Democratizing AI Accelerators for HPC Applications: Challenges, Success, and Support BoF@SC24

Cache Me If You Can: Outpacing Traditional Hardware

Hatem Ltaief, KAUST

Steering Customized AI Architectures for HPC Scientific Applications

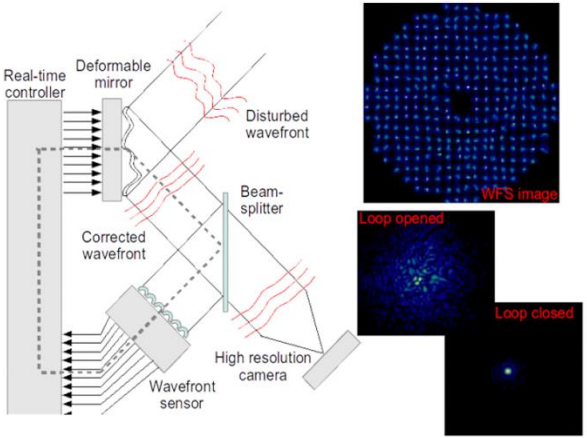
Steering Customized AI Architectures for HPC Scientific Applications

H. Ltaief[1], Y. Hong[1], A. Dabah[1], R. Alomairy[1], S. Abdulah[1], C. Goreczny[3], P. Gepner[4], M. Ravasi[2], D. Gratadour[5], and D. Keyes[1]

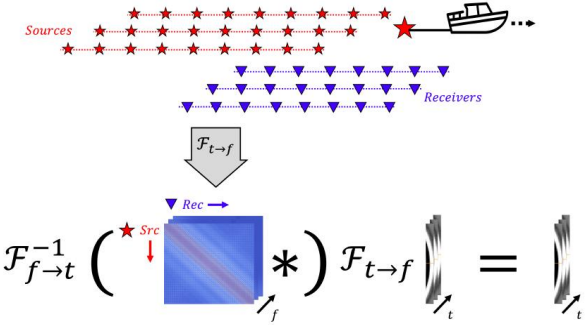
[1]Division of Computer, Electrical, and Mathematical Sciences and Engineering
[2]Division of Physical Sciences and Engineering
Extreme Computing Research Center
King Abdullah University of Science and Technology
Thuwal, Jeddah 23955 Saudi Arabia
Hatem.Ltaief, Yuxi.Hong, Adel.Dabah.1, Rabab.Alomairy, Sameh.Abdulah, Matteo.Ravasi, David.Keyes@kaust.edu.sa
[3]Graphcore, Poland
chrisgo@graphcore.ai
[4]Warsaw University of Technology, Poland
pawel.gepner@pw.edu.pl
[5]Paris Observatory, France
damien.gratadour@obspm.fr

Abstract. AI hardware technologies have revolutionized computational science. While they have been mostly used to accelerate deep learning training and inference models for machine learning, HPC scientific applications do not seem to directly benefit from these specific hardware features unless AI-based components are introduced into their simulation workflows, for instance, as a replacement of their numerical solvers. This paper proposes to take another direction in an attempt to democratize customized AI architectures for HPC scientific computing. The main idea consists in demonstrating how legacy applications can leverage these AI engines after a necessary algorithmic redesign. It is critical that the resulting software implementations map onto the underlying memory-austere hardware architectures to extract the expected performance. To facilitate this process, we promote the matricization technique for restructuring codes (1) by exploiting data sparsity via algebraic compression and (2) by expressing the critical computational phases in terms of tile low-rank matrix-vector multiplications (TLR-MVM) and batch matrix-matrix multiplications (batch GEMM). Algebraic compression enables to reduce memory footprint and to fit into small local cache/memory, while batch execution ensures high occupancy. We highlight how we can steer the Graphcore AI-focused Wafer-on-Wafer Intelligence Processing Units (IPUs) to deliver high performance for both operations. We conduct a performance benchmarking campaign of these two matrix operations that account for most of the elapsed times of four real applications in computational astronomy, seismic imaging, wireless communications, and climate/weather predictions. We report bandwidth and execution rates with speedup factors up to 150X/14X/25X/40X, respectively, on IPUs compared to other systems.

Adaptive Optics



Seismic Redatuming



Tile Low-Rank Matrix-Vector Multiplication

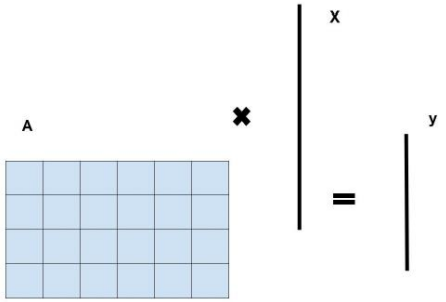


Fig. 1: Dense MVM.

