aiPIM 技術白皮書：透過記憶體內運算重塑 AI 邊緣運算格局

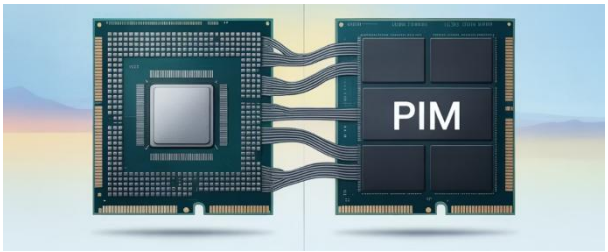
本技術白皮書詳細闡述了我們創新的 aiPIM（處理器內記憶體）解決方案，旨在解決生成式 AI 所面臨的記憶體瓶頸問題。aiPIM 透過將高效能 AI 處理器與標準 DRAM 整合，實現可直接替換現有記憶體的 AI 加速方案，特別針對 7B 參數以下的語言模型及各類視覺模型。本文將探討現有 AI 運算架構的挑戰、aiPIM 的技術實現、軟體生態系統整合以及市場應用前景，展現這項技術如何重新定義邊緣 AI 運算的可能性。

AI 時代的記憶體瓶頸與 PIM 解決方案

隨著生成式 AI 的快速發展，傳統運算架構面臨嚴峻挑戰。我們發現「記憶體牆」(Memory Wall) 已成為限制 AI 效能擴展的關鍵瓶頸。當代 AI 模型需要處理海量參數，導致處理器與記憶體之間的頻繁資料傳輸成為主要效能障礙。

為解決這一根本性問題，我們推出名為 **aiPIM** 的創新記憶體內運算 (Processing-In-Memory) 晶片。aiPIM 顛覆傳統運算與儲存分離的架構，將運算單元直接整合到記憶體結構中，從根本上消除資料傳輸瓶頸。

我們的方案專為 7B 參數以下的語言模型及各類視覺模型設計，提供高效能、低功耗的 AI 加速能力，同時保持與現有系統的完全相容性。這意味著透過簡單替換記憶體模組，即可為各類邊緣裝置帶來顯著的 AI 運算能力提升。



aiPIM 的核心價值在於將高效能 AI 處理器與標準 DRAM 整合，實現可直接替換現有記憶體的 AI 加速方案。透過消除處理器與記憶體之間的數據往返延遲，aiPIM 能大幅提升 AI 推論效率，同時降低功耗，為邊緣運算裝置帶來前所未有的 AI 處理能力。

核心問題

生成式 AI 的發展正受到傳統運算架構的「記憶體牆」瓶頸限制，導致高效能 AI 難以在邊緣裝置上實現。

我們的方案

推出創新的記憶體內運算 (Processing-In-Memory) 晶片 aiPIM，實現運算與儲存的緊密整合。

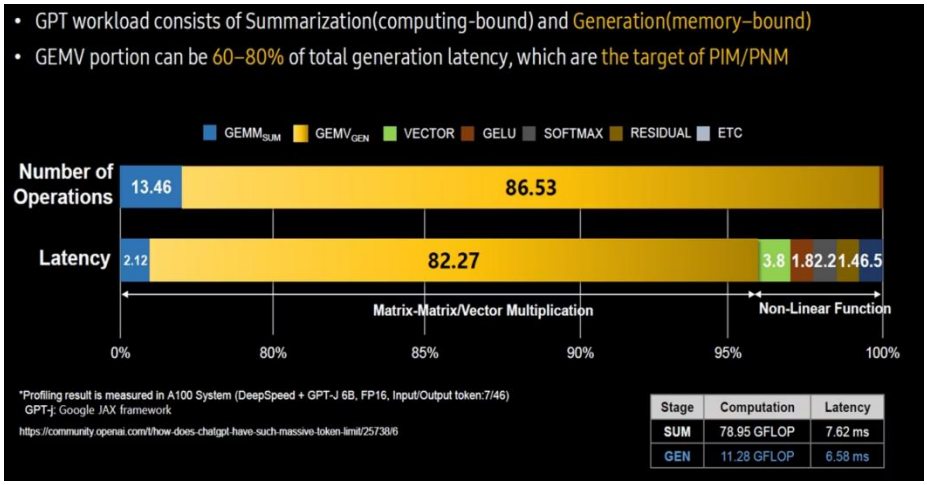
核心價值

將高效能 AI 處理器與標準 DRAM 整合，實現可直接替換現有記憶體的 AI 加速方案，無需修改系統架構。

問題分析：為何高階 GPU 在 AI 推論時效率低落？

現代 AI 工作負載，尤其是大型語言模型 (LLM) 的推論過程，存在一個關鍵問題：它們本質上是「記憶體受限 (Memory-bound)」的任務。這意味著處理器的運算速度遠超過記憶體提供資料的速度，導致處理器大部分時間處於等待狀態。

引用研究分析，AI 推論的延遲主要來自以下幾個方面：



分析顯示，在生成網路運算中，高達 82% 的延遲來自於記憶體存取與頻繁的資料傳輸，特別是大量讀取權重。此外，非線性函數計算也受到記憶體延遲的影響，佔用約 15% 的執行時間。由於資料頻寬不足，在生成 token 的階段，高階 GPU 的實際運算單元利用率甚至可能低至 10%。

記憶體瓶頸的本質

在生成網路運算中，運算晶片必須不斷從記憶體獲取大量權重參數。然而，記憶體和處理器之間的資料傳輸速度遠低於處理器的運算速度，導致處理器大部分時間處於閒置狀態，等待資料到達，記憶體延遲約佔總運算時間的 82%。

運算資源浪費

高階 GPU 的運算能力可能高達數百 TFLOPS，但在生成 token 階段，實際利用率往往低於 10%。這相當於購買了一輛法拉利，卻只能以自行車的速度行駛，造成巨大的資源浪費和經濟損失。

