Performance Evaluation of Multicore Systems: From Traffic Analysis to Latency Predictions (Embedded Tutorial)

Zhiliang Qian\textsuperscript{1}, Paul Bogdan\textsuperscript{2}, Chi-Ying Tsui\textsuperscript{1} and Radu Marculescu\textsuperscript{3}

\textsuperscript{1}Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong
\textsuperscript{2}Department of Electrical Engineering, University of Southern California, Los Angeles
\textsuperscript{3}Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh

ABSTRACT
As technology scaling down allows multiple processing components to be integrated on a single chip, the modern computing systems led to the advent of Multiprocessor System-on-Chip (MPSoC) and Chip Multiprocessor (CMP) design. Network-on-Chips (NoCs) have been proposed as a promising solution to tackle the complex on-chip communication problems on these multicore platforms. In order to optimize the NoC-based multicore system design, it is essential to evaluate the NoC performance with respect to numerous configurations in a large design space. Taking the traffic characteristics into account and using an appropriate latency model become crucially important to provide an accurate and fast evaluation. In this tutorial, we survey the current progresses in these aspects. We first review the NoC workload modeling and traffic analysis techniques. Then, we discuss the mathematical formalisms of evaluating the performance under a given traffic model, for both the average and worst-case latency predictions. Finally, the advantages of combining the analytical and simulation-based techniques are discussed and new attempts for bridging these two approaches are reviewed.

1. INTRODUCTION
With technology scaling down, more and more components can be integrated on a single chip. As a result, the modern computing systems led to the advent of Multiprocessor System-on-Chip (MPSoC) and Chip Multiprocessor (CMP) designs. Network-on-Chips (NoCs) have been proposed as a modular and scalable solution for future on chip communications [1]. The design of NoC begins with a specification of performance requirements together with some cost constraints. These performance metrics drive the choice of NoC design parameters, such as the routing algorithm, the buffer sizing and the flow control scheme etc [2]. When exploring the NoC design space, it is of the utmost importance to provide an accurate and fast evaluation with respect to different configurations. Both the traffic variation over time and the latency models are required to capture the router delay behaviors under various traffic scenarios. Traditional NoC performance evaluations are predominantly based on the Poisson traffic characteristics [3], i.e., the packet inter-arrival times are exponentially distributed. Typically, the packet service time at each router is also assumed to be exponentially distributed. However, recent researches have demonstrated these assumptions may not hold for many NoC applications [4, 5, 6]. New approaches incorporating these considerations are required to provide more accurate latency/throughput estimations.

The purpose of this embedded tutorial is to present recent advances in performance evaluations of NoC-based multicore systems. Towards this end, we first identify the NoC dynamics as being highly dependent on the traffic patterns variation [5]. Second, we discuss several types of workload models that are able to capture the burstiness and long range dependence (LRD) in traffic behavior [5, 7]. Then we discuss the NoC latency prediction models. According to the modeling methodology, NoC performance estimation tools can be categorized into analytical-based and simulation-based models. Analytical techniques for both the average [8] and the worst-case [9, 10] performance metrics are needed to cater for different application requirements. In addition to the analytical-based and simulation-based approaches, an open research direction is to combine these two techniques together to exploit the advantages of the high fidelity in simulations and fast speed in mathematical models. Towards this end, we survey several new attempts that utilize machine learning techniques [11] and hardware modeling [12] to accelerate the performance estimation.

2. NOC TRAFFIC ANALYSIS
2.1 NoC workloads
To evaluate an NoC design, the most appropriate traffic patterns to use are those generated directly by running the intended applications on the platform that models both the processor cores and the NoC infrastructure. Execution-driven workload is such an approach which emulates the processors in addition to the NoC itself. Although accurate, the execution-driven workload requires a full-system implementation and suffers from long evaluation time. Trace-based workload is an efficient alternative under realistic applications. It only evaluates the network model while treating the processor core as a “black-box” that only generates packets
queuing-theory-based models have gained the most popularity. Among these three techniques, the model the average latency in NoC: the probabilistic models, design constraints. Three types of techniques are used to provide the highest performance at a given cost, towards this end, a mean exponent, but rather by a series of interwoven time scales traffic variation cannot be characterized by a single Hurst

2.2 Statistical traffic models for LRD traffic

The statistical traffic models are important in two aspects: 1) They can help analyze and characterize NoC applications. 2) They can be used to generate new traffic traces which possess features that are not covered by existing applications.

Two types of statistical techniques have been developed to take the traffic self-similarity (i.e., LRD) into consideration. The first one utilizes Hurst parameter $H$ to capture the temporal dependencies [4, 7]. $H$ parameter of an application is fitted by applying the variance-time or R/S statistics [4] techniques on the application traffic traces, while Fractional Gaussian Noise (FGN) model is utilized [4] to generate a new synthetic trace. The second method is based on the characterization of the traffic using a Markov-modulated Poisson process (MMPP) [8, 16] or a more generalized Markov arrival process (MAP) which models LRD as the superposition of several simple processes.

