Design of Low-Power Coarse-Grained Reconfigurable Architectures

Yoonjin Kim
Rabi N. Mahapatra

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Design of Low-Power Coarse-Grained Reconfigurable Architectures

Design of Low-Power Coarse-Grained Reconfigurable Architectures

Yoonjin Kim
Rabi N. Mahapatra


CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-4398-2511-2 (Ebook-PDF)

Contents

List of Figures

List of Tables

Preface

Application-specific optimization of embedded systems becomes inevitable to satisfy the market demand for designers to meet tighter constraints on cost, performance and power. On the other hand, the flexibility of a system is also important to accommodate the short time-to-market requirements for embedded systems. To compromise these incompatible demands, coarse-grained reconfigurable architecture (CGRA) has emerged as a suitable solution. A typical CGRA requires many processing elements (PEs) and a configuration cache for reconfiguration of its PE array. However, such a structure consumes significant area and power. Therefore, designing cost-effective CGRA has been a serious concern for reliability of CGRA-based embedded systems.

As an effort to provide such cost-effective design, the first half of this book focuses on reducing power in the configuration cache. For power saving in the configuration cache, a *low-power reconfiguration technique* is presented based on reusable context pipelining achieved by merging the concept of context reuse into context pipelining. In addition, we propose *dynamic context compression* capable of supporting only required bits of the context words set to enable and the redundant bits set to disable. Finally, we provide *dynamic context management* capable of reducing power consumption in configuration cache by controlling a read/write operation of the redundant context words.

In the second part of this book, we focus on designing a cost-effective PE array to reduce area and power. For area and power saving in a PE array, we devise a *cost-effective array fabric* that addresses novel rearrangement of processing elements and their interconnection designs to reduce area and power consumption. In addition, *hierarchical reconfigurable computing arrays* are proposed consisting of two reconfigurable computing blocks with two types of communication structure together. The two computing blocks have shared critical resources and such a sharing structure provides efficient communication interface between them with reducing overall area. Based on the proposed design approaches, a CGRA combining the multiple design schemes is shown to verify the synergy effect of the integrated approach.

Audience for This Book

This book is intended for computer professionals, graduate students, and advanced undergraduates who need to understand issues involved in designing and constructing embedded systems. The reader is assumed to have had introductory courses in digital system, VLSI design, computer architecture, or equivalent work experience.

Chapter 1

Introduction

1.1 Coarse-Grained Reconfigurable Architecture

With the growing demand for high quality multimedia, especially over portable media, there has been continuous development on more sophisticated algorithms for audio, video, and graphics processing. These algorithms have the characteristics of data-intensive computation of high complexity. For such applications, we can consider two extreme approaches to implementation: software running on a general purpose processor and hardware in the form of Application-Specific Integrated Circuit (ASIC). In the case of general purpose processor, it is flexible enough to support various applications but may not provide sufficient performance to cope with the complexity of the applications. In the case of ASIC, we can optimize best in terms of power and performance but only for a specific application. With a coarse-grained reconfigurable architecture (CGRA), we can take advantage of the above two approaches. This architecture has higher performance level than general purpose processor and wider applicability than ASIC.

As the market pressure of embedded systems compels the designer to meet tighter constraints on cost, performance, and power, the application specific optimization of a system becomes inevitable. On the other hand, the flexibility of a system is also important to accommodate rapidly changing consumer needs. To compromise these incompatible demands, domain-specific design is focused on as a suitable solution for recent embedded systems. Coarse-grained reconfigurable architecture is the very domain-specific design in that it can boost the performance by adopting specific hardware engines while it can be reconfigured to adapt to ever-changing characteristics of the applications.

Typically, a CGRA consists of a main processor, a Reconfigurable Array Architecture (RAA), and their interface as Figure 1.1. The RAA has identical processing elements (PEs) containing functional units and a few storage units such as ALU, multiplier, shifter and register file. The data buffer provides operand data to PE array through a high-bandwidth data bus. The configuration cache (or context memory) stores the context words used for configuring the PE array elements. The context register between a PE and a cache element (CE) in configuration cache is used to keep the cache access path from being the critical path of the CGRA.

