解決方案簡介

Supermicro、AMD 與 Myrtle.AI 共同打造最佳化解決方案，突破金融 AI 微秒延遲障礙

消除機器學習在交易決策中的推論差距

Supermicro AMD 伺服器 - AS -2015CSTNR

目錄

執行摘要

在資本市場中，尋找超額報酬 (alpha) 已進入一個新時代。現代 AI 模型現在能夠產生超越任何傳統方法的市場訊號。然而，在金融領域成功應用 AI 仍面臨一個熟悉的障礙：**延遲**。對於交易公司而言，即使是最強的訊號，如果無法在微秒級的執行時間窗口內到達，也幾乎沒有價值。到目前為止，公司必須做出取捨：使用較簡單的模型來維持速度，或接受較慢的執行速度以使用更精密的智慧。

透過利用 AMD Versal™ Adaptive SoC 與 Silicom 的伺服器適配器技術，此解決方案已在 STAC-ML™ Markets (推論) 基準測試中創下新的世界紀錄，提供現代電子交易所需的確定性、超低延遲效能。

此解決方案簡報概述了在 Supermicro AS-2015CS-TNR 伺服器上執行的 Myrtle.AI VOLLO™ 加速堆疊的效能突破與商業價值，該成果已於近期發布的 STAC-ML 基準測試報告中展現。

產業龍頭矽廠推動 CPU 與 GPU 的創新，大幅提升了先進 AI 推論的效能，為一般金融工作負載顯著提高了「效能底線 (performance floor)」。然而，在極低延遲交易的最前端，依賴批次處理的工作負載所固有的「抖動 (jitter)」仍是重大障礙。一般用途加速器並未具備低延遲最佳化架構，且常在高市場波動時期遭遇「尾部延遲 (tail latency)」——無法預測的效能突增，可能導致滑價 (slippage) 與錯失成交 (missed fills)。

了解 STAC-ML™ 基準測試結果

了解 STAC-ML™ 基準測試結果: 這是首次有系統突破 LSTM 推論 99th percentile (99p) 低於 2 微秒的障礙。證券技術分析中心 (STAC®) 為金融產業提供客觀、經過審計的基準測試。STAC-ML Markets (推論) 基準測試套件 (SUT ID: MRTL260323) 專門測試系統對時間序列市場資料執行長短期記憶 (LSTM) 推論的能力，提供延遲、吞吐量與效率的「公平比較」。

- 三種不同複雜度的 LSTM 模型 (A、B、C)。
- 測量「tick-to-model」延遲——從接收資料到產生訊號的時間——在不同吞吐量下的表現。
- 受測系統在所有模型中均達成史上最低的 99th percentile (99p) 延遲，證明其在「極端」市場條件下的能力。

為何 STAC-ML™ 基準測試如此重要 —— 「推論差距」(Inference Gap)

當交易公司從簡單線性模型轉向深度學習時，常會面臨「延遲稅 (latency tax)」。傳統運算 (CPU) 或一般用途加速器 (GPU) 因批次處理需求，常引入抖動與較高延遲。此解決方案的重要性在於它消除了「推論差距」。交易公司現在能夠以簡單模型的速度部署 LSTM 等複雜模型的預測能力，實現更智慧、更具反應力的交易策略。

STAC-ML 的效能可直接轉化為真實金融工作負載:

- 演算法價格預測：分析限價委託簿 (limit order book, LOB) 以預測短期價格走勢。
- 流動性分析：識別「隱藏」流動性並最佳化訂單路由。
- 即時詐欺與合規：即時比對交易與合規模型，而不延遲執行路徑。
- 做市交易 (Market Making)：根據不同資產類別的相關性變化即時調整報價。

真實金融工作負載與 AI 驅動交易的轉型

金融服務產業正經歷從基於規則的系統，轉向機器學習驅動策略的典範轉移。此基準測試之所以重要，正是因為它解決了「推論差距」——即從接收市場 tick 到產生交易訊號之間的延遲。

使用 FPGA 於現代量化交易的效益

- 平行處理：FPGA 以真正的平行方式處理資料串流，非常適合 ML 推論。
- 直接網路到推論：Silicom Artena 卡 (搭載 AMD Versal™ VP1802 FPGA) 讓資料能從網路介面直接進入 ML 模型，幾乎無需 CPU 介入。
- 未來保障：隨著 STAC-ML 基準測試演進至更大、更複雜的模型，myrtle.ai VOLLO 堆疊的可程式化特性，讓金融機構無需更換硬體即可更新策略。

客戶需求

專為金融服務產業（FSI）中重視執行速度的技術與業務領導者設計:

