

開發日文版生成式 AI 與數位廣告服務轉型

CyberAgent, Inc. 運用搭載八個 NVIDIA® H100 Tensor Core GPU 的 Dell PowerEdge XE9680 伺服器來加速生成式 AI 並提升廣告效益。

業務需求

自 2016 年起，CyberAgent, Inc. 就一直積極研究和開發人工智慧，並將人工智慧與其廣告業務結合。該公司必須讓員工能夠以快速、經濟實惠的方式存取搭載最先進的 NVIDIA GPU 且效能極可靠的到場部署伺服器，以利進行生成式 AI 開發工作。

業務成果



與前一代搭載 PowerEdge XE9680 伺服器相比，大型語言模型 (LLM) 效能加速約 5.14 倍。



搭配 NVIDIA Transformer Engine 最佳化後，未來預期效能提升超過 10 倍。



可以依據最新資料集迅速微調機器學習模型。



與主流的 8U 相比，6U 外型規格能夠節省資料中心空間且冷卻效率佳。

解決方案簡介

- 搭載 NVIDIA® H100 GPU 的 Dell PowerEdge XE9680 伺服器
- Dell ProSupport

CyberAgent, Inc. 公司是國內網際網路廣告業著名的市場領導者，其投資的事業包括創新的電視平台 ABEMA。2016 年，該公司成立一個名為 AI Lab 的人工智慧研究組織，並且一直積極參與人工智慧的研究和開發。2020 年，CyberAgent 推出最先進的預測式 AI 技術，可改善高衝擊性橫幅廣告標語和影像組合的製作，大幅提升廣告效益。

CyberAgent 持續進行生成式 AI 開發，打造出含有 130 億個參數的獨特日文版專屬大型語言模型 (LLM)。該 LLM 設計為一般用途 AI 模型，適合在各種不同情況下使用，經微調後可創造出讓各個廣告平台的使用者產生共鳴的標語文案。CyberAgent 已經在 Kiwami Prediction AI、Kiwami Prediction TD 和 Kiwami Prediction LP 等人工智慧服務中使用日文版 LLM 來支援創意廣告製作和預測廣告效益。未來，CyberAgent 志在開發出能夠同時處理日文版 LLM 以及影像的多模態人工智慧。

“我們的內部研究人員能夠獲得數量更多的資源，並且能盡情使用而無須擔心成本；之前則無法在公有雲保障使用繪圖處理器 (GPU)，或是因為長期使用而必須支付更多費用。”

Daisuke Takahashi

CyberAgent, Inc. 集團 IT 部門 CIU 的解決方案架構師

2023 年 5 月，CyberAgent 發佈可供商業使用的開放原始碼日文版 LLM，稱為 OpenCALM (Open CyberAgent Language Models)，其中含有多達 68 億個參數。

相對於針對聊天功能微調的 ChatGPT，OpenCALM 比較偏向於一般用途的日文版語言模型，能夠依據使用者需求進行微調。CyberAgent 以開放原始碼專案的形式發佈 OpenCALM，是因為相對於在封閉式環境中開發日文版 LLM，開放原始碼對於公司而言好處更多，可以接收其他來源的意見回饋並與其他公司協同合作，進而為日本人工智慧技術之發展做出貢獻。

為 CyberAgent 的人工智慧創新提供技術支援的基礎結構

CyberAgent 在 2016 年成立 AI Lab 時，每位研究人員都有一部搭載繪圖處理器 (GPU) 的工作站進行研究。但是，在 2020 年疫情期間，因為必須進行遠端工作，無法讓每位研究人員都使用搭載繪圖處理器 (GPU) 的工作站。為確保研究人員擁有所需的運算資源，當最新的 NVIDIA® A100 GPU 推出時，該公司就開始考慮在其資料中心或公有雲中建置搭載繪圖處理器 (GPU) 伺服器的集中式機器學習 (ML) 平台。

CyberAgent, Inc. 集團 IT 部門 CIU 的解決方案架構師 Daisuke Takahashi 表示：「如果只是想要使用繪圖處理器 (GPU)，我們大可選擇公有雲；但使用公有雲時，永遠都無從得知最新的繪圖處理器 (GPU) 何時可供使用。此外，也無法保證在我們需要時隨時提供繪圖處理器 (GPU)，因此我們決定到場部署繪圖處理器 (GPU) 以利方便使用。為了實現基礎結構在公有雲和私有雲之間移動的彈性，我們設計一種盡可能接近公有雲規格的使用者介面。」CyberAgent 運用搭載四個 NVIDIA A100 GPU 的 Dell PowerEdge XE8545 伺服器，建置到場部署的第一個機器學習平台。

