白皮書

Supermicro X14 HGX B200 GPU 伺服器搭配 Intel® Xeon® 6 處理器，提供最高 1.44 倍的並行使用者容量

Intel® 先進矩陣擴展 (Intel® AMX) 透過 CPU-GPU 協同服務 (Co-Serving) 強化 vLLM 推論效能：結合 Supermicro HGX B200 系統的多代理 (Multi-Agent) 架構

Supermicro GPU 伺服器 SYS-822GS-NBRT

目錄

執行摘要

現代大型語言模型 (LLM) 服務系統通常將大部分推論工作負載執行於 GPU 上，而 GPU 系統中的主機 CPU 在完成排程、分詞 (tokenization) 與資料傳輸後，往往處於閒置狀態。本文提出一種異質架構，透過在 CPU 與 GPU 上共同執行 vLLM，並利用多代理 (multi-agent) 或 vLLM 語意路由 (semantic routing) 機制，在兩個推論端點 (CPU 路徑與 GPU 路徑) 之間分配工作負載，以提升整體系統效率。此架構不再將所有請求一律交由 GPU 上的大型模型處理，而是在不需要大型模型能力時，將較小型、部署於 CPU 上的 vLLM 用於研究、資料擷取、驗證與短回應等任務。如此一來，GPU 上的 405B 模型可專注於長文本生成與複雜推理，而搭載 Intel Xeon 6 與 Intel® AMX 的 CPU 則可平行處理適合 CPU 的工作負載。最終可達成更有效的運算資源利用、更高的使用者併發能力，以及在混合 LLM/SLM 生產環境中更具優勢的「每併發使用者成本」。

此架構由三個協同運作的代理 (agents) 組成: CPU 端代理 (Researcher 與 Reviewer) ：負責上下文準備、快速回饋與內容驗證。GPU 端代理 (Writer) ：專注於高品質回應生成。這種工作分工反映實務應用模式：使用成本較低的 CPU 推論處理事實查詢、摘要、評論與安全檢查 (guardrail) ，而將 GPU 的高價值運算資源保留給最能受益於大型 405B 模型的任務。此閉環 (closed-loop) 設計可同時維持 CPU 與 GPU 的高利用率，降低閒置資源，並提升單一系統可服務的總使用者數量。

從行銷與客戶價值定位角度來看，其價值不僅在於速度提升。相同的 GPU 伺服器可支援分層式推論架構：將 GPU 資源保留給高價值生成任務，並由 CPU 處理短時、重複性高或需大量驗證的工作。這有助於降低因每新增使用者需求而擴充 GPU 資源的必要性，並在包含 CPU 適用任務的工作負載組合下，改善每併發使用者的實際成本效益。

CPU-GPU 協同服務在 Supermicro 系統上提升併發能力 (Intel Xeon 6 + AMX)

本實作基於異質架構擴展 vLLM，將推論分散至兩個協同運作的端點: CPU 路徑：執行 Llama-3.1-8B 模型，用於輕量推理 (reasoning) 與計算密集型前處理任務，例如資料擷取、摘要、評論與回應驗證。GPU 路徑：執行 Llama-3.1-405B 模型，用於高計算量生成任務，包括長文本生成、深度推理與多步驟合成 (multi-step synthesis)。

系統技術特點如下：採用 vLLM 多進程執行器 (multiprocessing executor) 、支援 NUMA 感知 CPU 綁定 (NUMA-aware CPU binding) 、透過 OpenMP 調校最佳化 CPU 效能、利用 Intel Xeon 6 的 AMX 加速 BF16 推論與 KV cache 管理、NVIDIA Blackwell 架構 GPU (如 B200) 負責處理複雜推論任務。CPU 與 GPU 端點皆以獨立的 vLLM 服務形式部署，並由多代理應用進行協調控制。參考 GitHub 實作中，Researcher 與 Reviewer 對應至 CPU 端 8B 模型，而 Writer 則對應至 GPU 端 405B 模型。CPU 服務透過 Intel Xeon 6 AMX 執行 BF16 推論，GPU 服務則在 Supermicro HGX B200 系統上提供高吞吐量解碼能力。系統亦透過 CPU 綁定機制，自動生成 Docker Compose 覆寫設定，依據 Intel Xeon 6 CPU 拓樸進行最佳化配置。GPU 服務綁定於 NUMA 本地 CPU，並可將部分 PCT 核心分配給延遲敏感的 GPU 執行緒，其餘 CPU 核心則用於 CPU 推論任務。相同工作流程同時支援部署 (deploy) 與效能測試 (benchmark) 模式，適用於 CPU 與 GPU 服務。