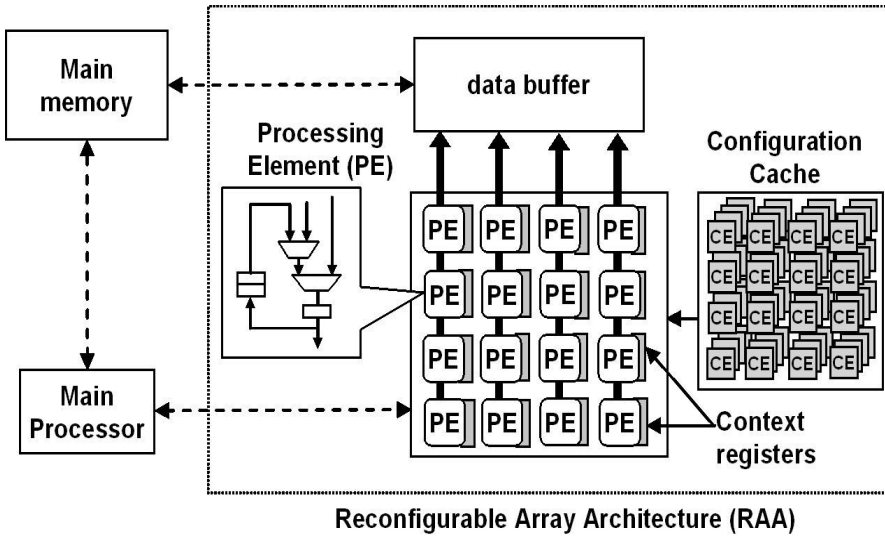

FIGURE 1.1: Block diagram of general CGRA.

Unlike FPGA (most typical of a fine-grained reconfigurable architecture), which are built with bit-level configurable logic blocks (CLBs), CGRA is built with PEs, which are word-level configurable functional blocks. By raising the granularity of operations from a bit to a word, CGRA can improve on the speed and the performance as well as the resource utilization for compute-intensive applications. Another consequence of this raised granularity is that whereas FPGA can be used for implementing any digital circuits, CGRA is targeted only for a limited set of applications, although different CGRAs may target different application domains. Still, CGRA retains the idea of "reprogrammable hardware" in the reprogrammable interconnects as well as in the configurable functional blocks (i.e., PEs). Moreover, since the amount of the configuration bit-stream is greatly reduced through the raised granularity, the configuration can be actually changed even at the runtime very fast. Most of the CGRAs feature single-cycle configuration change, fetching the configuration data from a distributed local cache. This unique combination of efficiency and flexibility, which is the main advantage of CGRA, explains an evaluation result [9] that under certain conditions CGRAs are actually more cost-effective for wireless communication applications than alternatives such as FPGA implementations as well as DSP architectures. It is worth mentioning that the improved efficiency of CGRAs in terms of the performance and flexibility is a result of the architecture specialization for compute-intensive applications.

In spite of the above advantages, the deployment of CGRA is prohibitive due to its significant area and power consumption. This is due to the fact that CGRA is composed of several memory components and the array of many processing elements including ALU, multiplier and divider, etc. Especially,

processing element (PE) array occupies most of the area and consumes most of the power in the system to support flexibility and high performance. Therefore, reducing area and power consumption in the PE array has been a serious concern for the adoption of CGRA.

1.2 Objective and Approach

This book explores the problem of reducing area and power in CGRA based on architecture optimization. To provide cost-effective CGRA design, the following questions are considered.

- How to reduce area and power consumption in CGRA? For power saving in CGRA, we should obtain area and power breakdown data of CGRA to identify area and power-dominant components. Then the components may be optimized for area and power by removing redundancies of CGRA wasting area and power. Such redundancies may depend on the characteristics of computation model or applications.
- How to design cost-effective CGRA with non-sacrificing or enhancing performance? Ultimately, the goals of designing cost-effective CGRA is that proposed approaches do not cause performance degradation with saving area and power. It means that the proposed cost-effective CGRA keeps original functionality of CGRA intact and does not increase critical path delay. In addition, the performance may be enhanced by optimizing the performance bottleneck with keeping the area and power-efficient approaches.

In this book, these central questions are addressed for area/power-critical components of CGRA and we suggest new frameworks to achieve these goals. The validation of the proposed approaches is demonstrated through the use of real application benchmarks and gate level simulations.

1.3 Overview of the Book's Contents

This volume is divided into 11 chapters as follows:

- **Chapter 1. Introduction**
  This chapter introduces general characteristics of Coarse-Grained Reconfigurable Architecture (CGRA). In addition we present the contribution and the organization of this book.