產品簡介

AI 加速 Intel® Xeon® 可擴充處理器

適用於 Intel® Xeon® CPU 的 Intel® AI 引擎可提升整個 AI 流程的效能

70%
資料中心 AI 推論
在 Intel® Xeon® 處理器上執行[1]

提升高達
10 倍
PyTorch 即時推論和訓練工作負載效能，使用內建 Intel® AMX (BF16) 相較於上一代 (FP32)[2]

即時推論效能

提升高達
6.2 倍
即時自然語言處理推論效能 (BERT)，使用配備 Intel® AMX (BF16) 的第 4 代 Intel® Xeon® Platinum 8480+ 相較於上一代 (FP32)[3]

AI 涵蓋廣泛的工作負載和使用案例，從資料預處理和傳統機器學習到語言處理及影像辨識等深度學習模型。Intel® Xeon® 可擴充處理器與 Intel® AI 引擎相結合，為整個 AI 流程提供強大的運算效能，並為機器學習、資料分析及深度學習中的特定 AI 工作負載提供內建加速器。

為整個企業提供 AI 的內建功能

AI 無處不在，涵蓋各種關鍵工作負載。從核心企業應用到自動化語音助理，傳統機器學習和深度學習模型已逐漸成為業務運作的基本組成部分。AI 的大規模應用取決於從資料預處理到訓練，再到部署的漫長開發過程。每個步驟都有專屬的開發工具鏈、架構和工作負載，這些都會產生獨特的瓶頸，對運算資源產生不同的需求。Intel® Xeon® 可擴充處理器具有內建加速器，開箱即可用於運行整個流程，全面提高 AI 效能。Intel® 加速器引擎是專門打造的整合加速器，支援最嚴苛的新興工作負載。

採用 Intel® Advanced Matrix Extensions (Intel® AMX) 加速深度學習

Intel® AMX 是 Intel 在第 4 代 Intel® Xeon® 可擴充處理器上，用於深度學習訓練的新一代進階技術。Intel® AMX 是自然語言處理、推薦系統及影像辨識等工作負載的理想選擇，將前代 Intel® Xeon® 可擴充處理器內建的 AI 加速功能加以延伸，同時提供顯著的效能提升[4]。

Intel® AMX 為 AI 模型提供工作負載提升，並可以透過將特定的 AI 工作負載聚合到 CPU 上，而不是將其卸載到獨立的加速器，進而協助客戶提高總體擁有成本 (TCO)。

與 CPU 核心上的 Intel® Advanced Vector Extensions 512 (Intel® AVX-512) 相比，Intel® AMX 還以更高的最大處理量 (運算/週期) 改進平鋪乘法效能[5]。

改進自然語言處理和推薦系統

第 4 代 Intel® Xeon® 可擴充處理器和 Intel® AMX 為自然語言處理提供大幅效能提升，而且無需額外的硬體。函式庫已整合到 TensorFlow 和 PyTorch 中，開發人員無需額外工作，即可享受內建 AI 加速的優勢，還可以更輕鬆地從不同的硬體環境中遷移程式碼 (這過程可能既漫長又昂貴)。

客戶成功案例：
於 Intel® Xeon® 可擴充處理器上體驗真實世界加速

騰訊雲使用 Intel® Xeon® 可擴充處理器提供即時語音合成。

瞭解詳情 ›

作為世界上最大的粒子加速器的所在地，**CERN** 使用內建的 Intel® DL Boost 來加快推論速度，而無需犧牲準確性。

閱讀案例 ›

透過加速深度學習推論和訓練，採用 Intel® AMX 的第 4 代 Intel® Xeon® 可擴充處理器可協助您在平衡 TCO 的同時自訂使用者體驗。該系統透過基於深度學習的推薦系統來實現這一點，並將即時使用者行為訊號和其他情境特徵（如時間和位置）考慮在內。

利用第 4 代 Intel® Xeon® 可擴充處理器和加速器引擎實現未來創新

無論是將 Intel® Xeon® 處理器用於處理本地工作負載，還是處理雲端或邊緣工作負載，Intel® 加速器引擎都能夠協助您的業務達到新高度。這些內建加速器具備一系列優勢，包括安全性方面的處理速度更快，數據保護力更強以及基礎設施利用得更充分。

Intel® 加速器引擎還可以協助提高虛擬和實體 CPU 利用率，並大幅減少每核心解決方案的授權。

除此之外，這些內建加速器還能夠提高應用效能、降低成本並提升平台層級效率。

Intel® Advanced Vector Extensions 512（Intel® AVX-512）實現更快的機器學習

Intel® Xeon® 核心可以對網站的 SSL 進行雜湊加密，處理大量資料庫，並為製藥研究、晶片設計或一級方程式引擎進行模擬。這些核心是全能的主力軍，但在 AVX-512 加速器的協助下，可以更快完成深度學習訓練工作負載。

經過多代改進，Intel® AVX-512 允許 Intel® Xeon® 可擴充處理器在每個時脈週期中包含更多操作，並提供可媲美平行處理的效能。Intel® AVX-512 中的擴充功能是指令集，會告訴 CPU 該做什麼以及該如何做。其運作方式非常複雜，但 AVX-512 的基本邏輯非常簡單。首先，盡可能將多個步驟壓縮為更少的操作。第二，協助 CPU 在每個時脈週期內進行更多的操作。

更少的步驟意味著更快的處理效率

數學可以非常聰明，並且非常簡練。Intel® AVX-512 使用大量智慧、優美的數學，將常見的運算操作壓縮、組合和融合到更少的步驟中。就拿一個簡單範例來說，您可以指示 CPU 計算 3 x 3 x 3 x 3 x 3，這需要五個時脈週期。或者，您可以為 3^5 建立一條 CPU 可以在一個週期內完成的指令。AVX-512 採用這種邏輯，將其應用於數百種具有特定工作負載的操作，包括 AI 中一些最艱難的操作。

