生成式人工智慧普及的機會：雲端算力服務與 Phison aiDAPTIV+

文／杜懿洵

生成式人工智慧（AI）自 2022 年底引起廣泛關注以來，不僅模型的數量在規模上年增數十、甚至數百倍，在架構上也從單一模型演化至多專家系統，展現出前所未有的多樣性和複雜性。然而，隨著各行各業積極探索將生成式 AI 技術融入工作流程，除了使得資訊安全和系統可控性的問題日益受到重視之外，生成式 AI 技術的普及更面臨著一個重大的挑戰：高昂的部署成本。

根據微軟研究報告指出，AI 模型的成長速度將會是 GPU 卡中的 DRAM 成長速度的 200 倍，這使得現行的 AI 運算硬體架構成長速度，逐漸地無法滿足 AI 應用的需求，而這將對許多組織採用生成式 AI 造成障礙，從而限制了生成式 AI 的廣泛應用。針對此一困境，出租算力的雲端服務商逐漸崛起，成為此波 AI 時代的另一個新商業模式。

除了 NVIDIA 不斷大力扶植 GPU 雲端託管商，並投資雲端串流服務業者，以推動各國雲端算力基礎建設之外，國際市場也大力挹注投資 AI 運算中心，像新創 CoreWeave 便計畫於今年底前興建 14 座資料中心，而其較為便宜的雲端租賃價格，也讓微軟選擇與其合作。其他像是 Lambda Labs、Together AI 也都獲得大量資金、擴建 GPU 機房，至於北美四大雲端巨頭，更開始自研 AI 晶片。而在台灣，除了有郭台強領軍的正崴集團，與在日本和台灣皆有自建機房的優必達合作，成立「優崴超級運算中心」，以第一階段 20 億的規模，向華碩採購千張 H100 顯示卡，並導入最新 G200 的 AI 運算基地，規劃建立全台最大 AI 算力中心之外，弘憶也獲得瑞昱集團約六億元的投資，向美超微採購，建立 55 台 H100 伺服器的算力中心。

除了出租算力的價格大戰儼然即將開打之外，深耕 NAND 控制晶片超過 23 年的群聯電子，則從自身優勢出發，通過利用 NAND Flash 技術擴充高帶寬記憶體（HBM）提升系統性能，藉由整合固態硬碟（SSD），提出自主研發的創新 AI 運算架構 aiDAPTIV+，以另一種模式降低生成式 AI 的部署成本，從而推動技術的更廣泛應用。

群聯 AI 研發團隊負責人暨國立陽明交通大學智慧科學暨綠能學院合聘助理教授林緯博士表示，群聯的本業與核心競爭力是 NAND 控制晶片研發，因此，如何擴大 NAND 儲存與 AI 應用的連結一直是群聯近幾年努力的方向。aiDAPTIV+ 此 AI 架構為透過群聯獨創整合 SSD 的 AI 運算架構，將大型 AI 模型做結構性拆分，並將模型參數隨應用時間序列與 SSD 協同運行，以達到在有限的 GPU 與 DRAM 資源下，最大化可執行的 AI 模型，預計能有效降低提供 AI 服務所需投入的硬體建構成本。

透過將 AI 技術導入 NAND 控制晶片與演算法裡，提升 NAND 儲存方案的運算效能與可靠度，群聯 aiDAPTIV+ 不僅能有效降低 AI 伺服器硬體建構成本，更能將此 AI 運算架構運用於各種 AI 應用場景，例如 aiDAPTIV+ AOI 光學檢測系統，首波應用場景便助力 SMT 工廠加速進入工業 4.0，並進一步提升檢測精準度，消除人力檢測所導致的不穩定性。

雖然生成式 AI 所引發的算力大戰正如火如荼地展開，但可預見的是，在應用成本越發降低之後，各種應用需求的研發也將越成為可能。本院張立平教授表示，之後將會引進教學課程，規劃作為學院老師的研究平台，也會與資訊學院老師成立 GAI(生成式 AI) 教學，讓同學在學校就能開始學習與嘗試生成式 AI 的各種創新應用。

Opportunities for the Rapid Adoption of Generative AI: Cloud Computing Services and Phison aiDAPTIV+

Since generative artificial intelligence (AI) gained widespread attention in late 2022, the number of models has grown exponentially, increasing by tens or even hundreds each year. Architecturally, AI has evolved from standalone models to multi-expert systems, showcasing unprecedented diversity and complexity. As industries actively explore integrating generative AI into their workflows, they face growing concerns about information security and system controllability. A significant barrier to its widespread adoption, however, remains the high cost of deployment.

