V okviru projekta EuroCC2 sta strokovnjaka s Fakultete za elektrotehniko, računalništvo in informatiko Univerze v Mariboru (FERI UM) aprila 2024 izvedla 8-urno izobraževanje z naslovom Pospeševanje aplikacij CUDA C++ z več grafičnimi procesorji. Predavanje s poudarkom na praktičnem delu je potekalo v sodelovanju z NVIDIA in IEEE. Izvedla sta ga doc. dr. Domen Verber in asistent Jani Dugonik s FERI UM.

Na izobraževanju so se udeleženci naučili, kako zagotoviti čim boljšo izrabo moči grafičnih procesorjev pri računsko intenzivnih aplikacijah.

Strokovnjaka sta predstavila proces izdelave aplikacije v programskem ogrodju CUDA C/C++, ki učinkovito izkorišča vse razpoložljive grafične procesorje, nameščene v lokalnem strežniku oziroma v enem od vozlišč visoko zmogljivega računalniškega centra. Programsko ogrodje CUDA C++ omogoča učinkovito izrabo grafičnih procesorjev za pospešitev delovanja računsko intenzivnih aplikacij na področjih kot so visokozmogljivo računalništvo, podatkovna znanost, bioinformatika, globoko učenje in drugih. S hkratno uporabo več grafičnih pospeševalnikov lahko bistveno povečamo prepustnost in skrajšamo skupni čas izvajanja opravil v primerjavi z uporabo ene pospeševalne enote. Dodatno lahko, s sočasnim prekrivanjem prenosa podatkov in računskih opravil, bistveno zmanjšajmo vpliv omejitve hitrosti prenosa podatkov med CPE in GPE. Kombinacija obeh tehnik nam omogoča maksimalno učinkovitost za strežnike z več grafičnimi pospeševalniki, bodisi v oblaku, omrežju ali lokalnih sistemih. Implementacija tehnik z več grafičnimi pospeševalniki na enem vozlišču je bistvenega pomena tudi za nadgradnjo aplikacij, ki se bodo izvajale na več vozliščih.

Udeleženci so se naučili, kako uporabiti tokove CUDA za prekrivanje prenosov podatkov z izvajanjem jeder na grafičnih procesorjih, kako učinkoviti uporabiti vse razpoložljive grafične procesorje na enem vozlišču, kako učinkovito združiti uporabo tehnike prekrivanja faz kopiranja podatkov in izračuna z uporabo več grafičnimi procesorjev, in kako uporabiti orodje NVIDIA Nsight™ Systems Visual Profiler za analizo in izboljšanje uporabe tehnik, obravnavanih na delavnici.

Po končani delavnici so udeleženci opravili kratek test in dodatno programsko nalogo ter pridobili uradni certifikat Inštituta za globoko učenje podjetja NVIDIA. Udeleženci lahko še pol leta po zaključku delavnice na NVIDIA portalu dostopajo do vseh delavnic, na katere so se prijavili, pa tudi do gradiv in testa za pridobitev certifikata.

As part of the EuroCC2 project, two experts from the Faculty of Electrical Engineering and Computer Science at the University of Maribor (FERI UM) delivered an 8-hour training course on Accelerating CUDA C++ Applications with Multiple GPUs in April 2024. The lecture, with a focus on hands-on practical work, was held in collaboration with NVIDIA and IEEE and was delivered by Associate Professor Dr. Domen Verber and Assistant Professor Jani Dugonik from FERI UM.

During the training, participants learned how to ensure the best possible use of GPU power in computationally intensive applications.

The experts presented the process of building an application in the CUDA C/C++ programming framework, which efficiently utilizes all available graphics processors installed in a local server or in one of the nodes of a high-performance computing centre. The CUDA C++ programming framework enables the efficient use of graphics processors to accelerate computationally intensive applications in areas such as high-performance computing, data science, bioinformatics, deep learning and others. By using multiple graphics accelerators simultaneously, we can significantly increase throughput and reduce overall task execution time compared to using a single accelerator unit. Additionally, by overlapping data transfer and computational tasks simultaneously, we can significantly reduce the impact of the data rate limitation between the CPU and the GPU. The combination of both techniques allows us to maximize performance for servers with multiple graphics accelerators, whether in the cloud, on the network or on-premises systems. Implementing techniques with multiple GPUs on a single node is also essential for upgrading applications to run on multiple nodes.

Participants learned how to use CUDA streams to overlay data transfers while running cores on GPUs, how to efficiently use all available GPUs on a single node, how to efficiently combine the use of overlapping data copy and compute phases using multiple GPUs, and how to use the NVIDIA Nsight™ Systems Visual Profiler tool to analyze and improve the use of the techniques discussed in the workshop.

After the workshop, the participants completed a short test and an additional programming assignment to obtain official an official certification from the NVIDIA Deep Learning Institute. Participants can continue to access all the workshops they registered for, as well as the materials and the certification test, on the NVIDIA portal for up to six months after the end of the workshop.