

行業概覽

本節及本文件其他章節所載資料及統計數據摘錄自我們委託灼識諮詢編製的報告、多份政府官方刊物及其他公開來源。我們委聘灼識諮詢就[編纂]編製獨立行業報告灼識諮詢報告。我們相信這些資料的來源都是適當的資料來源，並已合理審慎地摘錄及轉載該等資料。我們沒有理由相信這些資料為虛假或具誤導性，或遺漏了任何事實而導致這些資料為虛假或具誤導性。我們、聯席保薦人、[編纂]、[編纂]或參與[編纂]的任何其他人士，並無對來自官方政府來源的資料進行獨立核實，且概無就其準確性作出任何聲明。

資料來源

我們委聘灼識諮詢(一家提供行業諮詢服務、商業盡職審查及戰略諮詢的獨立市場研究及諮詢公司)就中國AI推理芯片相關產品及服務行業以及智能設備行業進行詳細研究及分析。我們同意向灼識諮詢支付人民幣869,200元(含稅)的費用，以編製灼識諮詢報告，我們相信該費率符合市價。我們認為支付有關費用不影響灼識諮詢報告所達致結論的公平性。

編製灼識諮詢報告時，灼識諮詢進行了一手及二手研究，並收集有關目標研究市場行業趨勢的知識、統計數據、資料及見解。一手研究包括與主要行業專家及領先的行業參與者訪談。二手研究包括分析來自各種公開來源的數據，如國家統計局。

灼識諮詢報告的編製基於以下假設：(1)在預測期內，中國整體政治、經濟及社會環境預計保持穩定；(2)在預測期內，相關主要行業推動因素很可能推動中國AI推理芯片相關產品及服務行業以及智能設備行業持續增長，包括優惠政策；及(3)在預測期間，不會發生可能對市場造成重大或根本性影響的極端不可抗力事件或出現不可預見的行業法規。

董事確認

經作出合理查詢後，董事確認，就彼等所知，自報告日期起，灼識諮詢報告中所述的市場資料並無發生任何可能限制、違背或影響本文件所載資料的不利變動。

大模型應用帶動AI推理的快速增長

AI、大數據和雲計算的快速發展正推動著一場席捲行業的智能化轉型浪潮，從根本上重塑各行各業的生產模式和競爭格局。AI正日益解決交通、互聯網搜索和製造等領域的實際問題，從而加快社會的智能化轉型。

大模型參數量持續增長，訓練千億參數模型的成本已突破千萬美元級別，邊際效益開始遞減。同時，推理成本的顛覆性下降也引發了推理應用的爆炸性增長。DeepSeek等國產開源模型的出現，無需完整的訓練週期即可進行特定場景的優化，進一步降低開發壁壘。此外，國產芯片替代的進展亦為大規模落地創造了更可靠的硬件基礎。

行業概覽

AI已從迭代算法改進發展到計算基礎設施的穩健進步，推動了基礎硬件的創新。在這一演變過程中，作為實現智能計算的核心硬件，AI芯片正成為價值鏈中的關鍵組成部分。AI的發展亦重新定義硬件產品。大模型正在催化兩條並行的硬件演進路徑，包括由AI快速發展催生的全新產品類別的AI原生產品及通過整合AI能力對現有產品進行升級改造而產生的AI賦能產品。

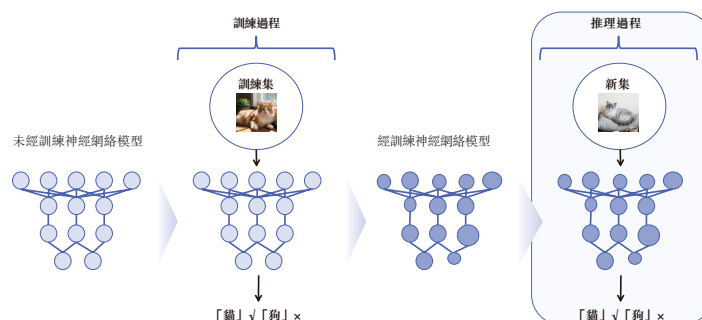
AI芯片的定義及分類

AI芯片已成為支持大規模運算任務方面不可或缺的硬件，在AI模型開發和部署的兩個基本階段——訓練和推理中，均發揮著至關重要的作用。*訓練*涉及利用大規模數據集和機器學習算法開發AI模型。作為AI發展的基石，此階段著重於設計模型架構及優化參數。其計算需求隨著模型複雜度和訓練規模的指數增長而增加。*推理*則與之相對，其將應用訓練好的模型來處理新數據，可執行影像、語音、文字識別、分類及預測等任務。其效能取決於併發處理及響應速度，針對特定部署情境進行定製。

