

業 務

我們的願景

讓AI重構世界。

我們的使命

構建智算超級系統，突破AI性能邊界。

概覽

我們是誰

我們是中國規模最大的獨立AI算力集群提供商，引領AI基礎設施領域的顛覆性創新。AI技術的快速演進推動了社會對高性能AI基礎設施日益增長的需求。設計及交付高性能AI基礎設施，其瓶頸不僅在於高性能計算和網絡芯片本身，更在於能否通過網絡系統、編排調度機制與並行計算技術，將成千上萬處理器變成一台高性能超級計算機。憑藉超過二十年的分佈式系統與計算機體系結構領域的專業經驗，以及我們在硬件級架構融合與系統級跨層優化方面的獨特設計理念，打造並提供全棧AI算力集群產品及運營服務，充分挖掘並釋放AI基礎設施潛能，支撐客戶獲取卓越的AI生產力。

我們以*Galaxy*產品品牌構建的端到端AI算力集群產品，由AI算力設備和AI算力軟件構成。AI算力設備包含計算、網絡及存儲等組件，其中我們自研的AI算力網絡系統*Mercury*實現了各類設備組件的超大帶寬互聯互通，支持同構、異構、多域等硬件資源的高效協同通信。我們將AI算力軟件中研發的諸多核心技術統一打造為AI算力操作系統*Venus*。*Venus*包括執行AI工作負載任務的運行時和公共組件接口，以及面向預訓練、後訓練、推理和智能體的高速服務引擎。同時，我們以*Galaxy Service*的品牌提供一系列的AI算力集群運營服務，主要包括集群運維與技術支持服務，以及算力運營服務，使更多客戶得以通過便捷和靈活的方式獲得高質量的算力。

據弗若斯特沙利文資料，我們是中國少數具備AI算力集群全棧產品及服務能力的廠商之一，涵蓋從軟硬件研發、架構設計、設備及系統部署，至性能優化、故障檢測、持續運營的整個AI基礎設施生命週期。我們持續服務重點客戶以滿足它們不斷演進的各類AI工作需求，包括大模型公司、雲服務提供商、教育與科研機構及電信運營商，真實地幫助客戶解決問題，提升AI算力集群利用率與穩定性，得到客戶的一致認可與極高客戶忠誠度。我們也逐步通過內置智能體應用的AI算力集群，將業務邊界從大模型產業拓展至千行百業。

對於客戶需求的精準洞察以及對於技術的不懈堅持讓我們取得了諸多成就，根據弗若斯特沙利文資料：

- 我們是中國首家成功參與部署RoCE千卡集群的獨立AI算力集群提供商，且擁有自研的AI算力網絡系統以及多個萬卡集群部署運維的成熟體系；

業 務

- 在國內AI算力集群提供商中，我們在集群擴展性和可靠性方面的表現領先，集群MFU提升30%且平均年化SLA達99.97%；
- 以2025年收入計，我們位列中國獨立AI算力集群提供商榜首；
- 截至2025年12月31日，我們在中國萬卡集群服務市場佔有率約10%，其他頭部廠商主要為大型互聯網企業及電信運營商。

以下為我們的業務亮點概述：

業務成果	4個 由我們服務的萬卡規模AI算力集群	超90,000張 由我們技術支持的AI算力集群所包含的GPU卡數 ¹	超34,000PFLOPS 具備SLA保障的在線運維算力規模 ¹
集群性能	30% MFU提升比例 ¹	99.97% 平均年化SLA ¹	超96% 萬卡規模集群算力利用線性度 ¹
經營數據	304% 收入複合增長率 2023-2025年	超80% 主要客戶復購率 ²	連續2年 扣除非經常性損益後盈利

附註：

- 1 截至最後實際可行日期。
- 2 於往績記錄期間及直至最後實際可行日期。

我們的市場機遇及價值主張

大規模AI工作負載通常涉及龐大的參數規模、海量的數據和密集的算力需求，因而需要超大規模的海量數據並行處理能力以及高度高效的參數同步功能。傳統以通用算力為中心的計算架構，主要依賴以中央處理器為基礎，集中式和串行式的處理方式，已愈來愈難以滿足上述要求。AI算力集群中通過AI算力軟件將AI工作負載任務拆分成眾多併發子任務，在成千上萬張GPU上做並行計算處理，用AI算力網絡完成各GPU上每次迭代計算結果的高效同步，通過這種高併發模式實現AI工作負載任務的加速處理。

業 務

我們捕捉到AI算力集群行業正呈現以下發展趨勢：

- **高效擴展與穩定性升級：**儘管AI算力集群規模迅速擴大，AI算力建設不再單純追求集群規模擴張，而是轉向規模拓展與資源高效利用並重。根據弗若斯特沙利文的資料，AI算力集群當前模型算力利用率（「MFU」）為35%至45%，任務中大部分並行算力未獲充分利用，算力利用率提升成為核心關注點。同時，隨著工作負載對算力需求快速增長，需要並行協同處理的計算節點也越來越多，軟硬件故障易造成算力浪費與任務中斷，因而集群優化重心從單節點性能優化，轉向分佈式集群穩定性升級。
- **開放生態與全棧技術：**過往行業長期受制於單一廠商專有技術棧，架構兼容性弱、互通性不足，推高部署與升級成本，也制約算力性能釋放。因此，AI基礎設施公司正加速搭建高互操作性架構，封閉獨立組件逐步邁向開放互通，整合多元算力資源，銜接現有基礎設施，實現彈性高效擴容。

同時，隨著算力集群規模擴大、架構複雜度提升，僅依靠單一硬件或單項服務已無法滿足需求，行業亟需整合計算、網絡、調度等核心能力，打造覆蓋集群全生命周期的全棧解決方案，全面適配超大模型訓練推理等高端複雜場景。AI算力集群交付及服務模式從碎片化點對點解決方案，轉向軟硬件一體化賦能。

- **AI應用多元化發展：**AI產業快速迭代演進，業務鏈路已從基礎預訓練，逐步延伸至後訓練、推理、智能體等多元化場景，單一同質化的算力架構難以適配複雜需求。因此，行業急需建立可以支持和應對不同類型AI訓練、推理和應用場景的AI算力集群架構，精準匹配多場景資源部署、網絡傳輸與任務調度要求。
- **快速的技術演進：**當前算力架構逐步凸顯帶寬牆、功耗牆、時延牆及運維瓶頸，嚴重制約多節點協同效率與算力橫向擴展能力。伴隨大模型向萬億級參數規模快速演進，AI算力集群在多維度快速迭代，規模上集群向萬卡甚至十萬卡級別升級，技術上算力芯片、光互連和存儲技術不斷突破。因此，立足下一代算力網絡發展趨勢，依托光電混合互連、異構算力融合、前沿計算架構等多元技術路線，行業需要持續優化集群架構與資源調度模式以匹配最新技術。

業 務

我們的價值主張基於以下核心能力：

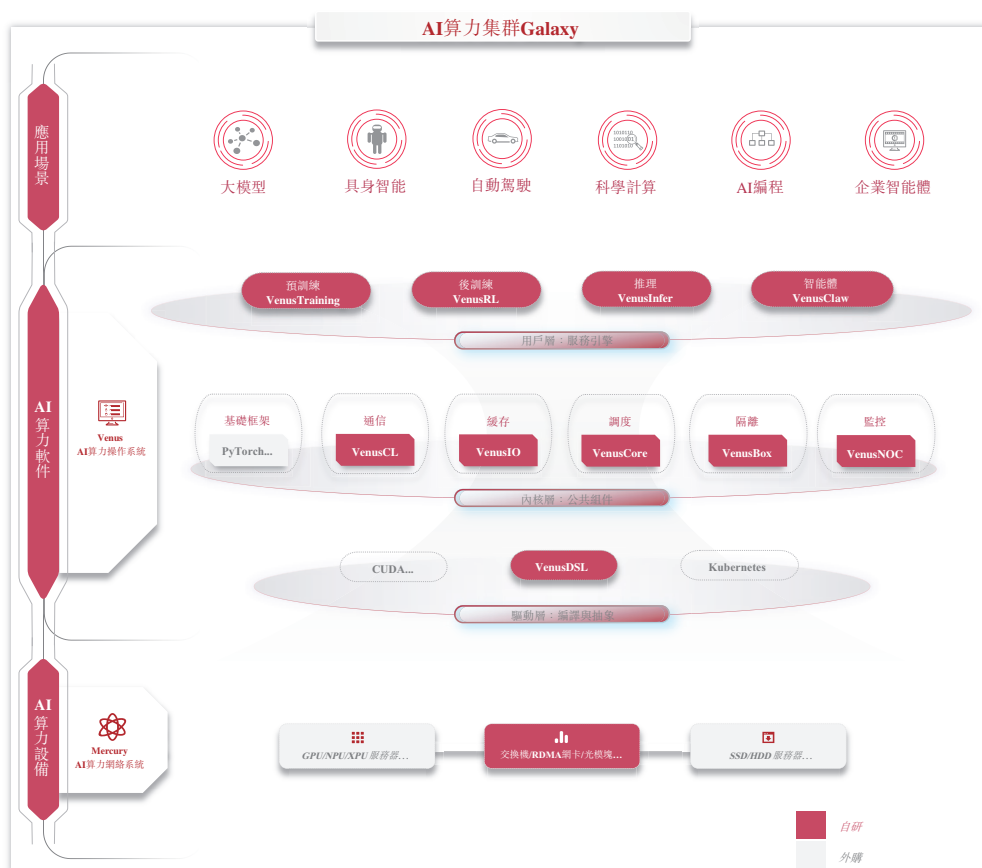
- **高性能AI算力集群產品和服務：**為響應上述不斷演變的需求，我們設計部署高性能AI算力集群產品，並提供運營服務，支撐客戶實現高性能AI計算。通過改善不同計算資源在AI算力集群內的互連、運行及編排方式，我們協助客戶實現更大並行度計算、更高資源利用率，以釋放更強大的集群性能。我們通過快速故障定位與層次化容錯恢復等方式，強化分佈式集群運行穩定性，同時搭建全指標實時監測與智能預警體系，大幅降低網絡抖動與硬件故障概率，保障萬卡級集群高穩定運行。
- **多元生態和全棧技術能力：**我們採用分層解耦的開放架構，通過基於客戶部署需求的模塊化擴展機制，實現對不同廠商硬件、主流AI框架的全面兼容，實現多廠商、多規格異構硬件資源的統一納管與協同運行。我們積極參與構建AI算力集群自主可控產品生態，提供對標國際領先解決方案的性能，以抓取國產AI算力芯片崛起的機遇。我們的全棧能力覆蓋垂直整合優化體系，針對大規模複雜AI工作負載進行協同優化，為客戶提供一站式全方位解決方案，實現一鍵啟動、開箱即用的良好客戶體驗。
- **AI算力操作系統支撐多元化AI工作負載：**我們的AI算力操作系統*Venus*支持涵蓋預訓練、後訓練、推理及智能體場景的多種AI工作負載。針對預訓練場景，提供拓撲感知智能調度、高速統一集合通信庫及多層次容錯機制，保障超大模型訓練的高效穩定運行；針對後訓練場景，通過任務分解與細粒度調度，聚合多個後訓練任務的不同階段，有效提升資源利用率；針對推理場景，支持基於流量負載的彈性擴縮容，以及計算網絡存儲融合的硬件加速機制，充分提升推理帶寬；針對智能體場景，在細粒度隔離組件中開發了智能體沙箱技術，滿足任務理解、規劃、工具調用及迭代驗證等流程對高效受控執行環境的需求。
- **多維度技術創新：**我們緊跟行業發展趨勢，致力於讓前沿的技術應用於我們的產品和服務中。我們的AI算力網絡系統*Mercury*，完成基於OCS光電融合交換驗證，落地NPO方案破解超高速傳輸瓶頸，兼顧高性能與易維護性。我們採用異構計算與前沿計算架構，整合不同芯片的算力優勢，實現計算資源的精準匹配與高效利用，大幅提升複雜AI任務的計算效率，破解單一芯片算力瓶頸。

業 務

我們的產品及服務

AI算力集群產品 – Galaxy

我們通過AI算力集群產品，將成千上萬張GPU變成一台超級計算機，以強大的可擴展性、可用性、可觀測性和安全性，提供統一的AI算力資源。我們的AI算力集群產品部署在高功率數據中心內，這些數據中心包含眾多機架，這種整體架構設計使AI算力集群能夠作為一體化AI基礎設施運行。我們的AI算力集群產品覆蓋AI算力設備和AI算力軟件，支持大模型、具身智能、科學計算等各類AI工作負載類型。下圖說明我們的AI算力集群產品Galaxy的架構及組件，包括內部開發及外部採購的組件以及其支持的AI應用：



- **AI算力設備：**我們在AI算力設備中自研的AI算力網絡系統Mercury包括超高帶寬的交換機、RDMA網卡和光模塊等硬件組件，以及負責調度與運維各硬件組件的開源網絡操作系統MercuryNOS。MercuryNOS旨在提升高併發負載下的網絡帶寬利用率、流量分發效率與傳輸穩定性，同時提供全量細粒度的監控可視性，以支撐AI任務的性能調優與故障診斷。同時，我們的算力集群產品和解決方案能力也覆蓋了集群設備中的算力和存儲單元，憑藉我們對於AI算力集群的深刻理解，以及與合作夥伴包括算力硬件廠商、存儲廠商、OEM廠商的緊密合作，我們具備構建、建設和優化計算以及存儲單元的能力。

業 務

- **AI算力操作系統：***Venus*作為AI算力集群的大腦，從下而上由驅動層，內核層、客戶層構成：驅動層為內核層提供統一的分佈式設備調用接口。在內核層中，除廠商自帶並行編程庫和開源雲原生資源管理平台外，*VenusDSL*可進一步提升系統遷移性和芯片適配性；內核層為客戶層提供AI工作負載任務處理所需的核心運行時組件，包括基於開源AI算力基礎框架、高速集合通信庫*VenusCL*、緩存輸入輸出*VenusIO*、智能任務調度*VenusCore*、資源與任務隔離*VenusBox*、多維多尺度監控*VenusNOC*；客戶層為各類AI工作負載提供優化的服務引擎，包括預訓練服務引擎*VenusTraining*、後訓練服務引擎*VenusRL*、推理服務引擎*VenusInfer*、智能體服務引擎*VenusClaw*。

為應對AI基礎設施的持續性能挑戰，我們的AI算力集群產品具備硬件級架構融合與系統級跨層優化，滿足AI算力集群效能提升和帶寬牆、能耗牆突破的需求。不同於其他廠商進行單點局部優化的思路，我們的AI算力集群產品破除單一組件優化的性能上限，針對大規模AI訓練推理等高端複雜場景進行算力利用率優化，可以在SOTA之上進一步提升MFU。以下為我們AI算力集群產品中核心自研組件的技術亮點：

- **AI算力設備：***Mercury*充分使用光電協同的設計思路，規模驗證OCS光電融合交換技術，落地3.2T近封裝光學技術，在破解超高速傳輸瓶頸的同時，兼顧了系統的高性能與易維護性，有效突破了傳統電互連的帶寬與功耗限制。進一步，在AI計算設備層面，針對推理過程中預填充和解碼階段的特性區別，以及解碼過程中注意力和專家計算階段的特性區別，我們採用優化且適配的硬件架構，實現軟硬協同，異構加速。
- **AI算力操作系統：***Venus*深度使用算網存協同和軟硬協同的機制，在多維度實現跨層優化。*VenusDSL*實現算子融合，提升計算效率；*VenusCL*實現計算和通信的無核調度與深度交疊，提升算力利用率，優化預訓練效率；*Venus RL*與*VenusCore*實現軌跡採樣與計算的部分交疊，提升算力利用率，優化後訓練效率；*VenusInfer*與*VenusIO*構建顯存－網絡直驅－閃存的分級緩存體系，擴充緩存容量，顯著降低長上下文推理成本；*VenusClaw*與*VenusBox*在保證安全隔離性下，通過動態優化沙箱起停，降低沙箱空轉時間，優化成本；*VenusNOC*通過主被動融合的監控定位方式，實現在線任務性能無感實時分析，補齊複雜問題追蹤能力。詳情請參閱「－我們的產品及服務－AI算力集群產品－AI算力軟件－Venus」。

憑藉我們的全棧AI算力集群產品及服務能力，我們能夠於整個集群生命週期為客戶提供服務，涵蓋架構設計、硬件選型、軟硬件適配、網絡互聯、系統調優及性能驗證。視乎客戶需求，

業 務

我們的交付範圍可涵蓋完整集群部署、特定軟硬件組件的針對性部署、集群級網絡及融合服務，或基於客戶現有硬件基礎設施的軟件與性能調優。

AI算力集群運營服務*Galaxy Service*

在提供AI算力集群產品的基礎上，我們推出*Galaxy Service* AI算力集群運營服務。*Galaxy Service*深度依托我們自研的AI算力網絡系統*Mercury*與AI算力操作系統*Venus*的底層能力，幫助客戶打破傳統AI算力集群「重建設、輕運營」的痛點，解決算力資源分散、運維門檻高、內部協作效率低、大額資本開支成本核算難等問題，讓客戶無需投入大量專屬運維人力，同時完整保障數據與業務安全。

*Galaxy Service*具備以下關鍵特性，可滿足客戶提升業務靈活性和降低運營開支的需求：

- **全面支持及SLA保障：***Galaxy Service*依托*Galaxy*原生全棧可觀測與彈性架構，提供覆蓋硬件、系統、任務的全面運維與業務適配服務，幫助客戶降低日常集群運維門檻。我們亦提供系統化的維護支援，以確保硬件的可用性。透過*VenusNOC*，我們能夠精準識別並診斷AI算力集群中的隱藏及複雜問題。
- **高性價比的優質算力：***Galaxy Service*通過智能負載調度、精細成本管理和投資保護機制，為客戶提供高效且經濟的優質算力。系統可動態分配資源、提升利用率，並支持精細化計費、成本核算與優化建議，幫助企業降低資本支出與營運支出。同時支持存量資產複用和集群平滑擴展，兼顧性能、成本與長期價值。

我們的業績

我們的業績於往績記錄期間錄得大幅增長。我們的收入由2023年的人民幣31.8百萬元增至2025年的人民幣520.3百萬元，年複合增長率為304.5%。我們於2023年、2024年及2025年的毛利分別為人民幣2.4百萬元、人民幣65.4百萬元及人民幣113.2百萬元。我們於2024年開始盈利，2024年及2025年的經調整淨利潤（非國際財務報告準則計量）分別為人民幣20.3百萬元及人民幣31.1百萬元。

競爭優勢

中國領先的高性能AI算力專家，持續引領行業創新

我們是中國頂尖的AI基礎設施團隊，匯聚了諸多AI算力集群專家，核心研發團隊在分佈式系統與計算機體系結構領域持續深耕超20年，團隊成員已經在USENIX網絡系統設計與實現研討會(USENIX Symposium on Networked Systems Design and Implementation)、ACM數據通訊專委會(ACM Special Interest Group on Data Communication)、ACM編程語言及操作系統架構支持國際會議(ACM International Conference on Architectural Support for Programming Languages and Operating Systems)、美國光纖通訊展覽會及研討會(Optical Fiber Communication Conference)等

業 務

國際頂級會議期刊發表論文超過90篇，具備國際學術視野與落地經驗，持續研發AI算力技術，引領AI基礎設施領域的創新發展。我們具備業界領先的可擴展性和可靠性：我們將萬卡級集群的稀疏大模型訓練的算力利用率提升了30%，萬卡級集群的算力利用率線性度超96%，我們運營的集群具備業界領先的穩定性，年化硬件故障率低於0.3%，平均年化SLA達99.97%。

我們構建了AI算力優化方向獨特的硬件級架構融合與系統級跨層優化理念。我們的研發團隊由對AI及算力系統技術始終保持專注的工程師和科學家組成。我們充分發揮團隊的經驗積累，推動算網存協同、軟硬協同、光電協同、異構加速等設計機制落地。儘管不同的AI應用各有不同特徵，支持該等工作負載的底層基礎設施通常可分為預訓練、後訓練、推理及基於智能體的應用。截至目前，我們就該等負載類型的AI算力設備及軟件方面已取得一系列研發成果：