能源效率低下

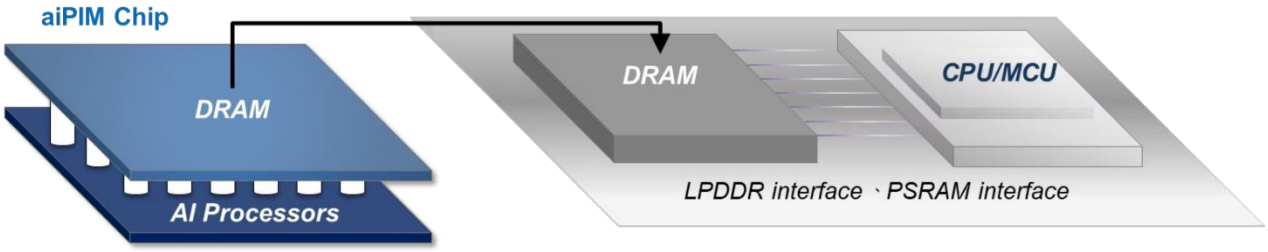
由於大部分能耗用於資料傳輸而非有效運算，AI 推論的能源效率極低。這不僅增加了營運成本，也限制了邊緣裝置上 AI 應用的可能性，因為這些裝置通常有嚴格的功耗限制。

這種效率低落的現象在處理大型語言模型時尤為明顯。例如，運行 7B 參數的 LLaMA 模型時，即使是最新的 NVIDIA A100 GPU，其運算單元的實際利用率也僅約 15%。換言之，這些高階 GPU 的大部分計算能力在 AI 推論過程中被閒置，造成嚴重的資源浪費。

因此，我們需要一種全新的運算架構，能夠從根本上解決記憶體與處理器之間的數據交換瓶頸，提高 AI 運算的效率。這正是 aiPIM 的核心價值所在。

我們的解決方案：記憶體內運算 (PIM)

面對「記憶體牆」的挑戰，我們提出了記憶體內運算 (Processing-In-Memory, PIM) 的創新解決方案。這一技術徹底顛覆了傳統的馮·諾依曼架構，打破運算與儲存分離的傳統，實現「就地運算 (Compute-in-place)」。



PIM 的核心優勢

- 從根本上消除處理器與主記憶體之間的資料往返延遲
- 大幅降低資料傳輸所需的功耗
- 顯著提高運算單元的利用率
- 減少記憶體存取帶寬需求，優化系統整體效能
- 降低系統整體功耗，提高能源效率

aiPIM 的市場定位

在眾多 PIM 技術路線中，我們選擇聚焦於市場潛力最大的 DRAM PIM 方案。這一選擇基於以下考量：

- 可利用成熟的 DRAM 製程，降低開發風險
- 能直接替代現有系統的記憶體，無需修改系統架構
- 兼具高效能與成本效益，適合大規模商業化

標準記憶體角色

aiPIM 首先是一個完全符合 JEDEC 標準的記憶體晶片，可直接替換系統中的現有記憶體

內建 AI 加速器

同時，aiPIM 內建專用 AI 加速處理器，可直接在記憶體內執行 AI 推論運算

實現邊緣 AI

透過消除資料傳輸瓶頸，aiPIM 能在邊緣裝置上實現高效能、低功耗的 AI 推論

aiPIM 的獨特價值在於，它旨在推動普惠型邊緣 AI (Edge AI) 的實現。透過提供一種可直接替換現有記憶體的 AI 加速方案，aiPIM 使得各類邊緣裝置無需更換主晶片或重新設計系統架構，就能獲得強大的 AI 運算能力。這大幅降低了邊緣 AI 的部署門檻，為萬物智能化開闢了一條務實可行的路徑。

aiPIM 核心理念：無縫整合，原地升級

aiPIM 的戰略核心是實現「無縫整合，原地升級」。我們的願景是讓所有物聯網 (IoT) 與工業電腦 (IPC) 裝置，無需更改主晶片設計，僅透過替換記憶體，即可獲得強大的 AI 加速能力。

這一理念基於一個關鍵洞察：全球數十億的邊緣裝置若要支援 AI 功能，最大的障礙不是技術本身，而是部署的成本與複雜度。透過提供一種「即插即用」的解決方案，aiPIM 可以大幅降低 AI 技術普及的門檻。

在快速成長的邊緣 AI 市場中，高效能、低功耗 AI 推論的需求日益迫切。傳統 CPU/GPU 在邊緣裝置上運行複雜 AI 模型時面臨功耗、延遲和成本挑戰。因此，整合了神經網路處理單元 (NPU) 的記憶體內運算 (PIM) 解決方案正迅速崛起，成為推動邊緣 AI 普及化的關鍵技術，其市場潛力巨大。

1

雙重角色定位

aiPIM 對系統而言，既是標準主記憶體，也是一個外掛式 AI 加速器。當系統需要普通記憶體功能時，aiPIM 表現為標準 DRAM；當需要 AI 運算時，aiPIM 可切換至加速器模式。

2

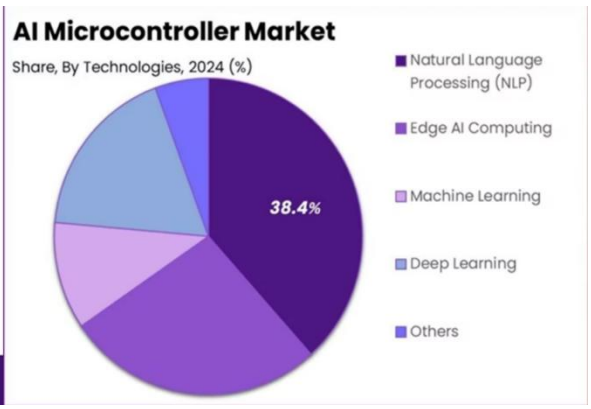
標準介面兼容

完全相容 LPDDR、PSRAM 等主流記憶體介面，確保即插即用。無需修改現有系統的硬體設計或主晶片，大幅降低採用門檻。

3

高效能 AI 加速

內建專用 AI 加速核心，針對 LLM 與 CNN 等主流模型進行優化，提供高效能、低功耗的 AI 運算能力。



隨著人工智慧技術持續滲透至終端應用，AI微控制器市場正快速擴張，預估從2024年的 61億美元 成長至2034年的 247億美元，年均複合成長率（CAGR）達 14.7%。傳統8-bit、16-bit、32-bit控制器正逐步升級為支援AI推論的 64-bit架構，以滿足智慧感測、語音辨識、邊緣推理等多樣化應用需求。