鑑於該等階段的運算強度、處理頻率及部署環境上的顯著差異，各階段都需使用專門AI芯片：*訓練芯片*擅長構建大規模AI模型，強調高運算精度、大內存帶寬及高吞吐量。該等芯片主要用於雲端數據中心，用於處理高負荷工作。*推理芯片*則優化效率，強調低延遲、高性價比、能效比及可擴展性。該等芯片可覆蓋雲平台至邊緣及終端等多樣化場景，滿足各種應用的即時需求。

AI推理指經訓練的模型運用其學習到的參數，分析先前未見過的新數據，進而產生分類結果或預測結果的過程。其代表AI模型從「學習」轉向「應用」的關鍵階段。

訓練及推理過程



資料來源：灼識諮詢報告

附註：訓練集中的標註數據輸入至未經訓練模型中，例如使用標註為「貓」的圖像作為訓練樣本。模型會根據其預測與真實標籤之間的差異，迭代調整其內部參數，逐步學習相關特徵並形成穩定的識別能力，最終成為經訓練模型。訓練完成後，模型進入推理階段，此時其接收來自訓練集之外的新數據集圖像，並運用訓練過程中學習到的特徵對輸入數據進行分析，並生成預測結果。

行業概覽

AI推理芯片的架構演變

隨着人工智能應用的快速普及，為滿足市場對極致推理性價比的高能效人工智能推理系統的需求，用於人工智能推理的芯片架構正在發生根本性變革，從CPU通用計算，GPU通用並行計算轉向NPU面向人工智能推理領域優化的並行計算。在人工智能的早期階段，人工智能模型的訓練和推理計算，採用的是面向通用計算的基於CPU的IT基礎設施，但CPU在並行處理能力和能效方面的局限，促使行業開始採用具備通用並行計算能力的圖形處理器(GPU)。當前，多數人工智能模型的訓練是利用GPU完成的。但是，人工智能模型的訓練和推理是兩類不同的任務，它們之間存在顯著的差異。儘管GPU具備更出色的並行計算能力，但其面向圖形處理的架構，對於張量運算(即大規模矩陣運算)，尤其是僅關注神經網絡前向計算的AI推理而言，並非最優選擇。隨着人工智能應用採用的深度學習模型日趨複雜，GPU在功耗和計算密度上的短板逐漸暴露，行業由此轉向專門設計的神經處理單元(NPU)。

NPU是人工智能推理加速計算領域的一次巨大飛躍，它原生支持張量運算，支持低精量化，並具備大規模並行處理能力。這種架構專用化的芯片在四個關鍵維度實現了前所未有的提升：(1)延遲降低(支持實時推理)；(2)能效提升(最大化每瓦吞吐量，將部署範圍擴展至邊緣設備和終端設備)；(3)計算密度提高(單位尺寸的芯片可容納更強的專用工作負載計算性能)；及(4)成本效益優化(推動其在各類應用場景中更廣泛、更可持續地落地。)通過軟硬件協同設計，NPU實現了算法感知層面的優化，將人工智能推理從一種功能性能力轉變為可擴展、高性價比且可投入生產的解決方案。

從通用CPU到半專用GPU，再到針對人工智能推理工作負載優化的融合架構GPNPU或專用架構NPU/ASIC，這種架構演進歷程映射出人工智能部署的成熟過程。場景特定化優化的趨勢將進一步強化，未來的芯片及系統會更精準地適配垂直領域應用的獨特計算模式。這種硬件專用化與算法優化的結合，正形成一個良性循環，推動人工智能從實驗室走向實際應用，產生真實世界價值。

AI推理芯片的主要部署地點及分類

AI推理芯片可按部署地點分為三類：雲端推理芯片、邊緣推理芯片及終端推理芯片。

- **雲端推理芯片**用於數據中心及公眾雲服務器，提供每秒100至超過1,000 TOPS的運算能力，用於高併發任務及大模型推理。雖然該等芯片具有優秀處理能力及可擴展性，但其高能耗帶來顯著能源管理挑戰。
- **邊緣推理芯片**用於雲服務器及終端，通常為邊緣網關、機器人等應用提供100-300 TOPS的計算能力。該等芯片以低延遲、高可靠度及均衡能效為設計核心，通過本地數據處理以降低對雲端的依賴及帶寬需求。該等芯片的中階運算效能使其非常適合需要快速響應及在數據來源附近穩定運作的場景。該等芯片亦能夠執行壓縮的AI模型，支持各種邊緣智能任務。

行業概覽

- **終端推理芯片**直接集成至智能手機(1至50 TOPS+)及可穿戴設備(0.1至0.5 TOPS)等輕量級商業產品，著重超低功耗及緊湊設計。該等芯片可進行毫秒級本地推理，在加強即時響應的同時優化電池壽命。隨著輕量級大模型逐漸增加在終端層面的應用，終端推理芯片在支持智能語音助手、實時翻譯和圖像識別等應用上扮演著重要角色。通過處理終端內的數據，該等芯片能大幅改善移動應用的用戶體驗，並大幅降低雲傳輸需求。

