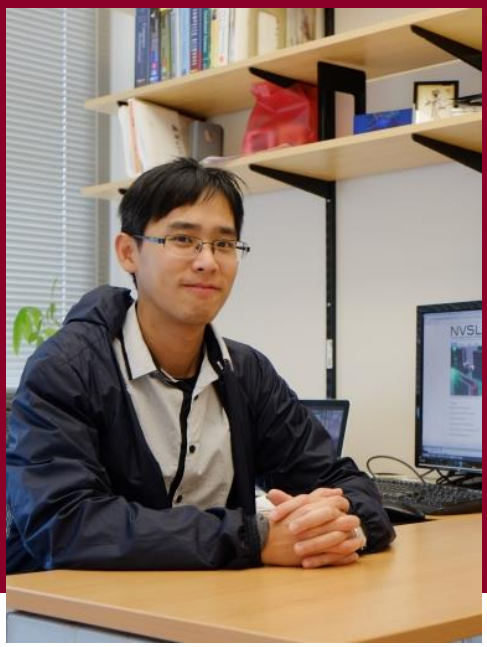
Invited Talk



Hung-Wei Tseng

Associate Professor

University of California, Riverside

Department of Electrical and Computer Engineering
Extreme Storage & Computer Architecture Lab (ESCAL)

일시: 2023. 12. 13 (수), 오후 2시

장소: 고려대학교 우정정보관 202호

General-Purpose Processing on AI/ML Accelerators

The significance of artificial intelligence (AI) and machine learning (ML) applications has changed the landscape of computer systems: AI accelerators start to emerge in a wide range of devices, from mobile phones to data servers. In addition to the direct contribution of performance gain in AI/ML workloads, the introduction of AI/ML accelerators brings a new flavor of computation model that any matrix-based algorithms can leverage. However, the highly application-specific designs of these accelerators place hurdles for a wider spectrum of workloads

In this talk, Hung-Wei will discuss state-of-the-art AI/ML accelerators and share his experience in transforming existing algorithms into AI/ML-specific functions. Hung-Wei's research group has demonstrated up to 288x speedup for database join operations by using NVIDIA's tensor cores. If we can extend the design of AI/ML accelerators to support various matrix operations, a set of matrix applications can achieve more than 10x speedup over conventional GPUs. Finally, Hung-Wei will talk about the new programming model, paradigms that new hardware accelerators enable, simultaneous and heterogeneous multithreading (SHMT). By using GPUs and TPUs at the same time, SHMT reveals 2x speedup over state-of-the-art GPU implementations. Hung-Wei will also discuss some extensions that are essential to make the upcoming revolutions of general-purpose computing successful.