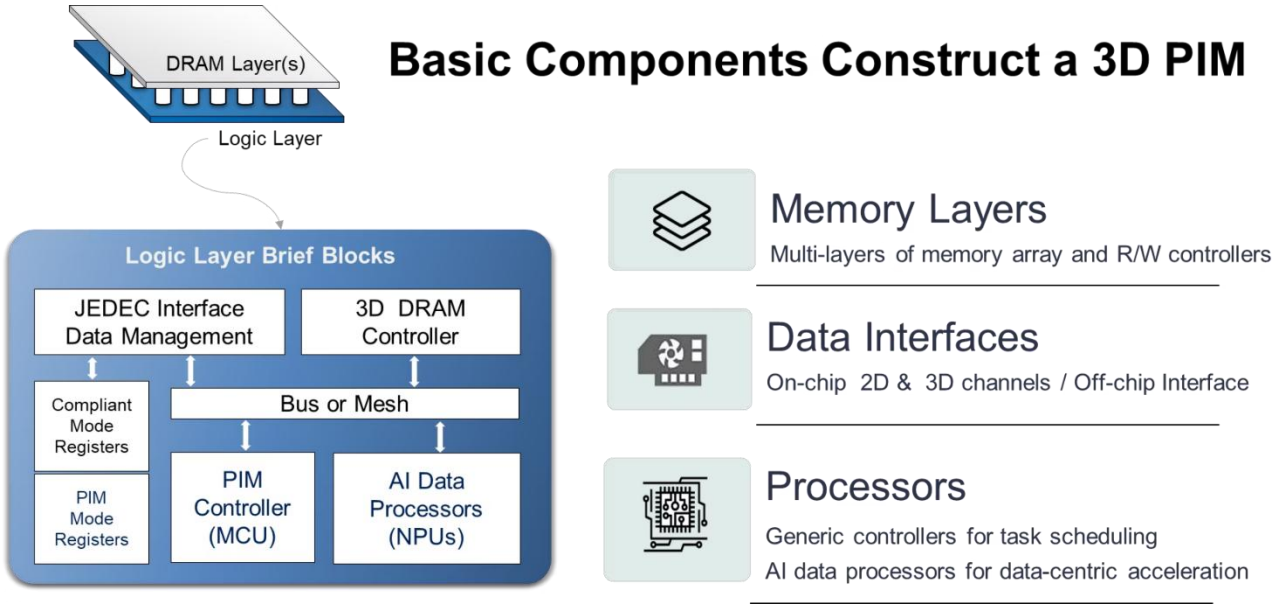
2024年應用分析顯示，自然語言處理（NLP） 佔整體AI微控制器應用比重達 38.4%，其次為 邊緣AI運算、機器學習 和 深度學習。此顯示AI不再只是雲端運算的專利，而正逐步進入記憶體受限、功耗敏感的終端設備。

在這波趨勢中，DRAM型PIM（Processing-In-Memory） 技術成為重要的關鍵突破。透過在DRAM內整合NPU或AI加速核心，PIM技術可大幅減少微控制器與外部記憶體之間的資料搬移瓶頸，實現下列優勢：

未來DRAM PIM將可作為AI MCU平台的重要搭配技術，特別是在32-bit或64-bit控制器搭配PIM作為內建AI引擎時，將形成低功耗、高效能、記憶體即算力的新型態AI微控制器架構，廣泛應用於智慧家庭、工業自動化、穿戴裝置與智慧感測節點等領域。

aiPIM 晶片通用架構

aiPIM 晶片採用創新的異質整合架構，將記憶體與邏輯處理單元緊密結合。從物理結構上看，aiPIM 可採用 3D 堆疊或 SiP 封裝，分為上層的「DRAM 層」和下層的「邏輯層」。



DRAM 層

- 採用標準 DRAM 製程，確保高密度儲存能力
- 支援標準的記憶體存取協議，確保兼容性
- 內建特殊訪問模式，支援高頻寬的 PIM 操作
- 可擴展至不同容量規格，滿足不同應用需求

邏輯層

- 採用先進邏輯製程，提供高效能運算能力
- 整合多種功能模組，實現智能記憶體管理
- 內建專用 NPU 叢集，加速 AI 運算
- 低功耗設計，最大化邊緣裝置的電池壽命

邏輯層核心模組詳解

記憶體介面 (Memory Interface)

負責與主系統通訊，支援標準 JEDEC 記憶體協議，同時整合 3D 控制器以實現 DRAM 與邏輯層之間的高效通訊。這確保了 aiPIM 可以無縫替代現有記憶體，同時提供額外的運算能力。

PIM 控制器 (PIM Controller)

採用低功耗 RISC-V 核心，負責整體調度與控制。它管理 aiPIM 的運作模式切換、任務分配與執行監控，並處理與主系統的通訊協議。這一模組確保了 aiPIM 能夠高效地協調記憶體操作與運算任務。

AI 資料處理器 (AI Data Processor)

由多個 NPU 叢集組成，針對 AI 推論任務進行優化。每個 NPU 包含張量處理單元、向量處理單元與資料重排單元，能高效處理矩陣運算、非線性函數計算等 AI 核心操作。這是 aiPIM 實現高效 AI 運算的核心。

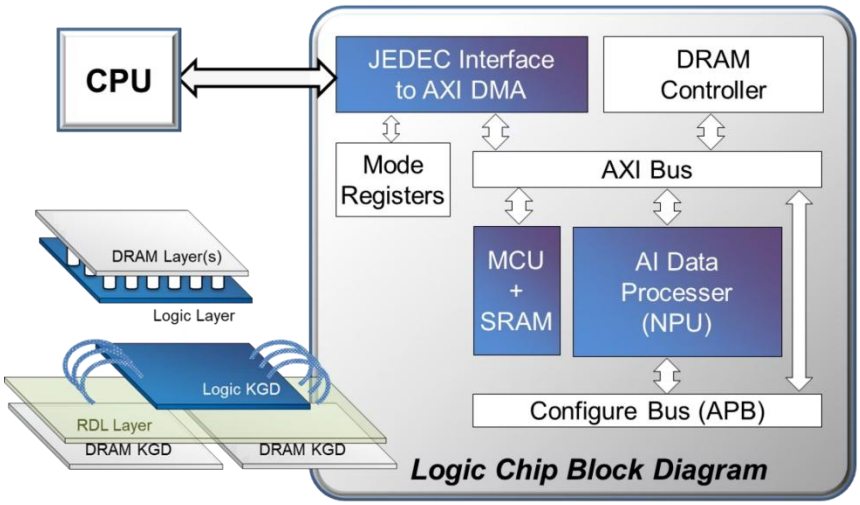
模式暫存器 (Mode Register)