AI推理芯片相關產品及服務的技術演進

AI應用在各行各業的快速擴展，正提高推理芯片的三項關鍵需求：更高的性價比、更高的能源效率以及更強的場景適應性。該等需求正加快四大關鍵技術領域的創新。

- **算法及芯片架構協同優化**。先進的共同優化方法現已整合整個開發鏈，從算法分析到工具鏈開發，創造出專門針對目標工作調整的硬件。該協同效應可提供更快更精準的推理，同時最大程度提升資源效率，特別是在面積及能耗受限的環境中。
- **神經網絡效率及能耗優化**。邊緣及終端部署需要完美平衡吞吐量與能源效率。前沿技術通過優化乘加單元(MAC)結構、實現稀疏計算以及精簡存儲訪問路徑，在保持模型精度的同時大幅降低能耗。
- **存算一體(「PIM」)技術**。傳統的芯片架構將計算與存儲分開，造成能源密集的資料傳輸，減慢AI處理速度。近存計算將計算單元放置在更靠近存儲的位置，部分解決了該問題。PIM技術將計算直接嵌入存儲單元，毋需移動即可處理數據。這項突破可大幅提升效率與速度，讓PIM架構成為未來AI芯片的必要條件。
- **基於Chiplet的異構集成**。AI複雜度不斷提升，單片芯片已難以滿足需求。Chiplet技術將系統拆解為模塊化的NPU、存儲器、I/O接口及存儲塊，允許在先進封裝載體上靈活組合集成。此模式轉變帶來了三個關鍵優勢，包括通過優化IP重用提高性能、通過更小的裸片提高產量以及通過並行開發縮短產品上市時間。Chiplet集成正在成為下一代高效能AI芯片設計的關鍵方針。

AI推理芯片相關產品及服務的應用場景

AI推理芯片相關產品及服務為企業、消費者及工業應用的重要基礎設施。在企業場景中，該等芯片相關產品及服務為互聯網公司、AI公司、電信營運商及研究機構提供AI推理服務器、邊緣網關及機器人等。該等平台提供實時大規模推理任務所需的運算能力，推動企業營運與創新。

對於消費者應用而言，該項技術可透過設備、雲端或混合應用展現。可穿戴設備採用專用芯片，著重超低能耗與緊湊設計，以提供反應迅速、注重隱私的AI體驗。同時，雲端部署著重支持高併發性和節能處理，以處理跨服務的動態用戶需求，例如兒童智能硬件、學習設備與智能家居硬件。

行業概覽

工業應用透過針對特定垂直需求打造的邊緣及雲端解決方案，充分利用該等芯片相關產品及服務。從智慧城市基礎設施到智慧交通系統，該技術可實現本地化處理、定製化功能及實時響應能力。該等芯片相關產品及服務在不同場景中的適應能力突顯其在整個AI價值鏈中的轉型潛力。

中國全場景AI推理芯片相關產品及服務的需求

中國全場景AI推理芯片相關產品及服務行業涵蓋了廣泛的硬件和軟件解決方案，旨在支持AI推理的多樣化應用場景——包括企業、消費者和工業應用。該產業可實現全面的全場景佈局，確保將AI功能無縫嵌入數字化生態系統的各個層級。

AI推理芯片相關產品及服務與智能設備形成緊密整合的系統，可實現高效率、適應性和智算。透過緊密的協同合作，可支持雲端、邊緣與終端之間的全面部署，確保智能處理能力可橫跨所有場景。AI推理芯片相關產品及服務作為計算骨幹，正逐漸嵌入智能設備，以滿足不同應用場景下對實時、設備端AI的日益增長的需求。智能設備透過收集寶貴的真實數據（包括視覺、音頻、語義和行為數據）及提供重要的表現回饋，積極推動AI創新。此外，已部署設備的運作指標與已識別的效能瓶頸，可為算法與芯片架構的共同優化提供重要的見解，創造出一個持續改善的循環，讓實際應用推動軟硬件的進步。

全場景生態系統的發展不僅推動了AI推理芯片的演進，亦促進了向智能設備等下游應用的擴展，為整個價值鏈帶來了顯著優勢。AI應用場景正在穩步擴大，從工業製造、智慧城市到消費電子，增強了智能設備的數據處理和自主決策能力。全場景部署不僅提高了AI推理芯片和智能設備的實用價值，也為行業釋放了更廣闊的市場潛力。在此背景下，具備全場景能力的公司將獲得更強的競爭優勢。

在全場景AI能力不斷擴展的過程中，NPU架構的發展不僅體現在芯片設計的進步，還推動了智能設備的變革。智能設備在此被單獨說明，因為它們不僅是計算平台、數據生成和智能用戶交互的重要載體，還具有自身的品牌價值。以下內容將分為兩部分進行分析：AI推理芯片相關產品及服務與智能設備，因這兩部分在各自的價值主張方面存在差異。

