On CUDA and its Applications

Steven Yu
Department of Electrical and Computer Engineering
California State Polytechnic University
Pomona, U.S.A.

***Abstract*—This paper will introduce CUDA and provide a list of applications that utilize the CUDA platform. A brief description CUDA will increase understanding and awareness of the technological impact that it brings. The overall use case of the platform will be explored along with some of the tools that it provides to developers who are looking to implement CUDA for their application.**

I. Introduction to CUDA

The GPU was first developed alongside the PC as a specialized processor to address the increasingly intensive demands of real-time high-resolution 3D graphics in video games. Continuous development of the GPU over the past few decades by companies such as NVidia has resulted in a powerful and highly parallel multi-core system that processes large blocks of data more effectively than general-purpose CPUs. In order to allow developers to take advantage of the efficient processing power of GPUs for applications beyond gaming, NVidia launched CUDA in 2006 as its own general computing solution [1]. CUDA (Compute Unified Device Architecture) is a parallel computing platform and programming model developed for general purpose processing on NVidia GPUs (GPGPU). Despite the introduction of competitors such as OpenCL by Apple and the Khronos Group, CUDA has continued to dominate the general computing space through superior performance and exclusive support. Today CUDA is utilized in numerous cutting edge applications that greatly expands the possibilities of technology.

II. Describing CUDA

CUDA allows developers to harness the parallel processing power of GPUs to accelerate compute-intensive applications. It makes this possible by providing a platform with direct access to the virtual instruction sets and parallel computational elements of the GPU. Moreover, CUDA is designed to incorporate into an expanding list of compatible programming languages including C, C++ and Fortran so that developers can easily access GPU resources with greater familiarity [2]. NVidia's CUDA toolkit also comes with a variety of libraries, debugging and optimization tools, a compiler, documentation, and a runtime library to help developers to utilize CUDA in any application [1]. As a result, CUDA's performance, support, and ease of use has allowed it influence to quickly spread and become dominant in several application areas.

III. Applications of CUDA on the GPU

Currently, GPUs are used to serve a variety of purposes such as mobile applications running on cloud servers, analyzing web data, and much more. CUDA helps to make all this possible on NVidia GPUs which is evident in its adoption in many fields that require high floating-point computing performance. NVidia's general computing solution has been applied to application domains including computational finance, climate and ocean modeling, data sciences and analytics, deep learning and machine learning, defense and intelligence, and manufacturing [1]. Other areas include media and entertainment in animation and processing, medical imaging, oil and gas, scientific research, safety and security, and tools and management [1].

In order to simplify access to GPU resources for the ever-growing list of applications, CUDA provides developers with a slew of libraries that support existing software in the industries. For an example, the cuDNN library is crucial for deep neural network computations in deep learning frameworks such as TensorFlow. CUDA's deep learning libraries also include TensorRT a deep learning inference optimizer and runtime, and Deepstream a video inference library [1]. Another example is CUDA's linear algebra and math libraries which include cuBLAS, a GPU-accelerated Fortran algorithm for high-performance matrix arithmetic on GPUs [1].

IV. Conclusion

Overall, the CUDA general-purpose parallel computing platform gives developers much greater access to the processing power of GPUs. Continued innovation and maturation of CUDA will allow it to reach many more application domains than imaginable. The acceleration of computation that is made possible by CUDA and its competitors serves as the foundation of modern technology and will be a crucial component to its advancement in the future.

References

[1] M. Heller, "What is CUDA? Parallel programming for GPUs," InfoWorld, 30-Aug-2018. [Online]. Available: https://www.infoworld.com/article/3299703/what-is-cuda-parallel-programming-for-gpus.html. [Accessed: 18-Apr-2021].

[2] S. 10 and M. Ebersole, "What Is CUDA: NVIDIA Official Blog," The Official NVIDIA Blog, 26-May-2018. [Online]. Available: https://blogs.nvidia.com/blog/2012/09/10/what-is-cuda-2/. [Accessed: 18-Apr-2021].