- 量化交易員：希望在不增加延遲預算的情況下部署更精密的 ML 模型。
- 電子交易部門：旨在競爭激烈的市場中降低滑價 (slippage) 並提升成交率 (fill rates)。
- 風險管理人員：需要能夠處理大量資料串流且無延遲的即時日內風險評估。
- CTO 與基礎設施架構師：設計高密度機房共置部署，特別注重電力與空間受限的環境。
- 極致確定性：即使在高百分位尾端 (99th 和 99.9th) 也能維持一致的延遲，確保在「閃電 (flash)」市場事件中效能可預測。
- 亞微秒級延遲：LSTM_A 模型可低至 1.5 微秒，讓使用者能夠在更廣泛市場處理相同資訊前就辨識訊號並做出反應。

Supermicro 解決方案與受測系統 (SUT)

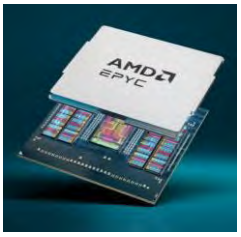
Supermicro CloudDC 伺服器:

- 針對「Tick-to-Trade」效能最佳化的 1U/2U 平台
- 外型規格：2U 機架式
- CPU：1 顆 AMD EPYC 9575F (64 核心)
- 記憶體：12 條 16GB DDR5 DIMM 6000MT/s (總計 192GiB)
- 擴充：雙 PCIe 5.0 x16 插槽，用於 Silicom Artena 加速卡
- PCIe Gen 5，提供 NIC 與加速器最大頻寬
- 散熱優勢：可選用 EVAC 散熱器，在不發生熱節流的情況下維持最高 Turbo 頻率

AMD Versal™ Premium 系列 Adaptive SoC

- 高邏輯密度，實現差異化、適應性與更快的洞察時間
- 超高頻寬網路，最高達 5 Tb/s，可處理市場資料爆量 (data surges)
- 低延遲 SERDES (Serializer/Deserializer)，支援快速訂單執行
- 可程式化邏輯，支援 AI/ML 網路智慧，例如異常偵測與自我佈建 (self-provisioning)
- PCIe Gen5，提供更低的系統延遲與更高的吞吐量
- 專為熱效率設計，具業界領先的每瓦效能

AMD EPYC™ 9005 系列處理器:



- 高頻率「F」型號 SKU 提供強大的單執行緒時脈速度，這對低延遲堆疊至關重要

AMD/ Silicom PCIe 加速卡:

- Silicom FBAP4@VP18-2L0S，搭載 AMD Versal™ Premium 系列 VP1802 Adaptive SoC

AMD Solarflare™ X4 網路介面卡
:

- 達成亞微秒級延遲，比前一代降低最高 40% 的延遲
- 將網路堆疊從 CPU 卸載，大幅降低延遲；具備 kernel-bypass Onload 技術，最大化 CPU 效率並減少額外負荷

軟體堆疊 – myrtle.ai VOLLO™: VOLLO 是連接 AI 研究人員與 FPGA 硬體之間的智慧橋樑。客戶可使用它將標準 ML 模型 (例如來自 PyTorch 或 TensorFlow 的模型) 編譯後，在高度最佳化的 FPGA 硬體上執行，而無需具備 FPGA 程式設計專業知識。

- 無需 RTL (Register Transfer Level)：VOLLO 讓量化分析師可以直接從 PyTorch 取得模型，並使用標準函式庫部署到 FPGA 上
- 架構最佳化：採用專為時間序列推論設計的資料流架構，最小化資料搬移並最大化平行處理能力
- 靈活性：支援多個模型實例與不同配置，讓單一 FPGA 可同時服務多種交易策略

硬體加速 – 此解決方案的硬體基礎建置於 Silicom Artena (FBAP4@VP18-2L0S) 平台上，搭載 AMD Versal™ Premium Adaptive SoC

- AMD Versal™ Premium VP1802：此 Adaptive SoC 結合傳統 FPGA 邏輯、AI Engines (AIE) 與硬化記憶體控制器，提供低延遲 LSTM 所需的大量記憶體頻寬
- Silicom 工程設計：Artena 卡提供高速 PCIe Gen5 連線與最佳化散熱設計，確保 AMD SoC 在 Supermicro 機箱中維持最高效能

