A Naive D3Q27 Lattice Boltzmann Implementation on the Cerebras CS-3: Comparison With GPUs

Kouki Sakai
ce255006@meiji.ac.jp
School of Science and Technology, Meiji University
Kawasaki-shi, Kanagawa, Japan

Saho Orihara
ce255032@meiji.ac.jp
School of Science and Technology, Meiji University
Kawasaki-shi, Kanagawa, Japan

Takashi Shimokawabe
shimokawabe@cc.u-tokyo.ac.jp
Information Technology Center, The University of Tokyo
Kashiwa, Chiba, Japan

Takaaki Miyajima
takaaki.miyajima@cs.meiji.ac.jp
School of Science and Technology, Meiji University
Kawasaki-shi, Kanagawa, Japan

ABSTRACT

The lattice Boltzmann method (LBM) is a stencil-based CFD method widely used across engineering fields, in which the domain is divided into a lattice, each element is computed independently, and data is exchanged with adjacent elements. Thus, it is expected to map well onto the Cerebras CS-3, which features large distributed on-chip memory and high-bandwidth inter-PE communication. In this work, we naively implement a three-dimensional LBM (D3Q27 Model) on the CS-3 by assigning a $1 \times 1 \times n$ lattice to each processing element (PE). To the best of our knowledge, this study is the first attempt made to implement 3-D LBM and measure its performance. We measure performance in lattice updates per second (LUPS) and weak scaling. On the largest domain, the CS-3 achieves 83.74 GLUPS—a 17.4× and 8.5× speedup over the NVIDIA A100 and GH200—while sustaining a parallel efficiency of 1.0 under weak scaling. These results indicate that wafer-scale computing can serve as a promising platform for memory-bound stencil computations.

ACM Reference Format:
Kouki Sakai, Saho Orihara, Takashi Shimokawabe, and Takaaki Miyajima. 2023. A Naive D3Q27 Lattice Boltzmann Implementation on the Cerebras CS-3: Comparison With GPUs. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 CEREBRAS CS-3 AND ITS ARCHITECTURE

Cerebras CS-3 is an AI accelerator equipped with the WSE-3, a single chip spanning a wafer that integrates approximately 900,000 processing elements (PEs) and 44 GB of on-chip SRAM[1]. As shown in Figure 1, each PE has 48 KB of local SRAM as its main memory, accessible in one cycle for a high bytes-per-FLOP ratio, and connects to four neighbors in a 2-D mesh with one-cycle communication. Without caches or cross-PE coherency, it is a distributed-memory system in which data is partitioned across PEs and exchanged through explicit, user-described communication. Prior work has implemented stencil computation on the CS-3 [2]. Our LBM program is written in the Cerebras Software Language (CSL) and run through the Cerebras Hardware SDK.

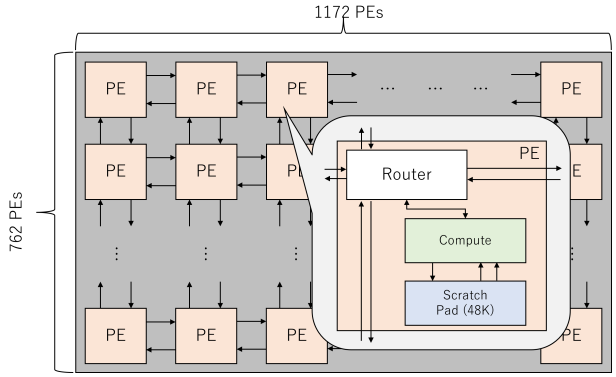


Figure 1: An overview of WSE-3.

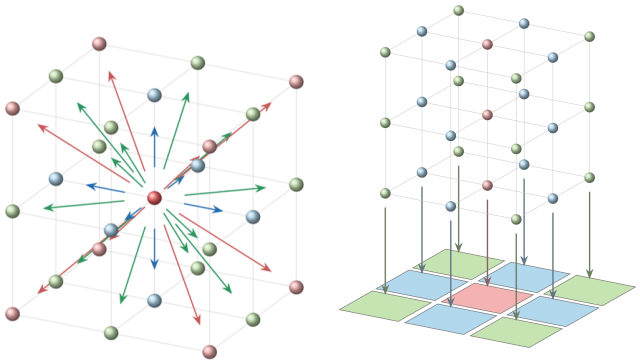
2 3-D LBM AND ITS IMPLEMENTATION ON THE CS-3

LBM approximates a fluid as many virtual particles with a finite set of velocities, advancing them through successive streaming and collision steps. We employed the D3Q27 model (Fig. 2a), in which each lattice point holds 27 discrete velocities: a rest particle and 26 directions to its neighbors. As a stencil method with local operations at each lattice point, it maps naturally onto the WSE-3's 2-D mesh. All computations use single precision.

We mapped the D3Q27 model so that each PE handles a $1 \times 1 \times n$ skinny-tall lattice along the z-axis (Figure 2b). Lattices adjacent along the z-axis reside in the same PE, so z-direction streaming requires no inter-PE communication. Communication is therefore confined to the x–y plane, where each point exchanges data with its eight neighbors (four axial, four diagonal). Each PE, however, can communicate directly only with its four axial neighbors.