- **網絡基礎設施：**通過光電協同，構建了光電混合的AI算力網絡，集合智能流量調度機制，實現了集合通信時間下降50.3%、帶寬利用率上升47.1%、大模型訓練能耗降低32.1%的效果；
- **預訓練基礎設施：**通過計算和通信的跨層優化，實現GPU無核通信，將通信任務卸載至CPU，並在流水線工作流中實現計算通信深度交疊，根據弗若斯特沙利文的資料，這在千卡集群訓練算力利用率SOTA性能之上提升2%-6%，按萬卡集群規模平均資本投入計算節省資本開支超億元；
- **後訓練基礎設施：**通過計算和輸入輸出的跨層優化，實現強化學習軌跡採樣與計算的部分交疊；與業內普遍使用的常規後訓練實施方法相比，我們的VenusRL架構將整體任務效率提升37%。通過計算和安全的跨層優化，實現智能體安全沙箱的快照共享，提升沙箱按需起停的效率，降低沙箱的空間運轉，將智能體沙箱運行成本降低超50%；
- **推理基礎設施：**通過計算、網絡和存儲的跨層優化以及異構硬件加速，構建GPU通過網絡直驅訪問閃存的分層緩存體系，提升10-30倍緩存容量，顯著降低長上下文推理的規模化成本，構建注意力與專家層計算的解耦，實現推理資源的細粒度動態伸縮，通過通信卸載將延遲降低49.6%；

我們已構建一套完善的產品創新與前沿技術規劃深度融合的研發體系，依托深度產學研協同，我們攜手國內多家頂尖科研機構、高校及雲服務商，落地多項前沿AI基礎設施研究項目，有效縮短了從技術研發到產品落地的周期。相關產研協同的顯著成效如下：

- 我們僅耗時8個月完成基於國產技術的自研交換機全流程開發與落地，顯著優於行業18個月的平均周期；聯合高校專家優化RDMA擁塞控制算法，將研發周期由8個月壓縮至4個月以內；

業 務

- 我們的智能運維系統 *VenusNOC*，最初源於學術研究場景下的概念原型，並已通過我們發表的研究論文完成進一步技術驗證。在此基礎上，我們已將該概念轉化為部署於生產環境中的運營系統。*VenusNOC*可精準識別並診斷萬卡級超大規模算力集群中的隱性複雜問題。其已獲多家頭部大模型公司採用，協助定位分析眾多線上問題，實現在線實時的功能級別監控，MFU下降少於0.1%。

AI基礎設施能力領域的先行者，不斷樹立行業標桿

作為國內算力集群行業的先行者，我們率先洞察市場機遇，依托先發優勢搶佔市場、沉澱用戶認可，持續積累一線行業洞察與實戰經驗，構築長期壁壘。我們深度參與行業標準制定、主導關鍵技術規範與項目技術門檻設定，構築深厚的護城河。我們已錄得一系列具有標誌性意義的相關往績，有力推動了中國AI基礎設施的發展，根據弗若斯特沙利文：

- 開創性的集群水平擴展功能：**我們是中國第一家完成不同規模AI算力集群標桿項目的獨立AI算力集群提供商，2023年落地首批RoCE千卡集群、2024年落地首批四千卡與萬卡集群。我們已建立超大規模部署與資源優化的核心能力，方案靈活度與落地經驗領跑行業。
- 開創性的集群跨域擴展功能：**我們是中國第一家部署RoCE架構長距離分佈式AI算力集群的獨立AI算力集群提供商，於2024年3月組網能力持續升級至30公里、50公里級，等效算力超過98%，於2025年7月完成**1500公里**跨地域異構分佈式訓練實測，兼具系統可靠擴展性與全棧國產適配能力。
- 開創AI網路自主性與突破性進展：**我們是中國首家自研量產基於國產25.6T交換芯片的全國產高通量以太網交換機的獨立AI算力集群提供商，並完成了AI算力集群的光電融合互連，在調試與部署驗證方面擁有深厚的經驗。此外，我們還基於國產NPO技術自主研發了100T超高速交換機，其整體交換容量可媲美國際領先的旗艦級設備。2025年，我們率先建成了配備國產CPU、GPU、交換機及RDMA網卡的全自主研發AI算力集群，並通過大規模部署，驗證了其商業可行性。
- 開創企業級AI集群部署：**我們是最早落地大規模企業級商用融合AI算力集群的企業，打造標桿智能體，構建全國產化AI基礎設施底座，為企業智能化轉型提供了核心驅動力。

業 務

構築業務擴張飛輪，深度助力行業標桿客戶

高性能AI算力集群部署運營技術門檻高且經驗累積漫長。憑藉先發優勢，我們服務的AI算力集群規模持續擴大。截至最後實際可行日期，我們累計交付及服務的AI算力集群由合共超過90,000張GPU組成，且透過具備SLA保障的AI算力集群營運服務提供超過34,000 PFLOPS的算力能力。我們已開發出一套成熟且高度可擴展的運營模式，持續為客戶創造更大價值：

- **成熟的落地經驗：**我們充分掌握AI算力軟硬件特性並規範標準交付流程，使萬卡集群可在硬件部署後兩周內完成軟件調優並投入生產，大幅縮短算力落地周期，有效支撐客戶業務快速落地，提升客戶資產利用效率與投資回報水平。
- **豐富的運營數據及智能體：**依托海量一線運維數據和完備故障數據庫，結合後訓練強化學習能力，我們構建的AI算力集群運維智能體，可提前識別潛在運行風險、精準定位故障根源、完成系統化深度分析、執行基礎自修復功能，全方位提升算力集群穩定性、可靠性與運營服務效率。
- **獨特的產業鏈價值：**伴隨運營持續擴張，公司在光模塊、白盒交換機、高速互聯線纜等核心網絡硬件領域形成規模化採購優勢，議價能力顯著提升，並充分讓利客戶，打造高性價比AI算力集群。

我們聚焦於對AI具有嚴格要求的標桿客戶，涵蓋大模型廠商、雲服務提供商、教科研單位以及電信運營商。該等客戶為中國AI基礎設施投資及行業增長的主要推動力。依託長期服務該等客戶積累的深厚的技術專業知識，我們能夠迅速適配多元應用場景，滿足更廣泛客戶持續擴張的算力需求。隨著行業集中度提升，我們與標桿客戶的長期戰略合作夥伴關係亦為我們帶來穩定及持續的增長機會。於往績記錄期間及直至最後實際可行日期，我們主要客戶的複購率超過80%。我們通過持續深化標桿客戶合作，穩步構建業務正向循環，全面驅動業務擴張飛輪高效運轉。

全棧AI算力集群，一站式高性能算力服務

當前AI算力集群建設面臨較高門檻，主要體現在技術壁壘高、涉及技術維度多、需要協同的合作方多、整體部署周期長等方面。一個高性能、可穩定運行的AI算力集群，不僅涉及硬件部署，還包括網絡互聯、智能調度、負載適配、運維保障等多個關鍵環節，同時需要硬件廠商、軟件供應商、部署服務商等多方協同配合。由於AI算力集群各環節之間高度耦合，任何一處銜接不

業 務

暢，都可能影響集群的性能釋放、交付效率和長期穩定性。因此，客戶在實際建設過程中，往往面臨系統融合複雜、上線周期長、後續運維難、整體性能難以充分發揮等問題，迫切需要一套覆蓋全鏈路、能夠快速落地並持續優化的一站式AI算力集群方案。

根據弗若斯特沙利文，我們是極少數擁有全棧AI算力集群產品及部署服務能力的獨立AI算力集群提供商。我們真正實現了一站式滿足客戶從軟硬件開發、架構設計及系統部署到性能調優、故障檢測及持續營運的全面需求。我們通過協同優化，充分釋放系統的性能潛力，真正破解了「效率、性能、成本」的三重難題。

不同於在初始交付後僅提供有限支援的傳統算力服務模式，我們將AI算力集群運營服務作為全棧能力的重要組成部分，保障客戶業務長期高效運行。依托全棧技術的統一數據感知能力，我們為客戶提供不斷迭代的和精細化算力集群運營服務，實現軟硬件問題自動定位與快速排查，配備7×24小時專屬運維支持。同時，我們AI算力集群服務持續推送算子優化、並行策略和加速技術，保障集群性能持續處於最佳，並通過成本分析與調度優化建議，幫助客戶降低整體資本開支及營運開支。

開放生態系統的踐行者，全面賦能AI產業鏈

當前AI產業價值鏈正面臨上下游割裂、生態封閉的行業痛點，上游硬件創新無法快速匹配前沿模型的負載需求，下游大模型也難以充分發揮多元硬件資源的性能潛力，制約了整體創新效率。我們以「產業橋樑」為核心定位，聯動上游硬件供應商與下游大模型開發商，將前沿AI工作負載需求與基礎設施創新緊密銜接，既鞏固自身全棧核心優勢，也推動整個AI生態走向開放協同。

針對上游硬件，我們產品和服務已完成主流計算設備廠商、RDMA網卡廠商、光模塊廠商以及交換機廠商的產品適配驗證和部署落地，客戶可自由選擇不同廠商的硬件組件，靈活構建符合自身需求的AI算力集群，無需被單一廠商的特定硬件架構綁定。我們充分挖掘國產硬件的性能潛力，幫助國產硬件集群達到國際一流的性能水準。

針對下游模型，我們緊跟大模型的快速迭代節奏，依托開放兼容的適配層，實現了不同架構模型的快速部署與高效運行。無論是基於CUDA生態的主流模型，還是基於國產生態自研的開源模型，都能無需修改核心代碼，即可在Venus平台上高效運行。Venus為主流開源模型提供即時推理適配。我們與國內頭部開源模型廠商緊密合作，助力大模型快速研發與迭代。

我們是AI基礎設施層的中堅力量。不同於傳統方案只做簡單的適配對接，我們精準洞察AI開發與集群部署中的瓶頸與需求，通過算力集群產品Galaxy賦能下游模型和上游硬件，真實構建

業 務

開放生態平台，實現AI生產力在商業世界的躍升。上游硬件廠商得以專注於芯片創新，下游模型廠商得以專注於算法創新，而我們負責打通各個環節的適配與優化，讓整個產業鏈的能力都能充分釋放，共同構建繁榮、開放的AI算力生態，推動整個產業的加速升級。

增長策略

持續迭代AI算力集群產品，鞏固市場領先地位

隨著AI智能體爆發式增長，市場焦點已從模型創新轉向推理服務，AI基礎設施亦從依托GPU搭建集群的科研裝置形態，逐步演進為工業化AI工廠。我們將圍繞AI基礎設施發展趨勢，持續鞏固並提升自身作為世界一流獨立AI算力集群提供商的行業影響力。

我們的*Galaxy*系列AI算力集群產品將以異構融合的趨勢，在未來持續迭代演進。下一代*Galaxy*集群架構將升級為光電協同互連支持的超節點，通過超節點聚合更大顯存的算力、光技術提升互連效率；未來我們的AI算力集群產品將升級為分層異構加速方案，引入更多的領域專用處理器，推動集群高效化、專業化升級，更好適配大規模落地級AI算力需求。

其中主要應用的核心技術和業務佈局包括：

- **硅光及光交換：**我們將圍繞互連技術持續推進戰略佈局，依托硅光引擎PIC、EIC等光電技術，打造適配GPU高帶寬scale-up的互連方案，攻克224G及下一代高速率SerDes信號完整性難題；同時佈局硅光大容量、超低延遲OCS光交換技術及光電融合高性能網絡調度優化技術，並結合NPO、OCS等先進光互連能力，構建超高帶寬、超低時延的高性能網絡，持續強化在AI算力網絡領域的核心優勢。
- **硬件加速：**我們將深度落地分層異構加速架構，完成集群全棧架構深度優化。借鑑大模型專屬加速架構設計思路，我們計劃與生態系統合作夥伴共同開發LPU，獨立承接大模型解碼推理、上下文處理、序列調度等專用輕量化任務，實現計算鏈路拆解與業務分層解耦，形成訓練、推理差異化算力分工；針對xPU異構協同體系，依據不同作業負載特徵動態分配算力單元，按需組合各類加速芯片能力，徹底打破單一硬件的性能瓶頸，全面提升集群整體吞吐、併發處理能力與資源利用效率。
- **算力軟件：**我們結合超節點中高密度的算力節點、scale-up/scale-out的海量互連高帶寬鏈路特性迭代研發新一代AI算力操作系統*Venus*，從集合通信庫、內存統一管理、

業 務

訓推任務彈性調度與故障切換等多個方面進行架構優化、功能完善，高效發揮超節點及超節點集群性能，保障集群的穩定性。

伴隨著上述技術能力的持續迭代與升級，我們亦將根據客戶的需求，進一步擴大自主設計、部署及自持的AI算力集群規模，持續提升高質量算力服務供給能力，更好地滿足客戶多元化的算力需求。

戰略性垂直整合及價值鏈擴張

在AI基礎設施領域，我們計劃圍繞AI產業鏈的關鍵環節，尤其是核心硬件及前沿技術系統設計領域，進行針對性地垂直整合。我們將通過戰略性投資與收購，強化對關鍵技術的自主掌控力，提升從核心設計、可擴展生產到項目落地的全流程協同能力。這些戰略舉措，預期將有效提升產品交付效率，鞏固全鏈路質量管控水平，並為長期成本優化提供有力支撐。

通過此類戰略性收購或投資，我們可快速獲取如光互連等前沿技術，液冷核心工藝與成熟產能，無需經歷漫長的技術研發周期與客戶認證流程，大幅縮短技術落地與產品量產的時間成本。同時，這也將進一步提升我們從系統設計到規模化生產的協同效率，持續完善AI基礎設施硬件能力，夯實自身在行業中的核心競爭力。

橫向拓展層面，我們擬通過對專注於智能體開發的企業進行定向收購或投資，穩步延伸至生產場景。我們將依托自身強大的算力基礎設施能力，推動產業整合與應用層創新深度融合，助力更多適配特定場景的企業級一體化AI算力集群產品及解決方案規模化部署，構建「算力+應用」的融合發展生態。

深化與核心客戶合作並擴大市場滲透率

憑藉與AI產業頭部客戶的持續合作，為頭部大模型廠商提供深度優化的AI工廠，應對持續指數增長的詞元流量。我們將深化與客戶合作，通過敏捷產品迭代、技術升級及長期穩定協作，進一步加深與核心客戶的互動黏性。同時，我們將更深入地參與更廣泛的標桿項目，例如下一代模型開發、超大規模算力部署等，積累具備更高戰略價值的實踐經驗，進一步強化與核心客戶的合作深度與廣度。

此外，我們還計劃擴大在場景客戶中的業務佈局，如科學計算、具身智能、自動駕駛、企業服務等創新場景。我們標準化及模塊化的產品架構、與主流計算網絡芯片兼容的開放式集群架構，以及精細化運營服務能力，均與這些客戶的需求高度契合。我們將依托為客戶提供本地化部署的AI算力集群產品與服務，精準匹配客戶需求，進一步推動與場景客戶的業務增長，拓寬市場覆蓋範圍。

業 務

拓展海外業務版圖及提升商業化能力

相較於國內AI基礎設施市場的快速發展，海外市場目前仍存在明顯的空白——獨立AI算力集群提供商相對稀缺，市場需求多依賴大型國際供應商。依托我們在國內長期積累的成熟運營經驗、全棧技術能力，以及標準化、模塊化的產品優勢，我們同步推進技術出海、產品服務出海、供應鏈出海，構建完備的全球化交付體系，全面支撐AI算力集群跨區域落地部署，穩步擴張海外業務版圖，持續強化全球化商業化運營能力。

目前，我們計劃開展東南亞地區的業務拓展工作，重點聚焦新加坡、馬來西亞及周邊鄰近地區，精準切入區域內AI算力需求市場。針對該區域內由中國合作夥伴主導的AI算力集群項目，我們將提供全方位的支持，涵蓋軟硬件部署、性能調優、資源管理及遠程運營等全流程服務，以專業能力夯實海外市場基礎、積累國際項目經驗。

展望未來，我們計劃進一步向歐洲及其他高增長海外地區拓展，持續擴大國際業務覆蓋範圍。同時，我們將依托為客戶提供本地化部署的AI算力集群產品，精準匹配海外不同區域、不同類型客戶的核心需求，進一步推動海外業務增長，助力全球商業化能力的持續提升，打造全球化的AI基礎設施服務體系。

我們的產品及服務

我們以*Galaxy*品牌提供AI算力集群產品，以滿足客戶在各類AI任務中對高性能計算不斷增長的需求。憑藉我們在驅動AI算力集群方面的豐富專業知識及實踐經驗，我們亦以*Galaxy Services*品牌提供一套AI算力集群運營服務，使更廣泛的客戶能夠獲得我們的技術專長及高質量的計算服務。我們的服務涵蓋AI基礎設施的完整生命週期，從軟件與硬件開發、架構設計、設備及系統部署，到性能調優、故障偵測與持續運營。下表載列我們於所示期間按業務線劃分的收益：

	2023年2月20日至 2023年12月31日期間		截至12月31日止年度			
			2024年		2025年	
	人民幣元	%	人民幣元	%	人民幣元	%
(人民幣千元，百分比除外)						
AI算力集群產品.....	31,802	100.0	307,717	94.8	436,624	83.9
AI算力集群運營服務.....	—	—	16,955	5.2	83,637	16.1
合計.....	<u>31,802</u>	<u>100.0</u>	<u>324,672</u>	<u>100.0</u>	<u>520,261</u>	<u>100.0</u>

作為一家具備全棧AI算力集群產品及服務能力的企業，我們能夠在AI算力集群的完整生命週期中，提供端到端的產品與服務交付。根據客戶需求，此類端對端交付可採取全面集群部署形式，或僅涉及選定硬件及軟件的更具針對性的部署，連同集群級網絡搭建及融合服務。此外，在

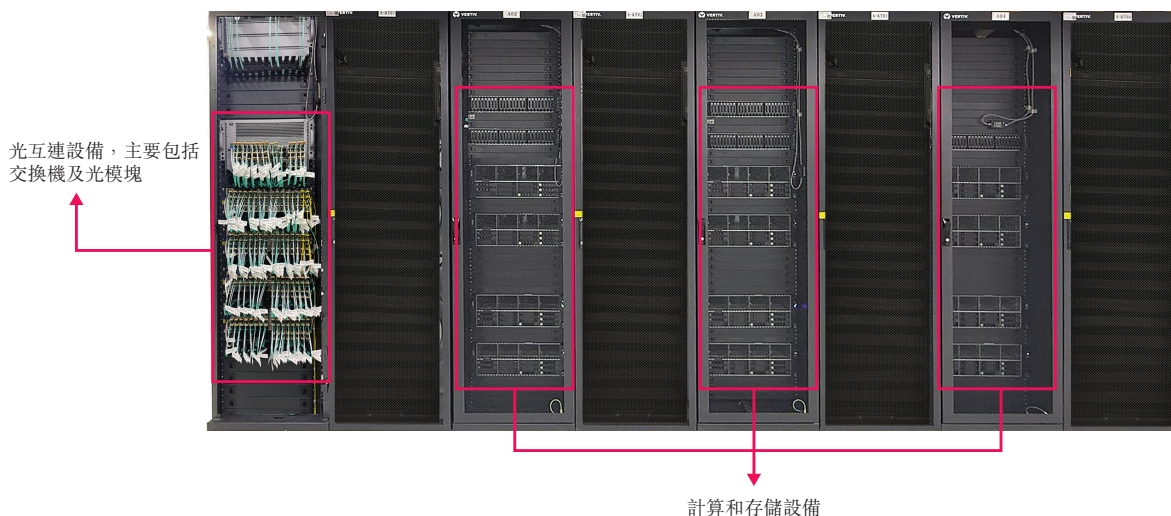
業 務

若干情況下，我們可能會基於客戶現有的硬件基礎架構，提供軟件及性能調優服務。不論採用何種交付模式，我們的目標及責任是透過我們的產品、技術專長及實踐經驗，提供更佳的AI算力集群效能。*Galaxy Services*協助客戶解決諸如算力資源分散、營運與維護需求高昂，以及與供應商合作效率低下等挑戰。*Galaxy Services*還能減少客戶維持龐大專職營運團隊的需求，同時實現精細化管理與成本優化。