一次計數八項比計數一項快得多

AVX-512 中的「512」指的是，這些指令在每個時脈週期增加 CPU 處理位元數的第二種方式。四十年前，16 位元 PC 令人嘆為觀止。不久後，32 位元的機器成為主流。現今，您的智慧型手機可以執行 64 位元。位元計數指的是，CPU 在每時脈週期內可定址的資料所在的記憶體插槽中的暫存器數量。AVX-512 將暫存器數量擴充到多少？您能猜到嗎？答案是 512 位元。採用 Intel® AVX-512 時，只需擴充暫存器數量，執行速度就能比 CPU 基本 64 位元速度快 8 倍。這就像從 1、2、3……數到 96，以及 8、16、24……數到 96 的差別。

Intel® Deep Learning Boost（Intel® DL Boost）：針對神經網路的智慧數學

訓練深度學習模型可能需要數小時或數天的運算能力。深度學習推論可能不到一秒就能完成，也可能需要花費幾分鐘，具體取決於模型的複雜性和結果所需的準確性。若將訓練或推論擴充到資料中心等級的運算，就會需要龐大的時間、精力和效能預算。

Intel® DL Boost 使用多條 Intel® AVX-512 指令，透過同時使用 INT8 和 BF16 來加速深度學習工作負載。該系統將三個操作組合成一個向量神經網路指令（VNNI）集，進而減少每個時脈週期的操作次數，同時提供 Intel® Xeon® 可擴充處理器的全部運算潛力。VNNI 使用 INT8 精度，加速深度學習推論。

第 4 代 Intel® Xeon® 可擴充處理器的推出，也有望帶來更大的效能提升。與第 3 代 Intel® Xeon® 可擴充處理器相比，Intel® AMX 在第 4 代 Intel® Xeon® 可擴充處理器上與 AVX-512 協同工作，平鋪乘法效能在整個過程中具有更大的最大處理量（運算/週期）[6]。

以更低功耗運行更強大 AI 的引擎

採用 Intel® AI 引擎的 Intel® Xeon® 可擴充處理器所需的硬體資源較少，因此可為運行 AI 工作負載提供更強大、更節能的解決方案。

具有內建加速器引擎的 Intel® Xeon® 可擴充處理器可以協助提供改進的工作負載結果，如降低 TCO 及為當今要求嚴苛的 AI 工作負載提供更好的投資報酬率（ROI）。

例如，平均而言，使用 Intel® Xeon® 可擴充處理器的系統，成本比需要 GPU 整合的同類系統低 17%[7]。

借助 Intel® Xeon® 處理器，輕鬆實現 AI 加速

Intel® Xeon® 可擴充處理器的 AI 加速，內建於 CPU 的指令集架構（ISA）中。這意味著該系統已經準備好支援任何適用的軟體，並助其發揮效益。Intel 軟體工程師持續打造更佳的開源 AI 工具鏈，並將這些最佳化的成果回饋給社群。例如，預設情況下，TensorFlow 2.9 配備 Intel® oneAPI Deep Neural Network Library（Intel® oneDNN）。下載最新版本，TensorFlow 便會自動利用 Intel® 最佳化功能。

對於 AI 管道中的其他應用，資料科學家和開發人員可以下載免費的開源 Intel® 發行版、資料庫和開發環境，利用我們適用於 Intel® Xeon® 可擴充處理器的 ISA 中的每個內建加速器。

資料科學家和 AI 開發人員不需要重新編寫工具並為 Intel® AVX-512 重新編譯，這件事我們負責完成。

如今的企業需要從其基礎設施中獲得更高的工作負載效能，並以更高的電源效率和更低的成本實現此目標。Intel® Xeon® 可擴充處理器中專門打造的 Intel® AI 加速器引擎，將協助您最大限度地利用對業務最重要的 AI 工作負載。

進一步瞭解內建 Intel® 加速器引擎的 Intel® Xeon® 可擴充處理器，可以為對您的業務最重要的 AI 工作負載實現哪些功能。

進一步瞭解

Intel® Xeon® 可擴充處理器上的 AI 和深度學習 ›

Intel® AVX-512 ›

Intel® Deep Learning Boost ›

Intel® AI 分析工具組 ›

立即透過 Intel 針對 AI 和機器學習的最佳化功能，開始在雲端或您自己的基礎設施上加速 AI 工作負載。

進一步瞭解 ›

intel® XEON®

[1]基於截至 2021 年 12 月執行 AI 推論工作負載的全球資料中心伺服器安裝基礎的 Intel 市場模型。

[2]請於 intel.com/processorclaims 查看 [A16] 和 [A17]：第 4 代 Intel® Xeon® 可擴充處理器。結果可能有所差異。

[3]請於 intel.com/processorclaims 查看 [A19]：第 4 代 Intel® Xeon® 可擴充處理器。結果可能有所差異。

[4]與上一代 (FP32) 相比，採用內建 Intel® AMX (BF16) 的第 4 代 Intel® Xeon® 可擴充處理器的 PyTorch 訓練效能提高 3.5 倍至 10 倍，請於 Intel.com/processorclaims 查看 [A16]：第 4 代 Intel® Xeon® 可擴充處理器。結果可能有所差異。

[5]https://edc.intel.com/content/www/tw/zh/products/performance/benchmarks/vision-2022/，#41 和 #42 基準測試。結果可能會有所差異

[6]https://edc.intel.com/content/www/tw/zh/products/performance/benchmarks/vision-2022/，#41 和 #42 基準測試。結果可能會有所差異

[7]請於 https://edc.intel.com/content/www/tw/zh/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/ 查看 [100]