產品簡介

Supermicro 基於 ARM 的機架規模解決方案提供高端效能，以支援 Agentic AI 工作負載的快速成長

新系統滿足對 token 密集型常時運作代理所需的高能源效率伺服器的日益增長需求

Supermicro 基於 Arm AGI CPU 的機架規模解決方案

目錄

執行摘要

AI 代理的快速普及正創造出對能源效率系統的需求，這些系統必須在預期時間範圍內提供所需的結果。雖然 AI 訓練近年來備受關注，但對推論與 Agentic 工作負載日益增長的需求，呼喚著能在多樣環境中運作的能源最佳化系統。Agentic AI 的興起將需要能夠協調調度與管理持續運作 agent 的 CPU。這一轉折點將需要在機架規模密度下部署大量 CPU，同時將能源消耗控制在現有企業資料中心的電力範圍內。

Supermicro 搭載 AI 導向 Arm® AGI CPU 的解決方案，正是為大規模 Agentic AI 協調調度時代所打造，提供高效能、高效率與高密度的特性，最大化機架規模部署的經濟效益。Supermicro 憑藉領先市場的實績、結合大規模資料中心建構區塊解決方案® (DCBBS) 的部署能力，以及自研熱管理技術，打造出專為現代 Agentic AI 優化的全新架構類別。

目標工作負載

Agentic 工作負載是指能夠代表使用者執行動作的系統，涵蓋廣泛的應用範圍。其中包括但不限於客戶支援自動化、程式碼撰寫以及供應鏈管理。在企業內，其他「代理」還可能涵蓋人力資源與新人入職流程，以及財務相關工作負載。創新的 Agentic 系統還能隨著時間學習、修正錯誤，並與其他代理進行溝通。

Supermicro 系統

Supermicro 的新系統採用最新的 Arm 技術，在兩款專為 Agentic AI 客製設計的系統中搭載 Arm AGI CPU。這兩款系統皆具備以下特性:

- **效能:** 每顆 CPU 最高搭載 136 個 Arm Neoverse® V3 核心，提供領先的核心、SoC、刀鋒伺服器以及機架級效能，每核心記憶體頻寬達 6 GB/s，延遲低於 100ns。Supermicro 系統支援雙插槽配置。

- **規模:** 基礎 TDP 為 300 瓦，搭配每個程式執行緒專用核心，能夠在持續負載下提供穩定的確定性效能，避免節流 (throttling) 與閒置執行緒。

- **效率:** 支援高密度 2U 氣冷伺服器機箱，可在氣冷部署下達到每機架 6528 核心，而液冷系統更可達到每機架 26,112 核心。

Supermicro 的全新超高密度系統專為需要極高密度運算的 AI 推論與 Agentic AI 環境所設計。此系統可單機訂購，或以機架規模配置部署。其系統組件如下:

系統 SKU: ARS-222H-NR - 通用運算 (包含 Agentic AI、邊緣運算、雲端與記憶體密集型工作負載)

- 2U 高度
- # of CPUs: 2
- Arm AGI CPU Neoverse V3 (64、128 或 136 核心)
- 記憶體:
  - 插槽數：24 個 DIMM 插槽
  - 最大記憶體容量 (1DPC)：最高 6TB ECC DDR5-8800 MT/s RDIMM
- 1 個 OCP 3.0 相容 AIOM
- 最多 8 個前置熱抽換 2.5” NVMe 磁碟槽
- 冗餘 2700W 鈦金級電源供應器
- 最高支援 2 張 GPU

系統 SKU: ARS-522GP-NR - Agentic AI 推論與 AI 訓練

- 5U 高度
- # of CPUs: 2
- Arm AGI CPU Neoverse V3 (64、128 或 136 核心)
- 記憶體:
  - 插槽數：24 個 DIMM 插槽
  - 最大記憶體容量 (1DPC)：最高 6TB ECC DDR5-8800 MT/s RDIMM
- 1 個 OCP 3.0 相容 AIOM
- 最多 8 個前置熱抽換 2.5” NVMe 磁碟槽
- 最高支援 6 組 N+N 冗餘 2700W AC 熱抽換電源供應器
- 最高支援 8 張雙寬度 GPU

系統 SKU: ARS-212HE-FNR

- 2U 高度
- # of CPUs: 1
- Arm AGI CPU Neoverse V3 (64、128 或 136 核心)
- 記憶體:
  - 插槽數：12 個 DIMM 插槽
  - 最大記憶體容量 (1DPC)：最高 3TB ECC DDR5-8800 MT/s RDIMM
- 1 個 OCP 3.0 相容 AIOM
- 最多 6 個前置熱抽換 2.5” NVMe 磁碟槽
- 1+1 冗餘 2000W 鈦金級電源供應器
- 最高支援 2 張 GPU

系統 SKU: ARS-242TP-QNR-LCC

- 2-OU 高度
- 4 個獨立節點
- 每個節點 CPU 數量：2
- Arm AGI CPU Neoverse V3 (64、128 或 136 核心)
- 每個節點記憶體:
  - 插槽數：24 個 DIMM 插槽
  - 最大記憶體容量 (1DPC)：最高 6TB ECC DDR5-8800 MT/s RDIMM
- 每個節點 2 個 OCP 3.0 相容 AIOM
- 每個節點 2 個前置熱抽換 NVMe 磁碟槽
- 液冷