允許系統在標準記憶體模式與 PIM 模式之間切換。當處於標準模式時，aiPIM 表現為普通 DRAM；當切換至 PIM 模式時，則啟動內部的 AI 加速功能。這種雙模式設計確保了與現有系統的完全兼容。

硬體實現：封裝技術與 NPU 規格

封裝技術選擇

aiPIM 的物理實現可採用兩種主要的封裝技術：SiP (System-in-Package) 封裝與 3D (WoW, Wafer-on-Wafer) 封裝。兩種技術各有優劣，適用於不同的應用場景與產品定位。



在初期產品中，我們將優先採用 SiP 封裝技術，以降低開發風險並加速產品上市。隨著技術成熟與量產規模擴大，我們將逐步導入 3D 封裝技術，以實現更高的性能與能效。

封裝方案	優點	缺點	適用場景
SiP 封裝	良率高、成本低、製程成熟	面積大、功耗較高、頻寬受限	成本敏感型應用、早期產品
3D (WoW) 封裝	面積小、功耗低、頻寬高	成本高、製造限制多、良率風險	高效能應用、空間受限裝置

NPU 核心規格 (以 PSRAM aiPIM 為例)

Tensor Core

支援 INT8/BF16 精度，專為 CNN 等視覺模型優化。採用脈動陣列 (Systolic Array) 架構，能高效執行矩陣乘法運算。每個核心可提供高達 256 MAC/cycle 的運算能力。

Vector Core

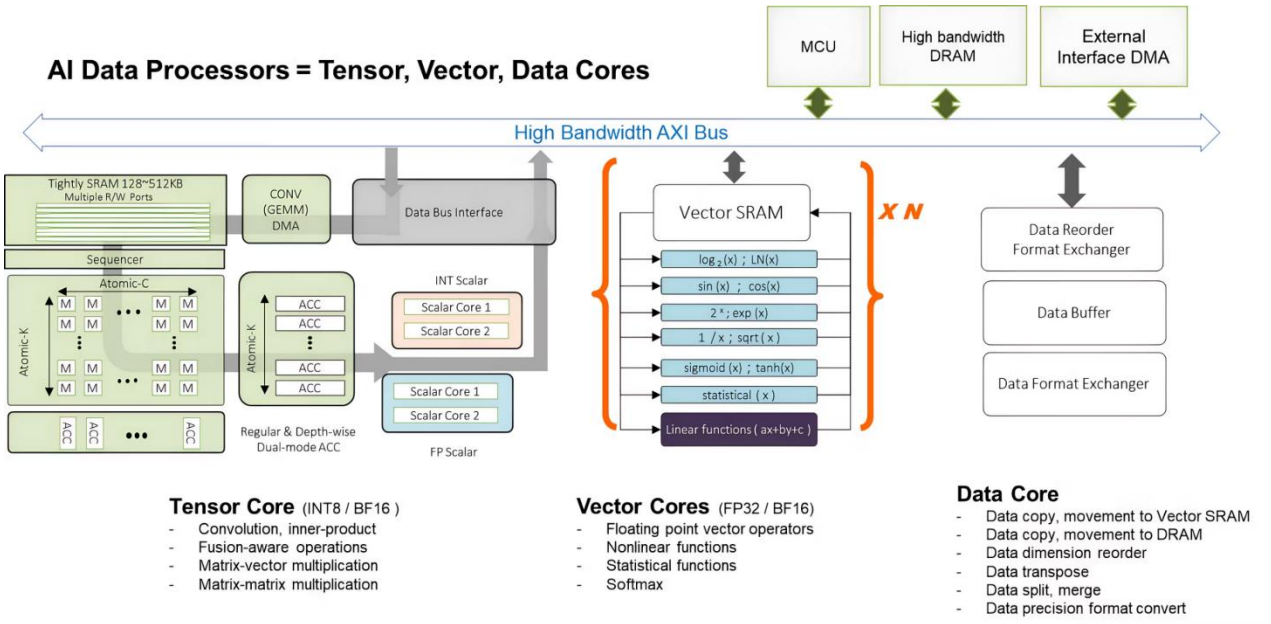
支援 FP16/FP32 精度，針對 Softmax、GELU 等非線性函數進行硬體加速。採用 SIMD 架構，可平行處理向量運算，大幅加速 LLM 推論中的關鍵操作。

Data Reshape Core

負責資料轉置與重排，優化記憶體存取模式。能執行高效的 im2col、矩陣transpose等操作，減少記憶體存取次數，進一步提升整體效能。

aiPIM 的 NPU 設計採用模塊化架構，可根據不同應用需求進行靈活配置。例如，針對視覺識別應用，可增加 Tensor Core 的比重；而針對語言模型，則可強化 Vector Core 的能力。這種靈活性使 aiPIM 能夠適應不同的 AI 工作負載，提供最佳的效能/功耗比。