運算引擎 – AMD EPYC™ 處理器： 雖然 FPGA 負責推論執行，但 AMD EPYC™ 9005 系列 (Turin) 處理器作為高速協調調度器

- PCIe Gen5 領先優勢：EPYC 處理器提供業界領先的 PCIe Gen5 通道數，確保網路、CPU 與 Silicom FPGA 卡之間零瓶頸

- 記憶體吞吐量：搭載 12 通道 DDR5 記憶體，確保市場資料的「前處理」與「後處理」能跟上 FPGA 亞微秒級的速度

受審計的伺服器為 Supermicro AS-2015CS-TNR，這是一款 2U 單處理器 H13 CloudDC 系統，設計上強調最大靈活性。此伺服器針對高效能邊緣運算與資料中心部署進行最佳化。其散熱設計允許 Silicom FPGA 卡在高頻率 bitstream 下運行而不發生節流，這對於維持 STAC-ML 報告中所強調的確定性延遲至關重要。

破紀錄的延遲表現

此系統在三個基準測試模型 (LSTM_A、LSTM_B 與 LSTM_C) 中，均達成史上最低的 99th-percentile (99p) 延遲。

- LSTM_A: 首個突破 2 µs 障礙的系統
- LSTM_B: 首個突破 3 µs 障礙的系統
- LSTM_C: 首個突破 8 µs 障礙的系統

確定性效能 - 與一般用途運算不同，此基於 FPGA 的解決方案即使在增加模型實例數量時，仍能維持一致且低抖動的效能。

轉型 - 從基於規則的系統轉向機器學習驅動策略，已成為演算法交易的當前戰場。雖然市場持續採用先進 CPU 與 GPU 的 AI 解決方案，但也逐漸認知到，針對工作負載固有抖動所造成的尾部延遲，專用解決方案才是最佳選擇。

將 myrtle.ai 的 VOLLO 與 Supermicro 伺服器整合，搭配 Silicom Artena 加速卡上的 AMD Versal™ Premium Adaptive SoC，成功突破微秒延遲障礙。Supermicro + myrtle.ai 解決方案提供三大決定性優勢:

- 競爭執行優勢：99th-percentile 延遲低至 2 微秒，讓企業能夠透過複雜機器學習模型處理訊號，並比依賴 CPU 或 GPU 推論的競爭者更快執行交易——在波動市場中減少滑價並捕捉更多超額報酬 (alpha)。
- 提升模型複雜度：交易員現在能夠以過去僅限簡單線性回歸的速度，部署精準的 LSTM、SSM、CNN、MLP 等模型——不再需要被迫妥協。
- 營運效率：在緊湊的 2U 空間內提供破紀錄的吞吐量，大幅降低昂貴機房共置環境中的機架空間與電力消耗。

更多資訊

- Supermicro Financial Services: https://www.supermicro.com/en/solutions/ai/finance
- AMD EPYC Processors: https://www.amd.com/en/products/processors/server/epyc.html
- myrtle.ai: https://myrtle.ai/products/vollo-for-capital-markets/

附錄

參考: SUT ID: MRTL260323 STAC-ML™ Markets (Inference) Benchmarks Tacana Suite

雖然此解決方案採用 FPGA，但 Supermicro 同時也為 GPU 與 FPGA 兩種架構的解決方案，均提供業界頂尖的 AI 推論效能基準測試成績。

	GPU 加速	FPGA 加速
應用案例	中頻率、Alpha 生成、情緒分析、風險	高頻交易/中頻交易、從 Tick 到交易、訂單執行
致勝指標	吞吐量與多功能性	確定性超低延遲
模型複雜度	大型模型 (LLM、Deep LSTM、Transformer)	較小型、高度最佳化的模型 (例如 SSM、RNN、LSTM、MLP、CNN)
開發依賴性	Python, PyTorch, CUDA	VOLLO SDK 可編譯來自以下框架的模型：PyTorch 和 ONNX
抖動 / 變異	最小化，受軟體影響	幾乎為零 (硬體層級確定性)
確定性模型效能特徵	高吞吐量、低延遲推論	超低延遲推論

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

AMD

過去 50 多年來，AMD 一直引領高性能運算、圖形與視覺化技術的創新。全球數十億人、財富 500 強企業以及尖端科學研究機構，每天都仰賴 AMD 技術來提升他們的生活、工作與娛樂方式。AMD 員工致力於打造領先的高性能與適應性產品，不斷挑戰技術的極限。

了解更多資訊，請至 www.amd.com

MYRTLE.AI

Myrtle.ai 是一家 AI/ML 軟體公司，致力於在全球主要 FPGA 供應商的平台上，提供世界級的推論加速器。憑藉深厚的神經網路專業知識，Myrtle.ai 已為金融科技、無線通訊、大型語言模型 (LLM)、語音處理以及推薦系統等應用領域，交付多款加速器解決方案。敬請蒞臨 myrtle.ai