截至2025年12月31日，我們已完成66個AI算力集群產品業務的項目，其中包括兩個超大規模AI算力集群，各集群規模的GPU數量均達到10,000張GPU級別。截至最後實際可行日期，我們為累計超過34,000 PFLOPs算力規模的集群持續提供具備SLA保證的運維服務，其中包括兩個萬卡級的AI算力集群。

我們的AI算力集群產品

AI算力集群是一種專為支持大規模AI工作負載而設計的融合AI基礎設施。它利用數以千計的GPU構建分佈式系統，將AI算力作為一種生產能力來交付。通過系統層面的工程設計，AI算力集群可實現計算、網絡、存儲和軟件等元素之間的深度協同，使用戶便捷地獲取高性能AI算力資源。一個典型的AI算力集群由AI算力設備及AI算力軟件組成。AI算力設備包括計算、網絡及存儲設備，而AI算力軟件則將底層硬件資源進行抽象化處理並向用戶提供訓練及推理服務。下圖展示由我們交付的AI算力集群：



鑒於AI基礎設施的複雜性，我們利用在分佈式系統與計算機體系結構方面的深厚專業知識，結合與上下游合作夥伴的緊密合作，提供了以*Galaxy*為品牌的AI算力集群產品。我們將該等產品與我們的實踐專長整合，作為AI算力集群的一部分提供給客戶，協助客戶有效提升集群的性能、效率、可擴展性、可用性及可觀測性。我們的產品及解決方案涵蓋AI算力設備及AI算力軟件，使我們具備全棧能力，得以服務AI算力集群的整個生命週期。

業 務

- **AI算力設備：**一個AI算力集群內的AI算力設備包括：計算、網絡及存儲設備。值得注意的是，對於AI算力集群而言，最關鍵的能力之一在於將其底層硬件組件（包括計算、網絡及存儲設備）有效連接成一個統一的系統。若無此等互連及協調，集群將無法充分發揮其性能潛力。因此，我們在網絡架構及系統互連方面開發了強大的能力，使該等硬件組件能夠在整個集群中高效通信並以協調的方式運作。這對釋放AI算力集群的整體性能、效率及可擴展性至關重要。就我們的AI算力集群產品內的AI算力設備而言，我們自主開發了網絡設備－即我們的AI算力網絡系統*Mercury*。憑藉我們對AI算力集群的深入了解，以及與包括計算硬件供應商、儲存設備供應商及網絡設備代工製造商在內的業務夥伴緊密合作，我們具備設計、部署及優化AI算力集群算力與儲存設備的能力。
- **AI算力軟件：**就我們的AI算力集群產品內的AI算力軟件而言，我們開發了多項核心技術並將其整合至我們的AI算力操作系統*Venus*中。作為我們AI算力集群產品的大腦，*Venus*由驅動層、內核層及用戶層組成。驅動層為內核層提供用於訪問分佈式設備的統一接口。內核層為用戶層提供處理AI工作負載所需的核​​心運行時組件。用戶層則為各類AI工作負載（包括預訓練、後訓練、推理及智能體工作負載）提供優化的服務引擎。

我們的AI算力集群產品建基於模塊化模式，結合了多種現有的硬件選擇及採購渠道、自主開發的軟件產品以及一套針對主流應用場景及客戶需求而設的成熟的設計與部署方案庫。此模塊化模式讓我們能夠評估每位客戶的算力規模、應用場景及性能要求，釐定整體集群配置，並挑選合適且具成本效益的AI算力設備及軟件組件組合。

AI算力設備

AI算力設備包括計算、網絡及存儲設備，該等設備被整合為一個統一系統，共同構成執行大規模AI工作負載的底層硬件基礎設施：

- **計算設備**通常包括GPU服務器，提供大規模AI工作負載所需的核​​心處理能力。GPU的效能決定了集群效能的上限。
- **網絡設備**通常包括交換機、光模塊及RDMA網卡以及網絡操作系統。它們作為高帶寬、低延遲的互連組件連接計算設備，實現支撐大模型訓練及推理等AI工作負載的分佈式通信模式。網絡設備的性能及架構直接決定了整個AI算力集群的效率、吞吐量及性價比。
- **存儲設備**通常包括SSD和HDD服務器及分佈式存儲系統。它們為AI算力集群提供數據基礎，支持AI工作負載所需的大規模數據集（包括模型參數、訓練數據及中間結果）的存儲及訪問。

業 務

計算

憑藉全棧AI算力集群能力，對於算力設備，我們的角色主要在於為每名客戶的預期AI工作負載設計及配置算力架構。我們考慮模型類型、規模需求及應用場景等因素，以在性能及成本之間取得適當的平衡。這包括確定加速器的適當組合、決定計算節點的組織及互連方式，並規劃可適應未來工作負載增長的可擴展架構。在確定針對客戶目標AI工作負載的算力組合時，我們會綜合評估GPU的算力能力、節點互連拓撲以及內存及功耗限制等因素。通過我們的算力設備設計以及硬件選擇與融合能力，我們的AI算力集群產品能夠幫助客戶構建與其AI工作負載相匹配的算力基礎架構，同時提供強大的性能及高效的資源利用率。

我們的算力設備架構設計支持異構算力配置，能根據不同的工作負載需求，靈活地選用並組合GPU、CPU及專用加速器。我們已針對多種模型與工作負載類型設計並交付了算力配置方案，包括密集型、MoE及多模態訓練，以及高吞吐量推理。我們的AI算力集群產品目前兼容國內外主流供應商的算力硬件，讓客戶能在不同的技術與監管要求下，更靈活地配置與擴展AI算力集群。

AI算力網絡系統 – Mercury

網絡對AI算力集群而言至關重要，它決定了計算資源如何有效地作為一個統一的系統進行通信、協調及運行，而不是簡單地將各個獨立的GPU連接起來。網絡層的性能直接影響數據傳輸效率、任務同步及集群整體利用率，因此是AI算力集群能否大規模提供穩定高性能的關鍵因素。鑒於其重要性，我們持續投入該領域，並開發了大量專有技術及研發成果。

針對我們AI算力集群產品的網絡設備，我們提供整體解決方案，涵蓋網絡設計與部署、性能調優以及持續監控、運營及運維。在設計與部署階段，我們為AI算力集群構建高性能網絡架構，包括交換機、光模塊及線纜等關鍵組件，以及設備互連技術。之後，我們針對不同的AI工作負載對網絡性能進行調優，並優化節點之間的數據交換及同步，提升大規模訓練與推理效率。通過該等能力，我們旨在為大規模AI工作負載提供高穩定性、低延遲及高可用性的網絡基礎設施，平均年化SLA達到99.97%。

彈性擴展亦是集群設計的核心考量，而大規模算力集群從數十顆加速器擴容至數萬顆的核心瓶頸集中在網絡互聯與全域可靠性層面。對此，我們依托網絡架構優化，採用模塊化資源組網、標準化通信邊界定義與精細化故障隔離域設計，有效解決大規模節點互聯痛點，實現集群無縫橫向擴容，無需重構底層網絡拓撲；同時依托優化後的網絡底座，支撐異構資源動態調度、集中管控，落地跨區域訓練協同、彈性推理調度及跨部門算力按需分配，最終保障AI算力集群在規模持續擴張過程中，網絡性能穩定、業務運行連續可靠。

為了滿足AI算力集群對高性能、低延遲、穩定性及可擴展性的網絡需求，我們開發了專有的AI算力網絡系統*Mercury*。該系統包含網絡操作系統*MercuryNOS*及硬件設備（包括交換機、光模塊及RDMA網卡），旨在實現整個集群的高效可靠通信。

業 務

MercuryNOS

*Mercury*亦在我們的專有網絡操作系統*MercuryNOS*的支持下運行，該系統作為協調與運行其硬件設備的底層軟件層。*MercuryNOS*用於提升高併發工作負載下的帶寬利用率、流量分配效率及傳輸穩定性，同時提供增強的監控可視性，以支持AI性能調校與故障診斷。其主要功能包括以下各項：

- **動態流量平衡**：*MercuryNOS*支持一系列流量負載平衡機制，可根據網絡狀況及交換芯片的特性選擇最合適的方式，包括自適應路由或基於流切片的動態流量平衡。這有助於更有效地利用網絡拓撲中的可用帶寬，並減少因多條鏈路流量不均所造成的局部擁塞。
- **增強擁塞管理**：*MercuryNOS*整合了先進的擁塞偵測與控制機制，用於減少數據包遺失與傳輸中斷。通過更有效地管理網絡流量與緩衝，其協助維持通訊穩定性，並減少因重複數據傳輸所造成的性能損失。
- **毫秒級監控可視化**：AI工作負載常伴隨網絡流量突然出現峰值。為支持這些情景下的性能調優與故障診斷，*MercuryNOS*與交換芯片協同運行，能以短至一毫秒的間隔捕捉流量狀況。相比之下，國際先進的AI算力網絡系統所支持的最小監控間隔通常為100毫秒。

*MercuryNOS*部署於我們自主開發的交換機上，並已成功適配國際主流白盒交換機。

*Mercury*硬件

*Mercury*的硬件組件包括自研交換機、光模塊及RDMA網卡，這些共同為AI算力集群提供實體互連基礎。

- **交換機**：我們自主開發的AI算力交換機為AI算力集群提供高容量、低延遲的互連。每台交換機支持400G連接，提供51.2 Tbps的總交換容量，數據包轉發率達21.6 Gbps。這有助於網絡在數千張GPU之間處理密集且高度並行的數據交換，而不會成為性能瓶頸。這些交換機支持集中式與分佈式環境部署，並可隨著計算需求的增加而擴展，以適應不同規模的算力集群。其用於在嚴苛的AI工作負載下維持穩定可靠的網絡性能，有助於減少數據包遺失，在流量高峰期支持更順暢的數據傳輸，並提升集群運行的整體穩定性。交換機還包括即時監控與遙測功能，可提供網絡狀態的可視性，支持更有效的流量管理以及與我們軟件系統的協調。下圖展示我們自主研發的AI算力交換機：



我們自主研發的AI算力交換機

業 務

- **光模塊**：我們的400G/800G高速光模塊，可實時監測誤碼率、估算光模塊剩餘使用壽命，同時為AI算力集群提供高速光連接，並支持網絡鏈路上的穩定數據傳輸，旨在滿足大規模AI工作負載對頻寬與可靠性的要求。其遙測功能可即時掌握底層實體鏈路的健康與穩定性狀態，能夠及早發現潛在異常，並對互連故障進行智能分析，從而支持更主動的營運與維護。
- **網卡**：我們的RDMA網卡整合了自主研發的擁塞控制演算法，為每個計算節點的數據傳輸提供高帶寬、低延遲的連接。每個網卡支持雙200 Gbps端口，提供400 Gbps的總帶寬，整體延遲低於10微秒。我們的RDMA網卡能讓數據在算力節點之間直接傳輸，最大程度減少CPU的參與，有助於降低通訊延遲，並提升AI工作負載的整體計算效率。此外，網卡還支持針對不同工作負載條件進行擁塞控制調整，以及硬件層加密，以提升網絡可靠性與安全性。

存儲

就存儲設備而言，我們的職責主要在於為每名客戶的預期AI工作負載設計存儲配置。我們決定儲存硬件設置、容量與性能參數並編排儲存設備，使其與計算及網絡設備協同運行。這有助於支持AI工作負載所需的帶寬、延遲及I/O性能。存儲設備亦設計為支持災難復原與靈活擴充容量，有助於按照工作負載需求的演變來維持集群的穩定、持續運行。

AI算力操作系統 – Venus

在大規模AI算力集群的架構中，硬件是承載算力的物理基礎，而AI算力軟件則是釋放硬件潛能、驅動整個集群高效協同運轉的核心樞紐——它不僅要實現對海量異構硬件資源的抽象與調度，更要打通算、網、存各層級的協同壁壘，讓成千上萬的計算單元能夠像單一系統一樣協同工作，充分釋放超級算力的性能潛力。過去，行業內常依賴零散的開源組件或多廠商軟件拼湊集群軟件棧來構建集群軟件系統，不僅難以發揮系統級的跨層優化能力，更帶來了硬件適配繁瑣、運維成本高、資源協同效率低等諸多痛點。為此，我們將自研的核心軟件技術統一為AI算力操作系統Venus，以統一的全棧軟件棧實現對整個集群的端到端管控，既打破了零散組件的性能瓶頸，也為用戶提供了一致的使用體驗與標準化的交互接口，簡化了複雜的大規模AI集群的部署及操作，使其變得簡單易用、高效可靠。

Venus包含驅動層、內核層及用戶層。驅動層為內核層提供訪問分佈式設備的統一接口。內核層為用戶層提供處理AI工作負載所需的核心運行時組件。用戶層為不同類型的AI工作負載（包括預訓練、後訓練、推理及智能體工作負載）提供優化的基礎設施服務。Venus使AI算力集群能夠作為由數以萬計的GPU組成的超級計算機運作，從而實現長期高效及穩定的營運。

業 務

Venus 具備以下主要特性：

- **全面的功能：***Venus* 提供一整套支持AI算力集群運行的功能系統能力，涵蓋調度、通信、緩存、隔離、監控、故障診斷、故障恢復及負載加速等能力，從而同時支持AI工作負載的執行與集群的長期穩定性。
- **用戶友好型交互：***Venus* 提供直觀且精簡的用戶介面，能使客戶在無需直接與底層硬件及系統元素交互的情況下管理和監控AI算力集群。
- **強大的適應性：***Venus* 支持多種AI工作負載與應用場景，並能通過其底層的組件編排、任務調度與抽象化能力，適應不同的模型、工作負載模式及異質硬件環境。

驅動層

驅動層主要負責編譯與抽象。這是與底層硬件協同運行最直接的軟件層，其將一般系統指令轉換為特定硬件設備可執行的命令。主要角色是為上層應用程序屏蔽硬件設備之間的差異，使其能通過統一介面使用計算資源，而無需處理底層硬件的技術細節。這使得硬件資源更易於抽象化、管理與協調。

在 *Venus* 的驅動層內，關鍵技術組件主要包括CUDA、Kubernetes及我們自主研發的*VenusDSL*。CUDA作為GPU算力的軟件接口，為廣泛的大模型訓練及推理工作負載提供並行計算API。Kubernetes將分佈式異構算力資源抽象為統一的資源池，隨後可分配至各種大模型訓練及推理任務，有助確保分佈式工作負載的穩定營運。我們自主研發的*VenusDSL*是用Python編寫的專有GPU計算接口，旨在實現大模型訓練及推理的高效並行計算。其關鍵優化機制包括：

- **計算與通訊重疊以加速推理：**在大模型推理中，需要進行張量並行(TP)組內的AllReduce等通訊操作，以及專家並行(EP)組內的詞元分發與聚合。透過並行處理兩個批次，系統將一個批次的計算與另一批次的通訊重疊，從而有效隱藏通訊延遲。為避免可能導致性能下降的GPU資源爭用，通訊任務被卸載至CPU，進一步提升推理性能。
- **預填充與解碼的算子優化：**預填充階段涉及計算密集型的大矩陣運算，而解碼階段則屬於內存密集型並專注於單詞元的注意力計算，兩者具有顯著不同的計算特徵。我們為每個階段設計專用算子，並透過線程束級專門化、流水線並行及以Triton實施的自定義內核等技術優化其性能，從而減少開銷並提升整體效率。
- **算子融合以減少內存開銷：**解碼階段受限於內存帶寬。我們應用算子融合以減少CPU開銷及不必要的內存訪問。中間張量直接傳遞至後續計算，而無需寫入高帶寬內存（「HBM」）並從中讀取，從而提升效率。採用Triton編程模型實現單內核融合，從而提升內存訪問效率。
- **高效MoE詞元路由與重新排序：**在MoE模型的推理過程中，詞元需根據Top-K路由決策重新排序。我們實施了一個融合的Triton算子，該算子將Top-K路由與詞元放置計算相結合，及並行執行詞元重新排序。整個過程在單個大規模並行內核內執行，從而實現高效執行，同時允許在資源使用與計算性能之間進行靈活權衡。

業 務

內核層

內核層是Venus的中央控制系統。內核層建基於驅動層之上，負責管理關鍵系統資源並支持AI算力集群的穩定運作。其基於PyTorch等多種基礎框架，為AI工作負載的高效可靠運行提供核心運行時環境。於Venus中，內核層最關鍵的組件包括：高速集合通訊庫(VenusCL)、緩存輸入／輸出(VenusIO)、多隊列任務調度器(VenusCore)、資源及任務隔離組件(VenusBox)以及多維度監控組件(VenusNOC)。

- **通信(VenusCL)：**Venus的通信組件由我們自主研發的統一集合通訊庫VenusCL提供支持，是分佈式大模型工作負載的基礎。它能夠實現跨多個計算節點高效和精確的數據同步，這對於訓練和推理任務的效能至關重要。與業界常用的解決方案相比，Venus透過計算與通訊的深度同步及零拷貝數據傳輸等機制，能夠提供更優異的同步效能。
- **緩存(VenusIO)：**Venus的緩存組件由我們自主研發的VenusIO提供支持。緩存是影響AI算力操作系統中AI工作負載運行效率的關鍵因素之一。於模型推理中，緩存用於減少重複的詞元計算並加速自回歸生成，使其成為提升推理效能的最重要手段之一。VenusIO採用大容量SSD緩存及GDS技術，以解決業內長上下文推理中常見的瓶頸問題，即KV緩存受限於GPU顯存容量。
- **調度(VenusCore)：**Venus的調度組件由我們自主研發的VenusCore支持，使我們的AI算力操作系統能夠通過Kubernetes運行時接口，將不同的AI工作負載分配至合適的GPU計算節點。除了模型與數據管理等基本功能外，其還支持拓撲感知調度、多種資源選擇策略及自動彈性擴縮，提升大規模AI算力集群的資源利用率，並根據推理需求的變化動態調整資源。
- **監控(VenusNOC)：**Venus的監控組件由我們自主研發的VenusNOC提供支持，作為AI算力集群的「健康檢查中心」與「儀表板」，收集計算、網絡及儲存資源數據，以支持性能分析、故障診斷及資源調度。其通過主動運維能力（例如光鏈路質量分析及實時在線任務效能分析），以及毫秒級流量監控，確保AI算力集群的高效及穩定運作。
- **隔離(VenusBox)：**Venus的隔離組件由我們自主研發的VenusBox提供支持，確保多個租戶與工作負載能在AI算力集群中共享資源而不互相干擾。其通過基於Kubernetes的配額與節點資源管理以及GPU虛擬化與分區來支持資源隔離。我們亦為需要大量虛擬運行時環境的強化學習及智能體任務，開發更細粒度、輕量級及高併發的沙箱隔離技術，大幅提升任務執行效率及AI算力集群資源利用率。

