產品解決方案簡介

第 4 代 Intel® Xeon® 可擴充處理器
高效能運算

高效能運算需要高效能技術

高效能運算 (HPC) 工作負載利用第 4 代 Intel® Xeon® 可擴充處理器，享有較高的單位核心效能、較多的核心數、更先進的 I/O 與記憶體子系統技術，以及各種內建硬體加速器等優勢。

AI 與 HPC 工作負載日益融合，為常見的效能挑戰增添新的維度。要進行企業規模的推論和訓練，都必須耗費大量系統資源，必須有效滿足這些需求，才能提供理想的最終使用者體驗。為達到此目標，除了傳統 HPC 工作負載以外，還必須兼顧生命科學、材料科學、製造業、模擬/建模及金融等領域不斷成長的需求。在此趨勢下，全球 HPC 市場預計以 7.7% 複合年增率成長，到 2026 年將達到 592 億美元規模[1]。

7.7% 全球 HPC 市場成長幅度[1]
到 2026 年的複合年增率

592 億美元 全球 HPC 支出[1]
到 2026 年

維持 HPC 系統效能平衡

第 4 代 Intel® Xeon® 可擴充處理器提供突破性效能，有助於 HPC 工作負載快速實現價值。該平台採用全新架構，具有較高的單位核心效能，每插槽最多 60 個核心，每系統可配置 2 個、4 個或 8 個插槽，相當於每核心密度高達 120 個執行緒，較前代產品增加 50%。

為因應核心數目的增加，此平台搭載了更先進的記憶體與 I/O 子系統，以利維持效能平衡。DDR5 記憶體的頻寬與速度最高可達 4800 MT/s，為 DDR4 的 1.5 倍。此外，此平台每插槽配有 80 條 PCIe 5.0 通道，I/O 規格大幅優於早期平台，並且提供 CXL (Compute Express Link 1.1)，以支援高架構頻寬及外掛加速器效率。

更高速的效能表現，適用於成長最快的工作負載。在 CPU 市場上，第 4 代 Intel® Xeon® 可擴充處理器擁有最多內建加速器，可提高 AI、分析、網路、儲存及 HCP 效能，包括下列各類工作負載：

- **選擇權定價。**增加人工智慧的運用，藉此因應緊迫的決策時限、複雜且要求繁多的應用程式，以及瞬息萬變的市場需求。
- **生命科學應用。**不斷微調改善模型，並執行大規模的計算，藉此提高模擬準確度，以利快速發現並改善研究成效。
- **電腦輔助工程。**支援電腦輔助工程應用程式迅速得出結果，協助您降低成本、改善產品安全性與設計，以及加快上市時間。

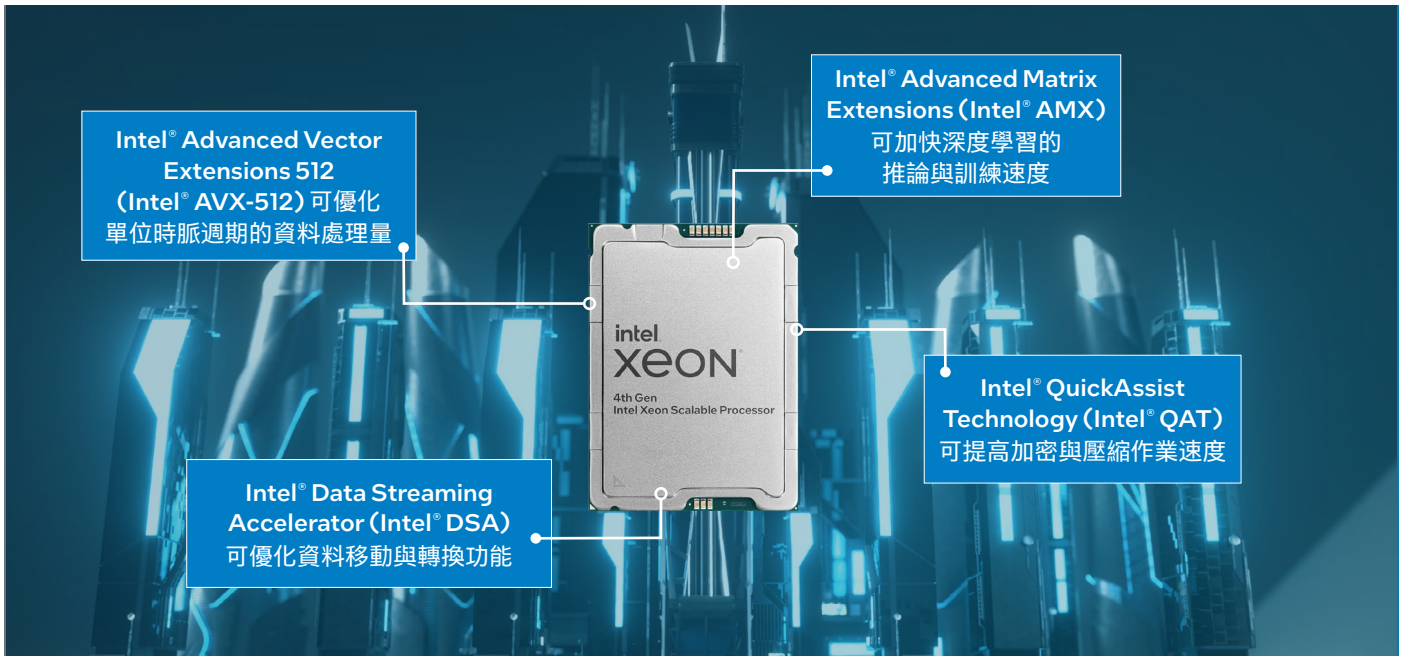
第 4 代 Intel® Xeon® 處理器以內建硬體加速器為基礎，造就全新效能典範，包括適用於高效能運算的 Intel® HPC 引擎。

