Accel-Bench: Exploring the Potential of Programming using Hardware-Accelerated Functions

Abenezer Wudenhe, Yu-Chia Liu, Chris Chen, and Hung-Wei Tseng

Department of Electrical and Computer Engineering, Bourns College of Engineering, University of California, Riverside

CNS-2007124
CNS-2231877

Abstract

This poster presents Accel-Bench, a benchmark suite that aims to capture the performance of accelerator-intensive programming.

Accel-Bench is the first benchmark suite that:
- Utilizes applications that can invoke different domain kernels
- Quantifies the potential performance gain of using hardware-accelerated functions to compose programs agnostic to their domain.

The Case for Accel-Bench

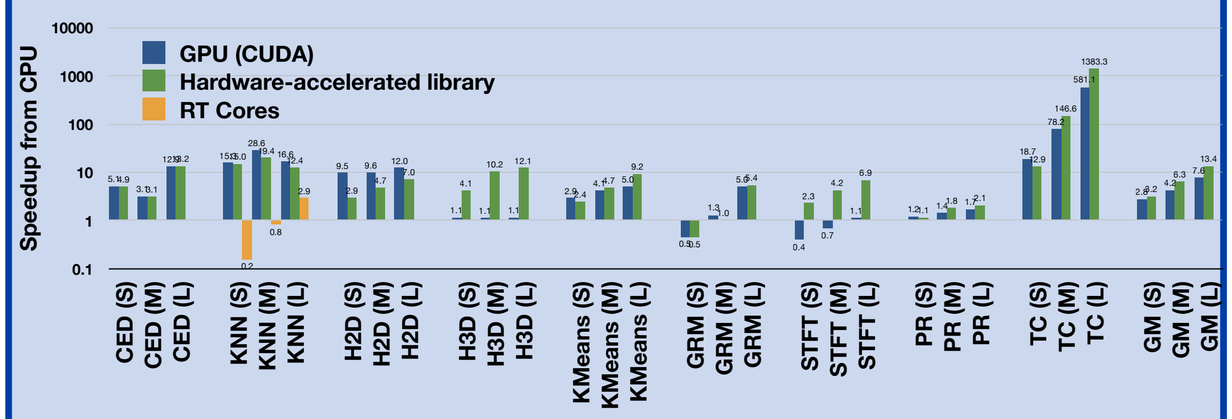
Several key technology trends push the development of Accel-Bench:

1. The discontinuation of Dennard scaling and the adoption of hardware accelerators.
2. The trend of applying AI/ML accelerators on a broader spectrum of applications.
3. The absence of benchmark suite targeting hardware-accelerators in more general domains.

The Accel-Bench Benchmark Suite

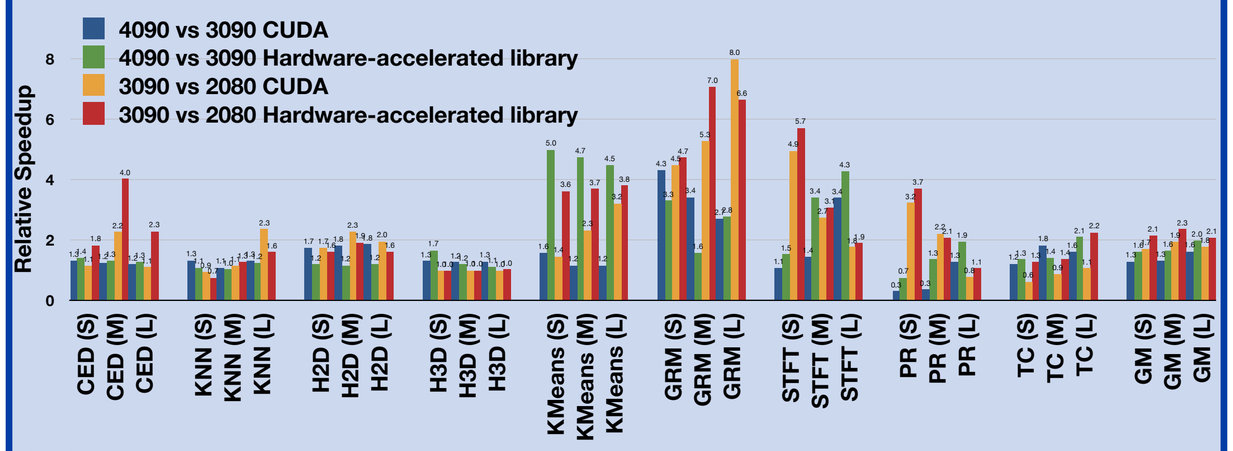
Benchmark Applications			
Benchmarks	**Dwarf**	**Application Domain**	**Hardware Accelerated Function**
Canny Edge Detection (CED)	Dense Linear Algebra	Image Processing	CONV
K Nearest Neighbor (KNN)	Dense Linear Algebra	Data Mining	GEMM
Heat (Heat2D/Head3D)	Structured Grid	Physics Simulation	CONV/FFT
KMeans (KM)	Dense Linear Algebra	Data Mining	GEMM
Genomic Relationship Matrix (GRM)	Dense Linear Algebra	Bioinformatics / Genomics	GEMM
Short-Time Fourier Transform (STFT)	Spectral Methods	Digital Signals Processing	FFT
PageRank (PR)	Graph Traversal	Web Mining	GEMV
Triangle Counting (TC)	Graph Traversal	Social Network Analysis	GEMM

Performance on default GPU



On average, using hardware accelerated API as the programming paradigm can speed Accel-Bench applications by 1.13× to 1.77× compared to the state-of-the-art GPU implementations, despite that GPU implementations are already 2.8× to 7.6× faster than CPU implementations.

Performance between GPUs Generations



2080 vs. 3090
- Performance gain has geometric means between 1.69× and 1.92×.
- With hardware accelerated library, speedup is 2.05× to 2.34×.
- 3090 also has 3.57× CUDA Cores & <1× Tensor Cores than 2080.

3090 vs. 4090
- Performance gain has geometric means between 1.29× and 1.62×.
- 4090 also has 1.56× CUDA Cores & 1.56× Tensor Cores than 3090.

Evaluation Platform

	Component	Notes
CPU	Core i5 12600K	3.7 GHz
Main Memory	64 GB	DDR4-3200
GPU (Default)	RTX 2080	2944 CUDA Cores 368 Tensor Cores 8GB Device Memory
GPU (3090)	RTX 3090	10496 CUDA Cores 328 Tensor Cores 24GB Device Memory
GPU (4090)	RTX 4090	16384 CUDA Cores 512 Tensor Cores 24GB Device Memory

Conclusion & Future Work

Using the hardware accelerated library, the same benchmark achieves an average speedup, even when the amount of tensor cores remain about the same. The result reveals that the architectural innovation of hardware accelerators like Tensor Cores would power more performance gain than relying on increasing conventional GPU cores.

We also anticipate Accel-Bench can help identify and design architectures to optimize the potential performance issues in such models with the current work of increasing the number of applications, providing architecture simulations, and expansion into other accelerator eco-systems such as SYCL & Intel's OneAPI.