業 務

內核層的主要功能包括：

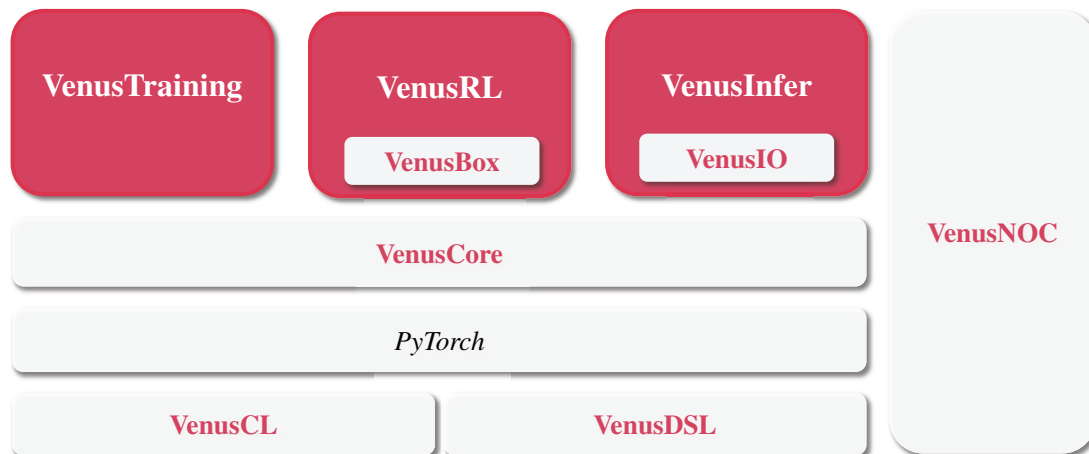
- **拓撲感知式工作負載部署：***Venus*根據跨計算節點網絡拓撲及節點內連通（包括GPU及RDMA網卡的互連）部署工作負載，以優化跨節點數據同步路徑、緩解瓶頸並提升資源利用率。這有助於減少閒置時間，並顯著提升超大規模算力集群的MFU。在客戶部署中，減少了GPU閒置時間，並降低了單次訓練的計算成本。*Venus*還支持多種調度模式，以更靈活地適應不同的工作負載需求。
- **跨區域計算調度：***Venus*支持對分佈於不同地點的AI算力集群資源進行統一管理及調度。通過彙總可用資源，並根據可用性、延遲和成本調度工作負載，其幫助客戶更好地利用自身部署中的閒置容量，進而提高了資源利用率，減少了延遲和響應時間，並支持多租戶環境和多樣化的產業AI用例。於客戶部署中，*Venus*將跨團隊協作效率提升高達45%。
- **多租戶工作負載隔離與調度：***Venus*允許多個團隊、業務單位和外部用戶共享同一個AI算力集群，同時減少工作負載之間的干擾。其通過資源配額、基於優先級的調度、工作負載隔離控制以及容器化來實現此目標。這有助於為每個租戶提供更可預測的資源可用性和服務品質，同時減少共享環境中的資源爭用和不公平調度。
- **容錯任務連續性：***Venus*持續監控網絡流量、設備狀態和工作負載執行情況，以偵測故障和性能問題。當數據同步鏈路出現問題時，其會實時切換至預先建立的備份路徑，以確保任務執行不受干擾。當節點出現問題時，其會隔離受影響的節點或GPU，將工作負載轉移到備份資源，並使用檢查點來支持更快速的復原。利用該設計，MTTR縮短至數分鐘內，訓練中斷的復原時間最多縮短70%。
- **以任務為中心的監控與故障管理：***Venus*支持整個AI任務執行過程中的監控與故障管理。在重要任務開始前，其可以執行集群範圍的系統檢查，涵蓋節點健康狀態、網絡穩定性、儲存吞吐及GPU狀況，以確認資源就緒狀態並提前識別潛在風險。在任務執行期間，*Venus*持續監控任務狀態、GPU健康狀況、網絡性能及儲存響應能力，以及早發現異常，降低長時間執行的任務在大量計算後失敗的風險。當問題出現時，*Venus*運用其專家知識庫和分析算法來識別根本原因並提供修復建議。*Venus*還支持自動和手動復原機制，包括基於檢查點的復原、節點隔離和硬件快速更換流程，以最大限度地減少中斷並有效恢復系統能力。

用戶層

在用戶層，*Venus*的核心能力旨在支持廣泛AI工作負載的執行。儘管模型訓練、自動駕駛及具身智能等不同的AI應用各有不同特徵，但底層基礎設施通常可分為預訓練、後訓練、推理及基於智能體的應用。儘管預訓練、後訓練及推理的服務引擎在其特定任務執行工作流程上有所不同，但其整體架構框架及核心組件大體一致。

業 務

訓練及推理服務引擎



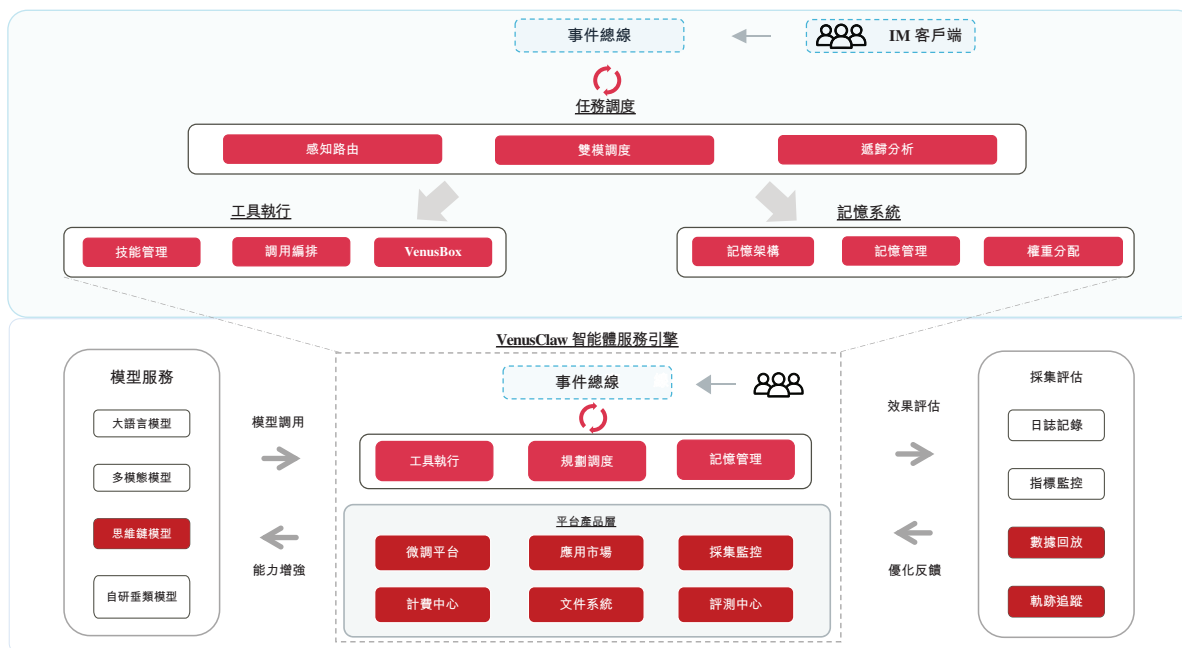
VenusCore 作為我們AI算力集群的統一入口，能夠根據集群狀態將預訓練、後訓練和推理工作負載動態調度至合適的資源，並協調三個專用服務引擎，即 *VenusTraining*、*VenusRL* 和 *VenusInfer*。這些引擎基於 *VenusCore* 構建，分別提供預訓練、基於強化學習的後訓練以及推理的基礎能力。這三個引擎均由共享內核組件提供支持，包括用於資源調度的 *VenusCore*、用於集合通信的 *VenusCL*、用於算子優化的 *VenusDSL* 以及用於監控和運維的 *VenusNOC*。此外，*VenusBox* 為 *VenusRL* 中的強化學習 workflow 提供專用沙箱執行環境，而 *VenusIO* 通過分層緩存支持 *VenusInfer*，以提升推理效率。

- VenusTraining***：Venus 的預訓練服務引擎支持各種大模型的預訓練。預訓練任務主要用於基礎模型的訓練（包括 Dense 模型、MoE 模型、VLA 模型及世界模型），以應用於具身智能及自動駕駛等高級應用場景。此等工作負載的特徵為極大的模型規模、海量訓練數據集及高度密集的分佈式計算，通常需要數以千計甚至數以萬計的 GPU，對分佈式協同及強大的故障監控與容錯能力有極高要求。我們的 *VenusCL* 高性能集合通訊、Mercury 高性能網絡及 *VenusNOC* 組件有效滿足了該等要求。
- VenusRL***：Venus 的後訓練服務引擎支持透過微調或強化學習進行的各種後訓練任務，例如醫療及金融等行業的垂直行業模型微調，或使用強化學習方法的模型量化及蒸餾。後訓練任務通常涉及兩個緊密相連的階段。首先，使用相對小量的特定任務數據透過微調進一步調整預訓練的基礎模型。其次，使用經微調的模型對指定數據集進行推理，並根據相應的驗證數據評估輸出結果。透過反覆進行的微調、推理及結果驗證，模型逐步優化以達到所需的推理質量及應用性能。由於後訓練涉及大量驗證過程，其極度依賴輕量級、高效率及高併發的隔離環境。我們的 *VenusRL* 透過軌跡採樣及調度與 *VenusCore* 協同，連同 *VenusBox* 輕量級及高併發的沙箱能力，提供高效的後訓練基礎設施。

業 務

- **VenusInfer**：Venus的推理服務引擎支持各種模型的推理任務。推理任務主要用於支持已訓練模型在一系列現實世界推理場景中的部署及營運，包括低延遲、高併發的實時交互、長上下文及邏輯密集的決策與規劃、內容生成及多模態推理。為應對該等工作負載的特徵（包括顯著的流量波動及異構負載模式），採用了預填充與解碼解耦的層次化架構，結合計算與緩存，其提供高彈性、低延遲的推理服務，有效適應波動的請求流量及日內負載變化。

智能體服務引擎



VenusClaw作為我們AI算力操作系統的智能體服務引擎，充當支持AI智能體大規模部署及穩定運行的基礎設施。與傳統推理服務不同，VenusClaw不僅為模型推理提供算力，亦在整個智能體工作流程中提供編排、執行及優化能力，支持從任務理解至工具調用、結果驗證及迭代改進的閉環流程。VenusClaw的控制面板實現智能體的聲明式配置、版本管理及健康監控。同時，VenusClaw透過其網關及策略模塊提供統一入口點，以處理請求路由及分發、流量管理、安全認證以及協議適配。

VenusClaw提供先進的推理及規劃能力，包括任務理解及分解、思維鏈推理、長程分層規劃及反思驗證，從而為複雜任務作出有效決策。其亦透過工具註冊及發現、動態調度及編排以及多模態感知等能力，支持智能體與外部環境之間的高效交互。此外，VenusClaw實現任務調度、多智能體協作及記憶增強，支持跨會話的上下文連續性及知識積累。

業 務

在此基礎上，*VenusClaw*建立分層智能體平台架構。憑藉*Venus*內核層的核心能力，其為智能體提供全面支持，並實現多種模型類型（包括大模型、多模態模型及領域特定微調模型）的統一編排。透過統一的模型接入層，其支持負載均衡、流式輸出及容錯後備機制。在中層，*VenusClaw*提供關鍵能力，包括智能體調度、低代碼編排、高性能安全沙箱、記憶管理、詞元計費、API網關服務及事件總線。特別是，其毫秒級安全沙箱機制為代碼執行及工具調用提供隔離且受控的執行環境。在上層，*VenusClaw*支持在性能、安全性及成本維度的全面評估。結合軌跡捕獲及回放能力，其實現智能體行為的閉環優化，並支持業務應用的持續迭代。

*VenusClaw*架構與基於智能體的應用程序的核心特徵高度契合。為應對智能體任務的迭代性質及波動的流量模式，*VenusCore*透過彈性擴展以及同步及異步任務的統一調度實現動態資源分配，從而大幅提高高併發場景下的系統吞吐量及穩定性。同時，透過標準化接口及全棧可觀測性，*VenusClaw*大幅降低智能體開發及運營的門檻，使企業能夠快速構建、部署及運營特定行業的智能體應用程序，並全面釋放自動化帶來的生產力提升。

我們已為企業服務部署嵌入辦公智能體及*VenusClaw*智能體服務引擎的融合AI算力集群，大幅提高員工的日常工作效率。先前需要長達一週人工處理的任務（如分析報告編製），現可於半天內完成。我們的融合AI算力集群建基於開源模型及全國產硬件組件（包括CPU、GPU、網卡及交換機），以滿足企業對私有化部署的要求。

用戶介面

*Venus*提供了一套易於使用且直觀的圖形化界面，用於管理AI工作負載（包括訓練、推理及智能體任務），以及監控AI算力集群的狀態和各個任務的實時執行狀態。

- **預訓練任務管理：***Venus*提供高度可配置的任務管理框架，允許用戶靈活定制不同模型及工作負載的訓練過程。具體而言，於任務初始化期間，用戶可根據模型規模指定擬分配的GPU數量，配置共享存儲路徑以高效管理大規模訓練數據集，並從多種容器鏡像中作出選擇，以使環境適應不同模型。此外，*Venus*提供拓撲感知智能化調度，其可透過專用介面進行配置，以優化整個計算集群的資源分配。用戶亦可啟用高效的*VenusCL*統一集合通訊庫，並透過在任務配置介面中設定相關環境變數以啟動備用通道，從而實現多層容錯。此外，系統預設啟用主動及被動監控機制，以對集群提供持續監督。綜合而言，該等功能有助確保超大規模AI算力集群的高效穩定運行，並使*Venus*能夠滿足預訓練工作負載的嚴苛要求。以下截圖說明*Venus*的預訓練任務調度功能：

業務

The screenshot displays the 'Resource Configuration' (资源配置) section of the Venus platform. It is divided into two main panels: 'Resource Configuration' (资源配置) and 'Environment Configuration' (环境配置).

Resource Configuration (资源配置):

- Instance Configuration (实例配置):** Includes tabs for MPI, PyTorch, Simple, CPU, and Deployment. Below these are toggle switches for 'Application MIG Mode' (应用MIG模式), 'Machine Learning' (整机训练), and 'vGPU'.
- Roles (角色):** A table with columns for Roles, Instance Count (实例数量), Resource Type (资源规格), GPU Count (GPU数量), CPU (核), and Memory (内存(GiB)).

Roles	实例数量	资源规格	GPU数量	CPU(核)	内存(GiB)
master	1	选择资源规格	1		1500
worker	1	选择资源规格	1		1500
- Priority (优先级):** A row of buttons labeled 6, 5, 4, 3, 2, 1. Below is a radio button for 'Queue Length, Priority Decreases' (队列越长, 优先级越低).
- Collection Policy (回收策略):** A toggle switch for 'Not Collectible' (不可回收).
- Automatic Restart (自动重试):** A toggle switch.
- Storage (存储):** Buttons for '+ 存储卷', '+ 临时卷', and '+ S3'. Below are explanatory notes about storage types and their lifecycle.

Environment Configuration (环境配置):

- Image (镜像):** Buttons for 'Predefined Image' (预置镜像) and 'Custom Image' (自定义镜像). Below is a text field for 'Select Image Address' (选择镜像地址) with a '+' icon.
- Working Directory (工作目录):** A text field for 'Enter Command Execution Directory' (输入命令执行目录).
- Command Mode (命令模式):** Buttons for 'Custom Mode' (自定义模式) and 'Shell Mode' (shell模式).
- Entrance Command (入口命令):** A text area containing a sample command:

```
1 ./bin/bash
2 ./c
3 sleep inf
```

Annotations on the right side of the image:

- A red dashed arrow points from the 'Instance Configuration' section to a box containing the text: '用户指定任务类型，GPU参数等' (User specifies task type, GPU parameters, etc.).
- A red dashed arrow points from the 'Environment Configuration' section to a box containing the text: '用户可自定义扩展环境变量' (User can customize and extend environment variables).

© 自定义模式可输入 Linux 命令（参数以空格分隔，按回车运行），Shell模式支持直接粘贴完整的 Shell 脚本。

預訓練任務調度

- **後訓練任務管理：***Venus*旨在透過綜合及可配置的平台能力應對相關後訓練要求。*Venus*使客戶能夠從模型庫中挑選合適的預訓練基礎模型作為後訓練的基礎，並支持多種主流微調方法（包括LoRA及RLHF），以及為所選方法量身定制的參數配置。其亦允許客戶分配各項任務所需的計算資源，並配置自定義環境變數，以支持特定運行時要求及功能選項。此外，*Venus*支持協調管理後訓練所需的推理數據集及相應的驗證數據集，從而提供微調、推理、驗證及迭代優化的綜合框架。通過相關功能，*Venus*助力客戶提升後訓練工作流程的效率、靈活性及可靠性，並更好地實現下游應用的目標模型性能。以下截圖說明*Venus*的後訓練任務調度功能。

業 務

< 创建模型微调任务 · 常规模式

基础信息

任务名称 *

描述

0 / 1024

模型名称 *

指定基础模型

训练配置

微调方式 *

参数配置

超参数	参数名	参数值
截断长度	cutoff_len	1024
学习率	learning_rate	0.0000001
日志间隔步数	logging_steps	5
最大样本数	max_samples	1000
NEFTune降噪参数	neftune_noise_alpha	0.00
训练轮数	num_train_epochs	1.0
批处理大小	per_device_train_batch_size	1
学习率预热比率	warmup_ratio	0.00

選擇微調模式及指定微調配置參數

资源配置

实例配置 * ☐ 启用MIG模式 ☐ 整机调度 ☐ vGPU

资源规格

GPU数量

CPU核

内存(GiB)

环境变量

数据集配置

训练集 *

验证集 *

指定後訓練數據集及驗證數據集

後訓練任務調度

- 推理任務管理：Venus提供一套全面的配置及管理功能，以實現推理算力資源的動態及高效分配。客戶可從集中式模型庫中挑選已訓練模型以部署為推理服務，並挑選最切合推理任務要求的推理框架。客戶亦可配置是否採用分佈式推理，從而允許單一推理實例利用多個協同工作的計算節點。此外，客戶可指定擬部署的推理副本數量。由於單一推理實例僅能處理數量有限的推理請求，故並行運行多個副本可按比例擴展推理吞吐量。關鍵而言，Venus支持彈性擴展，以應對波動的推理請求動態調整活躍推理實例的數量；於日間高峰時段部署更多實例以滿足較高的請求量，並於夜間非高峰時段縮減規模，以減少資源消耗並避免不必要的浪費。透過該等功能，Venus使客戶能夠高效部署及管理推理服務，同時針對不同的需求模式優化資源利用率。以下截圖說明Venus的推理任務調度功能。

業 務

環境配置

分布式推理 ☐

◎ 每个实例都在多台机器上进行部署。例如需要 4 个实例，单实例部署的机器数为 2，则一共需要 4*2=8 台机器来部署该服务

推理框架 *

预置框架 自定义框架

请选择推理框架

模型仓库 *

请选择模型仓库

环境变量

+ 新增环境变量

指定推理模型相關信息

资源配置

实例配置 *

启用MIG模式 ☐

整机调度 ☐

vGPU ☐

资源规格

GPU数量 CPU(核) 内存(GiB)

请输入资源规格

1 0 0

副本数

- 1 +

◎ 设置或调整部署的副本数量，可提高部署的吞吐量响应速度

存储

+ 存储卷

+ 临时卷

+ S3

◎ 存储卷、临时卷、S3：每个挂载目录不能重复。
◎ 存储卷：可通过“库 → 存储”页面管理的持久卷。
◎ 存储路径：持久卷有存储在内部挂载的路径。
◎ 临时卷：由 EmptyDir 提供的临时卷，其生命周期与 Pod 绑定。
◎ 主机路径：由 Host Path 提供支持的卷，用于将宿主节点文件系统中的目录或文件挂载到 Pod 内。

指定多實例數以滿足性能目標

高级配置

扩缩容 ☒ 开启

选择方式

定时伸缩 自动扩缩

规则

扩缩容规则

目标副本数

每 天 时 分 每 天 时 分

+ 扩缩容规则(/10)

