分析師簡報

贊助者：Dell

人工智慧的基礎架構投資正飛快成長。好消息是，截至 2024，超過 50% 的人工智慧應用不需要加速器，可以直接在標準伺服器和乙太網路上運行。

隨著企業展開生成式人工智慧旅程，生成式人工智慧將以超乎想像的速度發展

2024 年 2 月

作者：網路與通訊支援技術團隊研究總監 Brandon Hoff，以及雲端與資料中心網路研究副總裁 Vijay Bhagavath

概覽

重點摘要

開始規劃生成式人工智慧的基礎架構：

- 要比其他公司更快整合人工智慧，首先需要制定計畫，而不是單純的將資料儲存在雲端，至少得將一份資料副本儲存在就地部署環境。
- 投入資源瞭解生成式人工智慧、人工智慧、機器學習和數位分身技術演算法的商業價值，並根據商業價值確定優先順序。
- 三個步驟：以標準伺服器和乙太網路維運生成式人工智慧應用。運用標準乙太網路，並因應需要提升生成式人工智慧技術以完善企業級工作負載。重新平衡就地部署和雲端基礎架構的工作負載，在未來三到五年內將資本支出和營運支出降到最低。

簡介

IDC 預測會影響 2024 年與以後 IT 投資的驅動力，供企業營運業務單位（LOB）的技術領導者和類似職務者擬訂策略的參考。

營運團隊一直在收集資料、建立資料湖，並利用雲端儲存資料。隨著 ChatGPT 掀起熱潮，生成式人工智慧就像曾經的 iPhone 一樣受到矚目，營運團隊也知道如何運用資料集。每個人都知道生成式人工智慧的好處，營運團隊也面臨來自投資者、主管和市場的額外壓力，希望團隊可以採取有效的生成式人工智慧策略。無論是生成式人工智慧、機器學習還是數位分身，當今世上都有許多技術選項，有利於改善企業營運和提升員工生產力。有鑑於此，能否採用適當技術將成為營運團隊乃至整個企業的基本關鍵績效指標（KPI）。

瞭解生成式人工智慧的現況

人工智慧需求呈現爆炸增加趨勢，雲端服務供應商（SP）和企業正加緊腳步建立基礎架構。雲端服務供應商將絕大多數資源投入人工智慧加速器，並建立自己的人工智慧基礎架構，但這些加速器的價格昂貴，如繪圖處理器（GPU）、TPU、FPGA、ASSP 和 ASIC 等，導致供應商的人工智慧服務成本增加。為了滿足各式各樣的公司的需求，這些雲端服務供應商打造人工智慧工廠以應付大量工作負載，並主要部署在全球最大的 IT 環境中，而這樣的公司大約有九家。

另一方面，企業級生成式人工智慧應用可以直接在標準化系統環境執行，無須加速器。IDC 預測，截至 2024 年，超過 50% 生成式人工智慧系統不用透過加速器加速，任何人都可以標準伺服器和網路建構生成式人工智慧基礎架構。如有需要，也有 GPU 可以使用。部署人工智慧基礎架構的方式很多，生成式人工智慧、人工智慧、機器學習和數位分身技術種類也很多元，可以為不同公司帶來各種優勢。在標準伺服器上執行生成式人工智慧有許多好處，因為一般環境都支援生成式人工智慧軟體堆疊。投資就地部署型標準基礎架構的公司將比其他公司更快推動生成式人工智慧計畫。IT 團隊必須開始評估各種生成式人工智慧、人工智慧、機器學習和數位分身演算法，藉此確定哪些演算法對業務影響最大。

優勢

生成式人工智慧和其他基礎模型正在改變世界，不僅能將數位科技提升到新的層次，不熟悉技術的使用者也可以使用功能強大的人工智慧技術。生成式人工智慧不僅有助於提高效率和生產力、創造新的成長機會，還能降低成本，並且讓使用這項技術的公司在市場上掌握競爭優勢。