A Microsoft research report indicates that the growth rate of AI models will surpass that of DRAM in GPUs by a factor of 200. This significant disparity is putting pressure on current AI hardware architectures, making it increasingly challenging to meet the demands of AI applications. As a result, this creates obstacles for many organizations looking to adopt generative AI on a large scale. In response to these challenges, cloud service providers have emerged, offering rentable computational power and introducing a new business model for the AI era.

In addition to NVIDIA's ongoing and robust support for GPU cloud hosting providers and its investments in cloud streaming services to bolster global cloud infrastructure, international investments in AI computing centers have also increased significantly. For instance, the startup CoreWeave plans to build 14 data centers by the end of this year with competitive prices of cloud rental services which has prompted Microsoft to partner with the company. Similarly, companies such as Lambda Labs and Together AI have secured substantial funding to expand their GPU data centers. Meanwhile, North America's four leading cloud giants have started developing their own AI chips. In Taiwan, Foxlink Group, led by T.C. Gou, has teamed up with Ubitus, a company operating data centers in Japan and Taiwan, to establish the "Ubilink Supercomputing Center." With an initial investment of 2 billion NTD, Ubilink plans to acquire 1,000 H100 GPUs from ASUS and integrate the latest DGX GH200 AI computation platform to establish the latest AI computing infrastructure in the project's first phase. This initiative aims to establish the largest AI computing center in Taiwan. Additionally, G.M.I. Technology has secured an investment of approximately $600 million NTD from Realtek and is purchasing 55 H100 servers from Super Micro Computer to build its own AI computing center.

Beyond the impending price war in the cloud computing market, Phison Electronics Corp., with over 23 years of expertise in NAND controller chips, is leveraging its core strengths to improve system performance by utilizing NAND Flash technology to expand high-bandwidth memory (HBM). Phison has also developed its innovative AI computing architecture, aiDAPTIV+, by integrating solid-state drives (SSDs). This approach provides an alternative model to reduce the deployment costs of generative AI, thereby facilitating its broader adoption and application.

Dr. Wei Lin, Head of Phison's AI R&D team and Assistant Professor at the College of Artificial Intelligence at National Yang Ming Chiao Tung University, explained that Phison's core business and competitive advantage lie in NAND controller chip development. As a result, the company has increasingly focused on strengthening the integration of NAND storage with AI applications in recent years. Phison's innovative aiDAPTIV+ AI architecture, which integrates SSDs with AI computing framework to optimize performance, structurally decomposes large AI models and coordinates the operation of model parameters with SSDs according to application time sequences to maximize the efficiency of AI models within the constraints of limited GPU and DRAM resources. This solution is expected to substantially reduce the hardware cost associated with deploying AI services.

By integrating AI technology into NAND controller chips and algorithms to enhance both computational performance and reliability, Phison's aiDAPTIV+ not only effectively reduces the hardware costs of AI servers but also makes the AI computing architecture applicable to a wider range of AI use scenarios. For example, the aiDAPTIV+ AOI optical inspection system helps accelerate the transition to Industry 4.0 in SMT factories. Additionally, it enhances inspection accuracy and reduces the instability caused by manual inspection.

Although the competition for computing power driven by generative AI is intensifying, it is expected that the research and development of a wide range of applications will become increasingly feasible as application costs continue to decrease. Professor Li-Pin Chang from the College of Computer Science at NYCU stated that the college will introduce training courses and establish a research platform for faculty members. Additionally, the faculty will collaborate to develop a GAI (Generative AI) curriculum, providing students with opportunities to learn about and experiment with innovative generative AI applications.