Computers and Electrical Engineering 62 (2017) 1–16

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Research on tridiagonal matrix solver design based on a combination of processors☆

Jingmei Li[a], Zhigao Zheng[b,*], Qiao Tian[a], Guoyin Zhang[a], Fangyuan Zheng[a], Yuanyuan Pan[a]

[a] *College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China*
[b] *School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 470074, China*

A R T I C L E I N F O

Article history:
Received 28 February 2017
Revised 18 July 2017
Accepted 20 July 2017

Keywords:
Tridiagonal matrix solvers
Heterogeneous systems
Central processing unit
Graphics processing unit
Extended algorithms

A B S T R A C T

Large-scale tridiagonal matrix solvers based on heterogeneous systems currently cannot balance computational efficiency and numerical stability when solving a non-diagonally dominant matrix. A tridiagonal solver combined central processing unit with graphics processing unit is proposed, based on $SPIKE^2$ as a solver framework, a simplified SPIKE algorithm as a central processing unit component, and a diagonal pivot algorithm as a graphics processing unit component. The solver performance is further improved by using a data-layout-transformation mechanism to obtain continuous addresses, reducing memory communication using constant memory to store unchanged data in the calculation process, and employing a kernel-fusion mechanism to reduce power consumption of graphics processing unit. For a diagonally dominant matrix, extended Thomas algorithms and cycle reduction to replace the graphics processing unit component are proposed in the solver. Experimental results show that the tridiagonal matrix solver in this paper can effectively consider both numerical stability and computational efficiency, and reduce total power consumption while improving memory efficiency.

1. Introduction

In many scientific computing and engineering applications, the final numerical calculation is summed up in the form of linear equations to solve one or more large-scale sparse matrices. Most typical is a tridiagonal matrix, which plays an important role in these applications. A tridiagonal matrix solver is used in computer graphics [1–3], fluid mechanics [2,4,5], Poisson solvers [6], semi-coarsening multigrids [7,8] and atmospheric simulations [9], where the matrix scale can reach the order of millions or even higher. Past algorithms for solving linear equations are mostly executed on a serial system. However, the computations are becoming increasingly complex, and the associated matrix scale is growing. Dramatic increases in calculations and storage requirements have resulted in rapidly expanding consumption of computer resources and computing time. Consequently, solving a tridiagonal matrix based on a serial system cannot meet the demand, hence parallel devices and parallel computing are developing rapidly.

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Associate Editor Dr. R.Varatharajan.

* Corresponding author.
E-mail addresses: zhengzhigao@hust.edu.cn (Z. Zheng), zhangguoyin@hrbeu.edu.cn (G. Zhang).

http://dx.doi.org/10.1016/j.compeleceng.2017.07.014

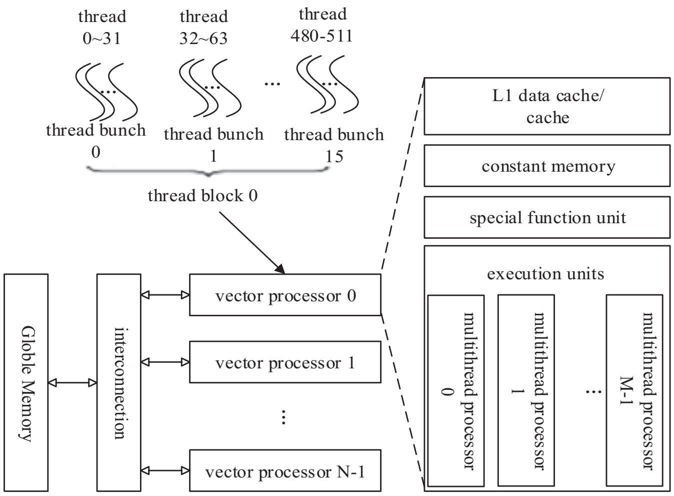

Fig. 1. The structure of a GPU.

Research of graphics processing units (GPUs) provide a new idea for solving a tridiagonal matrix in parallel. Due to the limitations of GPU architecture and memory-access methods, the parallel execution of a tridiagonal matrix solver must break the inherent linear dependency of the equations. In 2011, Davidson et al. proposed a "register packing" approach that can reduce communication of shared memory [10]. Daniel Egloff proposed a parallel cyclic reduction (PCR) algorithm for solving the finite difference partial differential equation (PDE) solver [11]. Nikolai Sakharnykh proposed a thread-level parallel Thomas algorithm combined with PCR to form a PCR-pThomas algorithm, and implemented it on a GPU [12]. Although the serial computation method for solving linear algebraic equations is mature, the serial algorithm cannot be directly used for parallel computation. Therefore, parallelism is a key factor when selecting an algorithm for a GPU.

When a central processing unit (CPU) and GPU develop steadily and independently, their combination can more practically exploit their respective advantages. As current solvers based on heterogeneous systems cannot consider efficiency and stability when solving a large-scale tridiagonal matrix, this paper proposes a strong, comprehensively performing tridiagonal solver combined central processing unit with graphics processing unit (T-SCG) solver that integrates a $SPIKE^2$ framework, a simplified SPIKE algorithm, and a diagonal-pivot algorithm to ensure numerical stability. In addition, three optimization mechanisms are proposed for the solver's GPU component to improve its overall performance.

This paper includes five parts. The first part explains the purpose of the research. The second part introduces the structure of the GPU and typical tridiagonal solving algorithms, such as Thomas, cyclic reduction (CR), and SPIKE, laying the groundwork for the improved solver. The third part shows how the proposed T-SCG solver combines the $SPIKE^2$ framework with the simplified SPIKE algorithm. The fourth part covers the optimization of the T-SCG solver with three different methods. The last part outlines the experiments used to prove the performance of the T-SCG solver.

2. GPU structure and matrix algorithm

2.1. GPU structure

The typical structure of a modern GPU is shown in Fig. 1. The GPU consists of N vector processors, each containing M multithread processors, one or more special function units, on-chip level-one (L1) data cache/cache [13], and read-only constant memory.

The function executed on a GPU device is called a kernel, which is the main function of parallel computing. The kernel is written in single-instruction multiple threads (SIMTs), runs on compute unified device architecture (CUDA) or OpenCL, where one thread executes an instance of the kernel. L1 cache is the trend in modern GPUs, which are shared by all thread blocks in a streaming multiprocessor.

2.2. Tridiagonal matrix algorithm

The tridiagonal matrix is a special band matrix; its form is similar to that of matrix A in Eq. (1). The main function of a tridiagonal matrix solver is to solve the equation $AX = F$.