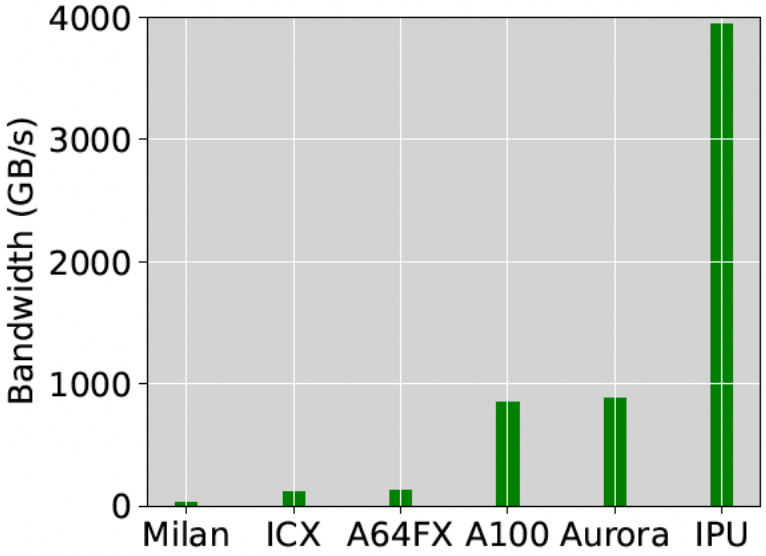
Hardware Settings

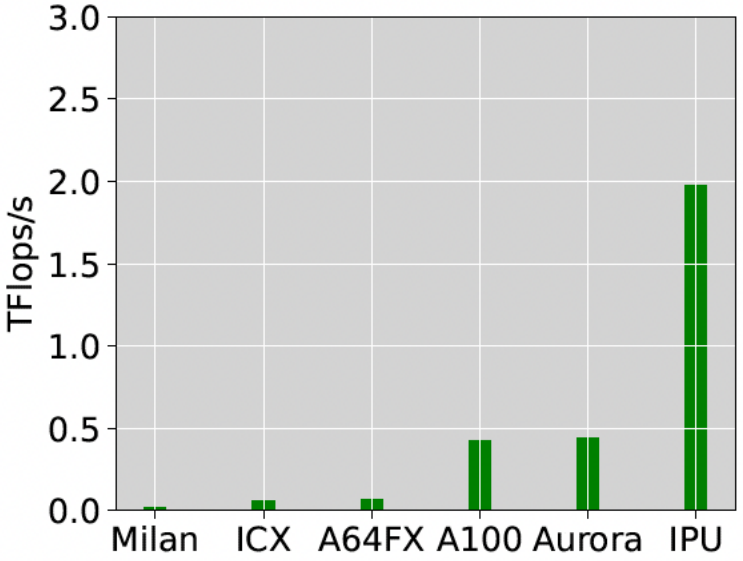
Vendor	Intel	AMD	Fujitsu	NEC	NVIDIA	Graphcore
Family	Cascade Lake	EPYC Milan	Primergy A64FX	SX-Aurora TSUBASA	Ampere GPU	IPU
Model	6248	7713	FX1000	B300-8	A100	Bow
Node(s)/Card(s)	1	1	16	8	1	1
Socket(s)	2	2	4	N/A	N/A	1
Cores	40	128	48	8	6912	1472
GHz	2.5	2.0	2.2	1.6	2.6	1.85
Memory Sustained BW	384GB DDR4 232GB/s	512GB DDR4 330GB/s	32GB HBM 800GB/s	48GB HBM2 1.5TB/s	40GB HBM2e 1.5TB/s	3.6GB 261TB/s
LLC Sustained BW	27.5MB 1.1TB/s	512MB 4TB/s	32MB 3.6TB/s	16MB 2.1TB/s	40MB 4.8TB/s	N/A
Compiler	Intel 19.1.0	GCC 7.5.0	Fujitsu 4.5.0	NEC 3.1.1	NVCC 11.0	POPLAR 2.6
BLAS library	Intel MKL 2020	BLIS 3.0.0	Fujitsu SSL II	NEC NLC 2.1.0	cuBLAS 11.0	N/A
MPI library	OpenMPI 4.0.3	OpenMPI 3.1.2	Fujitsu MPI 4.0.1	NEC MPI 2.13.0	NCCL 2.0	N/A