硬體結構



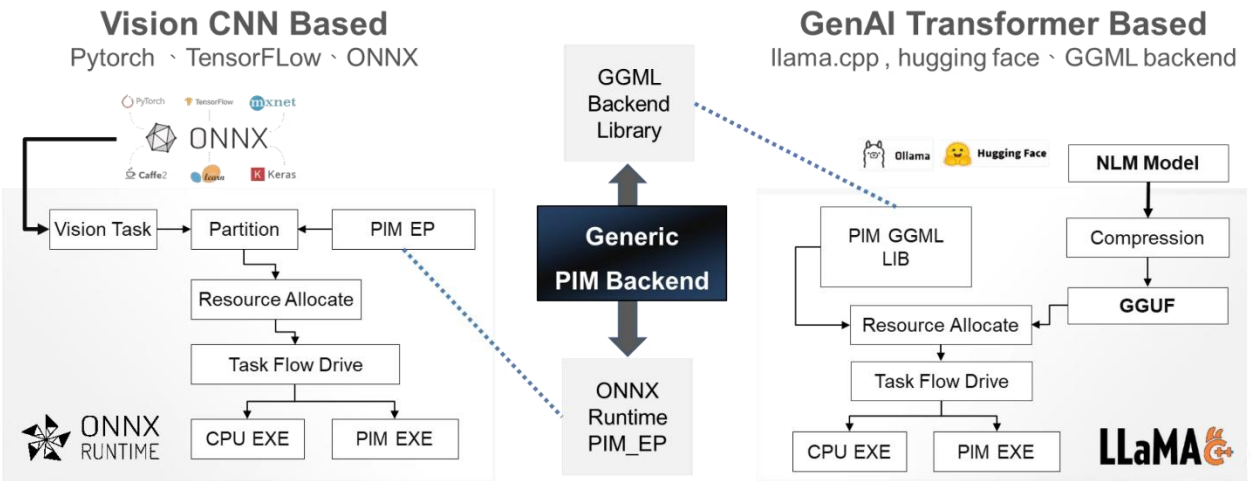
主要硬體特色可分為以下幾個部分：

1. **雙核心運算引擎 (Dual-Core Computing Engines):**
   - **張量核心 (Tensor Core / CONV/GEMM Engine):** 此單元是整個處理器的運算主力，專為卷積 (Convolution) 與通用矩陣乘法 (GEMM) 等密集的張量運算而設計。它擁有自己的 DMA 控制器 (CONV (GEMM) DMA)，能直接從高頻寬的 AXI 匯流排或緊密耦合的 SRAM (Tightly Coupled SRAM) 中高速抓取權重與特徵圖資料。其內部由大量的乘加器陣列 (MAC Array) 組成，並透過一個專用的序列器 (Sequencer) 來控制複雜的運算流程，以實現最高的算力密度。
   - **向量核心 (Vector Core):** 此單元專門處理 AI 模型中的非線性函數與各種向量/純量運算。其最大的特色是內建了大量的硬體特殊函式單元 (Special Function Units, SFUs)，能以極低的延遲和功耗執行如 sigmoid, tanh, exp(x), 1/x, sqrt(x) 等傳統上需要用軟體或查找表 (Look-up Table) 實現的複雜函數。架構中的 x N 標示意味著此向量單元可根據效能需求進行多倍擴展，提供平行的向量處理能力。
2. **高效的記憶體子系統 (Efficient Memory Subsystem):**
   - 卷積架構極度依賴本地 SRAM 來餵飽運算核心。左側的 **Tightly SRAM** 具有多個讀寫埠 (Multiple R/W Ports)，確保張量核心在進行大量運算時，資料供應不會中斷。同時，向量核心也配有專屬的 **Vector SRAM**，形成了分層式的記憶體架構，大幅減少對外部高頻寬 DRAM 的存取次數。
3. **專用的資料處理單元 (Dedicated Data Manipulation Units):**
   - 右側的 **Data Reorder** 與 **Data Format Exchanger** 等單元，扮演了資料物流中心的角色。它們專門負責在資料進入運算核心前後，進行必要的預處理與後處理，例如資料維度的轉置 (Transpose)、排序 (Reorder)，或是不同數值精度之間的轉換 (例如Q4~Q8量化格式)，這使得運算核心可以更專注於純粹的數學計算。

總體而言，這個架構的設計精髓在於**任務切分**與**硬體特化**。它將 AI 推論中計算量最大的張量運算、複雜的非線性函數運算，以及繁瑣的資料格式整理，分別交由最高效的專用硬體來執行，並透過高頻寬的 AXI 匯流排與高效的內部 SRAM 進行串聯，最大化了整體運算的能效比 (Performance per Watt)。

軟體堆疊與生態系整合

一個創新的硬體解決方案若要成功，軟體生態系統的支援至關重要。aiPIM 的軟體策略核心是「無縫整合」，我們的目標是讓開發者能夠輕鬆地將現有 AI 應用遷移到 aiPIM 平台，無需大幅修改代碼或重新訓練模型。



底層控制機制

aiPIM 採用記憶體映射 (Memory-mapped) 方式，主 CPU 透過讀寫特定位址來控制 PIM。這種方式與傳統驅動程式類似，使系統能夠輕鬆識別並使用 aiPIM 的功能。

高階框架整合

視覺模型 (CNN)

提供 ONNX Runtime 的客製化後端 (PIM_EP)，支援主流視覺模型如 ResNet、MobileNet、YOLO 等。

語言模型 (LLM)

提供 GGML 後端函式庫，支援 llama.cpp、Ollama 等流行框架，實現 LLaMA、Vicuna 等模型的高效執行。

第一階段：基礎驅動

1

開發核心驅動程式，實現基本的控制與管理功能，支援簡單的矩陣運算 API。這一階段已完成，為後續開發奠定基礎。

第二階段：框架整合

2

開發 ONNX Runtime 後端與 GGML 支援，實現主流模型的遷移與優化。目前正在此階段，已完成關鍵 API 的設計與實現。

第三階段：生態拓展

3

提供更廣泛的框架支援，開發模型自動優化工具，構建開發者社區。計劃在產品量產前完成此階段。

第四階段：應用加速

4

針對特定垂直應用領域提供優化解決方案，如智慧監控、智慧工廠等，最大化 aiPIM 的應用價值。

市場定位與商業模式

市場定位

aiPIM 的市場定位是「普惠型邊緣 AI 加速器」。我們關注的不是超高端的資料中心市場，而是更廣闊的邊緣運算市場，包括但不限於：

- 智慧家電與消費電子
- 工業自動化與智慧工廠
- 智慧城市與公共安全
- 醫療設備與健康監測
- 零售與物流自動化

這些領域共同的特點是：需要本地化的 AI 推論能力，但受到功耗、成本與部署複雜度的嚴格限制。aiPIM 正是為解決這些痛點而設計。

製程優勢

使用成熟邏輯製程打造 NPU，效能更高。相較於 DRAM 製程中實現 PIM，我們的方案可提供更複雜的邏輯設計與更高的運算效率。

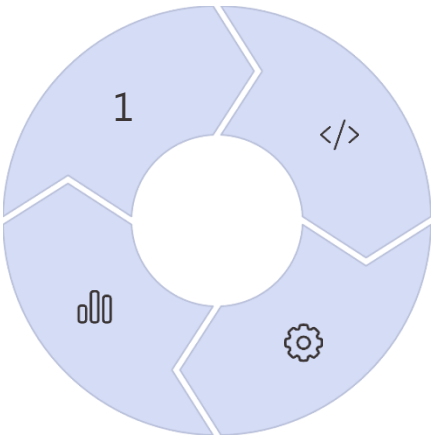
彈性與成本

聚焦主流 LPDDR/PSRAM 規格，適用性廣。無需客製化系統設計，大幅降低採用門檻與成本，加速市場滲透。

混合精度

原生硬體加速支援多種精度，特別是針對 LLM 的 Q4~Q8等 量化格式，以及 CNN 的INT8 量化格式。這確保 aiPIM 能高效處理複雜 AI 模型，實現最佳效能與功耗平衡。