CyberAgent 為何選擇搭載 NVIDIA H100 GPU 的 PowerEdge XE9680 伺服器

CyberAgent 持續密切注意繪圖處理器 (GPU) 創新，尤其是最新的 NVIDIA H100 GPU。「我們認為它非常吸引人，不僅因為效能提升，也因為其擁有可加速特定運算演算法的機制，比如 Transformer Engine。」Takahashi 先生解釋，「據 NVIDIA 指出，與前一代 NVIDIA A100 GPU 相比，Transformer Engine 能夠使 LLM 的人工智慧訓練加速高達九倍，並讓人工智慧推斷加速最多 30 倍。」

CyberAgent 選擇搭載八個 NVIDIA H100 GPU 的 PowerEdge XE9680 伺服器模型。Takahashi 先生說明：「我們得知搭載 NVIDIA H100 GPU 的 Dell PowerEdge XE9680 伺服器推出時，就決定盡快加以採用。我們當時與 Dell Technologies 密切溝通，針對即將推出的 PowerEdge XE9680 伺服器和繪圖處理器 (GPU) 的可能組態進行討論。我們希望盡可能減少裝置數量來提高連續運作時間，因此 Dell Technologies 能夠以合理價格提供高層級的維護 (包括四小時現場服務) 讓我們很滿意。」



目前使用含有 130 億個參數的 LLM 加速 5.14 倍，
未來可望超過 10 倍。

Takahashi 先生繼續說明：「我們之所以選擇 PowerEdge XE9680 伺服器，也是因為先前安裝的 PowerEdge XE8545 伺服器提供穩定的效能且容易維護。此外，我們很重視 Dell iDRAC 管理工具的易用性，讓本機和遠端伺服器管理安全可靠。」

對於在 2023 年 3 月下訂單並一個多月之後，即 5 月中完成交付一事，Takahashi 先生深表讚同。「在供應鏈因為疫情而陷入混亂之際，Dell Technologies 也向我再三保證他們擁有相對穩定的供應鏈，我也很高興知道他們能夠在這麼短的時間內交付。」

在交付後的建置流程中，該公司亦進行了各項創新。據 Takahashi 先生回憶：「擁有大量參數的 LLM 必須使用多個繪圖處理器 (GPU)，因此我們每個伺服器都安裝八個 400 Gbps 網路介面卡 (NIC)，並使用 RDMA (遠端直接記憶體存取) 技術，在伺服器之間建立高速互連。繪圖處理器 (GPU) 伺服器會產生極高溫度，因此其設計必須能夠有效地冷卻。提供固態冷卻的 PowerEdge XE9680 伺服器 6U 外型規格也值得稱許。除此之外，資料中心也移動到可使用後門熱交換器的新位置，以便在機架後方安裝水冷式後門熱交換器即可進行有效冷卻，而不是冷卻裝設資料中心的整個房間。」

運用 Transformer Engine 最佳化來改善標語的準確性

安裝 PowerEdge XE9680 伺服器後，CyberAgent 獲得各種不同的效益。「因為效能大幅提升，我們希望日文版 LLM 的更新能夠更快速且更頻繁。」Takahashi 先生說道：「日文版 LLM 的演進速度也會提升。此外，與搭載四個 NVIDIA A100 GPU 的 PowerEdge XE8545 伺服器相比，搭載八個 NVIDIA H100 GPU 的 PowerEdge XE9680 伺服器使效能提升約 5.14 倍。我們也期望，未來 NVIDIA Transformer Engine 最佳化後，效能會提高超過 10 倍。我們也能夠依據最新的資料集對機器學習模型進行高速微調，讓我們更輕鬆回應各種請求以提供更好的服務、改善標語的準確度，並提供更有效的內容。」

搭載 PowerEdge XE9680 伺服器的機器學習基礎結構獲得使用者的高度讚賞。「我們收到內部研究人員的回饋，表示他們現在可以保障取得數量更大的資源，且盡量使用而無須擔心成本；而之前他們無法在公有雲保障使用繪圖處理器 (GPU)，或是因為長期使用而必須支付更多費用。」Takahashi 先生表示：「另一個優點是，我們能夠提供包括互連在內的高規格基礎結構，因此使用者能夠發揮業務影響力。」