中國AI推理芯片相關產品及服務行業概覽

AI推理芯片相關產品及服務的定義

AI推理芯片相關產品及服務融合專門的軟硬件部件，提供實時AI處理能力。此外，該類產品通常配備算力服務，進一步滿足多元化的推理工作負載需求，構成更全面的端對端解決方案。硬件的核心特色是採用高能效的芯片架構，搭配經優化的部署框架。軟件棧構成智能骨幹，當中包含算法模型、編譯器工具鏈、指令集及處理器架構等關鍵元素，所有元素協同運作，透過就特定場景最佳化將硬件潛力發揮到極致。這種整體性的方針可讓開

行業概覽

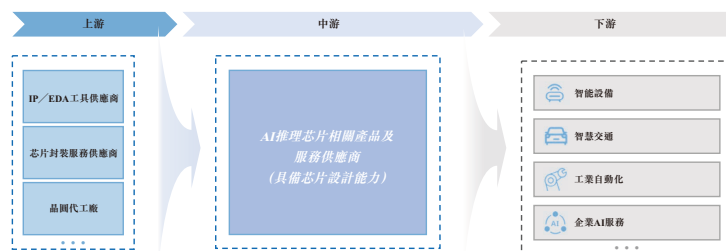
發人員充分利用芯片功能，同時大幅降低部署複雜性，縮短AI理論模型與實際高效能運作之間的差距。

AI推理芯片相關產品及服務價值鏈分析

AI推理芯片相關產品及服務的行業價值鏈包括三個核心部分。上游部分包括IP供應商（設計及授權可重複使用的半導體知識產權模塊，如處理器核心、存儲控制器及接口標準）、EDA工具供應商（提供用於芯片設計、驗證及測試的電子設計自動化軟件）以及芯片封裝及測試公司和晶圓代工廠，均提供必要的技術基礎能力及製造能力。中游部分圍繞AI推理芯片相關產品及服務供應商，利用彼等的創新能力及累積的技術知識進行電路設計及驗證，同時審慎地平衡效能、耗電量及成本。下游部分涉及於廣泛的應用場景中部署AI推理解決方案。下游需求的多樣性推動芯片設計公司建立豐富的產品組合，令其可透過靈活的產品組合滿足不同的客戶需求。

在企業級及行業級場景中，我們的產品及服務主要運作於行業價值鏈的中下游部分。

AI推理芯片相關產品及服務價值鏈



資料來源：灼識諮詢報告

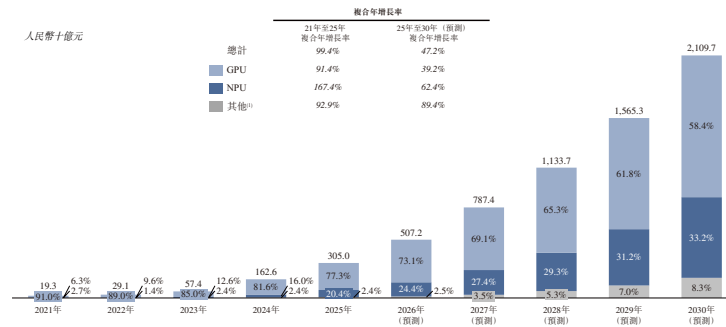
中國AI推理芯片相關產品及服務行業的市場規模

中國的AI芯片需求快速增長。中國AI芯片相關產品及服務的市場規模由2021年的人民幣284億元增至2025年的人民幣4,273億元，複合年增長率為97.0%，並預期將於2030年達至人民幣26,375億元，2025年至2030年的複合年增長率為43.9%。

中國AI推理芯片相關產品及服務行業正高速增長。市場規模由2021年的人民幣193億元增至2025年的人民幣3,050億元，複合年增長率為99.4%，並預期將於2030年達至人民幣21,097億元，2025年至2030年的複合年增長率為47.2%。NPU驅動的市場規模由2021年的人民幣12億元增長至2025年的人民幣621億元，複合年增長率為167.4%，預計2030年將攀升至人民幣7,012億元，反映2025年至2030年的複合年增長率為62.4%。

行業概覽

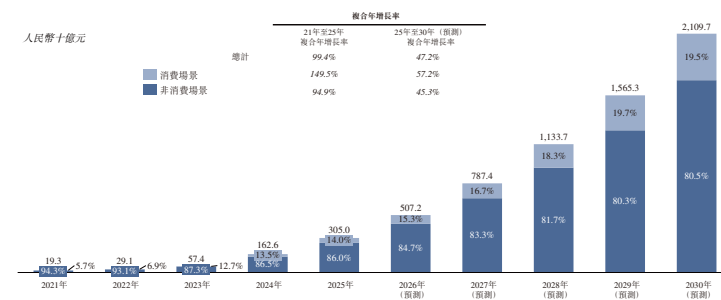
中國按收入計的AI推理芯片相關產品及服務行業的市場規模， 2021年至2030年（預測）