效能測試結果顯示，在結合 CPU 與 GPU 運算資源的異質架構中，搭載 Intel® AMX 的 Intel Xeon 6 處理器展現出顯著優勢。此基準測試採用官方 vLLM Docker 映像檔與 vLLM 測試工具套件，分別針對 CPU 與 GPU 進行測試，並依照 vLLM 手動測試流程執行，包括自適應併發搜尋 (adaptive concurrency search) ，以找出最接近服務等級門檻的最大併發數。在測試配置中，CPU 端點執行 Llama-3.1-8B-Instruct (128 輸入 / 128 輸出 tokens) ，用於輕量推理、流程協調與驗證；GPU 端點 (Supermicro HGX B200) 則執行 Llama-3.1-405B-Instruct (2048 輸入 / 2048 輸出 tokens) ，用於大規模生成與複雜推理。在 Intel 測試中，CPU 與 GPU 的 vLLM 容器皆在持續處理使用者請求的狀態下進行量測，符合實際全系統運作模式。在僅使用 GPU (B200) 的情境下，系統使用 12 個 CPU 核心，支援 127 位併發使用者；當進一步利用剩餘的 Intel Xeon 6 核心執行 CPU 端 8B 模型時，可額外支援 56 位使用者，使總併發數提升至 183 位，達到最高約 1.44 倍 的併發提升，相較於僅 GPU 的基準顯著增強。

需注意的是，此測試為 LLM/API 端點層級的服務測試，而非完整多代理應用流程的端到端測試。此外，基準測試亦顯示 CPU 綁定 (CPU binding) 的重要性：若未進行 CPU 綁定，讓 CPU 模型在所有核心間競爭資源，CPU 端可支援的使用者數將由 56 降至 51，約下降 9%。

此種平行執行架構可有效在整體流程中分配工作負載，並最大化 CPU 與 GPU 的資源利用率。透過 Intel Xeon 6 搭配 Intel® AMX 加速，系統可同時在 CPU 與 GPU 上執行工作負載，使整體併發能力最高提升至 1.44 倍，展現更高的吞吐量與服務效率。本評估基於 Supermicro SYS-822GS-NBRT HGX 平台 (搭載 2 顆 Intel Xeon 6 6776P 處理器 (支援 AMX) 與 8 張 NVIDIA HGX B200 GPU)，驗證 CPU-GPU 協同服務在可擴展 vLLM 推論場景中的效益。若需重現測試，可使用 Intel AI TCE vLLM CPU binding 分支的部署腳本，並採用 Benchmark 模式。vLLM 測試流程說明：CPU 測試需使用 CPU 容器映像並設定 ON_CPU 環境變數，並可透過 ENABLE_ADAPTIVE_CONCURRENCY=1 啟用自適應併發功能。測試流程將產生 benchmark_results.md 與 benchmark_results.json，用於比較不同併發數、TTFT (首次回應時間) 與 TPOT (每 token 輸出時間) 等指標。

在多代理應用情境中，Researcher (CPU) ：最先啟動，負責資料準備，Writer (GPU) ：在取得部分上下文後即開始生成，Reviewer (CPU) ：於初稿產生後立即進行驗證。這種分階段且重疊的執行方式，使各代理可進行管線化 (pipeline) 處理，較傳統完全序列化流程顯著降低整體延遲。

沒有管線化 (Pipelining) 的情況下，執行過程是完全循序的，因此端到端的延遲會在各個階段累加:

Total latency = Research + Write + Review

在多代理管線化設計下，CPU 與 GPU 階段會並行執行，並且具有重疊的執行過程:

Total latency = MAX (Research, Write, Review)

這使得系統能夠採用管線化執行模型：CPU 代理持續進行情境的準備、擷取與驗證，而 GPU 代理則專注於生成任務，從而實現跨階段的重疊執行，而非循序處理。因此，端到端的延遲從原本的累加式管線 (additive pipeline) 轉變為平行化限制 (parallelized bound)，大幅提升硬體利用率、減少 GPU 閒置時間、穩定尾部延遲，並在生產規模的 vLLM 部署中提高整體並行處理能力。

Block Diagram:

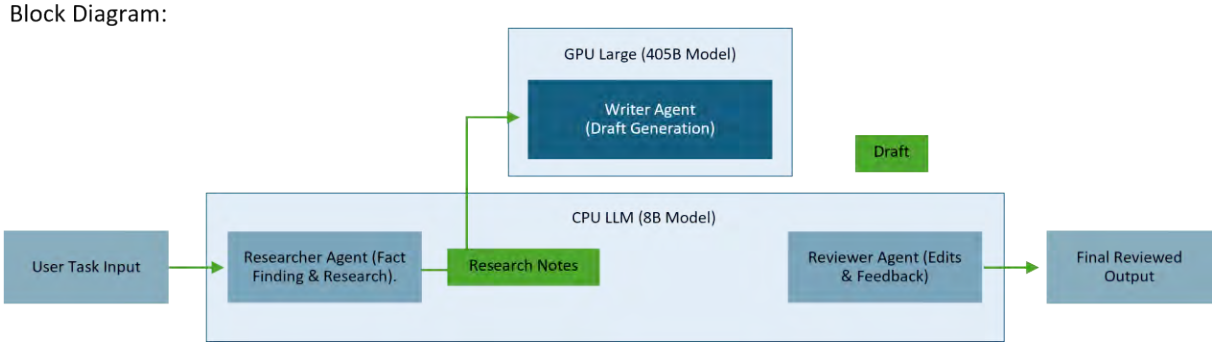

圖 1 - 硬體設計方塊圖

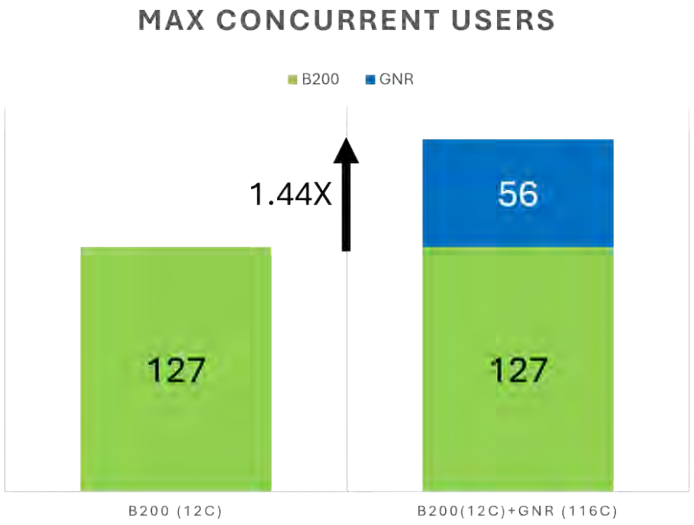

圖 2 - 使用者並行處理效能比較：純 GPU vs. CPU-GPU 共同服務[1]

什麼是 Intel® Advanced Matrix Extensions (Intel® AMX)

Intel® AMX 是整合於 Intel® Xeon® 6 處理器 (搭載 P-cores) 的兩大 Intel® AI 引擎之一，能幫助您充分利用 CPU 資源，大規模執行 AI 訓練與推論工作負載，帶來提升效率、降低推論、訓練與部署成本，以及減少整體擁有成本 (TCO) 等優勢。作為每個 CPU 核心內建的加速器，且靠近系統記憶體，Intel® AMX 通常比獨立加速器更容易使用，從而加快實現價值的速度。

Intel® AMX Works 的運作原理

Intel® AMX 是 Intel® Xeon® Scalable 處理器核心中的專用硬體區塊，能夠優化並加速依賴矩陣運算的深度學習訓練與推論工作負載。

Intel® AMX 讓 AI 工作負載得以直接在 CPU 上執行，而無需卸載至獨立加速器，從而帶來顯著的效能提升。其架構支援 BF16 (訓練/推論) 與 int8 (推論) 資料型態，並包含兩個主要組件:

- Tiles: 由八個二維暫存器組成，每個暫存器大小為 1 KB，可儲存大量資料區塊。
- Tile Matrix Multiplication (TMUL): TMUL 是連接至 Tiles 的加速引擎，負責執行 AI 的矩陣乘法運算。

這兩個組件共同作用，讓 Intel® AMX 能在每個核心儲存更多資料，並在單一操作中計算更大的矩陣。此外，Intel® AMX 的架構設計具備完整的可擴展性與延展性。

用於測試的 Supermicro 系統

Supermicro SYS-822GS-NBRT 是一款 8U GPU 系統，專為將八張 Supermicro HGX B200 GPU 與雙 Intel Xeon 6 處理器整合至單一平台而設計。在本次測試中，採用 CPU 綁定技術，為 GPU 上的 405B 服務分配所需 CPU 資源，同時保留其餘核心用於 CPU 上的推論。測試所使用的軟體堆疊包括官方 vLLM Docker 映像檔、vLLM 基準測試套件、vLLM 0.17.0 公開版本、來自 Intel AI TCE vLLM CPU_binding 分支的 CPU 綁定腳本，以及公開的 vLLM CPU 版本容器。

System	SYS-822GS-NBRT
CPUs	Dual Intel Xeon 6776P processors (64 cores, 350W TDP)
Memory	2048GB (32x64GB DDR5 6400MT/s [5200MT/s])
GPUs	NVIDIA HGX B200 8-GPU with 5th Generation NVLink® 1.8TB/s, 2.3TB of HBM3e GPU memory per system

