Computers and Electrical Engineering 40 (2014) 260–275

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

User satisfaction aware routing and energy modeling of polymorphic network on chip architecture ☆

Swamy D. Ponpandi, Akhilesh Tyagi *

Dept. of Electrical & Computer Engg., Iowa State University, Ames, Iowa, USA

ARTICLE INFO

Article history:
Received 3 March 2014
Received in revised form 21 August 2014
Accepted 26 August 2014
Available online 2 October 2014

Keywords:
User satisfaction
Sigmoid function
Virtual channels
Energy model
Communication flow
Routing

ABSTRACT

In mobile devices, multiple applications contend for limited resources in the underlying embedded system framework. Application resource requirements in mobile systems vary by computation needs, energy consumption and user interaction frequency. Quality of service (QoS) is the predominant metric of choice to manage resources among contending applications. Resource allocation policies to support static QoS for applications do not reflect the changing demands of the user in contemporary network on chip (NoC) based embedded architectures. User satisfaction with the user interactions and user interface design ought to be the primary design driver. Some recent research has integrated a saturating, non-linear user satisfaction function in the application thread scheduler. The application and operating system level user satisfaction research assumes that the throughput of inter-thread edges is limited only by the computational constraints of the nodes. With NoC, however, NoC resource allocation policies play an important role in determining the inter-thread communication flow's throughput and the resulting application level user satisfaction. In this paper, we filter down the user satisfaction from an application layer attribute to a router level attribute to improve the resource and energy utilization for routing in order to leverage the user satisfaction at the application and system level. We demonstrate that this technique improves the user satisfaction of audio (MP3) application by 10% while maintaining the user satisfaction of video (MPEG-2) application. Experiments also show that a fixed energy source can be extended for an average of 18% of the time using the NoC user satisfaction based energy optimization proposed in this research.

1. Introduction

Mobile devices are becoming ubiquitous. Their primary role is to serve as an enhanced user interface (UI) device to cloud based services. UI is the primary driver of application design in mobile embedded systems. UI transactions are at the heart of a modern mobile operating systems (OS) such as Android or iOS. These transactions are subject to strict constraints. Android and iOS enforce UI response constraints by terminating applications that fail to adhere to these strict constraints. For example, the UI should respond within 5 s for any user input event to avoid getting terminated in iOS mobile framework. It is important to note that these constraints actually reflect the nature of a broad class of mobile users. Such constraints are implemented with smart resource allocation policies by the underlying framework (App development, Mobile OS, Hardware

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Associate Editor Dr. Maurizio Palesi.
* Corresponding author.
E-mail addresses: swamy@iastate.edu (S.D. Ponpandi), tyagi@iastate.edu (A. Tyagi).

http://dx.doi.org/10.1016/j.compeleceng.2014.08.012
0045-7906/© 2014 Elsevier Ltd. All rights reserved.

architecture, Communication network) to guarantee conformity. Most users will be annoyed with an unresponsive app, though a relative minority may not have a similar response. Implicitly, this constraint models the behavior of a class of users to the response time of an app in the mobile system.

Consequently, *user satisfaction* with the UI interactions becomes a more dominant design goal than the traditional parameters such as computing time, bandwidth or even energy. Although this user satisfaction also depends on the traditional computing parameters of computing time, bandwidth and energy; due to human nature the relationship between user satisfaction and resources is non-linear and saturating. For instance, allocating additional resources beyond a threshold does not necessarily enhance the MP3 playback quality for the user. Prior works [1,2] have modeled the end user as a design parameter and used non-linear, saturating function for user satisfaction in the operating system scheduler and other system layers. Modern mobile devices deploy rich and complex systems on chip (SoC) that include or are likely to include NoC [3,4] for inter-core communication. For fine-grain threading, NoC bound traffic represents critical edges in computation. Hence, NoC performance is strongly coupled with user satisfaction. User satisfaction can be asymmetric along multiple applications, and within an application it can be saturating. However, at NoC layer, we tend to allocate resources to improve overall bandwidth or latency until resources such as virtual channels (VC) are exhausted. If the user satisfaction model were to permeate into the NoC layer, research has to address the two issues – (1) translate user satisfaction metric at application level into user satisfaction metrics at NoC layer (2) develop routing heuristics to use the user satisfaction metrics at NoC layer. The focus of this paper is in exploring a holistic user satisfaction optimization at the NoC layer.

