Advances in Applied Mathematics and Mechanics
Adv. Appl. Math. Mech., Vol. **18**, No. 5, pp. 1421-1452

DOI: 10.4208/aamm.OA-2025-0046
October 2026

Optimizing Thermal Lattice Boltzmann Method on an MT-3000 Processor with Neo-Heterogeneous Strategies

Qingyang Zhang[1,2], Lei Xu[3,4], Rongliang Chen[3,4], Hang Zou[1,2], Bo Yang[1,2], Jingzhi Li[5,*] and Jie Liu[1,2,*]

[1] *Science and Technology on Parallel and Distributed Processing Laboratory, National University of Defense Technology, Changsha, Hunan 410073, China*
[2] *Laboratory of Digitizing Software for Frontier Equipment, National University of Defense Technology, Changsha, Hunan 410073, China*
[3] *Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China*
[4] *Faculty of Computer Science and Control Engineering, Shenzhen University of Advanced Technology, Shenzhen, Guangdong 518106, China*
[5] *Department of Mathematics & National Center for Applied Mathematics Shenzhen & SUSTech International Center for Mathematics, Southern University of Science and Technology, Shenzhen, Guangdong 518055, China*

Received 25 February 2025; Accepted (in revised version) 21 October 2025

Abstract. The *Lattice Boltzmann Method* (LBM) is a computational fluid dynamics method for simulating fluid flows with the benefits of locality and simplicity, making it ideal for parallel computing and complex flow simulations. This study focuses on developing a specialized *Double Distribution Function* (DDF) LBM software framework optimized for the MT-3000, a novel heterogeneous processor, to facilitate thermal incompressible flow simulations. To improve LBM's performance on the complex multi-zone architecture of MT-3000, this paper introduces several innovative strategies. Firstly, a temporal fusion optimization strategy is implemented. This strategy involves postponing the temperature field calculations during time steps, efficiently decreasing the time overhead. Furthermore, we present "Pencil-H", a novel pipelined algorithm meticulously designed to harness the unique capabilities of the MT-3000, thereby enhancing computational efficiency and communication effectiveness. Additionally, an architecture-aware multi-level parallelization algorithm is proposed, tailored to maximize the computational capabilities of the MT-3000. The effectiveness of these optimization strategies has been thoroughly validated through extensive benchmarking tests. These validations have shown remarkable performance enhancements, including a significant acceleration factor of 32.02X when compared to using 16 CPU cores. Notably, the optimized code demonstrated high-fidelity simulation capabilities

*Corresponding author.
Emails: li.jz@sustech.edu.cn (J. Li), liujie@nudt.edu.cn (J. Liu)

for thermal incompressible flows, achieving 61.61% of the theoretical maximum performance defined by the roofline model.

AMS: 65Y05, 68W10

Key words: Lattice Boltzmann method, thermal incompressible flow, heterogeneous processor, parallel algorithm.

1 Introduction

Computational Fluid Dynamics (CFD) plays an indispensable role in understanding and analyzing fluid behavior across a spectrum of scientific fields, such as thermodynamics, biomechanics, aerodynamics, and environmental science. Within the scope of CFD, the LBM has emerged as a robust computational approach, especially adept at addressing challenges related to incompressible flows [1–10]. In recent years, the lattice Boltzmann method (LBM) has witnessed a substantial rise in its application for thermal fluid simulations, particularly in low-Mach number scenarios [11–14]. It has made significant advancements like two-phase flows [15], reactor safety analysis [16], electrothermal convection [17], and radiative convection [18]. However, as simulation complexity increases, the demand for computational power and memory capacity also rises. These challenges requires advanced optimization techniques to effectively manage the increased computational load, which are crucial for maintaining the practicality and effectiveness of LBM in complex thermal fluid problems [19–23].

High Performance Computing (HPC) technologies provide an excellent platform for efficiently executing LBM simulations [24]. Efficient discretization of complex equations is a core challenge in high-performance computing [25]. The inherent particulate nature of LBM, along with its focus on local dynamics, makes it suitable for parallelization, especially on multi-core or many-core architectures. Using GPUs for LBM simulations has gained popularity, resulting in notable performance improvements compared to conventional CPUs. Various strategies have been investigated, encompassing increased GPU multiprocessors, memory-efficient techniques, and optimized data access patterns. Kraus et al. [26] optimized the compute-intensive part of LBM code for multi-GPU systems, breaking the double-precision Tflops barrier on a single-host system equipped with two GPUs. Tran et al. [27] parallelized LBM on GPUs, employing memory-efficient techniques and optimization strategies to minimize register usage. Herschlag et al. [28] examined GPU data access patterns in complex LBM simulations, achieving near-optimal strong results for LBM simulations with arterial geometries. Vardhan and Gounley (2019, 2021) [29,30] introduced a novel implementation of regularized LBM that reduces memory usage by storing only macroscopic moment-based data, demonstrating its efficacy on GPU architecture. Mahmoud et al. [31] proposed a GPU-optimized implementation of the nonuniform grid refinement technique within the context of the LBM, enabling the simulation of unprecedentedly large domains on a single GPU. Feichtinger et