposium on Networks-on-Chip (NoCs), 2008*, pages 161–170, Washington, DC, USA, 2008. IEEE Computer Society.
- [8] Byungjae Kim, Jong Kim, SungJe Hong, and Sunggu Lee. A real-time communication method for wormhole switching networks. In *Parallel Processing, 1998. International Conference on*, pages 527–534, 1998.
- [9] S.L. Hary and F. Ozguner. Feasibility test for real-time communication using wormhole routing. *Computers and Digital Techniques, IEE Proceedings -*, 144(5):273–278, 1997.
- [10] Jong-Pyng Li and Matt W. Mutka. Real-time virtual channel flow control. *Journal of Parallel and Distributed Computing*, 32(1):49 – 65, 1996.
- [11] S. Balakrishnan and F. Ozguner. A priority-driven flow control mechanism for real-time traffic in multiprocessor networks. *Parallel and Distributed Systems, IEEE Transactions on*, 9(7):664–678, 1998.
- [12] Partha Kundu. On-die interconnects for next generation cmps. In *2006 Workshop on On- and Off-Chip Interconnection Networks for Multicore Systems*, Stanford, CA, USA, December 2006.
- [13] G. Michelogiannakis, D. Sanchez, W.J. Dally, and C. Kozyrakis. Evaluating bufferless flow control for on-chip networks. In *Networks-on-Chip (NOCS), 2010 Fourth ACM/IEEE International Symposium on*, pages 9–16, May 2010.
- [14] Hany Kashif, Hiren D. Patel, and Sebastian Fischmeister. Using link-level latency analysis for path selection for real-time communication on nocs. In *Proceedings of the Asia South Pacific Design Automation Conference (ASPDAC)*, pages 499–504, Sydney, Australia, February 2012.
- [15] Yue Qian, Zhonghai Lu, and Qiang Dou. Qos scheduling for nocs: Strict priority queueing versus weighted round robin. In *Computer Design (ICCD), 2010 IEEE International Conference on*, pages 52–59, Oct 2010.
- [16] Zhonghai Lu, Axel Jantsch, and Ingo Sander. Feasibility analysis of messages for on-chip networks using wormhole routing. In *Proceedings of the Asia and South Pacific Design Automation Conference*, pages 960–964, Shanghai, China, Jan. 2005.
- [17] S. Chakraborty, S. Kunzli, and L. Thiele. A general framework for analysing system properties in platform-based embedded system designs. In *Design, Automation and Test in Europe Conference and Exhibition, 2003*, pages 190–195, 2003.
- [18] Hany Kashif and Hiren Patel. Bounding buffer space requirements for real-time priority-aware networks. In *Proceedings of the Asia South Pacific Design Automation Conference (ASPDAC)*, SunTec, Singapore, January 2014.
- [19] W. J. Dally and B. Towles. Route packets, not wires: On-chip interconnection networks. In *Proceedings of the 38th Design Automation Conference*, pages 684–689, Jun. 2001.
- [20] Evgeny Bolotin, Israel Cidon, Ran Ginosar, and Avinoam Kolodny. QNoC: QoS architecture and design process for network on chip. *Journal of Systems Architecture, Special Issue on Networks on Chip*, 50(2-3):105–128, Feb. 2004.
- [21] P. Poplavko, T. Basten, M. Bekooij, J. van Meerbergen, and B. Mesman. Task-level timing models for guaranteed performance in multiprocessor networks-on-chip. In *Proceedings of the 2003 international conference on Compilers, architecture and synthesis for embedded systems*, pages 63–72. ACM, 2003.
- [22] Yue Qian, Zhonghai Lu, and Wenhua Dou. Analysis of worst-case delay bounds for best-effort communication in wormhole networks on chip. In *Networks-on-Chip (NoCs), 2009. 3rd ACM/IEEE International Symposium on*, pages 44–53, May 2009.
- [23] Gaoming Du, Cunqiang Zhang, Zhonghai Lu, Alberto Saggio, and Minglun Gao. Worst-case performance analysis of 2-d mesh nocs using multi-path minimal routing. In *Proceedings of the Eighth IEEE/ACM/IFIP International Conference on Hardware/Software Code-sign and System Synthesis, CODES+ISSS '12*, pages 123–132, New York, NY, USA, 2012. ACM.
- [24] Sunggu Lee. Real-time wormhole channels. *J. Parallel Distrib. Comput.*, 63(3):299–311, March 2003.
- [25] D. Rahmati, S. Murali, L. Benini, F. Angiolini, G. De Micheli, and H. Sarbazi-Azad. Computing accurate performance bounds for best effort networks-on-chip. *Computers, IEEE Transactions on*, 62(3):452–467, 2013.
- [26] Ciprian Seiculescu, Dara Rahmati, Srinivasan Murali, Luca Benini, Giovanni De Micheli, and Hamid Sarbazi-Azad. Designing Best Effort Networks-on-Chip to Meet Hard Latency Constraints. *ACM Transactions on Embedded Computing Systems*, 12(4), 2013.
- [27] Zheng Shi. *Real-Time Communication Services for Networks on Chip*. PhD thesis, The University of York, 2009.
- [28] Jean-Yves Le Boudec and Patrick Thiran. *Network Calculus: A Theory of Deterministic Queuing Systems for the Internet*. Springer-Verlag, Berlin, Heidelberg, 2001.
