台達電子技術長 郭大維博士：Data-Centric Computing

文／朱昱瑋 資訊科學與工程所碩士生

講者郭大維教授於 1986 年和 1994 年分別獲得國立臺灣大學和德克薩斯大學奧斯汀分校的計算機科學學士和博士學位。他目前是台達電子的首席技術長（自 2024 年 2 月起自）和國立臺灣大學計算機科學與資訊工程系的特聘教授。他曾擔任國立臺灣大學代理校長（2017 年 10 月至 2019 年 1 月）及學術研究副校長（2016 年 8 月至 2019 年 1 月）。郭教授亦曾是阿聯酋穆罕默德·本·扎耶德人工智慧大學的兼任 / 訪問教授及高級顧問（2023 年 2 月至 2024 年 1 月），以及香港城市大學的信息工程 (Information Technology) 講座教授、校長顧問及工程學院創院院長（2019 年 8 月至 2022 年 7 月）。他的研究領域包括嵌入式系統（Embedded Systems）、非揮發性記憶體程式設計（Non-volatile Memory Software Designs）、類神經網路計算（Neuromorphic Computing）和即時系統（Real-time Systems）。

20 多年前，快閃記憶體（Flash Memory）開啟了計算機領域的新世界的大門。自那時起，儲存裝置（Storage Devices）在性能、能耗甚至存取行為方面都獲得了巨大進步動力。在最近的幾年中，儲存裝置的性能提升已經超過 1000 倍，這引發了另一波在計算機設計中挑戰 --- 消除傳統 I/O 瓶頸的問題。在本次演講中，郭大維教授介紹了一些在類神經計算中的解決方案，這些方案賦予記憶體晶片新的計算能力。特別是，在其中探討了應用協同設計（Application Co-designs）在內存計算方面面臨的挑戰，並展示如何利用非揮發性記憶體的特性來優化深度學習。

演講內容主要分為兩個方案：電阻式隨機存取記憶體（Resistive Random Access Memory，ReRAM）以及相變記憶體 (Phase Change Memory；PCM) 作為優化深度學習的解決方案，並且針對計算深度學習時 ReRAM 產生的準確度問題和 PCM 產生的耐久性問題提供解法。

首先講者探討了物聯網時代下，深度神經網絡（DNNs）在嵌入式系統中的應用，特別是在邊緣設備中的影像和語音識別。為了提高 DNN 計算效率引入一種新興的技術——內存中處理（Processing in Memory, PIM），它將計算和記憶單元結合在一起，顯著降低了功耗。

近年來，帶有電阻式隨機存取記憶體（ReRAM）的交叉條加速器（Crossbar accelerators）成為研究熱點，尤其是作為物聯網設備的潛在解決方案。ReRAM 通過調整單元的電阻來存儲數據，同時實現計算功能，這使得它在物聯網和邊緣應用中具有重要的應用價值。然而，ReRAM 的編程變異誤差問題限制了其在大規模應用中的擴展性，特別是在多位元 ReRAM 設計和交叉條的可擴充性方面。郭教授主要專注探討如何通過創新的自適應數據操作策略來解決這些挑戰，從而降低 ReRAM 交叉條加速器中的模擬變異誤差。郭教授介紹了三個主要設計：權重捨入設計（WRD）、輸入子週期設計和位線冗餘設計（BRD）。這些設計不僅減少了重疊變異誤差，還提高了推理準確性。

至於相變記憶體（PCM），因其出色的性能、高密度和幾乎零漏電功率的特性，成為了一個神經網絡中極具潛力的幾決方案。然而，PCM 的寫入次數限制和讀寫性能不對稱等挑戰，使得在神經網絡中運用它面臨著諸多困難。郭教授主要探索如何在保持精度的同時，最佳地利用基於 PCM 的系統來訓練神經網絡。訓練和推理是神經網絡運行的兩個關鍵階段。訓練階段需要龐大的計算資源和主要存儲裝置容量，以進行反向傳播和梯度下降等操作。而推理階段則是應用神經網絡進行任務如分類等操作。雖然過去十年，研究人員通過縮減模型結構或優化數據流和數據內容等方法，努力解決了計算和存儲能力的挑戰，但在 NVM 上的應用研究仍然有限。郭教授提出了數據感知的編程設計，旨在優化 PCM 的寫入操作，以降低訓練過程中的內存訪問延遲，同時提升 PCM 的使用壽命，而這一切不會影響神經網絡的精度。並且透過實驗結果表明，所提出的方法能夠大幅改善訓練過程中的性能，並且提高 PCM 的生命週期達到 3.4 倍，同時保持神經網絡的準確性。