資料來源：IDC、灼識諮詢報告

(1) 「其他」指並非嚴格分類為GPU或NPU但亦用於AI推理界面場景的其他具備AI能力的硬件部件

中國按收入計的AI推理芯片相關產品及服務行業的市場規模， 按下游行業分類，2021年至2030年（預測）



資料來源：IDC、灼識諮詢

消費場景乃中國AI推理芯片相關產品及服務市場中成長最迅速的領域，2021年至2025年複合年增長率達149.5%，而非消費場景（主要指企業及行業場景）的複合年增長率為90.0%。然而，預計兩大領域在2025至2030年均將保持強勁增長勢頭，其中消費場景的市場份額有望進一步擴大。

中國AI推理芯片相關產品及服務行業的主要驅動因素及趨勢

中國AI推理芯片相關產品及服務行業的主要驅動因素及趨勢包括以下各項。

- AI推理需求急速增長。AI大模型的爆發性增長已成為推理需求的主要驅動因素。AI生成內容、虛擬助手及虛擬人等應用所產生的推理請求正在以幾何級數增長。

行業概覽

- **企業數字化轉型。**企業通過廣泛採用AI驅動的自動化和分析推動需求增長。企業對公共雲和私有AI部署均表現出強烈需求。結合提升運營效率的壓力，成本效益高的AI推理芯片相關產品及服務成為企業IT基礎設施的重要選擇。
- **消費電子革命。**消費者市場為推理芯片的部署帶來獨特的挑戰。耳機與手錶等智能可穿戴設備需要結合高效能與超低能耗的芯片，以克服嚴格的尺寸、電池及散熱限制。消費者對實時、個性化AI體驗及強隱私保護的期望日益提升，進一步加速了高效設備端推理解決方案的採用。
- **政府帶動基礎設施需求。**在行業層面，國家政策正積極推動以數字化、互聯網絡及智能系統為特徵的新一代城市發展。例如，國務院辦公廳提出全面建設城市運行管理服務平台及城市信息模型(「CIM」)平台等重點項目。此類大型項目需龐大運算基礎設施，使運營成本控制成為關鍵課題。因此，AI推理芯片相關產品及服務必須具備卓越的能效比，方能滿足城市規模部署及智能產業園區的需求。
- **持續技術進步。**供應端正以芯片架構與製程上的快速創新做出回應，配套工具鏈和模型壓縮技術的同步提升也顯著降低了部署門檻。該等技術迭代正在提升推理芯片從能源效率到計算密度的各個層面。
- **顯著的性價比優勢。**與訓練解決方案相比，推理芯片在成本效益和部署靈活性方面具有明顯的優勢。隨著生產規模擴大、供應鏈提升以及實現彈性共享資源分配，單位計算能力的成本持續下降。此優越的性價比使得AI推理芯片相關產品及服務在所有市場領域的商業部署中日益具有吸引力。
- **應用場景的擴展。**AI推理芯片相關產品及服務正擴展至新的垂直應用，拓展至傳統用途之外，為自動駕駛、機器人和AI智能體等進階領域提供動力。這些前沿應用對性能要求顯著提升，包括亞毫秒延遲、故障安全可靠性和自主決策能力。隨著推理解決方案深入整合至專門的行業工作流程中，其正從基本功能工具演進為最佳化、高效率的系統，以實現廣泛的商業部署。
- **NPU滲透率上升。**將更適合AI推理的NPU整合到各種終端中已成為明顯的行業趨勢。這一趨勢體現在NPU在AI推理芯片中的滲透率穩步上升，由2021年的6.3%升至2025年的20.4%，預計於2030年達到33.2%。NPU驅動的硬件正迅速普及，重塑了部署路徑，加速擴展AI推理芯片生態系統。

行業概覽

中國智能設備行業概覽

智能設備的定義

智能設備作為連接雲平台、邊緣節點和終端用戶環境的關鍵基礎設施，具有相當大的規模。目前上述AI增強型終端整合了先進的感知、即時處理和自主決策能力，從簡單的控制單元轉變為精密的認知系統。這一架構轉變使得設備不僅能夠收集數據，還能理解環境語境並主動採取適當的行動，標誌著從程式化操作到真正環境智能的根本性轉變。

下一代智能設備專為深度領域整合而設計，結合了專門的傳感器陣列和處理模組，可針對特定案例量身打造。從精確的影像識別到自然語音互動，這些專門打造的系統在提供定製化智能的同時，也維持高效率的運作。此趨勢反映了嵌入式AI的成熟，智能設備正從單純接收信息和被動響應，轉向主動感知並智能響應。設備日益具備感知、分析和適應環境的能力，實現更加自然和前瞻性的交互體驗。