建立您自己的生成式人工智慧基礎架構，將這項顛覆性技術運用到真實企業環境，並且將企業內的專業知識轉換成生成式人工智慧技術堆疊。優先投資技術，並以標準伺服器和乙太網路建立生成式人工智慧的基礎架構，將讓善用顛覆性技術的企業搶佔市場先機，享有競爭優勢。

考慮事項

立即行動，運用生成式人工智慧

ChatGPT 和其他模型的亮眼成效讓大家對生成式人工智慧應用的期待加深。生成式人工智慧確實可以創造價值，但價值取決於專有資料的來源和部署的演算法。董事會、投資者和主管們將提出問題，並積極了解生成式人工智慧可以如何幫助業務。

對於已經收集大量、非結構化專有資料的企業來說，生成式人工智慧可以根據既有的專有資料建立原創內容，這將有助於改造組織，以及持續推動創新。像小孩子學走路一樣，循序漸進的方法有助於瞭解生成式人工智慧可以如何幫助企業，以及企業如何繼續發展。

在乙太網路上建立初始生成式人工智慧基礎結構

對於企業級工作負載，標準系統足以支撐生成式人工智慧所需效能。此外，以標準伺服器和乙太網路打造的生成式人工智慧基礎架構將能直接使用企業的作業系統、管理工具和網路管理工具。釐清有助於企業的 LLM 所需的運算資源後，即可選擇適當的生成式人工智慧和人工智慧加速器以提升運算效能。關鍵是人工智慧基礎架構需要有良好的架構，因為，一個架構良好的網狀架構可以支援數十到數千個人工智慧運算節點。

雖然生成式人工智慧工作負載的網路選擇很多原，但從普及性和開放式的角度思考，乙太網路是多家廠商的首選。目前，適用於生成式人工智慧叢集的標準乙太網路即可支援初步部署生成式人工智慧。

以超級乙太網路建構生成式人工智慧基礎架構

隨著企業進入生成式人工智慧的開發成長期，建立可橫向擴展的生成式人工智慧基礎架構是有意義的。可橫向擴展的生成式人工智慧基礎架構有兩個關鍵元素：資料中心人工智慧加速器和人工智慧網路。在常見的橫向擴展生成式人工智慧基礎架構中，每一台伺服器都搭配了 8 個資料中心等級 GPU，且每一個 GPU 都搭配一個高速網路介面卡或 DPU 以提供高效網路服務。

可橫向擴展的生成是人工智慧基礎架構的關鍵核心之一是高效網路，因為，LLM 的執行瓶頸是資料的網路傳輸時間。對某些工作負載來說，資料的網路傳輸時間可能佔 LLM 執行時間的 60%，因此，當資料在運算叢集之間移動時，運算基礎架構會呈現閒置狀態。所幸，超級乙太網路聯盟 (Ultra Ethernet Consortium) 提供的網路服務沒有這個問題，聯盟承諾提供的網路架構效能跟超級運算網路相似，可以擴展到雲端資料中心且跟乙太網路一樣普及且符合成本效益。人工智慧網路是左右生成式人工智慧和 HPC 網路能否大幅成長的關鍵，好消息是，絕大多數乙太網路交換器廠商都支持超級乙太網路聯盟。

為了提高效能，需要高速 SerDes、PHY 和光纖等三項關鍵技術。無論是乙太網路還是其他網路皆採用這三項技術，因此，沒有哪種網路有效能優勢。為了將乙太網路效能最大化，InfiniBand 交易協會 (InfiniBand Trade Association) 啟動了 RDMA over Converged Ethernet (RoCE) 計畫，並定義 RoCE 通訊協定。標準資料中心的交換器不僅支援 RoCE 通訊協定，更進一步做了其他調整以提高效能，例如即將上市的高基數乙太網路交換、直通交換、負載平衡，以及高達 800GbE (4 個 200GbE) 的更高頻寬。

IDC 的 Vijay Bhagavath 表示：「生成式人工智慧資料中心的乙太網路交換器市場規模將從 2023 年 4190 萬美元增長到 2027 年的 10 億美元，年複合成長率高達 158.2%。」