商業模式



1 銷售 aiPIM 晶片

直接提供高效能、低功耗的 PIM 晶片，面向系統整合商與終端產品製造商

</> 授權 Logic Layer IP

向 DRAM 大廠授權邏輯層 IP，推動行業標準形成，擴大生態系統

提供開發工具

開發並銷售優化工具與解決方案，幫助客戶加速 AI 應用部署

垂直應用解決方案

針對特定行業需求提供完整解決方案，創造更高附加價值

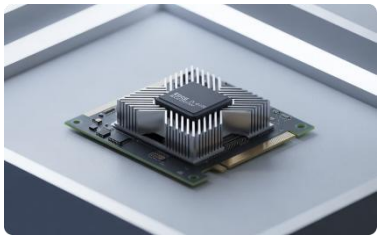
我們的商業策略採取雙軌並行的方式：短期內專注於提供完整的 aiPIM 晶片產品，實現快速市場滲透；長期則通過 IP 授權模式，與記憶體大廠合作，推動 PIM 技術的行業標準化，擴大生態系統。這種策略既確保了短期收入，又為長期技術領導地位奠定基礎。

競爭分析與技術優勢

當前 AI 加速器市場競爭激烈，各種解決方案各有優勢。為了清晰說明 aiPIM 的市場定位與競爭力，我們進行了詳細的競爭分析。

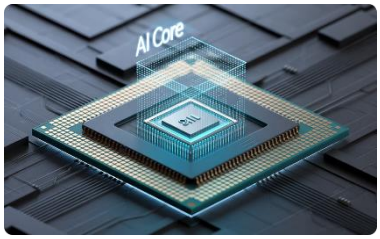
高階 GPU

如 NVIDIA A100/H100，提供極高的理論運算能力，但價格昂貴、功耗高，不適合邊緣部署。aiPIM 雖然絕對性能不及高階 GPU，但在能效比與部署便利性上具有明顯優勢。



邊緣 AI 加速器

如 Google Coral、Intel NCS 等，專為邊緣設備設計，但需要額外的硬體、介面與空間，與系統整合較為複雜。aiPIM 通過替換記憶體方式實現，大幅簡化部署流程。



整合式 AI 加速

如 ARM 的 NPU、Apple Neural Engine 等，與 CPU 整合提供 AI 加速，但受限於主晶片更新週期。aiPIM 無需更換主晶片，可為現有系統帶來立即提升。

解決方案類型	性能	功耗	部署難度	成本	靈活性
高階 GPU	極高	極高	高	極高	高
邊緣 AI 加速器	中等	中等	中等	中等	中等
整合式 AI 加速	中低	低	低（新系統）	中等	低
aiPIM	中高	低	極低	低	高

aiPIM 的關鍵競爭優勢 (預估)

1. **無縫升級路徑**：現有系統無需重新設計，僅替換記憶體即可獲得 AI 加速能力，大幅降低採用門檻。
2. **最佳能效比**：通過消除記憶體牆瓶頸，aiPIM 在單位功耗下提供更高的 AI 推論性能。
3. **廣泛兼容性**：支援主流記憶體介面，適用於各類邊緣設備，實現技術的普惠化。
4. **成本效益高**：相較於獨立加速器，aiPIM 避免了額外的接口、PCB 空間與供電設計，降低總體成本。

1

部署複雜度降低

相較於獨立 NPU，aiPIM 將硬體整合複雜度大幅降低

2

總體擁有成本降低

考慮硬體、整合成本，aiPIM 較HBM方案降低約 60%

3

與現有系統兼容性

aiPIM 與 95% 的現有邊緣裝置架構兼容，無需硬體修改

應用場景

aiPIM 的應用潛力廣闊，可為多個領域的邊緣 AI 應用帶來革命性提升。以下我們將探討幾個代表性的應用場景，展示 aiPIM 如何解決現有挑戰並創造新的可能性。

智慧工廠

在工業自動化環境中，aiPIM 可為現有 IPC (工業電腦) 提供實時缺陷檢測、設備預測性維護等 AI 能力，無需更換主控系統，大幅降低智慧升級成本。

醫療設備

為便攜式醫療設備提供本地 AI 診斷能力，如心電圖分析、皮膚病變識別等，同時確保資料隱私與低延遲處理。

智慧監控

使現有監控系統獲得人臉識別、異常行為檢測等 AI 能力，無需升級攝像機主晶片，顯著降低部署成本。

消費級機器人

為低成本機器人平台提供先進的視覺導航與語音互動能力，在嚴格的功耗預算內實現更複雜的 AI 任務。

智慧零售

為 POS 終端與自助設備增加客戶識別、商品辨識等 AI 功能，提升購物體驗與運營效率。

車載子系統

在現有車載子系統基礎上，增強駕駛員監控、車內語音交互等 AI 功能，提升安全性與用戶體驗。

使用者透過 Model Zoo / Open API 使用 PIM

AIoT 產品下載 PIM Model Zoo

常用模型固化、客製化後使用

使用者特性：B2B
系統整合商、系統服務商。

應用場景：智慧製造、智慧場域、自動維運、報告生成、摘要事件、邏輯判斷。

使用方式：韌體層封裝Model Zoo，下載燒錄後重開機自動執行。

消費者透過櫃台終端機/手機

即時反饋/收到商品各種訊息

AI-IPC 產品安裝 PIM Backend API

開源高階通用API，另外下載低階驅動

使用者特性：B2C
一般消費者、應用工程師。

應用場景：通用於攜帶式智慧裝置、也適用於各類型IPC。

使用方式：安裝標準框架，呼叫 PIM 加速、自行載入與執行模型。

使用者從公開網站下載API，使用熱門AI後台服務器，將AI直接指派給PIM來加速。

產品規格與技術路線圖

產品規格(預估)

aiPIM 產品系列計劃覆蓋多種記憶體規格與容量，以滿足不同應用場景的需求。目前我們的產品線規劃如下：

產品型號	記憶體介面	記憶體容量	AI 運算能力	功耗	主要應用場景
aiPIM-LP4-8G	LPDDR4	8Gb	2 TFLOPS	1.5W	中高階智慧設備
aiPIM-LP4-4G	LPDDR4	4Gb	1.5 TFLOPS	1.2W	中高階智慧設備
aiPIM-PS-2G	PSRAM	2Gb	0.8 TFLOPS	0.6W	低功耗邊緣設備
aiPIM-PS-1G	PSRAM	1Gb	0.5 TFLOPS	0.4W	低功耗邊緣設備

