

Packet-Level Static Timing Analysis for NoCs

Evgeni Krimer^{1,2}, Isaac Keslassy², Avinoam Kolodny², Isask'har Walter², Mattan Erez¹

¹Department of Electrical and Computer Engineering, University of Texas at Austin

²Department of Electrical Engineering, Technion — Israel Institute of Technology

{krimer, mattan.erez}@mail.utexas.edu, {isaac@ee, kolodny@ee, zigi@tx}.technion.ac.il

Abstract

Networks-on-chip (NoCs) are used in a growing number of SoCs and multi-core processors, increasing the need for accurate and efficient modeling to aid the design of integrated systems. A methodology for packet-level static timing analysis in NoCs is presented. It enables quick and accurate gauging of the performance parameters of a virtual-channel wormhole NoC without using simulation techniques. The network model can handle any topology, link capacities, and buffer capacities. It provides per-flow analysis that is orders-of-magnitude faster than simulation while being both significantly more accurate and more complete than prior static modeling techniques. Our methodology is inspired by models of industrial flow-lines. Using a carefully derived and reduced Markov chain, the model can statically represent the dynamic network state and closely estimate the average latency of each flow. Usage of the model in a placement optimization problem is shown as an example application of the method.

1 Introduction

Networks-on-chip (NoCs) are increasingly used instead of buses and dedicated signal wires in large-scale processors and, even more so, in modern systems-on-chip (SoC) [4]. In NoC-based systems, data transmission takes the form of multi-packet flows routed through the NoC over multiple links and routers. The purpose of this paper is to rigorously derive a delay model for packet-level static timing analysis (STA) for NoC-based SoCs. Static timing analysis in a shared network is a non-simulation-based technique to estimate the average delay of each flow in the network, given the network topology, link capacities, router architecture, and the bandwidth requirements and characteristics of all flows.

The motivation for a per-flow STA technique is to enable a range of design optimizations that can rely on accurate and fast network analysis. Methods such as module placement and resource allocation [1, 34] require a large number of iterations, and thus the evaluation of network performance within each iteration must be very efficient. Until now, an accurate and complete modeling of advanced NoCs has only been possible with detailed and time-consuming simulations. The main reason is that network resources, including links, routers, buffers, and ports, are shared between several information flows. Thus, contention can arise inducing *statistical uncertainty* in the delay of each packet.

Detailed simulation, however, is too slow to be effective within an optimization inner loop because all internal buffers and states must be modeled on a cycle-by-cycle basis.

Contributions

We present a rigorous analytical model that relies on a carefully constructed and reduced Markov chain to represent network state, including the occupancy of all buffers. Our model is inspired by industrial work-flow modeling techniques and, to the best of our knowledge, is the first that can accurately account for arbitrary network topology, link capacities, and buffering, when using wormhole routing with virtual channels. We rely on the well-developed theory of stochastic processes and show that our technique faithfully predicts network queuing delay for both synthetic and real-world SoC traffic scenarios. In this paper we limit the analysis to packets that have random arrival times according to a Poisson distribution. We present results and validate the model for the delay analysis of flows with fixed-length packets that are composed of a large number of flits. We discuss extensions to these assumptions as future work.

To summarize our contributions:

- We present the first rigorous NoC model that is based on stochastic theory and show how to represent and solve for the network state using a Markov chain.
- We show how to account for arbitrary and finite buffering, as well as support wormhole routing and virtual channels. We use network delay analysis as an illustrative example of the modeling technique
- We validate our model using synthetic and real-world scenarios, and discuss why it is more complete and more accurate than prior analytical models.
- We demonstrate that our model can serve at the core of a design optimization method by showing that it can faithfully choose between multiple placement options in a real-world SoC example, and do so while requiring *orders of magnitude* less time than simulation. We also show that the most advanced prior-art model fails to make the correct optimization decision.

The rest of the work is organized as follows. We start by discussing the related work in Section 3. Then, in Section 4, we establish a general analytical model for the average delay of each flow in a general NoC topology. We evaluate the delay model in Section 5 by comparing it with accurate simulation results and