指定潮汐效應下的動態擴縮容配置

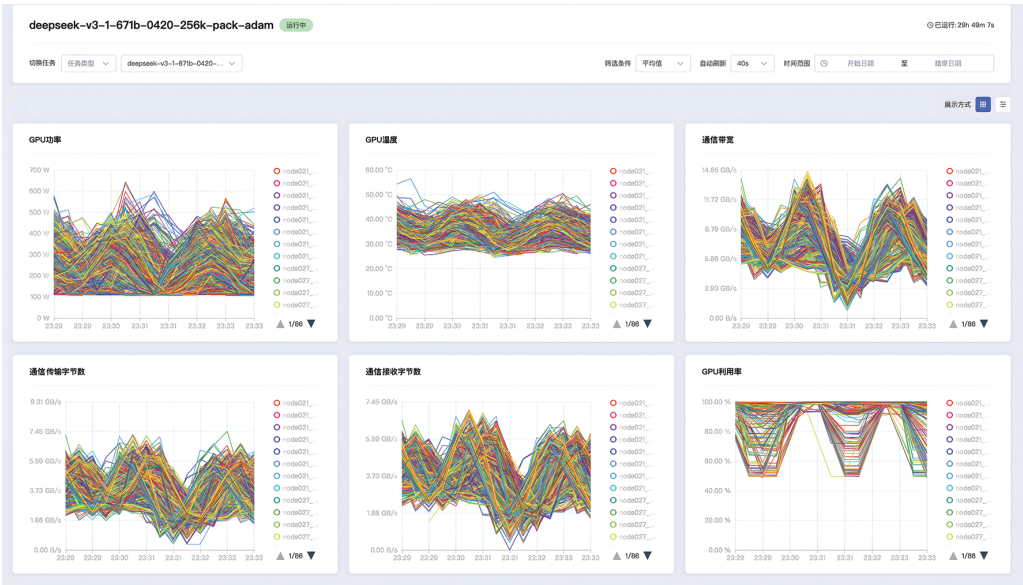
推理任務調度

- 可視化監控：Venus能夠監控AI算力集群的整體資源狀態，並以高度可視化即時監控各任務所使用的資源，例如GPU使用率與溫度，幫助用戶評估運行性能並在工作負載執行過程中識別異常情況。



集群信息概覽

業 務



任務運行狀態監控

案例研究

案例研究A

於2024年，我們為一家國內大模型公司部署及交付AI算力集群，以支持其模型訓練、推理及應用部署的需求。由於該客戶已成立專責團隊負責大模型開發，因此需要一套量身打造的AI算力集群，該集群須具備充足的計算能力、系統穩定性及可擴展性，以支援其不斷演進的開發藍圖。

於項目初期，我們與客戶緊密合作，瞭解其具體需求，包括研發中模型的參數規模、預期訓練周期、集群的預期使用模式、需接入的團隊數量、不同工作負載之間的隔離要求，以及偏好的計算資源共享模式。基於有關需求分析，我們提出整體解決方案，包括一個兩千卡規模的AI算力集群，以及相應的存儲配置、AI算力網絡架構、數據中心空間規劃，還有為集群部署配套的電力機櫃及冷卻系統安排。在軟件方面，該集群搭載了我們的AI算力操作系統Venus，以支持統一管理、資源調度及集群運營。客戶批准有關建議方案後，我們組織並完成相關硬件及軟件的採購、部署及性能調優，並將整個AI算力集群作為一體化產品交付。交付後，我們亦負責集群的持續運維，保障其部署後穩定運行。在其後集群使用期間，我們進一步通過Venus協助客戶於集群上部署推理服務及智能體應用，使客戶可將集群的使用場景從模型訓練延伸至下游應用場景。於最後實際可行日期，該集群一直保持穩定運行，性能表現優異。於2025年，該集群實現低於0.3%的年化硬件故障率及高達99.97%的平均年化SLA。

業 務

案例研究B

於2025年，我們為一家領先的大模型公司所使用的萬卡規模的AI算力集群，提供了軟件及性能調優服務。在模型訓練期間，該集群出現嚴重且意外的效能下降，並伴隨訓練效能的劇烈波動。訓練流程也經常陷入無回應狀態，導致整體訓練效率低下，使得該集群在執行預定工作負載時，幾乎無法正常運作。

為此，我們首先於集群上部署了我們的AI算力操作系統Venus。透過其調度及監控功能，我們觀察到在模型訓練期間服務器上發生了數據同步錯誤，並確定根本原因為集群內一台服務器上特定GPU的程式故障。基於我們在大規模集群優化方面積累的經驗及既有的故障排除專業知識，我們透過運行功能級線上性能分析工具，系統地識別並更換了性能欠佳的GPU，並更新了相關服務器參數配置。

經過我們的優化後，該客戶集群在後續的訓練執行中展現出穩定的表現。以MFU為衡量標準，整體訓練表現相較於優化前的水準提升了超過30%。

案例研究C

於2025年，憑藉我們的VenusBox技術，我們為一家領先大模型公司部署了我們的智能體沙箱產品。該項目專注於兩個核心場景：大規模後訓練及智能體應用。該解決方案利用VenusBox的輕量級沙箱實例及彈性資源池機制，在計算、網絡及存儲元素實現優化，並在高可用性的部署架構層面實現了當控制節點發生故障時，流量可進行無損熱切換。

該沙箱產品專為涉及高頻調用、批量執行及實時交互的智能體應用場景而設計。其採用的調度架構支持快速本地查詢資源可用性及高效更新資源分配，從而提升高併發環境下的調度效率與集群級負載均衡。經由對API調度模塊進行關鍵算法升級後，集群負載不平衡變異系數由0.941顯著下降至0.021，接近近乎完美的平衡。此外，其基於模板的沙箱預熱系統有效減少了冷啟動帶來的延遲波動，提升了首次響應速度及服務連續性。最終，我們的產品穩定支持逾10,000個併發沙箱實例，並協助客戶將每月平均沙箱成本降低40%，且期間未發生任何重大系統故障或服務中斷。

案例研究D

於2026年，我們為中國一家大模型公司部署了一個算力為250 PFLOPs的AI算力集群，並為其提供持續的營運及維護，該集群配備了我們的推理服務引擎VenusInfer。隨著客戶逐步推出AI智能體，推理流量快速增長，從而產生了建立穩定且具成本效益的推理服務的需求，以提供其自主研發的模型API。該基於客戶專有模型的推理服務已連接至OpenRouter。

VenusInfer提供涵蓋網關路由、負載均衡、推理引擎管理、使用計量、節點管理、任務調度及故障遷移的推理服務能力。VenusInfer針對AI智能體應用程式中典型的「長輸入、短輸出」工作負載進行優化。其透過我們自主研發的VenusIO實現動態KV Cache管理，並透過VenusCore實現高性能算力資源調度，即使在流量大幅波動的情況下，亦能實現強大的彈性及高資源利用率。

業 務

秉承跨層優化及異構加速的設計原則，我們引入了兩種類型的算力硬件選擇：一種具備高算力及大緩存，另一種具備低算力及高GPU內存帶寬，分別適用於推理的預填充及解碼階段。與其他使用同構GPU的解決方案相比，我們軟硬件協同及異構加速的方法在相同資本開支水平下，將支持的推理流量提升了100%。

我們的AI算力集群交付

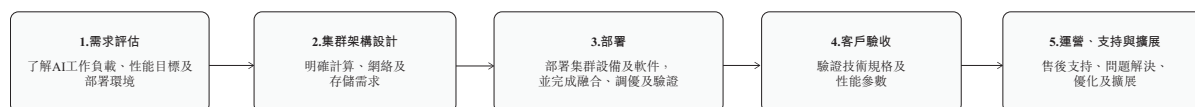
作為一家具備全棧式AI算力集群產品及服務能力的公司，我們可承接貫穿AI算力集群整個生命週期的端到端項目，包括集群架構設計、硬件選型、軟硬件適配、網絡互連、系統調優及性能驗證。基於模塊化模式，我們的AI算力集群產品結合了廣泛且隨時可用的硬件選擇及採購來源、自主研發的軟件產品，以及針對主流應用場景及客戶需求的成熟設計與部署解決方案庫。此模塊化模式讓我們能夠評估各客戶的計算規模、應用場景及性能要求，釐定整體集群配置，並挑選合適且具成本效益的AI算力設備與軟件組件組合。

根據客戶需求及現有基礎設施，我們的端對端產品交付可採用全面的集群部署形式，或僅包含選定的硬件及軟件組件，並配套集群級網絡搭建及整合服務的更具針對性的部署。此外，在若干情況下，我們可基於客戶現有硬件基礎設施，提供軟件優化、網絡部署及性能提升服務。於所有該等交付場景中，無論特定項目的具體交付範圍，我們均利用我們的產品、技術專長及實踐專有技術，以改善集群級互連、通訊效率及整體系統性能。此靈活交付模式使我們能夠滿足廣泛的客戶需求，同時持續對集群的整體可用性、性能及網絡性能負責。

於往績記錄期間，該業務線的全部收入均來自交付及服務AI算力集群。我們於客戶驗收時確認收入。於2023年、2024年及2025年，我們分別為4名、10名及48名客戶完成4個、10個及52個有關AI算力產品業務的項目。於往績記錄期間，AI算力集群產品項目的數量增加，主要受客戶對AI算力資源及AI算力集群相關解決方案的需求不斷增長所帶動。隨著AI應用的持續擴展，客戶不僅需要算力硬件本身，亦需要將算力、網絡及存儲資源有效連接並協調成一個能夠可靠、高效運作的統一集群。在此背景下，我們產品組合的關鍵價值並非僅在於個別項目中所交付的特定硬件或軟件組件，而在於實現集群級別互連、通訊效率及整體系統性能的能力。因此，不論項目涉及軟件、硬件或兩者的組合，亦不論具體交付範圍的規模，我們的交付均構成使客戶能有效使用AI算力集群的重要一環。

業 務

以下為從客戶需求評估到交付後運維的一般流程：



- **需求評估與架構設計：**交付流程通常以初步需求評估開始，期間我們與客戶合作了解其預期的AI工作負載、目標性能指標及部署環境。基於該評估，我們定義整體集群配置（包括計算、網絡及存儲要求），並提出符合該等要求的融合式集群架構。
- **集群部署：**在確認集群配置後，我們著手AI算力集群產品的部署。我們負責部署所需的所有AI算力設備組件，確保為集群配置滿足特定工作負載及性能要求所需的計算、網絡及存儲資源。我們亦部署及配置AI算力操作系統，並進行產品融合、調優及驗證，以確保AI算力集群作為統一系統運行。
- **客戶驗收：**部署完成後，AI算力集群產品由客戶進行驗收及驗證，以確認所交付的集群產品符合協定的技術規範及性能參數。驗收後，AI算力集群將根據客戶的部署計劃投入運營使用。
- **售後支持：**產品交付後，我們提供售後及運維支持，協助客戶隨時間推移運營及演進其AI算力集群。我們的售後服務包括響應客戶諮詢及運營要求、支持問題調查及解決，並在客戶工作負載、運行條件或部署要求變化時協助進行系統調整及優化。我們亦在初步部署後支持客戶進行容量規劃、系統擴展及配置更新，使AI算力集群產品能在維持穩定運行的同時適應不斷變化的使用模式。

我們AI算力集群產品的交付週期約為一個月，惟須視乎硬件採購週期、集群規模、整體部署複雜程度、客戶現有基礎設施、軟硬件適配要求、網絡互連安排及現場部署條件等因素而有所變動。

我們的AI算力集群運營服務

憑藉我們在設計及交付AI算力集群方面的廣泛專業知識及實踐經驗，我們以*Galaxy Services*品牌提供AI算力集群運營服務。我們的AI算力集群運營服務使更廣泛的客戶能夠獲取我們的技術專長及高質量算力服務。*Galaxy Services*由我們自主研發的AI算力網絡系統*Mercury*及AI算力操作系統*Venus*提供支持，針對傳統AI算力集群部署的局限性（其通常側重於初始部署，但缺乏持續運營支持）提供解決方案。其協助客戶應對算力資源碎片化、運營及維護要求高、內部協作效率低，以及龐大前期資本開支及持續運營成本負擔等挑戰。*Galaxy Services*亦在保障數據及運營安全的同時，降低客戶維持大型專職運營團隊的需要。我們AI算力集群運營服務的客戶數目由2024年的2名增加至2025年的16名。

業 務

Galaxy Services 具備以下主要特點，旨在協助客戶提升業務敏捷性並降低運營開支：

- **全面營運支持：**憑藉 *Galaxy* 的全面可觀測性及彈性系統架構，*Galaxy Services* 提供涵蓋集群全生命週期內硬件、系統及任務的運營及業務適配框架，協助客戶降低日常集群運營的複雜性。
 - **故障自恢復及主動式運營：***Galaxy Services* 具備內置任務容錯恢復及自動硬件故障轉移能力。透過運營智能體，其可進行主動式硬件健康檢查以及多維度、多尺度監控及告警，協助提前識別潛在風險，並在問題導致服務中斷前予以解決。運營智能體亦可自動處理節點宕機及網絡抖動等常見問題，實現99.97%的平均年化SLA。
 - **7×24小時技術專家支持：***Galaxy Services* 由專屬AI算力集群專家運營團隊提供支持，提供全天候技術支持。
- **提升成本效率：***Galaxy Services* 提供全面的成本管理能力，協助客戶更精確地追蹤算力資源使用情況、優化成本並更好地滿足企業需求。
 - **智能成本優化：***Galaxy Services* 可為不同項目團隊生成使用賬單，使客戶能夠對算力資源進行詳細的內部核算。透過分析歷史使用數據，其可自動識別閒置資源及浪費，並提供優化建議，協助企業降低隱性計算成本。
 - **兼容性及適配性：***Galaxy Services* 支持客戶現有算力基礎設施的整合及重用，並實現集群的漸進式升級及擴展。隨著客戶人工智能業務增長，彼等可在無需更換現有基礎設施的情況下擴展集群容量，從而提升長期資源利用率並保留其基礎設施投資價值。

運維及技術支持服務

我們提供運維及技術支持服務，以確保AI算力集群能有效運作，包括由我們交付的集群，以及由其他客戶部署且需要專業運維支持的集群。該等服務乃基於 *Venus* 的底層功能建構，尤其是其在監控、故障偵測、診斷及恢復方面的內核層功能，能強化我們識別問題、分析根本原因及支持及時修復的能力。我們的維護服務涵蓋系統監控、例行檢查、效能優化、故障排除及問題解決，並根據客戶營運需求提供技術支持。

業 務

案例研究

我們為一家領先研究機構集群規模超過20,000張GPU的AI算力集群提供運維及技術支持服務。我們的服務包括全天候運維支持以及針對容錯、故障遷移、智能運維、性能分析及監控的定制化軟件開發與維護。該等服務提升集群穩定性及服務質量，將每周因硬件故障導致的訓練中斷由超過15次減少至約3次，平均年化SLA於2025年達到99.97%。該集群支持超過700名用戶，最大單一任務使用數千張GPU，並連續運行長達三天，整個部署集群規模超過28,000張GPU。

算力服務

我們透過由我們設計、建設及運營的AI算力集群的算力提供算力服務。該等服務以我們AI算力集群產品所體現的相同技術、專業知識及運營專長為支撐，包括我們的AI算力設備能力及Venus的軟件功能，例如調度、資源協調、監控及人工智能工作流適配。因此，我們的算力服務使客戶能夠獲取與其人工智能工作負載需求高度匹配的高質量及高效算力資源。該項服務尤其適合尋求獲取高性能算力資源，而毋須擁有GPU或承擔建設及運營專用AI算力集群所涉及的重大資本開支、營運複雜性及持續維護負擔的客戶。根據該模式，客戶可以更靈活且更具成本效益的方式獲取算力資源，而我們則根據所提供算力資源的數量收取固定費用。我們的算力服務主要面向對算力資源及服務有穩定需求的大模型公司、雲服務提供商以及教育及研究機構。

我們的技術

我們已建立一套涵蓋硬件、通信協議、調度算法及系統軟件的全面技術體系，使我們的融合式AI算力集群能在大規模環境下以高性能、高可靠性及高擴展性運行。我們的核心技術涵蓋多個關鍵領域。在AI算力設備層面，該等技術包括光電集成技術、高通量交換硬件及網絡擁塞控制技術。在AI算力操作系統層面，我們的技術包括內核層的高性能集合通信、大容量緩存、超輕量級、高效及高併發的沙箱，以及多維度、多尺度的主動與被動監控。在用戶層面，我們亦開發高效的後訓練框架。展望未來，我們將繼續透過軟硬件協同設計及跨層優化，推動下一代AI基礎設施的發展，從而提升硬件可編程性、軟件智能化及系統級協同。透過該等努力，我們旨在提供更高質素的AI算力集群營運服務，並進一步鞏固我們在AI算力集群行業的領先地位。

業 務

AI算力設備技術

在AI算力設備層面，我們專注於硬件及*MercuryNOS*，其為我們的AI算力網絡系統*Mercury*的功能及性能提供底層能力支持，涵蓋AI算力集群網絡硬件的關鍵環節，並針對AI算力集群場景設計多項核心功能，包括基於NPO的光互連、OCS光交換技術及流量擁塞控制，旨在支持大規模AI算力集群內高效、可擴展且可靠的通信。

光電集成技術

隨著AI交換芯片的SerDes速率迅速提高，加上GPU擴展互連鏈路的數量及帶寬不斷增加，基於可插拔光模塊並結合高速SerDes PCB佈線的普遍採用架構越來越受到信號完整性問題及功耗限制的約束。基於NPO及CPO的光互連技術，連同OCS光交換技術，透過縮短電信號傳輸距離及以光交換取代電交換，解決了該等限制。該等方法可從根本上降低系統功耗、延遲及信號衰減，並正成為新一代AI算力集群的首選解決方案。

就OCS而言，我們已在AI算力集群內對結合電交換機與OCS光交換機的光電融合架構進行技術研究及驗證。此包括透過Venus對光及電交換路徑之間的實時流量調度進行驗證，以及能耗的比較分析。就NPO而言，我們專注於評估不同供應商的硅光解決方案之間的差異，包括其與AI算力集群架構的兼容性及其對運行及監控能力的支持。具體而言，我們已取得如下進展：

- **OCS光電集成**：經與我們的合作夥伴合作，我們已完成AI算力集群內光電融合架構的驗證，將基於MEMS的OCS光交換機與我們自主研發的高通量交換機相結合。這使大模型訓練的AllReduce時間減少50.3%，帶寬利用率提高47.1%，能耗降低32.1%。相關研究已獲接納於OFC 2026發表。我們亦已啟動基於硅光的OCS交換技術的進一步研究。
- **NPO**：我們已選定最先進的3.2T NPO組件，並已開始在我們自主研發的100T高通量交換機上進行部署。

展望未來，我們將繼續增加對硅光技術的投入，並加速將AI算力集群的需求整合至硅光產品中，以更好地滿足AI算力集群不斷變化的需求。

交換機硬件技術

AI算力集群依賴高速網絡將其分佈式計算節點連接成一個融合式系統。該高速網絡乃透過提供高帶寬及低延遲流量轉發及互連的AI交換機實現。交換機的總帶寬決定了AI算力集群可支持的計算節點規模，從而決定該集群的整體算力。交換機的總帶寬主要由兩個因素決定：(i)交換機內所用每個獨立交換芯片的轉發帶寬；及(ii)集成至單個盒式交換機的交換芯片數量。隨著交換芯片數量增加，功耗按比例增長，對散熱設計提出更嚴格及可預測的要求。每個交換芯片的更高帶寬乃透過增加芯片支持的轉發通道數量以及將每個通道的SerDes速率翻倍來實現。然而，這會大幅增加芯片的功耗。此外，一旦SerDes速率翻倍，為維持信號完整性，在硬件設計上需要採用更高性能的PCB材料並縮短傳輸距離。

業 務

我們已開發出針對高吞吐量AI網絡的專有交換機硬件技術。透過正交背板設計，我們的交換機硬件支持集成多個高帶寬交換芯片，以增加每台交換機的總帶寬。我們亦透過優化熱管與高密度鰭片的組合，解決由高功耗引致的風冷瓶頸。展望未來，我們計劃進一步迭代我們的液冷解決方案，以支持具有更高帶寬及功耗要求的交換機硬件。我們亦採用先進的3.2T近封裝光學技術，以應對超高速信號傳輸的挑戰，並支持下一代AI算力集群所需的高帶寬光電互連性能。在網絡操作系統層面上，我們採用主從架構更高效地管理多個交換芯片，這有助控制在多芯片部署場景中可能出現的資源需求大幅增加。

