統計科學研究所
INSTITUTE OF STATISTICAL SCIENCE

統計所學術演講

中研院統計所

學　術　演　講

講　題：Rethinking (parallel) programming on AI/ML accelerators

講　者：Prof. Hung-Wei Tseng
(Department of Electrical and Computer Engineering, University of California, Riverside)

時　間：2023年12月25日(星期一)，10:30-12:00

地　點：統計所B1演講廳

Abstract

The significance of artificial intelligence (AI) and machine learning (ML) applications has changed the landscape of computer systems: AI accelerators have started to emerge in a wide range of devices, from mobile phones to data center 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, the matrix processing model, that any matrix-based algorithm can leverage in theory. 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. compared with modern CPUs. If we can extend the design of AI/ML accelerators to support more matrix operations, a set of matrix applications, including dynamic programming-based algorithms, 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, and simultaneous and heterogeneous multithreading (SHMT). By using GPUs and TPUs at the same time, SHMT reveals a 2x speedup over state-of-the-art GPU implementations. Hung-Wei will also discuss some extensions that are essential to make the upcoming revolution of general-purpose computing successful.

※ 茶　會：10：10開始。

※ 實體與線上視訊同步進行。