技術規格詳情

記憶體特性：

- 符合 JEDEC 標準，與 xSPI/LP4參數相容
- 支援所有標準記憶體操作與時序要求
- 提供與標準記憶體相同的容量與頻寬

AI 加速能力：

- 支援 INT8/BF16/FP32 多精度運算
- 支援GGUF Q4等先進量化壓縮格式
- 優化的矩陣-向量乘法硬體加速
- 高效 Softmax、GELU 等非線性函數實現
- 支援注意力機制等 Transformer 核心運算

功耗特性：

- 標準記憶體模式下功耗與普通 DRAM 相當
- AI 加速模式下提供業界領先的性能/功耗比
- 支援多種省電模式，適應不同應用場景

技術路線圖

1

2025 Q4

第一代 aiPIM-PS 系列晶片設計與工程樣品完成，開始提供開發套件。

2

2026 Q2

第一代 aiPIM-PS 系列晶片完成測試，第二代aiPIM-LP4 系列的設計與模擬。

3

2026 Q4

第二代 aiPIM-LP4 系列完成投片與測試，推出工程樣品。

4

2027 Q2

第二代 aiPIM-LP4 系列正式量產，並開始評估下一代架構與新興記憶體標準。

1

第一代產品 (2025-2026)

聚焦 PSRAM 介面，採用 SiP 封裝，面向低功耗與成本敏感的邊緣 AI 應用。重點支援輕量級視覺識別與小型語言模型推論。

2

第二代產品 (2026-2027)

擴展至 LPDDR 介面，導入更先進的封裝技術，大幅提升處理能力。增強對中大型語言模型的支援，擴展至中高階邊緣運算市場。

核心技術與專利規劃

核心專利規劃

規劃在全球範圍內申請多項關鍵專利，建立起堅實的知識產權保護網。核心專利涵蓋以下幾個方面：

記憶體內運算架構

涵蓋 PIM 晶片的基本架構設計，包括 DRAM 與邏輯層的異質整合方法、高效能互連機制等。異質整合型記憶體內處理器架構及其控制方法，正在申請專利中。

智能模式切換機制

針對 aiPIM 在標準記憶體模式與 AI 加速模式間的無縫切換技術。具備多模式運作能力的記憶體內運算裝置正在申請專利中。

高效能 AI 加速引擎

針對 NPU 設計的創新架構，包括混合精度運算、能效優化技術等。針對模型優化的記憶體內矩陣運算加速器相關技術，已有多國專利獲證。

軟硬體協同優化

涵蓋 aiPIM 的軟體堆疊與硬體協同設計方法，提升整體系統效能。基於記憶體內運算的神經網絡推論優化方法，正在申請專利中。

知識產權戰略

除了專利保護外，我們的知識產權戰略還包括：(1) 核心技術的商業機密保護；(2) 與主要半導體廠商的交叉授權談判；(3) 積極參與相關標準制定，推動行業規範形成。這一多層次的知識產權保護策略，為 aiPIM 的長期技術領導地位提供了堅實保障。

合作夥伴與生態系統

aiPIM 的成功不僅依賴於自身的技術創新，還需要建立廣泛的合作網絡與強大的生態系統。我們積極與半導體廠商、系統整合商、軟體開發者以及主要的邊緣 AI 設備製造商建立夥伴關係，共同推動 PIM 技術的普及與應用。

製造與供應鏈合作夥伴

晶圓代工

與台積電 (邏輯) 和力積電、南亞 (DRAM) 建立戰略合作，確保先進製程與穩定產能。

封裝測試

透過力積電、日月光，提供 SiP 與 3D 堆疊封裝的技術支持與產能保障。

軟體與應用生態合作夥伴

AI 框架支持

連結 ONNX Runtime 、Llama.cpp等主流框架，實現 aiPIM 的原生支持。

應用開發者

通過開發者計劃，支持第三方開發團隊在 aiPIM 平台上創建創新應用。

開放標準與互操作性

我們積極參與半導體行業標準組織的工作，推動 PIM 相關標準的形成。同時，我們承諾將核心 API 與軟體接口開源，確保生態系統的開放性與互操作性。

開發者社區建設

我們計劃推出 aiPIM 開發者計劃，提供開發板、技術文檔、培訓課程與社區支持，培養活躍的開發者生態。同時設立創新應用基金，鼓勵基於 aiPIM 的創新應用開發。

產業聯盟推動

我們正在籌建「邊緣 AI 加速聯盟」，聯合產業上下游夥伴，共同推動邊緣 AI 技術的創新與應用。通過行業展會、技術論壇等方式，擴大 aiPIM 的行業影響力。

商業策略與市場機會

商業模式與收入來源

晶片銷售

直接面向系統整合商與終端產品製造商銷售 aiPIM 晶片，提供完整的硬體解決方案。

IP 授權

授權 aiPIM 的邏輯層 IP，收取授權費與銷售提成，擴大市場覆蓋。

解決方案銷售

針對特定行業需求，開發並銷售完整的 AI 解決方案，提高附加值。

市場進入與拓展策略

1 第一階段：技術驗證 (現在-2025 Q4)

與關鍵客戶共同驗證 aiPIM 技術，在 2025 年末推出 POC 晶片 (XSPI 介面) 與 FPGA 驗證平台，建立成功案例，培養早期採用者社區。

2 第二階段：垂直滲透 (2026-2027)

於 2026 年開始布局 LPDDR 介面產品。聚焦 2-3 個高價值垂直市場（如工業自動化、醫療設備），提供專業解決方案。建立品牌知名度與客戶信任。

3 第三階段：水平擴張 (2027-2028)