網絡設備流控技術

對於基於RoCE的AI算力網絡，無損組網乃透過基於反饋的機制實現。在該機制下，數據接收方透過數據包中的ECN標記檢測網絡擁塞，向發送方發送CNP通知，而發送方在接收到該等CNP通知後降低其傳輸速率。根據此機制，接收方基於特定時間內接收到附有ECN標記的數據包數量及數據包間隔等因素，決定發送CNP數據包的比例及頻率。發送方則基於多項參數（包括接收到的CNP數據包數量及該等數據包之間的間隔），決定適當的速率降低幅度。由於該機制依賴於發送端及接收端的多種因素，故選擇合適的參數組合尤為重要。我們已開發一套先進的擁塞控制技術，以解決大規模分佈式AI訓練中所產生的網絡擁塞問題：

- **分佈式自適應擁塞控制算法：**我們採用具備改良的自適應調優算法的分佈式架構，結合流量均衡編排，在整個AI算力集群中動態調整流量。該方法優化系統設計、算法效率及流量調度策略，以減少RoCE相關的數據包擁塞、降低通信延遲並維持高有效帶寬使用率，從而直接提升大模型訓練性能。
- **大模型感知擁塞控制（「LLM-CC」）：**LLM-CC是專為基於RoCE的AI算力集群網絡設計的擁塞控制通信協議，支持數據中心內及跨數據中心的統一流量管理，包括跨多個AI數據中心的長距離部署。由於RoCE對擁塞高度敏感，尤其是在大規模及長距離部署場景中，LLM-CC與我們交換機上的網絡操作系統協同運行，以監控網絡擁塞狀況並調節整個算力集群的流量。通過動態控制分佈式節點的流量速率及擁塞響應，該協議幫助降低長距離部署中延遲及擁塞增加對性能的影響。因此，在高達30公里距離內部署的AI算力集群能實現可比非長距離部署的約99%的有效計算性能，而在長達1,500公里距離內部署則能維持約95%的有效計算性能。

內核層技術

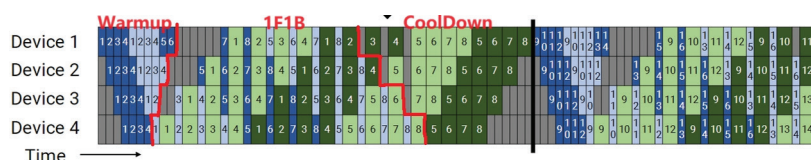
基於我們對AI算力集群需求的深厚理解及對性能優化的專注，我們採用跨層優化方法，並在集合通信、大容量緩存、沙箱隔離及主動運行等領域開展針對性的技術開發。

業 務

VenusCL集合通信庫

在AI算力集群中，計算節點及網絡交換機等AI算力設備共同決定集群的整體算力及網絡帶寬能力。在AI工作負載的實際執行過程中，該等工作負載通常表現出以下特徵：(i)其通常在多個分佈式計算節點上並行執行；(ii)每個節點在執行過程中通常在計算與數據同步之間交替進行；及(iii)儘管PyTorch等通用框架支持計算與數據同步的重疊，但該等重疊並不能消除同步開銷。

下圖展示跨四個計算節點的典型大模型分佈式訓練過程。如圖所示，儘管框架層已支持計算與同步之間的重疊，但由於數據同步效率等因素，各節點之間仍存在閒置等待期（如圖中灰色部分所反映），表明仍有進一步改善的空間。



我們已開發一套高性能GPU集合通信庫，即VenusCL，通過高速網絡通信系統支持GPU之間的計算結果同步。該通信庫旨在提升跨GPU數據交換的效率、可靠性及可觀測性，從而支持大規模分佈式AI工作負載：

- **異構統一通信層：**我們的通信庫採用封裝式設計，整合多種GPU通信庫，並向上層框架提供統一接口。其亦支持不同類型GPU之間的點對點通信，使得在同一AI算力集群中可基於混合硬件架構實現分佈式訓練及推理。
- **GPU通信性能優化：**我們通過將部分數據同步任務由GPU轉移至CPU，以提升GPU通信效率，使更多GPU資源用於計算。我們亦採用拓撲感知路由以降低通信延遲，運用先進的負載均衡算法以更有效利用可用網絡帶寬，並通過優化並行通信策略降低GPU通信開銷。在客戶部署中，該等技術使整體GPU利用率提升約2%至6%。
- **GPU通信可靠性及容錯能力：**我們在通信庫層面實施輕量級容錯及恢復機制，使通信問題可於數秒內自動檢測及恢復。該方法有助於將由網絡相關問題引致的任務中斷減少約50%。
- **高精度GPU通信測量：**我們在關鍵通信路徑中嵌入測量探針，以收集詳細的時序數據，包括GPU內存拷貝時間、GDR傳輸延遲及網絡傳輸延遲。該等實時遙測數據有助於我們更精準識別性能瓶頸，並作出更優化的決策。

業 務

VenusIO大容量推理緩存技術

隨著推理工作負載日益需要支持更長的上下文，對KV緩存的需求持續增長。現有（通常以內存為中心的架構為代表）KV存儲及管理系統受限於容量有限、成本高昂及在分佈式環境中相對較高的故障率。

為應對該等挑戰，我們已開發我們專有的VenusIO技術。VenusIO採用以存儲為中心的KV緩存管理架構，利用本地NVMe SSD的容量及成本優勢作為DRAM的替代方案，同時利用GDS技術加速GPU與存儲之間的數據傳輸。在引擎層面，VenusIO採用存儲預取及計算-I/O重疊流水線以掩蓋存儲I/O延遲。因此，VenusIO能夠實現與DRAM相若的延遲性能，同時顯著提升容量及成本效益，並在預填充階段支持高吞吐量計算。我們預期VenusIO將預填充階段的跨請求KV緩存命中率提升約30%，並將吞吐量提升約20%。

VenusBox高效沙箱技術

在高併發智能體環境中，大量沙箱請求可能同時到達。若資源信息更新存在延遲，過多請求可能被分配至少數節點，導致短期過載並降低調度效率。為此，我們已開發我們的專有VenusBox技術並升級了調度算法，通過維持更即時的本地資源使用視圖，並在作出調度決策後即預先調整可用容量，而非等待監控數據更新。這有助於避免沙箱請求集中分配至同一節點，從而減少因指標延遲引致的負載偏斜及羊群效應。

此外，在智能體環境中，許多請求需先獲得可用沙箱後方可執行工具操作。若沙箱僅於請求到達後才創建，首個回應可能出現延遲。為應對此問題，我們開發了基於本地影子狀態及樂觀資源更新機制的調度算法，在生產環境中將集群負載不平衡度（以變異系數或CV衡量）由0.941降至0.021。我們開發了基於模板的沙箱預熱框架。系統根據預期需求及使用模式預先準備一組可直接使用的沙箱實例。當請求到達時，可直接從該資源池分配實例，從而縮短等待時間，提升響應速度並消除與冷啟動相關的延遲變異。在實例被取用後，系統在後台自動補充並根據需求變化動態調整資源池規模。這有助於使我們的系統在流量高峰期間仍可維持低延遲表現，同時提升吞吐穩定性及資源利用率。

VenusNOC故障監測與主動維護技術

在大規模AI算力集群中，互連鏈路及光模塊的數量可達數以萬計。光模塊是計算節點互連中不可或缺的組件，任何鏈路或光模塊故障均可能導致鏈路不穩定或停機，從而影響在集群上運行的訓練及推理等AI工作負載。基於我們在大規模集群交付及營運方面的豐富經驗，我們發現涉及光模塊及線纜的鏈路故障是系統故障的主要來源之一。

為解決該等問題，我們為光模塊加入了增強功能支持（包括實時誤碼率及累計運行時間等指標），並在系統層面上建立了相應的主動檢測及監控能力。這使我們能夠在較早階段識別潛在

業 務

問題並提前規劃補救措施，從而透過更具前瞻性的維護方法減少任務中斷及其他干擾。結合主動運維技術、協同故障監控及調度容錯能力，我們的AI算力集群產品已實現低於0.2%的每月網絡抖動率，使萬卡或以上規模的AI算力集群能夠實現高達99.97%的平均年化SLA。

用戶層技術

於後訓練及基於智能體的場景中，強化學習已成為常用的基礎能力。為提升其效率，我們的VenusRL技術於強化學習框架中引入創新，大幅提升整體訓練效率。

VenusRL後訓練框架技術

後訓練乃大模型開發的關鍵階段。其涉及使用相對少量的標註數據對預訓練模型進行微調，以提升特定能力。隨著AI應用的迅速發展，基於人類反饋的強化學習(RLHF)已成為最廣泛採用的方法之一。以AI編程中的RLHF為例，工作流程通常遵循一個迭代循環，從標註數據訓練、推理以生成新數據、數據驗證及選擇，再到進一步訓練，從而不斷提升模型的編碼能力。於此過程中，最耗費資源及時間的階段為推理輸出的生成與驗證。由於通常需要外部工具以驗證及篩選生成的結果，工作流程涉及推理、結果驗證、結果未達標、更新輸入及重新推理的多個迭代循環，直至輸出符合預期標準為止。經驗證的輸出隨後被用作後續輪次的訓練數據，而該過程於不同任務中重複進行。因此，驗證階段需要大量隔離的執行環境，而用於推理的GPU於該等步驟中通常處於閒置狀態。該等因素共同導致後訓練工作流程的效率相對較低。

針對後訓練工作負載（其特徵為訓練與推理之間的反覆交替，且每次迭代的訓練時間相對較短），我們開發了VenusRL後訓練框架，作為旨在提高資源利用率的高效引擎。在我們的架構框架下，我們運用接口抽象化及調度編排技術，將整合的訓練及推理任務分解為可獨立編排及調度的較小工作流單元。透過預執行算法以及在線分析與追蹤技術，我們計算每個工作流階段的資源需求，並相應地進行細粒度的調度與編排。這使我們能夠將多項後訓練任務的訓練階段調度至相同的計算資源上，從而減少推理階段訓練資源的閒置產能，並提高整體資源利用率。

此外，推理階段的工作流程通常遵循一個迭代過程，即生成輸出、驗證輸出並優化結果以產生更佳的輸出。驗證通常使用CPU等通用算力資源進行，在此期間，分配予推理的GPU可能未獲充分利用。透過我們的異步輸入／輸出編排技術，我們能夠於驗證階段將優先級較低的計算任務調度至推理GPU，並於驗證完成後優先取代及移除該等任務。此舉能夠更高效地利用GPU算力資源，而不會阻礙推理任務。同時，就推理驗證過程中所用的隔離執行環境的管理而言，我們利用輕量且高效的VenusBox技術以提升環境創建及拆除的效率，從而縮短驗證時間。透過該兩項技術，與業內普遍使用的常規後訓練實施方法相比，我們的VenusRL架構將整體任務效率提升37%。

業 務

研發

我們的研發重點

於往績記錄期間，我們的研發主要側重於提升AI算力集群產品的技術能力。請參閱「— 我們的技術」。展望未來，我們擬繼續增加對研發的投入，以進一步提升我們一體化AI算力集群產品的技術能力。請參閱「— 策略 — 持續迭代AI算力集群產品，鞏固市場領先地位」。

我們已建立將工業應用與學術研究相結合的研發框架。通過研究與產品團隊的緊密合作，我們系統性評估新興技術需求，並在發現市場潛力時，將解決方案從概念開發推向商業部署。同時，我們積極與國內領先的硬件製造商及研究機構開展符合新一代硬件架構及性能要求的前瞻性聯合研究。該產學融合模式有助於加速創新並提高開發效率。例如，我們在八個月內完成自主研發交換機由核心研究至產品部署的完整開發周期，大幅縮短典型的行業時間線。此外，通過與學術夥伴的聯合研究計劃，我們縮短RDMA網卡擁塞控制算法的開發周期。我們亦參與國內大學、研究所及雲服務提供商的合作研究計劃，包括涉及智能模擬平台及開源操作系統的項目。該等舉措支持AI基礎設施技術的進步，同時加強我們的技術能力、創新效率及行業協作生態系統。

研發流程

我們已建立一個綜合研發框架，涵蓋需求評估、概念開發、架構設計、技術開發以及持續優化。

- **需求評估及技術評估：**我們系統地評估新興技術趨勢與市場需求，包括識別目標AI工作負載、性能基準及部署場景。我們的管理層及研發團隊會釐定優先次序。
- **核心研究及概念設計：**我們通過自主研發或與領先的硬件製造商、研究機構及學術合作夥伴進行具前瞻性的聯合研究，將研究見解轉化為初步技術概念及設計框架。於此階段，我們制定初步設計方案，並評估其於關鍵技術領域（包括高性能網絡、集合通信、調度算法及系統軟件）的可行性。該概念設計工作確立了基礎技術方向，藉以指導後續的研發工作。
- **架構設計及詳細規劃：**基於已評估的需求，我們界定整體技術架構。我們制定詳細計劃，以於整個研發流程中分配資源、設定里程碑及緩解風險。
- **技術開發及工程實施：**我們執行開發計劃並進行內部整合測試、調優及驗證。此舉包括嚴格的壓力測試及性能基準測試，以驗證正在開發的技術符合目標規格。
- **持續優化及迭代：**內部驗證後，我們將反饋及性能數據重新導入研發流程中，以推動持續的技術迭代、系統優化及能力提升。

業 務

我們的研發團隊

AI算力集群行業呈現多層級產業鏈架構，涵蓋芯片、硬件到集群調度及應用服務，每個層級涉及不同且複雜的技術要求。整個AI算力集群行業價值鏈技術迭代周期極短，且每個細分環節都築有極高的技術壁壘。我們擁有中國頂尖的AI基礎設施團隊之一，匯聚了諸多AI算力集群專家。我們在分佈式系統與計算機體系結構領域持續深耕超20年，研發團隊成員已經在國際頂級會議及期刊發表論文超過90篇。

截至最後實際可行日期，我們擁有一支由80名成員組成的研發及產品開發團隊，佔我們僱員總數的66.1%。我們的研發團隊在系統架構、軟件工程及高性能計算解決方案方面擁有紮實的教育背景及豐富的行業經驗，支持我們在AI算力集群領域的持續創新。

我們專注於通過實習計劃及校園招聘，構建及維持龐大的人才庫以推動研發工作。作為全球最前沿行業之一的參與者，該領域內具備直接相關經驗及專長的專業人士寥寥可數，因此我們高度重視通過實習計劃培養及發展自主人才庫。在招聘過程中，我們強調我們的技術雄心、工程挑戰及長期增長願景，以吸引與我們對創新懷有同樣熱情的候選人。

銷售及營銷

銷售

我們的AI算力集群產品及運營服務的技術實力及先進能力構成我們客戶獲取及擴展策略的基礎。通過提供高性能及可擴展的產品，滿足企業及公營機構不斷演變的計算需求，我們與客戶建立深厚的技術信譽及長期信任。我們透過招標方式取得若干項目。在對標桿客戶成功部署的基礎上，我們通過展示結果及可衡量的性能提升，繼續擴大市場滲透率，這不僅驗證我們所提供產品及服務的可靠性與競爭力，也可作為關鍵的示範案例，吸引相鄰領域的新客戶。

我們採用直銷模式，使我們的銷售及技術團隊能在整個銷售周期中（從技術諮詢及產品設計到部署及交付後支持）與客戶保持密切互動。在客戶互動及談判階段，我們的銷售團隊收集客戶規範並將其傳達給產品團隊，作為編製方案及改進產品的基礎。在整個項目部署過程中，駐場服務人員及技術專家與客戶保持溝通，以收集有關工作流程、客戶體驗及新功能需求的反饋。此類緊密且持續的互動，不僅能強化客戶黏性，更能提升我們發掘後續需求的能力，協助客戶從訓練階段拓展至推理與智能體應用場景，並創造持續性的商業機會。

定價

就AI算力集群產品而言，我們向客戶收取一次性固定費用，並於客戶驗收時確認收入。我們的AI算力集群產品定價基於多種因素，包括客戶工作負載要求的範圍、規模及複雜性、產品配置、交付模式、客戶狀況、特定硬件組件的採購成本及當前市況。

針對運維與技術支持服務，我們根據所維護的AI算力集群的規模、工時及所分配人力資源按年度向客戶收費。針對算力服務，我們根據提供的計算資源量向客戶收取固定費用。

業 務

營 銷

我們的市場擴張策略為資源驅動、口碑主導。在公司成立之初，我們憑藉管理團隊及投資者的廣泛校友及行業網絡，通過戰略合作夥伴的推薦及介紹，接觸到潛在客戶。在市場開發的早期階段，我們重點獲取成功部署後可作為標桿案例的代表性知名客戶。我們積極宣傳該等項目，以展示我們的技術能力及產品的可複製性，從而提升市場知名度及信譽。隨著客戶群不斷增長，我們已利用與標桿客戶的現有關係，尋求更廣泛的業務開發機會。通過該等已建立的網絡，我們能夠接觸到其他具類似AI計算需求的企業、政府實體及公共部門機構，促進規模化的客戶獲取及更深入的市場滲透。

客戶及客戶支持

客戶

我們的客戶主要包括大模型公司、雲服務提供商、教育與科研機構及電信運營商。

於2023年，我們為四位客戶提供了AI算力集群產品。於2024年及2025年，我們來自往績記錄期間各年度前五大客戶的收入分別為人民幣321.1百萬元及人民幣294.2百萬元，分別約佔我們有關年度總收入的98.9%及56.6%。於2023年、2024年及2025年，我們來自最大客戶的收入分別為人民幣15.5百萬元、人民幣191.5百萬元及人民幣86.3百萬元，分別約佔我們同期／年總收入的48.6%、59.0%及16.6%。我們給予客戶（包括於往績記錄期間各期間／年度我們的前五大客戶）的信貸期通常為30天以內。我們的客戶主要通過銀行轉賬向我們付款。

下表載列於往績記錄期間各期間／年度我們前五大客戶的詳情：

客戶	收入金額 (人民幣千元)	佔總收入 百分比 (%)	背景	購買的 產品及服務	業務關係 開始年份
<i>2023年2月20日至2023年12月31日期間</i>					
客戶A.....	15,460	48.6%	一家總部位於中國浙江省的私人公司，主要從事於提供軟件開發、雲計算及AI相關的技術服務及解決方案	AI算力集群產品	2023年
客戶B.....	15,419	48.5%	一家於香港聯交所上市的上市公司，主要從事於研發認知大模型，並提供相關技術支持及應用服務	AI算力集群產品	2023年

業 務

客戶	收入金額 (人民幣千元)	佔總收入 百分比 (%)	背景	購買的 產品及服務	業務關係 開始年份
客戶C.....	883	2.8%	一家於納斯達克交易所上市的上市公司，主要從事於運營數據中心、智能計算中心及提供綜合雲計算服務	AI算力集群產品	2023年
客戶D.....	40	0.1%	一家總部位於中國北京的私人公司，主要從事影像解決方案的研發、生產及銷售	AI算力集群產品	2023年
總計	31,802	100.0%			
<i>截至2024年12月31日止年度</i>					
客戶A.....	191,548	59.0%	一家總部位於中國浙江省的私人公司，主要從事於提供軟件開發、雲計算及AI相關的技術服務及解決方案	AI算力集群產品	2023年
客戶E.....	53,294	16.4%	一家於上海證券交易所上市的上市公司，主要從事於部署通訊系統及開發通訊技術，以及其一家附屬公司	AI算力集群產品及 AI算力集群運營 服務	2024年
客戶B.....	52,311	16.1%	一家於香港聯交所上市的上市公司，主要從事於研發認知大模型，並提供相關技術支持及應用服務	AI算力集群產品	2023年

