

建設人工智能創新發展第一城

技術攻堅

人工智能（AI）應用近年呈爆發式增長，底層算力芯片迎來前所未見的黃金發展期。由商湯科技分拆獨立而成的國產圖形處理器（GPU）新星曦望 Sunrise 憑借 S1 及 S2 兩款 AI 芯片的量產以及今年 1 月發布的針對推理場景設計的 S3，完成多輪融資，資本市場關注度持續升溫。曦望 Sunrise 聯席 CEO 王湛日前接受《經濟論》專訪時表示，公司戰略性聚焦於天花板更高的「推理算力」市場，並以「為中國 AI 發展鑄算力底座，讓 AI 推理算力如水電般普惠可得」為目標，銳意以更低廉的推理成本，為各行各業提供穩固的算力基石，推動 AI 算力未來邁向普惠化。

●文：香港文匯報

記者 黎梓田

圖：香港文匯報

記者 黃宇威

經濟論

走中國看經濟之杭州篇

王湛在訪問中指出，曦望前身為香港主板上市企業、亞洲 AI 龍頭商湯科技旗下的大芯片部門。他透露，商湯在拓展 AI 業務過程中深有感觸，意識到算力是 AI 發展最核心的底層支柱；為突破此關鍵瓶頸，商湯早年已投入龐大資源進行自主研發，並在充裕資金與技術支持下，順利完成 S1 及 S2 兩款 AI 芯片由研發至量產的整個流程。

推理計算需求將達訓練 5 倍

與市場上多數國產芯片商選擇「訓練與推理一體化」路線不同，曦望從成立之初便堅定「All-in 推理」賽道，專注於高性能 GPU 及多模態場景推理芯片的研發與商業化。此戰略布局建基於對 AI 產業趨勢的深度研判，曦望認為，AI 訓練屬少數巨頭主導的資本密集型遊戲，而推理需求則具備幾近無上限的增長潛力。曦望董事長徐冰此前亦指出，AI 產業正由「訓練驅動」全面轉向「推理主導」；預計 2026 年 AI 推理計算需求將達訓練需求的 4 至 5 倍，而推理算力租賃價格在半年內已上漲近 40%。王湛更直言：「誰的推理成本更低，誰就是最終贏家。」

讓 AI 成為像水電的普惠設施

早前，徐冰曾表示，隨著 AI 的發展，未來十年，地球上預計會出現上百億個智能體，成為新的「數字人口」，扮演人類的「同事」和「助理」角色。Token 作為智能時代的基礎能源，而推理 GPU 則是 AI 時代「發電廠」中最關鍵的部位。曦望的目標是持續降低推理成本，使 AI 像水電一樣成為普惠的基礎設施。為此，公司將持續加大對近存計算、光電共封裝（CPO）等前沿技術的探索投入，並通過開源策略攜手合作夥伴，共同構建萬億級推理基礎設施。

曦望選擇放棄訓練能力，專為大模型推理進行原生設計，其核心理念在於：AI 訓練決定的是模型能力的上限，而 AI 推理則直接創造商業價值，並決定 AI 技術能被多少用戶及企業實際應用。王湛認為，當前訓練芯片市場已由英偉達、華為等龍頭企業深耕，推理芯片則被視為更具成長空間的藍海市場。憑藉八年技術積累，以及兩代量產芯片的工程化驗證，曦望已逐步確立其國產 GPU 領域的核心力量。公司致力為各行業提供成本降低十倍、能效比顯著突破的智能算力基礎設施。

構建開放生態 與業界模型無縫對接

隨著大模型從「炫技」階段邁向「落地」應用，推理場景正由簡單的問答對話，擴展至即時影片生成、多模態互動、邊緣端智能等多元化應用。王湛指出，曦望的原生推理架構無需為訓練算子妥協面積與功耗，在同等製程下，推理能效比較通用 GPU 高出數倍，成為其敢於喊出「為中國 AI 發展鑄算力底座」的底氣所在。與此同時，曦望積極構建開放生態，透過統一編程框架與業界主流模型無縫對接，使開發者無需重新編寫代碼即可獲得極致推理效能，真正實現低成本、高效能算力的觸手可及。

誰就是 AI 時代贏家 誰的推理成本低

算力戰由訓練轉向推理
曦望芯片全速突圍



● S2-X1 PCIe 計算卡是推理/訓練一體 GPGPU 硬件產品。

迅速築三代芯片矩陣 精準切入 AI 推理賽道

香港文匯報訊（記者 黎梓田）自商湯科技分拆獨立營運以來，曦望 Sunrise 數年間迭代三代產品的芯片矩陣，全面覆蓋從多模態推理、大規模語言模型（LLM）至 AI 智能體的全場景推理需求，展現出清晰的技術演進路徑與市場布局策略。