系統 SKU: ARS-142TP-QNR-LCC

- 1OU 高度
- 每個節點 CPU 數量：2
- Arm AGI CPU Neoverse V3 (64、128 或 136 核心)
- 每個節點記憶體:
  - 插槽數：24 個 DIMM 插槽
  - 最大記憶體容量 (1DPC)：最高 6TB ECC DDR5-8800 MT/s RDIMM
- OCP 相容 AIOM
- 每個節點 2 個前置熱抽換 NVMe 磁碟槽
- Open Rack DC -48V 並搭配各節點 PDB
- 液冷

機架規模配置

雖然單一搭載 Arm AGI CPU 的伺服器效能已相當出色，但在機架規模下，其表現更是驚人。根據不同的機架配置，單一機架可為客戶提供超過 26,000 個 Neoverse V3 核心。

機架配置 1: (48U 機架，外部網路)

- 24 台 2U 伺服器
- 48 顆 Arm AGI CPU
- 每機架 6,528 核心

機架配置 2: (48U 機架，外部網路)

- 9 台 5U 伺服器
- 18 顆 Arm AGI CPU
- 每機架 2,448 核心
- 最高支援 72 張雙寬度 GPU

機架配置 3: (48U 機架，外部網路)

- 16 台 2U 4N (節點)
- 每個節點搭載雙 Arm AGI CPU，總計 128 顆 Arm AGI CPU
- 每機架 26,112 核心
- 液冷

機架配置 4: (ORW - 48U 機架，外部網路)

- 每機架 168 台伺服器
- 每機架 336 顆 Arm AGI CPU (42 台 1U 4N)
- 每機架 45,696 核心
- 液冷

Arm AGI CPU

Arm AGI CPU 採用先進的 3nm 製程技術製造。該處理器最高支援 136 個 Neoverse V3 核心、12 通道 DDR5-8800 記憶體，以及 96 通道 PCIe Gen 6 與 CXL 3.0。最高 CPU 頻率在 136 核心型號可達 3.5 GHz，64 核心型號可達 3.7 GHz，每核心配有 2 MB 專用 L2 快取。每核心搭載雙 128 位元 SVE2 (Scalable Vector Extension 2) 單元，支援 bfloat16 與 INT8 MMLA 指令。

AI 的演進與 Agentic AI 的興起正在改變組織部署 AI 基礎設施的方式。

資料中心電源供應的固有限制、大規模 AI 資料中心電力成本的上升，以及容納不斷成長的 AI 叢集所需的實體空間，都需要在原始運算能力、運算密度與效率之間取得平衡。隨著 AI 工作負載規模擴大，這種平衡變得越來越重要。

Supermicro 的系統產品組合涵蓋從 1U 運算伺服器到機架規模 AI 叢集的各種外型規格。透過利用新款 Arm AGI CPU 更高的核心密度與每瓦效能，Supermicro 能在相同的電力範圍內，提供比傳統機架高達 2 倍的每機架效能與 2 倍的核心密度。這能在持續電力消耗與實體機房空間兩方面，為資料中心帶來顯著的成本節省。

- Arm AGI CPU 密集的 136 核心微架構專為高效能而設計，能減少傳統負擔，並在每個週期完成更多工作，實現持續且不節流的效能。
- 低每核心基礎 TDP 結合 Supermicro 業界領先的氣冷與液冷熱管理技術，提升能源效率與系統密度，實現比傳統系統高達 2 倍的每機架核心數。
- 每核心 6GB/s 記憶體頻寬與延遲最佳化的記憶體存取，支援線性擴展。
- 擴展的記憶體容量與靈活的 I/O，打造能源效率高且可擴展的 Agentic AI 基礎設施，讓 CPU 能夠在分散式基礎設施中協調調度數千個平行任務。

更多資訊

Supermicro Arm AGI CPUs: https://www.supermicro.com/en/solutions/arm-agi
Supermicro Arm AGI CPU-based server 2U: https://www.supermicro.com/en/products/system/hyper/2u/ars-222h-nr
Supermicro Arm AGI CPU-based server 5U: https://www.supermicro.com/en/products/system/gpu/5u/ars-522gp-nr
Supermicro Arm AGI CPU-based server, 2U, Single Socket: https://www.supermicro.com/en/products/system/hyper/2u/ars-212he-fnr
Supermicro Arm SGI CPU-based server, 2U, 4Node: https://www.supermicro.com/en/products/system/hyper/2-ou/ars-242tp-gnr-lcc

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。我們的建構區塊解決方案® (Building Block Solutions®) 方法，讓我們能夠提供廣泛的 SKU，並根據您的需求打造與交付應用最佳化的解決方案。敬請蒞臨 www.supermicro.com

ARM

Arm 開發了推動全球 AI 革命的運算平台。我們的高性能、高效率 CPU 產品深受全球領先半導體公司的信賴，至今已部署超過 3,500 億顆晶片——其中包括超過 99% 的智慧型手機。

敬請蒞臨 www.arm.com