intel. XEON

效能證據

高達 **1.56 倍** 28 款常見 HPC 工作負載效能相較於前代產品的增幅[2,3]
幾何平均數

Intel® HPC 引擎



效能證據

高達 **1.68 倍** 幾何平均數　LAMMPS 工作負載效能相較於前代產品的改善幅度[2,3]

基於內建加速器的進階功能

隨著工作負載日趨複雜，以及對運算資源的需求不斷成長，我們得以從 CPU 核心卸載特定功能，並將這些資源保留下來，用於執行業務關鍵型工作負載任務。這類功能包括 AI、安全性以及通用儲存與網路功能。

直接內建在第 4 代 Intel® Xeon® 可擴充處理器晶片中的硬體加速器提供特定功能，有助於改善平台內資料移動與處理的成效。這些加速器內建於處理器中，比起在核心上執行的獨立式解決方案或軟體型解決方案，較不會對 PCIe 匯流排造成輸出延遲，且在能耗方面也較具節約效益。採用這些內建加速器的使用案例，有助於提升效能、節省資本支出與營運支出：

- **效能。**專門特製的加速器旨在顯著提高目標工作負載的資料處理量。
- **設備成本。**由於加速器內建於第 4 代 Intel® Xeon® 可擴充處理器中，因此不需另行挹注設備投資。
- **營運成本。**由於較不需為設備機架添置核心，因此內建加速器可以顯著節省能源。

Intel® Advanced Matrix Extensions (Intel® AMX)：加快深度學習速度

事實證明，機器學習可以有效調整 HPC 工作負載，提高其效率與成效。Intel® AMX 是一款內建硬體加速器，透過加速深度學習演算法核心的張量處理，大幅提升推論與訓練效能。該技術包含 TILE (每核心最多包含 8 個可擴充 2D 暫存器區塊，可儲存比前代產品更大份額的資料) 以及 TMUL (區塊矩陣乘法)，此為一組矩陣乘法指令，是 TILE 上的第一個運算子。Intel® AMX 可支援深度學習軟體在特定時段內完成更多推論或加速趨近解決方案，進而縮短實現價值的時間。

Intel® Advanced Vector Extensions 512 (Intel® AVX-512)：最新的 x86 向量指令集

在許多代技術中，越來越複雜的向量化有助於在更大的資料集上更快地進行計算。Intel® AVX-512 是最新的 x86 向量指令集，奠基於前代產品的向量處理功能，可加速完成資料密集型工作負載。HPC 運算應用程式可以在 512 位元向量內的每個時脈週期打包 32 次雙精度和 64 次單精度浮點運算，以及具有兩個 512 位元融合乘加 (FMA) 的 8 個 64 位元和 16 個 32 位元整數單元，適用於最苛刻的運算工作負載，以推動商業智慧。與 Intel® Advanced Vector Extensions 2 (Intel® AVX2) 相比，該技術使資料暫存器的寬度、暫存器數量及 FMA 單元的寬度新增一倍。