第一代芯片獲索尼小米採用

王湛表示，作為其進軍 AI 芯片市場的開篇之作，第一代 S1 芯片為專注多模態推理的芯片產品，成功適配市場主流模型。該芯片已實現 IP 授權予索尼及小米，分別應用於索尼 AI 攝像頭與小米智能手機，標誌着曦望在終端視覺推理領域取得初步商業化驗證。

2024 年，曦望推出 S2 芯片，主要採用通用圖形處理器（GPGPU）架構，性能對標市場主流的 A/H 系列 GPU，並高度兼容 CUDA 生態系統，顯示其進軍通用 AI 運算市場的決心。據公司透露，S2 已成功適配 DeepSeek、Qwen 等主流開源大模型，實測運算

性能接近國際頂尖產品。

隨之而來的曦望第三代旗艦產品啟望 S3 芯片於今年 1 月發布，為內地首款同時搭載 LPDDR6 並向下兼容 LPDDR5X 記憶體標準的推理專用 GPU。王湛稱，該產品在設計策略上作出重大取捨，放棄訓練功能，專注於大模型推理應用的原生深度定製。透過精簡訓練階段所需電路模組，S3 將節省下的電晶體及功耗資源集中投放於推理運算，使單位面積有效算力效率較前代產品提升逾 5 倍。

第三代芯片推理成本大降 90%

此外，S3 支援從 FP16 至 FP4 的多精度靈活切換，在典型推理場景下，整體性價比相較上一代產品實現超過 10 倍的躍升。曦望更宣稱，搭載 S3 芯片後，單位 Token 推理成本可較前代降低約 90%，為大規模 AI 應用落地提供更具競爭力的成本結構，進一步鞏固其在 AI 推理芯片市場的領先地位。

中國 AI 芯片業將洗牌 汰弱留強

香港文匯報訊（記者 黎梓田）隨着人工智能應用加速落地，中國 AI 芯片市場格局成為業界關注焦點。王湛認為，中國 AI 芯片市場未來會向頭部集中。半導體行業具有規模效應，但「贏家通吃」通常發生在同一細分市場、同一代技術路線之中。缺乏差異化能力、重複做同一件事的企業會面臨淘汰的風險。

王湛表示，隨着 AI 產業由訓練向推理延伸，推理

市場正在形成不同於訓練市場的競爭格局。訓練更強調生態、先發優勢和通用算力，而推理更加關注單位成本、顯存容量、能效及場景適配能力。

當前市場既有海外 GPU 廠商，也有訓練一體 GPU 和 ASIC 專用芯片等不同技術路線，各自在不同場景發揮優勢，未來競爭的關鍵在於形成差異化能力，而不是簡單重複同一條技術路線。

●曦望 Sunrise 數年間迭代三代產品的芯片矩陣，全面覆蓋從多模態推理、大規模語言模型（LLM）至 AI 智能體的全場景推理需求。



●曦望 Sunrise 聯席 CEO 王湛表示，公司聚焦於天花板更高的「推理算力」市場。

從商湯分拆一年多 已完成多輪融資

香港文匯報訊（記者 黎梓田）隨着人工智能技術於 2024 年底步入高速發展期，全球對高效能運算（HPC）資源的需求呈現幾何級數增長，AI 芯片研發進程亦隨之進入高資本密集投入階段。曦望 Sunrise 聯席 CEO 王湛指出，AI 芯片作為智能時代的核心基礎設施，其應用場景理應廣泛覆蓋多個行業。基於此戰略考慮，曦望於 2024 年底正式從商湯分拆獨立運營。

曦望自獨立營運至今僅逾一年，已獲得極高的市場關注度。王湛透露，公司在過去一年多已完成多輪融資，股東結構呈現三類戰略資本並行的格局，充分體現市場對其技術前景與商業模式的高度認可。

引入行業龍頭及知名基金

王湛在訪談中分別介紹了三類投資方，第一類為產業戰略股東，提供業務場景與市場渠道協同。除商湯科技持續支持外，曦望亦成功引入多家行業龍頭企業，涵蓋香港上市的人工智能公司、A 股上市公司如三一重工、美的集團、三七互娛，以及國際知名跨國集團正大集團。王湛補充指出，該等產業股東普遍具備強烈的 AI 轉型與算力需求，以三一重工為例，正積極規劃打造智能化與無人化的大型機械作業