To make the diagonal exchange possible, we employ two-step communication -north–south followed by east–west as illustrated in Figure 3. After the north–south exchange, each PE forwards the received data together with its own in the east–west direction, so that every PE obtains all eight in-plane neighbors for the collision step.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, July 2017, Washington, DC, USA
© 2023 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX



(a) 3-D LBM (D3Q27 Model) (b) Mapping 3-D lattice on mesh.

Figure 2: 3-D LBM and its mapping onto the 2-D mesh.

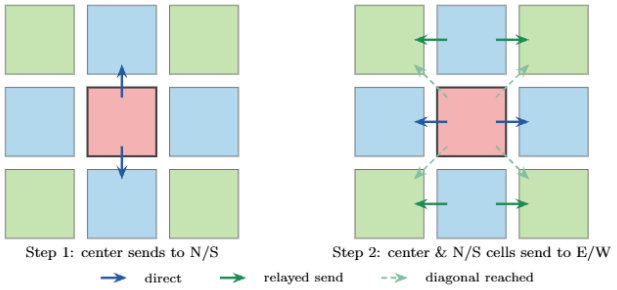


Figure 3: Two step diagonal communication on the 2-D mesh.

	CS-3	A100	GH200
GLUPS	83.74	4.81	9.84
Speedup of CS-3	(1.0×)	17.4×	8.5×

Table 1: Performance comparison on the largest domain.

3 MEASUREMENT RESULT

Measurements were performed on a CS-3 (SDK version 2.1.0) installed in the ALCF AI Testbed at Argonne National Laboratory, USA. Timing was measured with the SDK's cycle counter, and elapsed time was obtained by dividing the measured cycle count by the operating frequency of 750 MHz. For the GPU experiment, a server with an A100 (PCIe-40GB, memory bandwidth: 1555 GB/s, FP32 performance: 19.5 TFLOPS), and the GH200 (H200 GPU, memory bandwidth: 4000 GB/s, FP32 performance: 67.0 TFLOPS) from the University of Tokyo's Miyabi supercomputer system were used. The compilers are NVHPC 24.11 (nvcc 12.6, `sm_80`) and 25.9 (nvcc 13.0, `sm_90`), respectively. Compile option is `nvcc -O3 -Xcompiler "-fopenmp" -fast-math -std=c++17 -arch=sm_90`, while the `-arch` is changed for devices. We report the median of 10 runs.

The largest problem the CS-3 can hold is a $755 \times 1170 \times 66$ lattice, bounded by the PE count and the column height n that fits in a PE's local SRAM. As shown in Table 1. the CS-3 achieves 83.74 GLUPS, a 17.4x speedup over an NVIDIA A100 and 8.5x over an GH200.

We also measure weak scaling, increasing the number of PEs from 100×100 to the full 755×1170 while keeping the mapped lattice count per PE fixed in Figure 4. Taking the 100×100 configuration as the baseline, the parallel efficiency stays at the ideal value of 1.0 across all sizes. The maximum deviation is only about 3×10^{-5}.

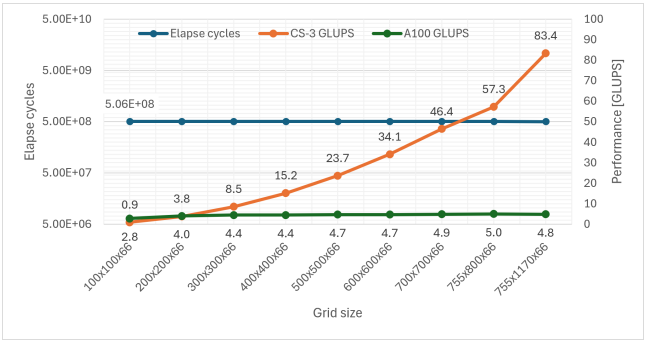


Figure 4: Weak scaling and processing cycle count on CS-3.

This near-perfect scaling reflects the locality of LBM, which maps onto the 2D mesh of the CS-3 without increasing communication overhead as the problem grows.

4 SUMMARY

We implemented a 3-D D3Q27 LBM on the CS-3, mapping a $1 \times 1 \times n$ column to each PE with two-step diagonal communication. It reaches 83.74 GLUPS while sustaining near-ideal weak scaling, indicating that wafer-scale computation is a promising platform for stencil computations such as LBM.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number 24K14972. This research used resources of the Argonne Leadership Computing Facility, a U.S. Department of Energy (DOE) Office of Science user facility at Argonne National Laboratory and is based on research supported by the U.S. DOE Office of Science-Advanced Scientific Computing Research Program, under Contract No. DE-AC02-06CH11357.

REFERENCES

[1] Sean Lie. 2024. Wafer-Scale AI: GPU Impossible Performance. Hot Chips 2024, IEEE Symposium on High-Performance Chips. https://hc2024.hotchips.org/assets/program/conference/day2/72_HC2024.Cerebras.Sean.v03.final.pdf Cerebras Systems.

[2] Ryuichi Sai, François P. Hamon, John Mellor-Crummey, and Mauricio Araya-Polo. 2024. Matrix-Free Finite Volume Kernels on a Dataflow Architecture. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis* (Atlanta, GA, USA) *(SC '24)*. IEEE Press, Article 28, 11 pages. https://doi.org/10.1109/SC41406.2024.00034