擴展至更多領域，推動 IP 授權模式。與主要記憶體廠商建立戰略合作，實現大規模市場滲透。

4 第四階段：生態結盟 (2028+)

建立以 aiPIM 為核心的邊緣 AI 生態系統，推動行業標準形成。拓展國際市場，實現全球化佈局。

競爭策略與差異化定位

在競爭激烈的 AI 晶片市場，我們採取以下差異化策略：

1. **低門檻採用**：強調 aiPIM 的「無縫升級」優勢，降低客戶採用風險與成本
2. **垂直整合**：提供從晶片到軟體的完整解決方案，簡化客戶的整合工作
3. **能效優先**：在同等性能下提供更低功耗，在同等功耗下提供更高性能
4. **靈活配置**：提供多種規格與形態，滿足不同應用場景的需求

這種差異化策略使 aiPIM 能夠避開與高階 GPU 廠商的正面競爭，在專注的邊緣 AI 市場建立獨特競爭優勢。

風險評估與管理策略

任何創新技術的商業化都伴隨著各種風險。我們對 aiPIM 項目可能面臨的主要風險進行了系統性分析，並制定了相應的管理策略。

技術風險

- **風險**：晶片設計複雜度超出預期，導致開發延遲或效能不達標
- **管理策略**：採用分階段驗證方法，先通過 FPGA 原型充分驗證設計；建立彈性架構，允許根據驗證結果進行靈活調整
- **降低措施**：招募資深晶片設計專家；與半導體設計服務公司建立合作，降低設計風險

2

製造風險

- **風險**：異質整合帶來的製程複雜性，可能導致良率問題或成本超支
- **管理策略**：首批產品採用較為成熟的 SiP 封裝技術，降低製造風險；與頂級代工廠建立早期合作關係
- **降低措施**：設計時考慮製造友好性；預留充足的良率學習曲線時間與成本預算

市場風險

- **風險**：市場採用速度低於預期；競爭對手推出類似解決方案
- **管理策略**：聚焦有明確痛點的垂直市場；建立典型應用案例與成功故事
- **降低措施**：積極建立專利壁壘；與關鍵客戶建立早期合作計劃，確保產品契合市場需求

財務風險

- **風險**：研發成本超支；量產資金需求大於預期；融資進程延遲
- **管理策略**：建立嚴格的預算控制機制；設計分階段的研發與商業化路線圖
- **降低措施**：保持精簡運營；積極尋求政府補助與產業合作，分散研發成本

競爭風險評估

AI 加速器市場競爭激烈，我們面臨來自多個方向的競爭壓力：

1. **大型 GPU 廠商**：如 NVIDIA、AMD 等可能將目光轉向邊緣 AI 市場
2. **專業 AI 晶片初創公司**：如 Hailo、Blaize 等專注於邊緣 AI 加速
3. **記憶體大廠**：如 Samsung、SK Hynix 等可能自行開發 PIM 解決方案
4. **SoC 廠商**：如高通、聯發科等可能在主晶片中增強 AI 加速能力

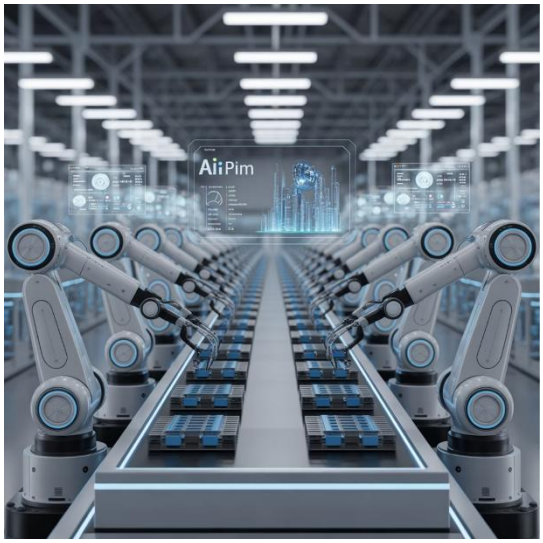
面對這些競爭，我們的核心防禦策略是聚焦獨特價值主張——「無縫升級」的 PIM 解決方案，避開與大廠的正面競爭，同時通過專利保護與快速市場滲透建立技術與市場壁壘。

結論：驅動傳統記憶體產業的 AI 轉型

aiPIM 代表了一種革命性的思維轉變——不再執著於提升運算核心的性能，而是從根本上重構運算與記憶體之間的關係，消除數據移動的瓶頸。這種方法不僅提供了更高效的 AI 推論能力，還開闢了一條普惠型邊緣 AI 的新道路。

透過異質整合技術，aiPIM 實現了高效能 AI 處理器與標準 DRAM 的完美結合，創造出一種可直接替換現有記憶體的 AI 加速方案。這一創新使得所有物聯網與工業電腦裝置無需更改主晶片設計，僅透過替換記憶體，即可獲得強大的 AI 加速能力。

aiPIM 不僅僅是一款晶片，更是推動傳統記憶體產業邁向 AI 新紀元的關鍵催化劑。在當前 AI 浪潮席捲全球的背景下，aiPIM 為記憶體產業提供了一條差異化競爭的新途徑，有望重塑整個產業的競爭格局。



1 完成晶片投片與原型板卡開發

我們將在未來 12 個月內完成第一代 aiPIM 晶片的投片與測試，並開發出可供客戶評估的原型板卡。

2 持續完善軟體後端

我們將加強與主流 AI 框架的整合，提供更完善的開發工具與優化方法，降低客戶採用門檻。

3 建立行業合作生態

我們將與半導體廠商、系統整合商、終端客戶建立廣泛的合作關係，共同推動 PIM 技術的普及與應用。

加入我們的願景

我們誠摯邀請各界合作夥伴與投資者加入 aiPIM 的技術創新與商業化旅程。透過攜手合作，我們能夠加速這一革命性技術的市場化進程，共同創造 AI 運算的美好未來。

隨著 AI 應用的普及，邊緣運算的重要性日益凸顯。aiPIM 有潛力成為這一領域的關鍵使能技術，為萬物智能化提供高效、低成本的運算基礎。

我們相信，aiPIM 不僅是一項技術創新，更代表著一種新的運算範式——讓運算走向數據，而非數據走向運算。這一範式轉變將為 AI 時代的運算架構帶來深遠影響。