摘要

在配備 2 顆 Intel Xeon 6 6776P 處理器與 8 張 Supermicro HGX B200 GPU 的 Supermicro SYS-822GS-NBRT HGX 平台上，Intel 測試報告顯示，CPU-GPU 共同服務配置的支援使用者並行處理容量最高可達純 GPU 基準測試的 1.44 倍。此結果來自一個簡單但重要的部署洞察：405B GPU 服務僅需使用主機 CPU 的一小部分資源，而剩餘的 Intel Xeon 6 核心則可執行 8B CPU 模型，進行有用的代理工作。透過 NUMA 感知的 CPU 綁定技術，以及可選的 Priority Core Turbo 指引，平台能在保護對延遲敏感的 GPU 服務的同時，利用閒置的 CPU 容量進行 CPU 推論。此模式最適合需要同時部署小型與大型語言模型的應用，包括多代理管線、LLM/SLM 共同服務，以及將合適請求路由至較小模型的語意路由架構。對於部署混合 LLM 與 SLM 應用的組織而言，這提供了一條極具吸引力的途徑，能在同一台伺服器上服務更多使用者，並提升成本效益，而無需為每個輕量級或驗證導向的任務額外增加 GPU。

更高的使用者並行處理容量證明，CPU–GPU 共同服務與管線化執行，能大幅提升單一 AI 加速系統中端到端 vLLM 服務的效率、可擴展性與資源利用率。整體而言，這些結果驗證了異質共同服務是一種實用且可擴展的策略，可優化生產環境中的 vLLM 工作負載，並展現 Intel® AMX 加速技術所帶來的效能優勢。

For More Information:

Supermicro SYS-822GS-NBRT: https://www.supermicro.com/en/products/system/gpu/8u/sys-822gs-nbrt
Intel® Advanced Matrix Extensions (Intel® AMX) Details:
https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/what-is-intel-amx.html

Intel AI TCE CPU Binding Deployment and Benchmark Mode: https://github.com/intel-ai-tce/vllm/tree/cpu_binding/benchmarks/cpu_binding
Intel AI TCE Multi-Agent Two-Endpoint Demo: https://github.com/intel-ai-tce/vllm/tree/cpu_binding_demo/benchmarks/cpu_binding/multi_agent_two_endpoints
vLLM Semantic Router Use Case Example: https://vllm-semantic-router.com/
vLLM Benchmark Manual: https://docs.vllm.ai/en/latest/benchmarking/dashboard/#manually-trigger-the-benchmark

SUPERMICRO

作為高效能、高效率伺服器技術與創新的全球領導者，我們開發並提供端到端的綠色運算解決方案，涵蓋資料中心、雲端運算、企業 IT、大數據、HPC 以及嵌入式市場。我們的 Building Block Solutions® 策略讓我們能夠提供廣泛的 SKU 選擇，並根據您的需求打造與交付應用優化解決方案。敬請蒞臨 www.supermicro.com

INTEL

Intel (Nasdaq: INTC) 是產業領導者，致力創造改變世界的科技，推動全球進步並豐富人類生活。受摩爾定律啟發，我們持續精進半導體的設計與製造，協助解決客戶最重大的挑戰。透過將智能嵌入雲端、網路、邊緣運算以及各類運算裝置，我們釋放資料的潛力，進而改善商業與社會。欲了解更多 Intel 的創新成果，敬請蒞臨

www.intel.com

[1] Results may vary based on system configuration, workload characteristics, hardware environment, service level objectives, and other operational factors. Performance improvements are dependent on proper setup and optimization. Configuration: 1-node, Supermicro X14DBG-LC+, 2x Intel(R) Xeon(R) 6776P, 64 cores per socket, 350W TDP, HT On, Turbo On, Total Memory 2048GB (32x64GB DDR5 6400MT/s [5200MT/s]), BIOS 1.5, microcode 0x1000405, 3x Unknown NIC, 2x Ethernet Controller X710 for 10GBASE-T, 1x 7T SAMSUNG MZTL67T6HBLC-00AW7, 1x 3.5T SAMSUNG MZTL23T8HCLS-00A07, Ubuntu 22.04.5 LTS, 6.8.0-90-generic. Software: official vLLM Docker images, vLLM benchmark suite, vLLM 0.17.0 public release, CPU-binding scripts from https://github.com/intel-ai-tce/vllm/tree/cpu_binding/benchmarks/cpu_binding, and public.ecr.aws/q9t5s3a7/vllm-cpu-release-repo:v0.17.0. Workload: Llama-3.1-8B-Instruct on Intel Xeon 6 CPU with BF16/AMX BF16, 128 input tokens and 128 output tokens; Llama-3.1-405B-Instruct on Supermicro HGX B200 with BF16, 2048 input tokens and 2048 output tokens. Benchmark method used adaptive concurrency search under SLA thresholds described in the vLLM benchmark manual. CPU and GPU endpoint measurements were taken while both CPU and GPU services were kept busy with user requests. The result is an LLM/API endpoint serving a benchmark, not an end-to-end multi-agent application benchmark. Test by Intel as of March 21, 2026.