Intel® Data Streaming Accelerator（Intel® DSA）：將串流資料移動最佳化

資料移動與轉換作業對於儲存、網路及資料密集型工作負載（如 HPC 中的分析應用）的效能至關重要。Intel® DSA 透過卸載在大規模部署中造成開銷最常見的資料移動任務，提高這些功能的效能。透過承擔幾乎所有的資料移動作業，包括校驗和、記憶體比較及檢查點，Intel® 減輕 CPU 核心在記憶體、儲存及網路子系統中移動資料的相關開銷。Intel® DSA 優化 CPU、記憶體和快取，以及所有連接的記憶體、儲存和網路裝置之間的串流資料處理。

Intel® QuickAssist Technology（Intel® QAT）：加速加密與壓縮

減少與加密和資料壓縮相關的開銷，可以在提高整體叢集效能方面發揮重要作用。Intel® QAT 內建為第 4 代 Intel® Xeon® 可擴充處理器中的硬體加速器，可實現更快的資料加密與解密，以及更高效的資料壓縮。與前幾代技術相比，該技術的最新版本可加速加密密碼、安全散列、公鑰加密及壓縮/解壓縮效能。透過從處理器核心卸載這些任務，Intel® QAT 為其他工作釋放資源，從而提高整體輸送量。Intel® QAT 有助於實現零信任安全性策略，在任何基礎設施的所有階段保護資料，無論是靜止、飛行還是使用中，都不會損失關鍵工作負載的效能。

關鍵技術

高達 60 核心	高達 8 條通道	高達 80 條
每個插槽	DDR5 高達 4800 MT/s	PCIe 5.0，每個插槽
增幅高達 50%[2]；每個系統 2、4 或 8 個插槽	記憶體、頻寬與速度最多新增 50%[2]	新增 I/O 容量[2]

第 4 代 Intel® Xeon® 處理器提供更高的每個核心浮點效能及一系列內建硬體加速器

開發人員啟用與支援

Intel® oneAPI 工具組是 Intel 長期致力於 HPC 軟體生態系統的進化結晶。此產品提供編譯器、資料庫及效能工具，可簡化開發路徑，打造出最適合 Intel® 架構的高品質軟體。該工具組為開發人員提供一條快速採用途徑，因為他們希望利用第 4 代 Intel® Xeon® 可擴充處理器內建加速器以及開放的、基於標準的軟體開發堆疊。開發人員可以使用 Intel® oneAPI 工具組來產生程式碼，從而在整個 Intel® 體系結構中提供加速效能，包括具有內建加速器、GPU 及 FPGA 的 CPU。

Intel® oneAPI 基礎工具組

用於開發以資料為中心的高效能應用程式核心編譯器、庫（包括 Intel® oneAPI 數學核心庫）及其他工具

Intel® oneAPI HPC 工具組

Intel® Fortran 編譯器、OpenMP GPU 卸載以及訊息傳遞介面（MPI）的可擴充性

Intel® AI 分析工具組

最佳化架構和 Python 庫，以利加速資料科學及分析管道

Intel® oneAPI 渲染工具組

渲染與光線追蹤庫，可實現高效能、高保真度的影像體驗

基於 Intel® oneAPI 的開放標準程式碼開發，得益於包括開源工具、API 及驅動程式的大型開放生態系統。這種彈性有助於企業降低將新服務與解決方案推出市場的複雜性、成本及時間要求，簡化新架構的採用，並使工程師與程式設計師能夠進行創新而不只是維護程式碼。

易於與現有系統整合

借助 Intel® 技術，企業可以加快他們所了解和使用的最大合作夥伴生態系統的部署時間。世界各地的硬體與軟體供應商以及解決方案整合商在 Intel® Xeon® 可擴充處理器上建置產品，提供最大的選擇與互通性，並保證數千種真實世界的實際應用。

效能證據

更高的 VASP 效能[2,3]