智能設備的分類

智能設備越來越多地採用雲端與設備端混合架構，利用雲端協同處理複雜的任務，同時透過改進的模型壓縮和推理芯片來推進本地化處理。這一演進為消費者帶來了重要優勢：更低延遲、更強大的隱私，以及透過邊緣運算進行更有效率的互動。

消費級智能設備⁽¹⁾可根據其形式和功能大致分為三個主要類別。智能音箱和智能攝像頭等智能家居設備，著重於語音互動、安全監控和家庭環境中的行為感知。其中有些硬件會先在本機執行初步處理，再將複雜的數據上傳至雲端進行深入分析。可穿戴設備包括AI驅動的耳機、智能眼鏡和智能手錶。其強調輕量設計與低能耗推理，因此適用於需要即時感知與回饋的行動環境。兒童智能硬件與學習設備為具備語音辨識與影像處理等功能的互動設備，可提供結合教育與娛樂的智能體驗。這些都是針對兒童早期學習和陪伴等特定需求量身打造。

智能設備價值鏈分析

智能設備的價值鏈由三個關鍵環節組成，包括：(1)上游元件供應商，生產定義設備功能和外形因素的關鍵元件，包括半導體製造商的核心IC芯片以及傳感器、顯示模組、電池及結構元件；(2)中游設計和製造專家，包括負責系統整合的獨立設計公司(Independent Design Houses, IDH)和負責生產的原始設計製造商(Original Design Manufacturers, ODM)；及(3)下游品牌擁有者負責產品定義、分銷和消費者參與，從元件到終端用戶體驗，形成了一個高效、專業的生態系統。

隨著AI整合的加速，智能設備價值鏈正在經歷重大轉變。芯片製造商及其他結構元件生產商等上游元件供應商正在運算密度和能源效率方面取得突破性進展，賦予設備日益強大的設備端AI能力。這些創新讓消費性產品具備更智能、更自主的功能。中游市場已成為策略戰場，IDH和ODM紛紛開發出獨特的價值主張。隨著對AI技術的日益重視，IDH正成為

⁽¹⁾ 此處所指的智能設備，特別是指專為特定場景設計的消費級產品，不包括智能手機、個人電腦、平板電腦和汽車。雖然這些被排除在外的類別亦整合了AI功能，但它們主要是設計為多功能運算終端，而非針對專門智能功能進行最佳化的AR眼鏡等智能設備。

行業概覽

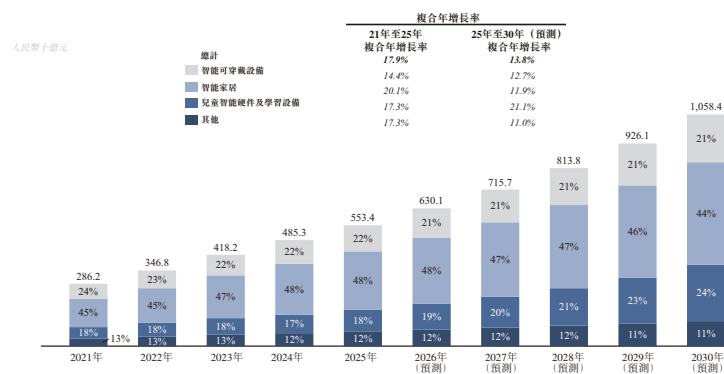
關鍵競爭力量，提供涵蓋硬件設計、算法優化和系統集成的全方位解決方案服務，提供比傳統ODM更深入的專業技術。其能夠為特定行業應用提供定製化的軟硬件協同開發解決方案，從而實現產品的顯著差異化，而ODM則專注於卓越製造和產品快速落地。品牌商則利用其市場地位，推動持續創新，通過深入的用戶洞察引導產品演進。在技術融合與需求分散的時代，行業發展愈發依賴無縫的跨鏈協作—整合上游技術能力、中游集成專長和下游市場洞察，共同打造成功的智能設備生態系統。

我們的主要產品包括同時服務中下游板塊的消費級場景智能設備。

中國智能設備行業的市場規模

中國智能設備行業正經歷快速增長，尤其是在消費場景方面。中國消費級智能設備的市場規模由2021年的人民幣2,862億元增至2025年的人民幣5,534億元，複合年增長率為17.9%，預計到2030年將達到人民幣10,584億元，2025年至2030年的複合年增長率為13.8%。