最後郭教授強調了記憶體性能對於類神經網路計算中扮演得中要角色，演講結束後，講者與聽眾做了一些問答，並針對聽眾的問題提出了一些建議。我非常感謝有這個機會，聆聽來自臺灣大學的郭大維教授所帶給我們的寶貴研究經驗。

Speech by Dr. Tei-Wei Kuo (CTO of Delta Electronics) Data-Centric Computing

Dr. Tei-Wei Kuo received his Bachelor's degree in Computer Science & Information Engineering from National Taiwan University in 1986 and his Ph.D. in Computer Science from The University of Texas at Austin in 1994. He is currently the CTO of Delta Electronics (since February 2024) and a Distinguished Professor in the Department of Computer Science & Information Engineering at NTU. He previously served as the Acting President of National Taiwan University (from October 2017 to January 2019) and Vice President for Academic Affairs (from August 2016 to January 2019). Professor Kuo was also an Adjunct/Visiting Professor and Senior Advisor at the Mohamed bin Zayed University of Artificial Intelligence (from February 2023 to January 2024), and the Lee Shau-Kee Chair Professor of Information Engineering, Advisor to the President (Information Technology), and Dean of the College of Engineering at the City University of Hong Kong (from August 2019 to July 2022). His research areas include Embedded Systems, Non-volatile Memory Software Designs, Neuromorphic Computing, and Real-time Systems.

Over two decades ago, flash memory transformed the computing industry. Since then, storage devices have made notable advancements in performance, energy efficiency, and access behaviors. In recent years, their performance has increased by over 1000 times, creating new challenges in computer design, particularly in eliminating traditional I/O bottlenecks. During his speech, Professor Kuo introduced various solutions in neuromorphic computing that endow memory chips with new computational capabilities. He specifically addressed the challenges of application co-designs in in-memory computing and demonstrated how the characteristics of non-volatile memory can be leveraged to optimize deep learning.

The lecture focused on two main strategies: using Resistive Random Access Memory (ReRAM) and Phase Change Memory (PCM) as the optimization solution for deep learning. It covered approaches to address accuracy issues with ReRAM and durability issues with PCM in deep learning computations.

Professor Kuo began by discussing the utilization of deep neural networks (DNNs) in embedded systems during the Internet of Things era. The discussion focused on image and speech recognition on edge devices. To enhance the computational efficiency of DNNs, he introduced a new technology called Processing in Memory (PIM), which integrates computation and memory units, substantially reducing power consumption.

In recent years, crossbar accelerators equipped with resistive random-access memory (ReRAM) have caught much attention as a promising solution for IoT devices. ReRAM stores data by modulating the resistance of cells and also performs computational functions, making it highly valuable for IoT and edge applications. However, programming variation errors in ReRAM hinder its scalability in large-scale applications, especially in multi-bit ReRAM design and crossbar scalability. Professor Kuo is dedicated to addressing these challenges through innovative self-adaptive data manipulation strategies aimed at reducing analog variation errors of ReRAM crossbar accelerators. He has introduced three key designs: Weight Rounding Design (WRD), Input Sub-cycle Design, and Bit-line Redundancy Design (BRD). These designs not only mitigate overlapping variation errors but also enhance inference accuracy.

Phase-change memory (PCM) is also a promising solution for neural networks due to its excellent performance, high density, and near-zero leakage power. However, challenges such as limited write cycles and uneven read/write performance hinder its application in neural networks. Professor Kuo is exploring the optimal use of PCM-based systems for training neural networks while maintaining accuracy. Neural network operation involves two crucial stages: training and inference. The training stage requires significant computational resources and main storage capacity for operations like backpropagation and gradient descent. The inference stage applies the neural network to tasks such as classification. Over the past decade, researchers have addressed computational and storage challenges by reducing model structures and optimizing data flow and content. However, research on non-volatile memory (NVM) applications remains limited. Professor Kuo has proposed a data-aware programming design to optimize PCM write operations, reduce memory access latency during the training process, and extend PCM's lifespan without compromising neural network accuracy. Experimental results indicated that this method significantly improved training performance and extended PCM's lifespan by up to 3.4 times while maintaining neural network accuracy.

In the end, Professor Kuo emphasized the significant impact of memory performance on neural network computations. During the Q&A session that followed, he provided valuable insights in response to questions from the audience. I am very grateful for the opportunity to learn from the invaluable research experience imparted by Professor Tei-Wei Kuo from National Taiwan University.