業 務					
客戶	收入金額	佔總收入百分比	背景	購買的產品及服務	業務關係開始年份
	(人民幣千元)	(%)			
客戶F	13,983	4.3%	一家總部位於中國上海的私人公司，主要從事於提供數字技術服務、數據處理服務及數字園區建設與運營服務	AI算力集群產品	2024年
客戶G	9,942	3.1%	一家總部位於中國上海的私人公司，主要從事於提供數據處理及存儲支持服務、雲計算設備服務及AI相關技術服務	AI算力集群運營服務	2024年
總計	321,078	98.9%			
截至2025年12月31日止年度					
客戶E	86,258	16.6%	一家於上海證券交易所上市的上市公司，主要從事於部署通訊系統及開發通訊技術，以及其一家附屬公司	AI算力集群產品及AI算力集群運營服務	2024年
客戶H	63,717	12.2%	一家總部位於中國廣西壯族自治區的私人公司，主要從事於提供高端製造、新能源及租賃服務	AI算力集群產品	2025年

業 務

客戶	收入金額 (人民幣千元)	佔總收入 百分比 (%)	背景	購買的 產品及服務	業務關係 開始年份
客戶I.....	51,096	9.8%	一家總部位於中國江蘇省上市公司的子公司，主要從事於提供軟件及信息技術服務、雲計算設備銷售及技術開發服務	AI算力集群產品	2025年
客戶J.....	48,894	9.4%	一家上市公司的子公司，總部位於中國北京，主要從事於建設及運營清潔能源發電廠	AI算力集群產品	2025年
客戶K.....	44,284	8.5%	一家總部位於中國上海的私人公司，主要從事於提供技術推廣及應用服務、信息系統集成及軟件開發服務	AI算力集群產品及 AI算力集群運營 服務	2025年
總計	294,248	56.6%			

我們過往的客戶集中度主要歸因於AI算力集群市場的特徵及我們於發展初期的客戶開發策略。對AI算力集群的需求通常集中於領先市場參與者，例如大模型公司、雲服務提供商、教育與科研機構及電信運營商，其計算需求通常規模更大、配置更複雜，且更具備先進AI算力集群部署場景的代表性。自起步階段起，我們即採取專注服務標桿客戶的客戶開發策略，此乃由於我們相信，與該等客戶合作將使我們能夠應對最苛刻的應用場景，驗證我們的產品及技術能力，並為進一步擴展市場建立參考案例。

由於我們於2023年方成立，故於運營初期亦處於完善產品及加強技術與交付能力的階段。鑒於我們策略性專注於標桿客戶，故投入大量資源為少數領先客戶進行產品適配、服務優化及項目執行，而服務及部署能力仍相對有限。因此，於運營首兩年，我們僅服務相對少數客戶，導致客戶集中度較高。

業 務

鑒於AI算力集群市場的特徵，我們並不認為過往的客戶集中度屬不尋常，此已獲弗若斯特沙利文確認。我們的主要客戶通常為領先的大模型公司、雲服務提供商、教育與科研機構及電信運營商，隨著彼等的AI工作負載增加及其對計算效能、集群效率及運營穩定性的要求不斷提高，彼等對AI算力基礎設施的需求持續存在且不斷發展。中國的AI算力集群用戶在獲取高效能算力硬件方面經常面臨限制，且須在具有不同架構及生態系統的硬件資源中運作。在此背景下，AI算力集群的其中一項最關鍵能力在於有效連接及協調算力、網絡及存儲資源，成為一個統一且高效運作的系統。我們的AI算力集群產品及AI算力集群運營服務旨在透過實現集群級別互連、通訊效率及整體系統效能以滿足該等需求。因此，我們與該等標桿客戶的關係屬互惠互利：彼等為我們提供服務要求嚴格及具代表性的應用場景的機會，而我們的產品及服務支持彼等對高效能及有效AI算力基礎設施的持續需求，繼而使我們能自該等客戶產生經常性收入。

同時，我們一直逐步實現服務及客戶群多元化。我們服務標桿客戶的經驗已轉化為實質裨益，包括進一步增強產品及技術能力、提高市場認可度及擴展客戶群。AI算力集群產品客戶的數量由2023年的4名增加至2024年的10名，並進一步增加至2025年的48名。此外，我們一直擴充產品組合及運營模式以回應客戶需求。儘管我們於2023年初步僅提供AI算力集群產品，惟於2024年開始提供AI算力集群運營服務。雖然此兩條業務線的交付形式不同，但兩者均旨在使客戶能夠可靠及高效地獲取AI算力資源。服務範圍的擴充使我們能夠應對更廣泛的客戶要求及部署場景，亦有助客戶群多元化。AI算力集群運營服務客戶的數量由2024的2名增加至2025年的16名。因此，我們認為自身並無依賴任何特定客戶。

截至最後實際可行日期，據董事所知，概無董事或其各自的緊密聯繫人，且概無擁有已發行股份5%以上的任何股東於往績記錄期間各年度的任何前五大客戶中擁有任何權益。

與我們客戶的主要條款

以下為我們AI算力集群產品合同中的主要條款概要：

- **產品及服務：**我們提供涵蓋硬件設備供應、專有軟件系統部署及／或性能調優的全面產品及服務。服務範圍在客戶合同中明確規定。
- **付款。**客戶通常需要在簽訂合同後支付部分合同金額，餘下款項則根據服務交付里程碑進行支付。
- **交付及驗收：**我們負責將產品及服務交付至客戶指定的地點，並按要求進行安裝、測試及試運行。客戶根據協定程序進行驗收。
- **知識產權：**我們擁有自主研發軟件及系統的專有知識產權。我們並無擁有自第三方採購的硬件的相關知識產權。

業 務

- **責任：**如客戶未能按時付款，其應承擔相應的違約責任。如我們未能如期交付或產品或服務未達協定標準，我們將負責維修、更換或賠償。如遇不可抗力，任何一方均可免除責任。
- **終止：**任何一方可在發生重大違約、不可抗力或經雙方書面同意時，通過書面通知提前終止合同。終止並不影響在終止前產生的責任。

我們與AI算力集群運營服務客戶的合約通常涵蓋有關服務範圍、各方責任、付款條款及終止條文等條款。運維及技術支援服務通常包括檢查、故障診斷及故障排除。就算力服務而言，我們為客戶提供專屬AI算力資源，客戶於合約指定期限內對該等資源擁有獨家使用權。客戶須支付預付款項及定期付款。我們有責任按照合約訂明的指定服務範圍、效能基準及SLA目標交付AI算力集群運營服務。倘發生重大違約、另一方無力償債或進入清盤程序、服務期限屆滿或適用法律規定，任何一方均可透過書面通知提前終止合約。

售後服務

我們的銷售及技術團隊保持緊密合作，以確保高質量的客戶體驗。我們針對我們的產品及服務提供全面的售後服務。詳情請參閱「我們的產品與服務－我們的AI算力集群產品－我們的AI算力集群交付」。我們的AI算力集群產品的硬件組件的保修期通常為三年。我們亦指派專職技術支持專業人員提供持續的諮詢和技術協助，以確保無縫的客戶體驗。

於往績記錄期間，我們並未遭遇任何重大服務中斷、合同終止或與項目交付或售後服務相關的客戶糾紛。

供應商及採購

於往績記錄期間，我們的供應商主要包括為我們的產品提供集成硬件組件（主要包括算力硬件、網絡設備及存儲單元）的提供商，以及我們光模塊、RDMA網卡及自主開發的交換機的同製造商。於往績記錄期間，我們的所有供應商均位於中國。

供應鏈管理

對關鍵組件及服務供應商的挑選及管理，我們維持嚴格標準。我們的選擇標準注重技術能力、生產穩定性及質量保證標準，關鍵供應商必須維持相關的認證，並展示一致的質量管理體系。我們的供應商管理框架包括定期績效評估，重點關注質量指標、交付可靠性及技術能力，對於發現的任何問題需立即制定糾正措施計劃。

為了降低潛在的供應鏈中斷風險，我們已實施全面的風險管理制度。我們與眾多合資格供應商保持密切聯繫，並在可行的情況下，確保每個關鍵供應商類別都有至少一個備用供應商可供選擇。隨著我們業務的持續快速擴張，我們認為預先保持充足的庫存水平以滿足對產品日益增長的需求並確保供應的連續性至關重要。於往績記錄期間及直至最後實際可行日期，我們未曾與主

業 務

要供應商發生關係中斷或終止，也未曾因供應鏈問題而遭遇任何重大糾紛或業務中斷，且與主要供應商的協議條款也未發生重大變更。通過這種全面的方法，我們與供應商群體保持牢固的關係，同時確保持續獲得滿足業務需求所必需的高質量組件及服務。

硬件

當我們採用端到端的方式交付融合式AI算力集群產品（其可採取全集群部署的形式，或涉及選定硬件及軟件組件連同集群級網絡及融合服務的更具針對性的部署形式），我們可能需要向合資格供應商採購算力硬件、網絡設備及存儲單元，以將其整合至我們的產品中，從而支持該等部署。以下為我們與主要硬件供應商訂立的合同中的關鍵條款概要：

- **產品及服務：**供應商負責提供相關採購訂單中指定的產品。
- **定價及付款：**價格及付款安排按照合同及相關採購訂單確定。對於硬件採購，通常根據雙方協定的交付及驗收進度分期付款。對於若干硬件組件，亦可在簽訂合同後規定時間內全額付款，供應商僅在全額付款結清後方交付產品。
- **交付及驗收：**供應商須於協定時間內將貨物交付至指定地點。我們於收到貨物後進行檢查及驗收，以核實數量、外觀及功能。若出現不符合規格、損壞或性能問題，供應商負責維修、更換或重新交付。
- **質量保證及售後服務：**供應商必須確保其產品符合適用國家及行業標準，並於特定期限內提供保修支持。於保修期內，供應商承擔因質量問題引起的維修、更換及相關維護費用。
- **知識產權及保密：**所提供產品及相關技術資料的知識產權仍歸供應商或原始製造商所有。我們獲授予在協定範圍內使用該等產品的權利。雙方受有關專有及商業信息的保密義務約束，該等義務在協議終止後仍然有效。
- **終止：**在發生重大違約、不可抗力或導致無法履約的其他情況時，任何一方可通過書面通知提前終止協議。

合同製造

我們委聘第三方獨立合同製造商生產我們的自研交換機、光模塊及RDMA網卡。由於網絡系統對AI算力集群至關重要，因此必須針對我們的AI算力集群產品客製化特定網絡硬件。據此，我們並非直接向供應商採購標準化產品，而是委託代工製造商為我們生產此類網絡設備。在往績記錄期間，我們曾委託七家代工製造商。

在選擇我們的合同製造商時，我們通常與領先製造商合作，以確保我們產品的質量與可靠性。我們亦對合同製造商實施質量控制措施。於發貨前，產品須接受測試，並保存相應的測試記錄。例如，我們通常要求交換機的合同製造商完成一系列裝運前測試（包括功能測試、全端口壓力測試及高溫老化測試），並在整個過程中保留全面的測試記錄。

業 務

為保障我們的知識產權，我們已實施一系列合同措施。具體而言，我們與合同製造商訂立的合同載有知識產權及保密條款，規定彼等有責任對我們的專有技術予以保密，且不得以任何方式將有關資料用於相關合同允許以外的任何用途。以下為合同製造商訂立的合同中的主要條款概要：

- **產品及服務：**供應商負責按照我們的技術要求及規格製造及交付硬件產品，並提供必要的系統集成及技術支持服務。具體產品及服務要求由個別採購訂單確定。
- **定價及付款：**價格及結算安排於相關合同或採購訂單中指定。一般在驗收相關產品後的協定期限內通過銀行轉賬付款。
- **知識產權及保密：**供應商須確保其產品及服務中涉及的所有產品及技術不侵犯任何第三方的知識產權。雙方受合作過程中獲得的專有技術及商業信息保密義務約束。該等義務在協議終止或到期後依然有效。
- **質量保證及售後服務：**供應商必須確保產品符合適用的國家及行業標準，並在驗收後提供特定質量保修期。於保修期內，如出現產品質量問題，供應商負責維修、更換及相關費用。
- **技術支持：**供應商須於合作期間提供必要的技術支持及維護服務，以協助我們確保相關產品及系統的穩定運行。
- **終止：**任何一方在發生重大違約、破產、監管變動或其他導致無法履約的情況時，可在符合合同規定的通知要求下終止協議。

主要供應商

於2023年、2024年及2025年，我們向往績記錄期間各期間／年度前五大供應商作出的採購分別為人民幣47.1百萬元、人民幣254.8百萬元及人民幣284.7百萬元，分別約佔我們相應期間／年度採購總額的90.7%、75.5%及54.2%。於2023年、2024年及2025年，我們向往績記錄期間各期間／年度最大供應商作出的採購分別約為人民幣28.3百萬元、人民幣122.4百萬元及人民幣151.6百萬元，分別約佔我們相應期間／年度採購總額的54.5%、36.2%及18.8%。我們主要通過銀行轉賬向供應商付款。於往績記錄期間，供應商給予我們的信貸期一般介乎2天至60天。

業 務

下表載列於往績記錄期間各期間／年度我們前五大供應商的詳情：

供應商	採購額 (人民幣千元)	佔採購總額 百分比 (%)	背景	業務關係 開始年份	提供的產品及服務
<i>2023年2月20日至2023年12月31日期間</i>					
供應商A	28,271	54.5%	一間總部位於中國北京的私人公司，主要從事提供算力硬件、網絡設備及技術服務	2023年	網絡設備
供應商B	8,173	15.8%	一間總部位於中國北京的私人公司，主要從事通信、電信及網絡設備相關的業務及技術服務	2023年	算力硬件及網絡設備
供應商C	6,097	11.7%	一間總部位於中國上海的私人公司，主要從事通信技術、網絡技術及芯片技術相關的技術開發及服務	2023年	網絡設備

業 務

供應商	採購額 (人民幣千元)	佔採購總額 百分比 (%)	背景	業務關係 開始年份	提供的產品及服務
供應商D	2,798	5.4%	一間總部位於中國北京的私人公司，主要從事提供信息技術專業服務、信息系統集成及端到端數據中心實施服務	2023年	網絡設備
供應商E	1,714	3.3%	一間總部位於中國廣東省的私人公司，主要從事提供軟件及信息技術服務及相關技術開發及諮詢服務	2023年	網絡設備
總計	47,053	90.7%			
<i>截至2024年12月31日止年度</i>					
供應商A	122,375	36.2%	一間總部位於中國北京的私人公司，主要從事提供算力硬件、網絡設備及技術服務	2023年	網絡設備及相關 保修服務
供應商F	39,679	11.8%	一間總部位於中國北京的私人公司，主要從事AI算力基礎設施技術服務及信息系統集成	2024年	網絡設備

業 務

供應商	採購額 (人民幣千元)	佔採購總額 百分比 (%)	背景	業務關係 開始年份	提供的產品及服務
供應商C	36,908	10.9%	一間總部位於中國上海的私人公司，主要從事通信技術、網絡技術及芯片技術相關的技術開發及服務	2023年	網絡設備
供應商G	33,496	9.9%	一間總部位於中國廣東省的私人公司，主要從事提供軟件及信息技術服務以及相關技術開發及諮詢服務	2024年	網絡設備
供應商E	22,292	6.6%	一間總部位於中國廣東省的私人公司，主要從事提供軟件及信息技術服務及相關技術開發及諮詢服務	2023年	網絡設備
總計	254,750	75.5%			

業 務

供應商	採購額 (人民幣千元)	佔採購總額 百分比 (%)	背景	業務關係 開始年份	提供的產品及服務
<i>截至2025年12月31日止年度</i>					
供應商H.....	151,649	18.8%	一間總部位於中國浙江省的私人公司，主要從事雲計算技術的研發，以及提供雲計算、大數據及人工智能相關的產品及服務	2025年	算力硬件、 網絡設備、安全 軟件與雲服務
供應商I.....	132,036	16.4%	一間總部位於中國上海的私人公司，主要從事提供信息系統集成、網絡技術服務及通信設備銷售	2025年	算力硬件
供應商J.....	63,009	7.8%	一間總部位於中國廣東省的私人公司，主要從事算力中心的運營，以及提供算力租賃、算力調度及相關技術服務	2025年	算力硬件
供應商K.....	48,451	6.0%	一間總部位於中國浙江省的私人公司，主要從事圖數據庫技術的研發，以及提供圖計算及數據智能相關的解決方案及服務	2025年	算力硬件

業 務

供應商	採購額 (人民幣千元)	佔採購總額 百分比 (%)	背景	業務關係 開始年份	提供的產品及服務
供應商L	41,239	5.1%	一間總部位於中國廣東省的私人公司，主要從事提供智能物聯網解決方案、軟件開發及資訊系統集成服務	2025年	網絡設備
總計	<u>284,735</u>	<u>54.2%</u>			

於2023年及2024年，我們向五大供應商的採購分別佔我們相應年度採購總額的90.7%及75.5%。2023年及2024年供應商集中度相對較高，主要歸因於我們決定向信譽良好的供應商採購，以確保業務初期產品品質的一致性。隨著業務成長，我們逐步擴展供應商基礎，可從供應商集中度從2023年的90.7%下降至2025年的54.2%中反映。針對我們採購的產品及服務，中國內地有多家替代供應商可供選擇，因此我們認為自身並未依賴任何單一供應商。

據董事所知，概無董事、其各自的緊密聯繫人，且概無擁有已發行股份5%以上的任何股東，於往績記錄期間各年度的任何前五大供應商中擁有任何權益。

主要客戶與供應商之間的重疊

於往績記錄期間，我們五大客戶中的若干客戶於往績記錄期間各期間／年度亦為我們的供應商，我們五大供應商中的若干供應商於往績記錄期間各期間／年度亦為我們的客戶，詳情如下：

客戶E於2024年及2025年為我們的五大客戶之一，亦為2025年我們的供應商。向客戶E提供AI算力集群產品及AI算力集群運營服務所產生的收入分別佔我們2023年、2024年及2025年總收入的零、16.4%及16.6%。作為我們AI算力集群產品供應的一部分，我們亦採購並提供若干配套集群管理系統，以支持客戶的集群運行環境。該等採購來自客戶E，分別佔我們2023年、2024年及2025年採購總額的零、零及0.2%。

我們的若干主要供應商亦向我們採購。作為AI算力集群價值鏈的參與者，該等供應商亦對AI算力集群產品及運營服務有需求。根據弗若斯特沙利文的資料，主要客戶與供應商重疊的情況於AI算力集群行業中屬常見現象。供應商A於2023年及2024年為我們五大供應商之一，於2025年亦為我們的客戶。我們主要向供應商A採購網絡設備及相關保修服務，採購額分別佔我們2023年、2024年及2025年採購總額的54.5%、36.2%及1.9%。向供應商A出售AI算力集群產品所產生的收入分別佔我們2023年、2024年及2025年總收入的零、零及7.4%。供應商C於2023年及2024

業 務

年為我們五大供應商之一，於2025年亦為我們的客戶。我們主要向供應商C採購網絡設備，採購額分別佔我們2023年、2024年及2025年採購總額的11.7%、10.9%及0.005%。向供應商C出售AI算力集群產品所產生的收入分別佔我們2023年、2024年及2025年總收入的零、零及0.003%。供應商E於2023年及2024年為我們五大供應商之一，於2025年亦為我們的客戶。我們主要向供應商E採購網絡設備，採購額分別佔我們2023年、2024年及2025年採購總額的3.3%、6.6%及4.7%。向供應商E出售AI算力集群產品所產生的收入分別佔我們2023年、2024年及2025年總收入的零、零及1.8%。

我們的董事確認：(i)該等重疊的客戶與供應商關係為日常及一般業務過程中的正常商業安排；(ii)該等交易的條款屬公平合理；(iii)我們向該等重疊客戶與供應商進行銷售及採購的條款乃按個別基準進行磋商，且銷售及採購既非相互關聯亦非互為條件；及(iv)該等安排並未對我們的運營獨立性、定價機制或業務安排產生任何不利影響。我們並不依賴該等重疊關係。