高達 **1.61 倍** 幾何平均數 在第 4 代 Intel® Xeon® 可擴充處理器上，與上一代相比

高達 **2.01 倍** 幾何平均數 在 Intel® Xeon® 處理器 MAX 系列上，相較於 2S 第 3 代 Intel® Xeon® 可擴充處理器

適用於所有類型的 HPC 使用案例

第 4 代 Intel® Xeon® 處理器透過使用 DDR5 的高效能、改進的記憶體頻寬以及使用 PCIe 5.0 與 CXL 1.1 的進階 I/O，加速一系列真實世界中的使用案例。使用 Intel 領先的最佳化軟體程式庫與編譯器，開發人員得以更快建置程式碼，使 HPC 應用程式開箱即有更出色的效能，無需額外調校或優化。程式碼與模型可以利用強大的 Intel® AVX-512 技術，每個核心有兩個 FMA 單元，用於最苛刻的運算工作負載。使用 Intel® MPI 程式庫，可以跨多個 HPC 叢集擴充工作負載。新增 Intel® Optane™ 持續性記憶體，支援從更大的記憶體佔用中受益的大型計算。

支援 HPC 工作負載的 Intel® 技術讓您事半功倍

極致優化頻寬。全新 Intel® Xeon® 處理器 Max 系列可解除受限於記憶體資源的工作負載瓶頸，大幅提升建模、AI、HPC 及資料分析等任務的效能，最高比僅使用 DDR5 的平台多出 4 倍。這是首款將高頻寬記憶體與加速器整合到處理器封裝中的 x86 CPU，具有高達 64 GB 的 HBM2e。本產品減少對 DDR 的依賴，具備最新軟體工具及出色的程式碼可重用性，進而改善 TCO 的表現。

極致影響力。旗艦級 Intel® 資料中心 GPU Max 系列採用 Intel 最先進的 IP 與封裝技術，旨在加速 Exascale 時代的 AI、HPC 運算及進階分析工作負載。本產品基於 Intel® X^{e} HPC 架構，提供 GPU 中最大的高頻寬快取。該 GPU 在具有 SIMT/SIMD 彈性的 Intel® oneAPI 開放生態系統支援下，在封裝內整合多種 IP 創新，包括高頻寬記憶體。

微秒級資料存取速度。DAOS（Distributed Asynchronous Object Storage，分散式非同步物件儲存）是一種開放原始碼的軟體定義擴充物件儲存區，可透過具成本效益的方式，在單一儲存層中為 HPC 和 AI 應用程式提供高頻寬、低延遲、高每秒 I/O 作業數（IOPS）的儲存容器。DAOS 原生支援結構化、半結構化及非結構化資料集，同時消除傳統分佈式儲存的限制。

進一步瞭解

www.intel.com/xeon/scalable
www.intel.com/hpc

[1] Intersect360 Research，2022 年 5 月 20 日。「Intersect360 Research 表示，2021 年 HPC 市場總營收成長 5.2%，達到 410 億美元。」https://www.hpcwire.com/off-the-wire/total-hpc-market-revenue-grew-5-2-to-41-0-billion-in-2021-says-intersect360-research/。

[2] 與 2S 第 3 代 Intel® Xeon® 可擴充處理器相比。

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

加速器供貨情況因 SKU 而異。如需額外的產品詳細資料，請造訪 Intel 產品規格頁面。

效能因使用情形、配置和其他因素而異。如要進一步瞭解，請至 https://www.intel.com/PerformanceIndex。

效能結果係依配置中所示日期的測試為準，且可能無法反映所有公開可用的更新。詳情請參閱配置信息披露。
任何產品或元件都無法提供絕對的安全性。

Intel 並不控制或審核第三方的資料。您應該參考其他來源以評估準確性。

您的成本和成果可能有所差異。

Intel 技術可能需要搭配合稱的硬體、軟體或服務啟動。

您不得使用或利用此文件進行與此處所述之 Intel 產品相關的任何侵權或其他法律分析行為。您同意將爾後所擬，包含此處揭露之主題的任何專利請求，授予 Intel 非獨家、免權利金之許可。

此處所述的產品可能包含設計瑕疵或錯誤（稱為勘誤內容），可能導致產品不符合先前發佈的規格。一經要求，我們將提供當前描述的非重要勘誤。

© Intel 公司。Intel、Intel 圖誌和其他 Intel 標誌是 Intel 公司或其子公司的商標。其他名稱與品牌可能業經宣告為其他所有者之財產。

1122/JAW/MESH/PDF 350500-003US