x86 - ARM - Vector **GPU**

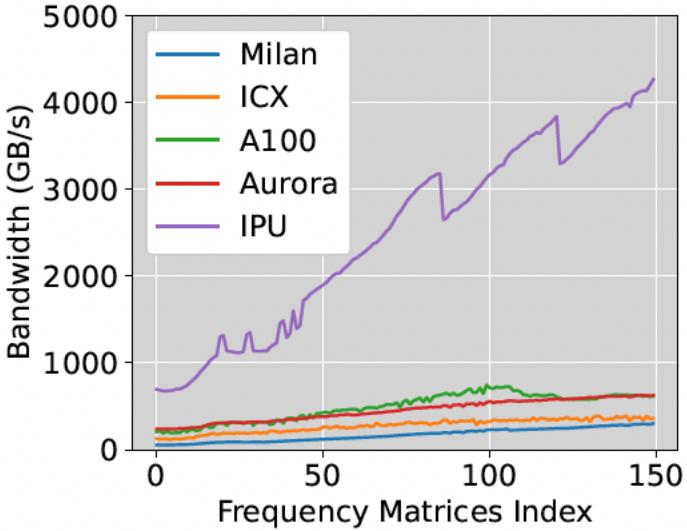
MPI + OpenMP CUDA

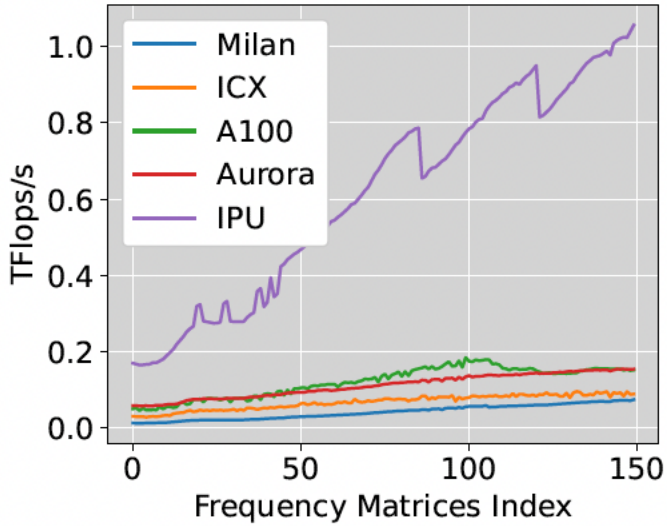
Performance Results for Astronomy





Performance Results for Seismic Processing





Scaling up on Cerebras CS-2 Wafer Scale (GB23 Finalist)

Scaling the "Memory Wall" for Multi-Dimensional Seismic Processing with Algebraic Compression on Cerebras CS-2 Systems

Hatem Ltaief
Yuxi Hong
Extreme Computing Research Center
Computer, Electrical and Mathematical Sciences & Engineering Division
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia
firstname.lastname@kaust.edu.sa

Leighton Wilson
Mathias Jacquelin
Cerebras Systems Inc.
Sunnyvale, California, United States
firstname.lastname@cerebras.net

Matteo Ravasi
David Keyes
Extreme Computing Research Center
Computer, Electrical and Mathematical Sciences & Engineering Division
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia
firstname.lastname@kaust.edu.sa

ABSTRACT

We exploit the high memory bandwidth of AI-customized Cerebras CS-2 systems for seismic processing. By leveraging low-rank matrix approximation, we fit memory-hungry seismic applications onto memory-austere SRAM wafer-scale hardware, thus addressing a challenge arising in many wave-equation-based algorithms that rely on Multi-Dimensional Convolution (MDC) operators. Exploiting sparsity inherent in seismic data in the frequency domain, we implement embarrassingly parallel tile low-rank matrix-vector multiplications (TLR-MVM), which account for most of the elapsed time in MDC operations, to successfully solve the Multi-Dimensional Deconvolution (MDD) inverse problem. By reducing memory footprint along with arithmetic complexity, we fit a standard seismic benchmark dataset into the small local memories of Cerebras processing elements. Deploying TLR-MVM execution onto 48 CS-2 systems in support of MDD gives a sustained memory bandwidth of 92.58PB/s on 35,784,000 processing elements, a significant milestone that highlights the capabilities of AI-customized architectures to enable a new generation of seismic algorithms that will empower multiple technologies of our low-carbon future.

November 12–17, 2023, Denver, CO, USA. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581784.3627042

1 JUSTIFICATION FOR THE GORDON BELL PRIZE

High-performance matrix-vector multiplication using low-rank approximation. Memory layout optimizations and batched executions on massively parallel Cerebras CS-2 systems. Leveraging AI-customized hardware capabilities for seismic applications for a low-carbon future. Application-worthy accuracy (FP32) with a sustained bandwidth of 92.58PB/s (for 48 CS-2s) would constitute the second-highest throughput from June'23 Top500.

2 PERFORMANCE ATTRIBUTES

Performance Attributes	Our submission
Problem Size	Broadband 3D seismic dataset (~ 20k sources and receivers and frequencies up to 50Hz)
Category of achievement	Sustained bandwidth

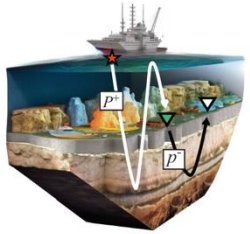


Figure 1: Schematic representation of the Multi-Dimensional Deconvolution problem. A red star indicates the source, a green triangle refers to the receiver, and the virtual source is represented by a white triangle.

48 Cerebras CS-2 systems, i.e., 35,784,000 processing elements

Scaling up on Cerebras CS-2 Wafer Scale (GB23 Finalist)

Strong scaling up to 48 CS-2 systems

Performance comparisons against the Top5 fastest Supercomputers

As per June 2023 Top500, 92.58 PB/s is:
- **2.3X > vs theo bw of Lumi #3**
- **3X > vs theo bw of Leonardo #4 / Summit #5**
- **35% > vs theo bw of Frontier #1**
- **close to est. sust. bw of Fugaku #2**
- **3X > vs theo bw of Oceanlite**