中國消費級智能設備行業的市場規模，2021年至2030年（預測）



資料來源：Statista、灼識諮詢報告

中國智能設備行業的主要驅動力及趨勢

中國智能設備行業的主要增長驅動力及趨勢包括以下方面。

- **AI重塑智能設備。** AI大模型的成熟從根本上重新定義了智能設備的功能，透過「端邊雲」協作，將先進推理從雲端延伸至邊緣。AI原生產品作為AI技術進步誕生的全新產品類別，圍繞大模型能力而建立，並提供自主決策等智能服務。此外，傳統設備也正在被AI賦能，使功能得以持續增強。這些設備正從被動硬件向能夠主動感應用戶意圖並提供回應式服務的智能代理轉變。
- **監管框架及行業發展。** 中國不斷擴大的智能設備市場見證著政策的逐步實施，包括頒佈《電子信息製造業數字化轉型實施方案》，該標準對數據隱私、安全性和互

行業概覽

操作性提出了嚴格的要求。於2024年9月，工業和信息化部發佈《關於推進移動物聯網「萬物智聯」發展的通知》，強調軟硬件一體化解決方案在推動計算和存儲服務需求方面的戰略價值，進一步加速行業轉型。這些措施正在提升產品品質、建立消費者信任，並為整個智能設備生態系統創造新的價值。

- **國產化創新與供應鏈韌性。**在強大的政策支持和國內市場需求的支持下，中國智能設備製造商正從進口替代過渡到全球技術領導。AI芯片、自主算法和系統軟件方面的突破，使中國企業獲得了關鍵的技術自主權，大大加強了供應鏈的穩定性。
- **多模態感知與邊緣智能化。**智能設備正經歷根本性的轉變，從基本的傳感器演變為精密的邊緣運算節點。藉助感知技術和人工智能算法的進步，現代設備現在已整合了多種輸入模式，包括視覺、聽覺、手勢和情感識別，實現了更自然的人機互動。這一從被動數據採集向主動環境理解與情境響應的轉變，標誌着設備智能的躍進，多模態感知成為新的行業標準。
- **一體化解決方案產品的興起。**行業正從單一硬件轉向採用整體的硬件—軟件—服務的商業模式。具有前瞻性的IDH公司正在開發自主算法平台、管理系統和分析後端，並將其與硬件專業技術相結合。這些組合方案通過提升產品功能、通過持續的服務收入增強客戶忠誠度，以及構建超越傳統單次硬件銷售的可持續商業模式，實現顯著價值提升。
- **生態系統驅動發展。**隨著硬件的複雜性和應用場景的成倍增加，無縫的生態系統整合變得至關重要。領先企業現重點關注三個維度：緊密的軟硬件協同、通用平台兼容性以及智能跨設備協作。從單一產品向互聯系統的演進，實現了設備網絡的統一管理和跨場景的協調回應，顯著改善了用戶體驗和運營效率，同時促進了更具活力的智能設備生態系統。

競爭格局

只有少數行業領先企業能夠提供涵蓋企業、消費者和工業應用的全場景解決方案。以收入計，我們躋身市場前三，擁有獨特的優勢，在中國市場提供全場景的AI推理芯片相關產品及服務。

AI推理芯片相關產品及服務行業大致分為兩類：具備全棧能力及商用產品的獨立廠商，以及提供內部業務需求的供應商。我們專注於第一類兼備AI推理芯片相關產品及服務的獨立廠商。

中國的AI推理芯片相關產品及服務行業為寡頭壟斷市場，到2025年，兩大領導廠商合共擁有約83.6%的市場佔有率。2025年，我們以收入計在中國市場所有AI推理芯片相關產品及服務供應商中排名第10。下表為業內名列前茅的公司。

行業概覽

2025年按收入劃分中國市場AI推理芯片相關產品及服務供應商排名⁽¹⁾

排名	公司	AI推理芯片 相關產品及服 務所得收入 (人民幣十億元)	市場份額 (%)	主要技術 路線圖
1	公司A ⁽³⁾	~200.0	~66.8%	GPU
2	公司B ⁽⁴⁾	~51.1	~16.7%	NPU
3	公司C ⁽⁵⁾	~5.5	~1.8%	NPU
4	公司D ⁽⁶⁾	~3.0	~1.0%	XPU ⁽²⁾
5	公司E ⁽⁷⁾	~1.6	~0.5%	GPU
6	公司F ⁽⁸⁾	~1.5	~0.5%	GPU
7	公司G ⁽⁹⁾	~1.0	~0.3%	GPU
8	公司H ⁽¹⁰⁾	~1.0	~0.3%	GPU
9	公司I ⁽¹¹⁾	~1.0	~0.3%	GPU
10	本公司	~0.7	~0.2%	NPU