測試生成式人工智慧 LLM 有助於企業瞭解生成式人工智慧帶來那些好處，制定相應的策略，以及確定需要哪種類型的基礎架構。從本質上來看，軟體堆疊有助於推動企業對生成式人工智慧應用的需求，從而帶動對半導體的需求。瞭解軟體堆疊將有助於建構最佳化的硬體基礎架構。

隨著半導體成本穩定下來，重新平衡內部就地部署與外部雲端部署的基礎架構

隨著資料中心 GPU 供給量增加，有更多的廠商提供資料中心 GPU、更多人工智慧加速器，進而在企業就地部署環境提供更強大的生成式人工智慧處理能力。與此同時，雲端服務供應商面臨的瓶頸將消失，成本有望穩定下來。如此一來，大約三到五年後，就能夠重新平衡內部就地部署和外部雲端基礎架構的生成式人工智慧工作負載，進而將資本支出和營運支出降到最低。

結論

生成式人工智慧是突破性的人工智慧技術。企業需要針對企業級工作負載制定生成式人工智慧策略/計畫，才能將這項顛覆性技術運用於企業營運。

強勁的需求將推動元件價格和雲端服務供應商的費用上漲。在這樣的大環境下，IDC 預測，截至 2024 年，超過 50% 的生成式人工智慧系統無須加速，任何人都可以透過標準伺服器和網路部署生成式人工智慧基礎架構。如有需要，也可以使用 GPU 。部署人工智慧基礎架構的方式很多，生成式人工智慧、人工智慧、機器學習和數位分身技術也有許多類型可選擇，可為不同公司帶來各種優勢。

IDC 預測，為了降低營運成本，企業會將數據資料從雲端帶回以利生成式人工智慧處理。企業將以標準的運算基礎架構、乙太網路硬體來開發和測試生成式人工智慧，進而瞭解 LLM 可以為企業帶來的益處，以及分析專有資料可以創造那些價值。

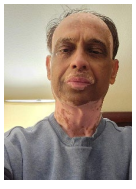
在既有的伺服器和企業乙太網路上建立生成式人工智慧 LLM 的測試基礎架構將有助於企業釋放生成式人工智慧價值。

IDC 的 Brandon Hoff 表示：「市場持續低估生成式人工智慧的成長，IDC 認為生成式人工智慧基礎架構和半導體領域將呈現蓬勃成長。」

分析師介紹

網路與通訊支援技術團隊研究總監 ***Brandon Hoff***

Brandon Hoff 在 IDC 支援技術團隊中主導 IDC 網路和通訊基礎結構領域的研究，他的研究涵蓋技術趨勢、工作負載、產品、廠商、供應鏈，以及企業 IT、網路資料中心、雲端資料中心、電信服務提供者資料中心的最終使用者採用策略等層面。



雲端與資料中心網路研究副總裁 ***Vijay Bhagavath***

Vijay Bhagavath 針對雲端和資料中心網路市場與技術提供實用的思考方向和務實的見解。他對整個網路市場、技術、產品藍圖、競爭差異化和部署策略都有深入的瞭解，因此能為廠商、雲端提供商、企業 IT 買家和從業者提供深入精闢的評論和指導。

贊助者訊息

將人工智慧融入您的資料

Dell Technologies 透過創新技術、全方位專業服務以及廣泛的合作夥伴網路，加速組織實現目標，將可能性轉換為經實證的成果。

- 輕鬆簡單：將策略指引和藍圖與經實證與驗證的解決方案結合，加快取得成果的速度。
- 量身打造：運用專為您的業務需求設計的基礎結構，充分發揮資料的價值。
- 值得信任：在安全的基礎上開創您的人工智慧未來，保護您的資料和智慧財產權。

提供最佳人工智慧效能，簡化專為生成式人工智慧時代設計的人工智慧基礎架構的採購、部署及管理流程。透過技術、創新及 Dell Technologies 優勢，提供更快速且出色的結果。

如需瞭解詳情，請造訪 www.dell.com/AI。

IDC Custom Solutions

本文內容改編自發表於 www.idc.com 的現有 IDC 研究。

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
電話：508.872.8200
傳真：508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com