Takahashi 先生對於 Dell Technologies iDRAC 管理工具也深表讚賞；該公司使用此工具已有一段時間，因為它能夠減少管理負擔。「我們不會一直固定在資料中心內作業，因此 iDRAC 是遠端工作的實用工具，例如可以檢查繪圖處理器 (GPU) 的溫度和狀態，同時也不需要存取作業系統即可更新韌體。」



提供固態冷卻的 PowerEdge XE9680 伺服器 6U 外型規格也值得稱許。」

Daisuke Takahashi

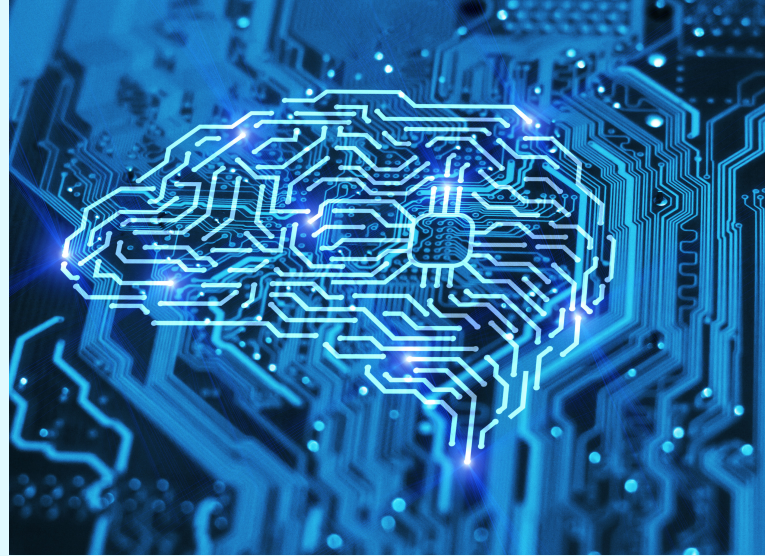
CyberAgent, Inc. 集團 IT 部門 CIU 的解決方案架構師



我們希望日文版 LLM 可以更快更新。搭載八個 NVIDIA H100 GPU 的 PowerEdge XE9680 伺服器，將效能提升約 5.14 倍。」

Daisuke Takahashi

CyberAgent, Inc. 集團 IT 部門 CIU 的解決方案架構師



專注於 LLM、GPU 和基礎結構

在未來，CyberAgent 計劃利用透過 OpenCALM 收集的意見回饋與學習內容來改善其員工使用的 LLM。透過 OpenCALM，CyberAgent 也會尋找與非屬廣告產業的公司與組織協同合作的機會。例如，CyberAgent 已開始與零售和金融界的業者進行討論，建置產業專屬 LLM，利用其產業特定的資料進行學習。

同時，Takahashi 先生也說明，他將持續關注最新繪圖處理器 (GPU) 及相關新技術的動態，思考如何將它們商業化。「我們也期待看到其他廠商能夠打造出與 NVIDIA 已實現的類似軟體生態系統。我也對 NVIDIA NVLink-C2C 的實作和 CXL (Compute Express Link) 這類連結 CPU 和繪圖處理器 (GPU) 的新標準感興趣；因為 PCIe 匯流排可能會成為繪圖處理器 (GPU) 效能的瓶頸。我期望 Dell Technologies 持續以快速的步調採用新技術，並設計效能出色的產品。」

使用最新且符合成本效益的繪圖處理器 (GPU) 後，CyberAgent 的人工智慧研究和開發團隊將為使用者提供所需的機器學習基礎結構，並藉持續進步。此外，隨著日文版 LLM 進一步發展，CyberAgent 在其中本身的廣告業務以及日本人工智慧市場都將持續受到眾人矚目。



我們希望盡可能減少裝置數量來提高連續運作時間，因此 Dell Technologies 能夠以合理價格提供高層級的維護 (包括四小時現場服務) 讓我們很滿意。」

Daisuke Takahashi

CyberAgent, Inc. 集團 IT 部門 CIU 的解決方案架構師

此內容由 Dell Technologies 從日文版本翻譯而成。

深入瞭解 Dell Technologies 生成式 AI 解決方案。

連結社群媒體。



DELLTechnologies

版權所有 © 2023 Dell Inc. 或其子公司。保留所有權利。Dell Technologies、Dell 與其他商標均為 Dell Inc. 或其子公司的商標。其他商標是屬於其各自擁有者之商標。本案例研究僅供參考。Dell 相信本案例研究中的資訊於 2023 年 9 月發佈時正確無誤。資訊如有變更，恕不另行通知。Dell 並未在本案例研究中作出任何明示或默示擔保。