資料來源：企業公開資料及年報、灼識諮詢報告

- (1) 不包括消費級場景智能設備。
- (2) XPU為GPU的一種形式。
- (3) 公司A成立於1993年，總部位於美國，是一家主要從事GPU設計與銷售的無晶圓廠半導體公司。A公司在納斯達克上市。
- (4) 公司B是一家成立於1987年的私人公司，總部位於中國。該公司為全球信息及通信技術(ICT)基礎設施及智能設備的供應商。
- (5) 公司C成立於2016年，總部位於中國，主要從事人工智能芯片產品的研究及創新。該公司於上海證券交易所上市。
- (6) 公司D是一家於2021年成立的私人公司，總部位於中國。該公司專門從事AI芯片及相關硬件解決方案的研究、開發及商業化。
- (7) 公司E是一家GPU芯片及解決方案供應商，總部位於上海，於2020年成立。該公司是一家上海證券交易所上市公司。
- (8) 公司F是一家全功能GPU供應商，總部位於北京，於2020年成立。該公司是一家上海證券交易所上市公司。
- (9) 公司G專注於GPGPU芯片及智能計算解決方案的研發，總部位於上海，於2019年成立。該公司是一家聯交所上市公司。
- (10) 公司H是一家GPGPU產品及AI計算解決方案供應商，總部位於上海，於2015年成立。該公司是一家聯交所上市公司。
- (11) 公司I是一家AI芯片及計算解決方案供應商，總部位於上海，於2018年成立。該公司是一家私人公司。

近年來，推理工作負載激增，成為AI應用的主要動力。與GPU相比，NPU作為專業、高效的技術解決方案，更適合推理任務。我們很早就NPU領域進行戰略佈局，並在市場佔據一席之地，躋身行業前三。較早進入NPU領域並已在市場上取得領先地位的企業，可能會因NPU加速滲透而更大幅獲利，在市場增長和擴大市場份額方面均有較大的潛力。

行業概覽

2025年按收入劃分中國市場NPU驅動的AI推理芯片相關產品及服務供應商排名⁽¹⁾

排名	公司	AI推理芯片 相關產品及服務 所得收入 (人民幣十億元)	市場份額 (%)
1.....	公司B	~51.1	~82.2%
2.....	公司C	~5.5	~8.9%
3.....	本公司	~0.7	~1.1%

資料來源：年報、灼識諮詢報告

(1) 不包括消費級場景智能設備。

中國的智能耳機模組市場仍然高度集中，前五大國內製造商在2025年合共佔有73.4%的市場份額，而我們則以37.9%的出貨量佔有率領先業界，反映出我們在這個重要的智能設備類別中的強勢地位，該品類支持語音交互、實時音頻處理及設備間的無縫連接。

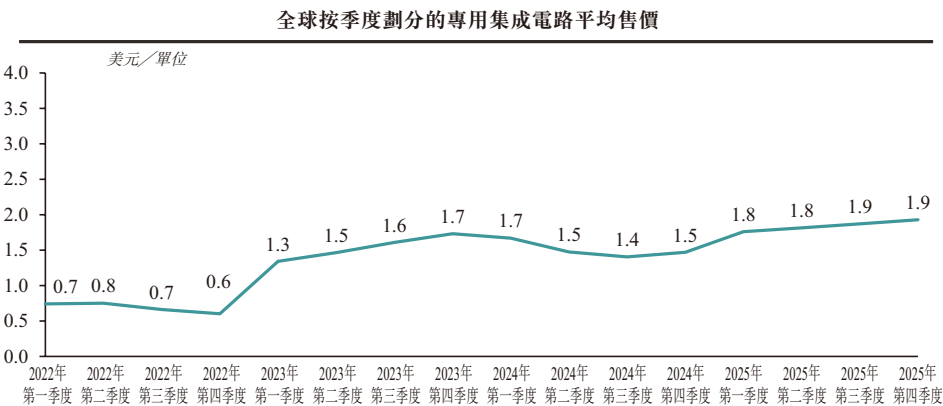
中國全場景AI推理芯片相關產品及服務行業關鍵成功因素

技術門檻高、行業知識沉澱週期長、人才優勢、客戶優勢、全國產產業鏈及全棧式解決方案及服務能力為中國全場景AI推理芯片相關產品及服務行業的關鍵成功因素。

專用集成電路的價格趨勢

消費級產品的主要原材料包含芯片及印製電路板(PCB)，此外還包含麥克風、電容器、電阻器等常見電子元件。下圖呈現全球專用集成電路(ASIC)自2022年第一季度至2025年第四季度的歷史價格走勢。中國市場的ASIC價格亦預計將保持整體穩定。

消費級產品使用的PCB通常根據特定產品架構進行定製，其成本受板面尺寸、層數及加工複雜度影響。採用國內採購PCB的企業往往能受益於本地化採購及更靈活的供應鏈所帶來的顯著成本優勢。因此，除非原料價格或產能出現重大波動，該等企業通常會發現PCB單位成本呈現穩定至輕微下降趨勢。結構性元件通常向國內供應商採購，價格依據重量及材質等級計價。該等零件多屬標準化量產產品，對全球商品價格波動敏感度較低，有助於維持行業整體成本趨勢的穩定性。



來源：世界半導體貿易統計組織，灼識諮詢報告