體系；美的集團則推進家電終端全面智能化升級。該等企業不僅參與產品定義，亦為曦望之重要直接客戶，形成產銷協同效應，有助加快產品落地及市場滲透。

第二類為知名投資基金，提供資本動能與生態聯動。參與本輪融資的機構包括中金資本、IDG 資本、高榕資本、心資本、普華資本及松禾資本等境內外頂尖財務投資者。該等機構旗下龐大的投資組合與產業網絡，預期將為曦望中長期發展提供豐富的業務拓展與跨域合作機會，進一步強化其於 AI 芯片生態體系中的戰略節點角色。

第三類則為浙江杭州地方資本，實現政策資源與戰略深度綁定。曦望今年正式將發展總部落戶杭州，並獲得浙江省、杭州市及濱江區三級政府的大力支持。王湛表示，地方國企與政府基金不僅參與股權投資，協助公司與杭州當地 AI 龍頭企業建立緊密戰略合作關係，進一步鞏固其在長三角人工智能產業鏈中的關鍵樞紐地位，並有助於獲取區域性產業集群的協同紅利。

整體而言，王湛期望曦望藉由多元化股東結構與地域政策優勢，加速推進 AI 芯片商業化落地，為下一階段規模化發展奠定堅實基礎。

憑三大因素 從行業競爭中突圍

香港文匯報訊（記者 黎梓田）在人工智能芯片行業經歷劇烈洗牌之際，曦望 Sunrise 如何站穩陣腳並脫穎而出，成為市場關注焦點。王湛在訪談中展現出高度自信，他認為將公司的核心競爭力歸納為三大內生優勢，構築其不可忽視的護城河。

他介紹指出，首先是「黃金創業團隊」的戰略協同效應。曦望董事長徐冰為商湯科技聯合創始人，亦為人工智能權威湯曉鵬教授之高足，曾深度參與商湯的融資布局、戰略規劃及業務管理，具備深厚的 AI 產業國際視野與資本運作經驗。徐冰自 2012 年起為香港中文大學多媒體實驗室博士候選人，專注深度學習與電腦視覺研究，並於 2017 年獲《麻省理工科技評論》評為 35 歲以下科技創新者，2019 年更入選福布斯亞洲 30 位 30 歲以下精英榜，學術與實戰背景兼備。與其搭檔的聯席 CEO 王勇，則擁有逾二十年芯片行業資歷，曾主導多款複雜芯片的架構設計與流片量產，熟悉半導體供應鏈及工程實現。王湛曾是百度的創始團隊成員之一，擁有超過 25 年的科技行業、產品戰略及市場開發經驗。三者互補，形成「視野引領方向，執行推動落地」的強勁組合。

軟硬一體化 實現協同優化

第二是「軟硬一體化」的 AI 算法基因，成為技術層面的核心壁壘。他表示，與傳統以硬件思維主導的芯片企業不同，曦望從創立之初便具備對大模型及 AI Agent（智能體）底層需求的深刻理解，實現算法與芯片的高度協同優化。此種協同並非僅止於硬件適配，而是在模型架構設計階段，即將算力分布、內存帶寬及延遲特性納入考量，確保神經

網絡每一層運算均能與底層電路「無縫貼合」，從而將核心計算環節利用率由行業平均的 60% 至 70% 顯著提升至 90% 以上。

他進一步指出關鍵點，是針對 AI Agent 推理過程中頻繁出現的動態分支、條件跳躍及上下文切換等複雜場景，曦望 Sunrise 在指令集層面預置專屬「Agent 加速原語」，有效將多輪對話中的首 Token 延遲降低約 40%，同時支援動態批處理與稀疏計算，確保芯片在處理長序列任務時仍能維持線性能效比。此外，該軟硬協同體系亦具備「自適應微調」能力，芯片內置輕量級監控單元，可實時採集運行數據，驅動上層調度策略動態調整內存分配與計算內核映射，無需重啟系統即可應對不同負載場景。基於此技術框架，曦望 Sunrise 不僅在標準 AI Benchmark 測試中表現突出，更在智能客服、代碼生成、多模態交互等真實 Agent 密集型應用場景中，展現出同級別芯片難以匹敵的響應流暢度與能效優勢，真正將「算法定義硬件」的理念轉化為可量產的競爭壁壘。

賽道定位精準 避與巨頭交鋒

最後是「精準的賽道定位」，他相信要體現其市場戰略的審慎與前瞻。曦望並未盲目投入高端訓練芯片市場，與國際巨頭正面交鋒，而是高度聚焦於 AI Agent 的「推理算力」領域。王湛強調，隨着全球 AI 應用生態加速爆發，市場對高性能、高能效比的推理算力需求正呈指數級增長。憑藉清晰的戰略眼光，和多代芯片迭代量產能力，曦望有信心在未來行業格局中，持續加大研發投入和技術創新，保持競爭優勢。