- [29] M. Gries, C. Kulkarni, C. Sauer, and K. Keutzer. Comparing analytical modeling with simulation for network processors: a case study. In *Design, Automation and Test in Europe Conference and Exhibition, 2003*, pages 256–261 suppl., 2003.
- [30] D.B. Chokshi and P. Bhaduri. Modeling fixed priority non-preemptive scheduling with real-time calculus. In *Embedded and Real-Time Computing Systems and Applications, 2008. RTCSA '08. 14th IEEE International Conference on*, pages 387–392, 2008.
- [31] Andrei Hagiescu, Unmesh D. Bordoloi, Samarjit Chakraborty, Prahla-davaradan Sampath, P. Vignesh V. Ganesan, and S. Ramesh. Performance analysis of flexray-based ecu networks. In *Proceedings of the 44th Annual Design Automation Conference, DAC '07*, pages 284–289, New York, NY, USA, 2007. ACM.
- [32] Lothar Thiele, Ernesto Wandeler, and Samarjit Chakraborty. Performance analysis of multiprocessor dsp: a stream-oriented component model. *Signal Processing Magazine, IEEE*, 22(3):38–46, 2005.
- [33] Ernesto Wandeler and Lothar Thiele. Real-time calculus (rtc) toolbox. <http://www.mpa.ethz.ch/Rtctoolbox>, 2006.
- [34] William James Dally and Brian Towles. *Principles and Practices of Interconnection Networks*. Morgan Kaufman Publishers, San Francisco, CA, USA, 2004.
- [35] N.E. Jerger and L.S. Peh. On-chip networks. *Synthesis Lectures on Computer Architecture*, 4(1):1–141, 2009.
- [36] W.J. Dally and S.G. Tell. The even/odd synchronizer: A fast, all-digital, periodic synchronizer. In *Asynchronous Circuits and Systems (ASYNC), 2010 IEEE Symposium on*, pages 75–84, May 2010.
- [37] E. Wandeler, L. Thiele, M. Verhoef, and P. Lieverse. System architecture evaluation using modular performance analysis: a case study. *INTERNATIONAL JOURNAL ON SOFTWARE TOOLS FOR TECHNOLOGY TRANSFER (STTT)*, 8(6):649–667, 2006.
- [38] Henrik Schiøler, Jan Jakob Jessen, Jens Dalsgaard Nielsen, and Kim Guldstrand Larsen. Network calculus for real time analysis of embedded systems with cyclic task dependencies. In *Proc. 20th International Conference on Computers and Their Applications, CATA 2005*, pages 326–332, 2005.
- [39] Lothar Thiele and Nikolay Stoimenov. Modular performance analysis of cyclic dataflow graphs. In *Proceedings of the Seventh ACM International Conference on Embedded Software, EMSOFT '09*, pages 127–136, New York, NY, USA, 2009. ACM.
- [40] Zheng Shi and A. Burns. Real-time communication analysis with a priority share policy in on-chip networks. In *Real-Time Systems, 2009. ECRTS '09. 21st Euromicro Conference on*, pages 3–12, July 2009.
- [41] Jia Zhan, N. Stoimenov, Jin Ouyang, L. Thiele, V. Narayanan, and Yuan Xie. Designing energy-efficient noc for real-time embedded systems through slack optimization. In *Design Automation Conference (DAC), 2013 50th ACM / EDAC / IEEE*, pages 1–6, May 2013.
- [42] S. Manolache, P. Eles, and Zebo Peng. Buffer space optimisation with communication synthesis and traffic shaping for nocs. In *Design, Automation and Test in Europe, 2006. DATE '06. Proceedings*, volume 1, pages 1–6, March 2006.
- [43] Elena Fersman, Leonid Mokrushin, Paul Pettersson, and Wang Yi. Schedulability analysis of fixed-priority systems using timed automata. *Theoretical Computer Science*, 354(2):301 – 317, 2006. Tools and Algorithms for the Construction and Analysis of Systems (TACAS 2003) Tools and Algorithms for the Construction and Analysis of Systems 2003.
- [44] Samarjit Chakraborty, Linh T. X. Phan, and P. S. Thiagarajan. Event count automata: A state-based model for stream processing systems. In *Proceedings of the 26th IEEE International Real-Time Systems Symposium, RTSS '05*, pages 87–98, Washington, DC, USA, 2005. IEEE Computer Society.
- [45] Nan Jiang, D.U. Becker, G. Michelogiannakis, J. Balfour, B. Towles, D.E. Shaw, J. Kim, and W.J. Dally. A detailed and flexible cycle-accurate network-on-chip simulator. In *Performance Analysis of Systems and*

- Software (ISPASS), 2013 IEEE International Symposium on*, pages 86–96, April 2013.
- [46] Zhonghai Lu and Axel Jantsch. TDM virtual-circuit configuration for network-on-chip. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 16(8):1021–1034, Aug. 2008.
- [47] Fahimeh Jafari, Zhonghai Lu, Axel Jantsch, and Mohammad Hossein Yaghmaee. Buffer optimization in network-on-chip through flow regulation. *Trans. Comp.-Aided Des. Integ. Cir. Sys.*, 29(12):1973–1986, December 2010.