We make three contributions in this research work. First, we propose user satisfaction as a metric to measure and optimize the utilization of router buffers using virtual channel abstraction. Second, we describe the modifications required for virtual channel allocation to support the user satisfaction based scheme. Third, we develop a simple energy model for thread communication in the NoC and apply user satisfaction to optimize energy utilization for a critical set of applications. Application layer uses the services of the NoC layer. The user satisfaction model provides a unified *QoS like* metric for design optimization at the application layer and the network layer. Most prior efforts reviewed in Section 10 related to virtual channel allocation have limited the focus exclusively at the NoC layer ignoring the effects of such solutions on the higher level abstraction layers such as the application layer. Our *QoS like* metric extends seamlessly to the network layer from the application layer. As far as we know, our work is the first attempt to integrate user model into the NoC layer resource optimization and energy modeling for communication flows.

The remainder of this paper is organized into ten sections. We start with a discussion of user satisfaction metric and its relation to QoS of network on chip framework in Section 2. A simple mathematical model to capture the behavior of a large class of users is described in Section 3. We further motivate the user satisfaction model for virtual channel allocation in Section 4. The throttling factor for communication flows and objective function for user satisfaction is derived in Section 5. We develop energy utilization model for application threads in Section 6. Section 7 contains discussion of our virtual channel allocation and arbitration policy heuristic based on the optimization of user satisfaction. Our polymorphic NoC simulation environment is described in Section 8. Results of experiments are discussed in relation to the proposed scheme in Section 9. Related research work is surveyed and reviewed in Section 10. Finally, in Section 11, we conclude by summarizing the contributions and future work.

2. User satisfaction metric versus quality of service in NoC

NoC based circuit design maps global design goals, of the whole system on chip, into local design choices at local domains. Domains may be a single logic block, a group of logic blocks, or more complex CPU, FPGA. NoC also raises the level of communication abstraction to transaction level enabling local domains communicate by passing messages [3]. Fig. 1 shows a 9 node NoC mesh communication system organized as a 3×3 grid.

Each node in Fig. 1 is a local domain which can be independently designed in the SoC. Communication between nodes is facilitated by routers which can buffer and exchange messages. The major elements of NoC router are first in first out (FIFO) queue buffers, virtual channels [5], arbitration protocols and output port buffers. A magnified view of the router on the right side of Fig. 1 depicts the various internal elements. The router in this example has two input virtual channels per input port which are shown in Fig. 1. Output ports are not shown for clarity. Each FIFO queue buffer has a depth of four. The simplified control logic consists of three stages – virtual channel allocation (VC), arbitration for output ports (ARB) and port allocation (PORT). Multiple routers lie on the critical path for SoCs using on chip communication networks.

Design policies for various stages in the control logic and allocation policies for buffers have been explored in several works reviewed in Section 10. Buffers are major energy hog in the on chip communication infrastructure. Maximizing utilization of buffers per unit energy spent is of utmost importance to achieve higher throughput of the SoC designs.

Static QoS policy for applications restricts virtual channel allocation flexibility and hence, under-utilizes NoC buffers in two major ways. First, by statically allocating VCs and associated buffers at application level routers are deprived of the flexibility to reallocate free buffers dynamically. Second, the static QoS policies do not reflect the user behavior in embedded systems using NoC paradigm. As compared in Fig. 2, static QoS allocation policies made at the application layer are difficult to permeate to the NoC layer due to the lack of appropriate model which can be used at both the layers. Existing QoS policy at application level do not provide any control knobs for the NoC layer to improve performance. In fact, each inter-thread communication edge is viewed as an atomic step at this layer. Fig. 2 shows a possible resource allocation among four