2.3 Non-stationary NoC traffic modeling

The single exponent (e.g., Hurst parameter $H$) or MMPP based traffic modeling usually focus on matching the first and second moments (i.e., the mean and variance) as well as the first-order autocorrelations (i.e., $\rho_1$) of the time series $X_t$ based on the assumption that $X_t$ is wide-sense stationary (WSS) [4]. These techniques correspond to monofractal traffic analysis. On the other hand, the state-of-the-art NoC traffic analysis [5] reveals highly nonlinear and nonstationary behaviors in the traffic due to the heterogeneous traffic sources and continuous changes in the pool of applications running on a multicore platform. Consequently, the traffic variation cannot be characterized by a single Hurst exponent, but rather by a series of interwoven time scales defined by several such exponents. Towards this end, a mean field (MF) traffic modeling technique [5] is proposed to analyze the multifractal spectrum by characterizing the higher order moments in the traffic.

3. NOC ANALYTICAL MODELS

3.1 Average-case performance models

For the NoCs with best-effort service, the design goal is to provide the highest performance at a given cost, i.e., maximizing the average-case performance metrics within the design constraints. Three types of techniques are used to model the average latency in NoC: the probabilistic models, the network calculus models [17], and the queuing-theory-based models [2, 8, 16]. Among these three techniques, the queuing-theory-based models have gained the most popularity in the average-case performance prediction because of its higher accuracy. A variety of queuing models have been developed to compute the NoC latency based on various traffic models, such as the $M/M/1/K$ [18], $M/M/c/K$ [3], $M/G/1/K$ [19], $G/G/1$ [8], $GE/G/1/K$ [20] and MMPP/G/1 [16] models. Some of these models target at homogeneous virtual-channel architectures [19, 21] while others are designed for routers with heterogeneous virtual channels [3]. Moreover, some models are developed for the cases where packet length is smaller than the buffer size [2, 19], while the others are derived for packets with extreme long length such that it can span a set of router buffers [3].

3.2 Worst-case performance models

In real-time systems with guaranteed service, the worst-case delay is of particular concern since it is essential to guarantee that tasks will always be finished before the deadline. Two types of mathematical formalisms are used to derive the delay bounds in QoS NoCs. The first type is based on the framework of network calculus (NC) [10]. The second type is based on real-time bound (RTB) formulations [9]. Network calculus is based on the min-plus algebra to obtain the output function from the convolution of the input arrival function and the system transfer function. On the other hand, the real-time bound (RTB) formulation is based on the fact that, the worst-case packet transfer scenario happens when all the intermediate buffers along the route are full, and the target flow loses arbitration at all routers against the other contending flows [9].

4. BRIDGING ANALYTICAL MODELS AND SIMULATIONS

4.1 FPGA-based acceleration engine

Instead of evaluating the NoC using software program, the recent trend is to use hardware for acceleration. However, different from the traditional approaches which implement the real NoC on hardware (ASIC or FPGA), these new approaches try to abstract the router model at a higher level (for example the mathematical models in [12] or the behavioral models in [22]) and implement the abstracted model on hardware so as to reduce the evaluation time.

4.2 Learning-based performance evaluation

Most NoC latency models are based on the classical queuing theory and treat each input channel as a queuing system. These models provide good estimations when the underlying assumptions, such as the Poisson distribution at all traffic sources [2] or a single flit buffer size [3], hold. However, in real applications, these assumptions may not hold and the accuracy of the analytical model is compromised. In order to relax some of the assumptions above, in [11], a machine-learning based technique is used to model the NoC latency. Given the target NoC platform, a learning-based method is used to create an estimation model by training with datasets obtained from the latency profiles first. Support vector regression (SVR) and cross validation (CV) are utilized to obtain the channel and source queuing models, respectively. In the prediction stage, the learned model is applied to the new traffic patterns to evaluate the latency performance. Compared to the analytical models, better accuracy is achieved in the learning-based model.
5. CONCLUSIONS

In this paper, we have summarized a few representative approaches for NoC performance evaluation ranging from traffic modeling to analytical and simulation-based models for latency predictions. We have also discussed the state-of-the-art approaches that attempt to bridge the analytical and simulation-based approaches so as to provide both fast and accurate estimations. As NoC has already become a practical solution for handling future multicore communications, the performance evaluation is essential in exploring the NoC design space to obtain the most suitable architecture. One of the promising research directions in future is to further combine the advantages of several performance evaluation methods for better speed and accuracy trade-offs.

6. REFERENCES