數據隱私及安全

於往績記錄期間及直至最後實際可行日期，我們的用戶是企業而非個人，我們的產品部署在客戶本身的本地基礎設施上，並完全在客戶控制的環境中運行，所有數據處理活動均在客戶的獨家管理下進行。鑒於我們的核心業務側重於產品交付及必要的技術支持，且我們的產品以私有化方式部署在客戶端，我們不會收集、存儲或使用客戶在使用已交付產品時輸入或生成的數據。我們在業務運營過程中收集及生成的數據主要包括：我們自主研發的源代碼；我們制定及交付的技術方案；我們內部的財務數據及人力資源信息；及客戶聯繫方式。

數據保護及網絡安全是我們的首要任務之一。就此而言，我們設計嚴格的數據保護及網絡安全政策，以確保嚴格遵守適用法律、法規及普遍行業慣例。我們實施關於保護數據隱私及安全的全面內部政策，旨在確保數據及信息安全、優化數據治理、保護客戶、僱員及其他第三方的利益，並確保在重大方面遵守所有適用法律及法規。

我們實施強大的內部認證及授權系統，以確保我們的機密數據僅供授權人員作授權使用。我們已制定明確且嚴格的授權及認證程序及政策。我們的僱員僅能訪問與其職責直接相關且必要的數據，且目的有限，並須在每次嘗試訪問時驗證授權。

我們的核心產品具有多維度的安全保護體系，包括：(i)基於角色的訪問控制機制，為不同最終用戶分配不同權限，嚴格限制其對客戶本地資源的訪問及操作，從而確保最終用戶數據及系統的安全；(ii)支持客戶系統與第三方認證系統之間的集成，以實現統一的最終用戶身份管理；對最終用戶數據進行嚴格的安全保護，支持加密數據存儲及傳輸，以防範數據洩露風險；(iii)限制集群網絡訪問權限，僅允許對外暴露必要的服務端口，以防範外部惡意攻擊；支持配置網絡白名單以進一步增強網絡安全；(iv)通過多租戶隔離架構實施用戶級權限注入及策略控制，結合資源配額的軟隔離及調度分區的硬隔離，建立多維、細粒度的資源安全邊界。

於往績記錄期間及直至最後實際可行日期，(i)我們未曾收到任何第三方根據任何適用法律法規以侵犯其數據及隱私保護權利為由向我們提出的任何申索；(ii)亦無主管政府部門或第三方

業 務

針對我們提出有關網絡安全、數據及個人信息保護的未決或潛在的調查或其他法律程序；及(iii)我們並無經歷任何與網絡安全、數據保護及個人信息保護法律法規相關的重大個人信息洩露事件。

基於上述情況以及(i)目前我們的主要業務不直接獲取或存儲客戶數據；(ii)我們實施有關保護網絡安全、數據安全及個人信息的內部政策；及(iii)我們將持續密切關注網絡安全、數據保護及個人信息保護方面的立法及監管發展，與相關政府部門保持持續溝通，並及時實施一切必要措施以確保確保持續遵守相關法律法規，我們的中國數據合規法律顧問認為，於往績記錄期間及直至最後實際可行日期，我們已在重大方面按照相關網絡安全、數據及個人信息保護法律法規的要求實施有關網絡安全、數據保護及個人信息保護的合規措施。然而，由於網絡安全、數據保護及個人信息保護方面的法律法規仍在發展中，我們無法保證我們能始終及時適應有關法律法規的所有方面。請參閱「風險因素－與我們業務及行業有關的風險－我們的業務受有關隱私及數據保護的各種法律、規則、政策及其他義務所規限。機密信息及個人數據的任何丟失或未經授權的訪問或發佈均可能使我們面臨重大的聲譽、財務、法律及運營後果」。

競爭

我們在中國AI算力集群行業面臨激烈競爭。我們行業的主要競爭因素包括產品的技術能力、功能、範圍及性能、服務的可擴展性及可靠性、交付專長、營銷及銷售能力、客戶體驗、定價、品牌認知度及聲譽。此外，新技術的出現及強化以及新市場進入者可能進一步加劇我們行業的競爭。我們相信我們在基於上述因素的競爭中處於有利地位。請參閱「－競爭優勢」。

然而，部分競爭對手可能擁有更高的品牌認知度、更長的經營歷史、更龐大的客戶群以及更充裕的財務、技術及其他資源。請參閱「風險因素－與我們業務及行業有關的風險－若我們未能有效競爭，我們的業務、財務狀況及經營業績可能受到重大不利影響」。有關我們行業競爭格局的更多資料，請參閱「行業概覽」。

知識產權

知識產權對於我們業務的基礎至關重要。我們目前持有多項與我們核心技術相關的知識產權，且我們投入大量時間及資源用於知識產權的開發及保護。我們依賴中國專利、商標、版權、域名、商業秘密及其他專有權利保護法律以及保密程序及合同規定的組合保護我們的知識產權。

截至2025年12月31日，我們在中國擁有七項註冊專利、39個商標、44項軟件著作權及一個域名。有關我們重大知識產權的詳情，請參閱「附錄四－法定及一般資料－有關我們業務的進一步資料－知識產權」。

我們已盡最大努力確保遵守適用的知識產權法律。我們的董事確認，於往績記錄期間及直至最後實際可行日期，我們並未涉及第三方提出的任何知識產權侵權訴訟。我們的董事進一步確

業 務

認，我們並未涉及任何可能單獨或共同對我們的業務、財務狀況及經營業績產生重大不利影響的知識產權侵權訴訟。請參閱「風險因素－與我們業務及行業有關的風險－我們或會面臨第三方提出的索賠，指控我們侵犯其知識產權」。

我們通過戰略規劃保護我們的知識產權，例如提交知識產權註冊及申請、防偽機制（特別針對商標及外觀設計）及防禦任何現有或潛在知識產權侵權的訴訟機制。儘管我們付出努力，但第三方仍可能未經授權獲取並挪用我們的知識產權。截至最後實際可行日期，我們並未發現任何該等挪用我們知識產權的情況。然而，第三方未經授權使用我們的知識產權及為保護我們知識產權所產生的費用可能對我們的業務及經營業績產生不利影響。有關相關風險的詳情，請參閱「風險因素－與我們業務及行業有關的風險－我們的知識產權對我們的成功至關重要，任何第三方侵犯我們的知識產權可能對我們的業務、財務狀況及經營業績產生重大不利影響」。

僱員

截至最後實際可行日期，我們擁有121名僱員。下表載列截至最後實際可行日期按職能劃分的僱員明細。我們所有僱員均位於中國。

職能	僱員人數	佔總數百分比
研發	80	66.1
銷售及營銷	16	13.2
一般及行政	25	20.7
總計	121	100.0

我們的成功取決於我們吸引、挽留及激勵合資格人員的能力，且我們相信高質量的人才庫是我們公司的核心優勢之一。我們在招聘中採取高標準及嚴格程序，包括校園招聘、線上招聘、內部推薦及高管獵頭招聘，以滿足我們對各類人才的需求。

按中國勞動法律的要求，我們與僱員訂立個人僱傭合同，涵蓋工資、獎金、僱員福利、工作場所安全、保密義務、不競爭及終止理由。為遵守中國法規，我們參與由適用地方市級及省級政府組織的各種僱員社會保障計劃，包括住房、養老、醫療、工傷及失業福利計劃。

截至最後實際可行日期，我們並無建立工會。我們相信，我們與僱員保持良好的工作關係，於往績記錄期間，我們並無經歷任何重大勞資糾紛或在招聘業務人員方面遭遇任何困難。

保險

我們認為我們的保險範圍足夠，因為我們已按中國法律法規的要求並根據我們行業的商業慣例維持所有強制性保單。我們與僱員相關的保險包括中國法律法規要求的養老保險、生育保險、失業保險、工傷保險及醫療保險。

業 務

與一般市場慣例一致，我們概無維持任何業務中斷保險或產品責任險，在中國法律下該等保險並非強制性。我們不維持關鍵人員人壽保險、涵蓋我們的網絡基礎設施或信息技術系統損壞的保單或針對我們財產的任何保單。請參閱「風險因素－與我們業務及行業有關的風險－我們的投保範圍有限，任何超出我們投保範圍的索賠均可能導致我們產生大量費用及分散資源。」於往績記錄期間，我們並無提出任何與我們業務相關的重大保險申索。

環境、社會及管治事宜

ESG管治

我們已制定全面的環境、社會及管治（「**ESG**」）政策，概述我們的環境舉措、社會責任承諾及內部管治框架。**[編纂]**後，我們將遵循ESG報告要求，並按照上市規則附錄C2的要求每年刊發ESG報告。

董事會對我們的ESG管治負有首要責任，負責監督重大ESG及氣候相關事宜、目標、策略、風險及機遇的識別、評估及批准，並確保ESG數據及對外披露的完整性。管理團隊負責監督日常ESG實施工作，並向董事會提供有關重大ESG事宜的年度更新，同時主動監控該等風險及機遇如何影響業務營運、財務表現及可持續發展承諾。

環境責任

空氣污染物及廢物管理

我們空氣污染物排放的主要來源為日常辦公活動所產生的廢物，我們的其他環境影響則主要包括購買電力所產生的溫室氣體排放。於往績記錄期間，我們並無直接溫室氣體排放。

我們所產生的廢物分為非危險廢物及危險廢物。非危險廢物主要包括一般辦公廢物，而危險廢物則包括廢電池、燈管及墨盒。我們已實施一系列措施以減少營運過程中所產生的廢物，包括預設雙面打印及節墨打印設定、優先選用可回收或可重複使用的消耗品、防止辦公用品過量囤積，以及鼓勵無紙化工作流程。

資源消耗

我們的能源消耗主要為辦公營運所購買的電力及熱力。為提高資源效率，我們定期維修空調系統，以盡量減少製冷劑洩漏，並在人流較少的區域限制辦公室照明。員工離開辦公室時須關閉空調、照明及電子設備。水消耗主要歸因於日常辦公活動，且水源的獲取符合所有適用的國家及地方法規。節水措施包括安裝紅外線感應水龍頭及節水裝置、降低水壓及定期檢查水錶以檢測及防止隱蔽洩漏。我們始終致力於提高資源效率。

業 務

下表載列我們於往績記錄期間的能源及水消耗：

	單位	2023年2月20日 至2023年12月 31日期間	截至12月31日止年度	
			2024年	2025年
總能源消耗量	兆瓦時	46.89	134.83	218.67
直接能源消耗量	兆瓦時	—	—	—
間接能源消耗量	兆瓦時	46.89	134.83	218.67
強度 — 總能源消耗量	兆瓦時／百萬元 人民幣收入	1.47	0.42	0.42
總用水量	立方米	191.63	369.56	1,478.25
強度 — 用水量	立方米／百萬元 人民幣收入	6.03	1.14	2.84

應對氣候變化

我們認識到氣候變化所帶來的風險。為最大程度降低日常營運中可能產生的環境及氣候相關風險，我們致力於在全集團推行綠色實踐。我們嚴格遵守中國規管空氣排放、廢水排放及廢物處置的適用環境法律法規。我們的環境管理策略以量化數據為依據，包括監測溫室氣體排放及資源消耗情況。

下表載列我們於往績記錄期間的溫室氣體排放：

	單位	2023年2月20日 至2023年12月 31日期間	截至12月31日止年度	
			2024年	2025
溫室氣體排放總量 (範疇二+範疇三)	噸二氧化碳當量	30.00	148.05	274.31
範疇二能源間接溫室氣體 排放量 ⁽¹⁾	噸二氧化碳當量	22.54	60.70	100.67
範疇三其他間接溫室氣體 排放量 ⁽²⁾	噸二氧化碳當量	7.46	87.35	173.64
強度 — 溫室氣體 排放總量	噸二氧化碳當量／ 百萬元人民幣收入	0.94	0.46	0.53

註：

- (1) 範疇二間接排放主要包括本公司使用所購買電力及熱力而產生的溫室氣體排放。
- (2) 範疇三其他間接排放包括商務差旅以及水及廢水處理用電所產生的溫室氣體排放。

我們已就四個關鍵方面制定一系列環境目標。就溫室氣體排放而言，我們以2023年為基準年，目標是到2030年將總溫室氣體排放強度降低5%。在耗電量方面，我們同樣承諾以2023年為基準年，到2030年將耗電強度降低5%。在水消耗方面，目標是以2023年為基準年，到2030年將水消耗強度降低5%。最後，就廢物產生而言，我們擬對辦公用品供應商嚴格實施辦公用品登記制度，並致力於減少辦公廢物的產生。

業 務

僱員責任

僱員是我們成功的基石。我們認可僱員在推動可持續增長中的重要作用，致力於在完善的制度政策及健全的管理體系支撐下，維持公平、安全及包容的工作環境。

我們嚴格遵守所有適用勞動法律法規，且我們的僱傭做法已正式載入標準僱員手冊。我們與每位團隊成員均簽訂正式僱傭合同，並在整個僱傭生命週期的每個階段均恪守不歧視、公平對待及機會均等原則。我們定期舉辦合規培訓課程。我們致力於構建多元化僱員團隊，截至2025年12月31日，35.2%的僱員為女性。

我們將職業健康與安全放在首位。所有新僱員及調任新職位的僱員均會接受全面的安全培訓，內容涵蓋安全規範及緊急應變程序。我們在辦公場所定期舉行消防演習。我們實行健全的職業健康計劃，其中包括定期體檢及意外保險保障，以保護僱員的整體福祉。於往績記錄期間，並無發生與本集團業務運營有關的重大職業健康安全事件、意外或索賠。

社區參與

我們致力於在經營所在社區創造正面影響。我們通過積極參與及有意義的貢獻，尋求支持當地發展，解決社會需求。展望未來，我們計劃鞏固當地合作關係、加強支持計劃，併發掘新的機會，深化社區參與，為週邊社區的福祉及可持續發展作出貢獻。

反腐敗

我們堅定致力於反腐敗及誠信經營行為，並以包括反欺詐、反賄賂及反洗錢管理制度在內的重要內部政策為基礎建立健全的合規架構，嚴禁賄賂、侵吞公款及未經授權披露機密資料等行為。我們要求所有僱員於入職時簽署反腐敗合規承諾書，同時設立保密舉報機制，讓舉報不當行為的人毋須擔心遭受打擊報復。

盈利路徑

於往績記錄期間，我們實現了顯著的業務增長。我們的收入由2023年的人民幣31.8百萬元增加至2024年的人民幣324.7百萬元，並進一步增加至2025年的人民幣520.3百萬元，年複合增長率為304.5%。我們於2024年錄得利潤人民幣11.5百萬元，而我們於2023年及2025年分別產生虧損人民幣2.2百萬元及人民幣355.6百萬元。作為一家於2023年成立的公司，我們於該年處於業務營運的起步階段，積極探索並完善我們的業務模式及產品供應。我們於2023年的虧損主要歸因於為奠定我們的技術基礎及產品供應而必需在研發活動方面作出的大量初期投資。自此，我們實現了顯著的收益增長，這得益於我們技術能力及產品與服務供應不斷成熟，成功向標桿客戶交付即為明證。這一堅實的基礎使我們能夠利用有利的市場順風，在人工智能行業快速發展的背景下，

業 務

客戶對我們的AI算力集群產品及AI算力集群運營服務的需求顯著增加。我們於2025年的淨虧損主要是由於附帶優先權股份的金融負債公允價值變動所致。於2024年及2025年，我們分別錄得經調整淨利潤（非國際財務報告準則計量）人民幣20.3百萬元及人民幣31.1百萬元。我們的附優先權股份將於[編纂]時自動轉換為普通股，而相關金融工具將由負債重新分類為權益。

展望未來，我們擬透過多項戰略舉措達致並維持盈利能力。我們將繼續完善我們的技術能力及進行產品迭代，以維持我們的市場地位並滿足不斷變化的客戶需求。我們計劃尋求針對性的垂直整合，以期提升產品交付效率、加強全流程質量控制並支持長期的成本優化。我們亦將繼續執行我們的銷售及營銷策略，重點為服務標桿客戶，同時透過展現我們能力的成功示範交付來擴展我們的客戶群。此外，我們計劃進一步鞏固我們的供應鏈地位並改善我們的供應鏈管理，使我們能夠維持具競爭力的成本結構。我們亦旨在透過在整個營運過程中採用數字化系統、自動化工具及標準化流程，以提升成本及營運效率。

物業

截至最後實際可行日期，我們通過七處租賃物業經營我們的業務，該等物業位於中國北京、上海、杭州及南京，總建築面積超過1,751.1平方米。該等物業主要用作我們的辦公設施。所有該等物業均已用於《香港上市規則》第5.01(2)條定義的非物業活動，並主要用作我們業務營運的辦公場所。

我們有關上述七處租賃物業的租賃協議的到期日通常介乎2026年5月20日至2028年11月30日。我們計劃在現有租約到期後續簽租約或談判新條款。所有出租人均為獨立第三方。於往績記錄期間及直至最後實際可行日期，我們在與業主磋商重續租約方面未曾遇到任何重大困難。我們相信中國有充足的物業供應。

根據中國適用法律法規，物業租賃協議必須向中華人民共和國住房和城鄉建設部地方分局辦理登記。該等租賃登記須由我們的出租人配合辦理。截至最後實際可行日期，我們的全部七處租賃物業尚未辦理租賃登記，主要原因為難以促使出租人配合辦理該等租賃登記。我們將採取一切實際可行及合理的措施，確保為該等租賃辦理登記。據我們的中國法律顧問告知，租賃協議未辦理登記不會影響該等租賃協議的效力。根據中國相關法律法規，有關政府部門或會責令我們於指定期限內辦理有關租賃協議登記，若未能辦理，我們或會就每項未登記租賃被處以人民幣1,000元至人民幣10,000元的罰款。截至最後實際可行日期，我們並未收到有關政府部門的任何相關要求，亦未遭受該等部門處以任何相關罰款。我們承諾，一旦收到有關政府部門的任何要求，將全力配合以促使租賃協議辦理登記。

截至最後實際可行日期，我們三處租賃物業的出租人未能提供房產證，這可能對我們日後繼續使用該等物業造成不利影響。該等業權存在瑕疵的物業的總租賃面積約950.2平方米。倘因該等物業存在產權負擔或政府行為而出現糾紛，我們或會在繼續租賃該等物業方面遇到困難，並可能需要搬遷。我們預期在附近物色可比較的替代物業及將業務搬遷至該等物業，不會耗費大量時間或產生重大開支。董事認為，搬遷不會對我們的業務、經營業績及財務狀況造成重大不利影響。

截至最後實際可行日期，我們並不知悉任何第三方或政府部門就該等租賃物業的業權提出任何可能對我們現時佔用造成重大不利影響的質疑。此外，如上文所述，董事預期就上述任何業

業 務

權存在瑕疵的物業物色可比較的替代場所，不會產生任何重大實際困難或重大開支。並無任何規則或法規規定承租人須取得業權證書，亦無規定承租人未取得該等證書會受到監管處罰。因此，我們的中國法律顧問已告知，我們作為承租人，不會因出租人未能履行糾正上述租賃物業的業權瑕疵義務而遭受任何重大行政處罰。此外，根據中國相關法律法規，若因出租人過錯導致租賃協議無效，承租人有權要求賠償。倘我們繼續租賃該等物業的能力受到第三方異議影響，我們可根據中國相關法律法規向出租人追討賠償。因此，董事認為該等業權瑕疵不會對我們的業務、經營業績及財務狀況造成重大不利影響。

截至最後實際可行日期，我們用於業務經營的一處租賃物業，其實際土地用途與土地使用權證載明的批准土地用途不符。若該物業的業主被政府部門責令整改該等土地用途，我們可能須搬遷，並承擔搬遷成本及其他額外開支。截至最後實際可行日期，我們並未獲悉政府部門發出任何該等整改要求。據我們的中國法律顧問告知，根據中國相關法律法規，因出租人未能履行確保實際用途與批准用途一致的責任而須接受行政處罰或罰款者為出租人，而非承租人。我們不會就此受到任何處罰，但可能無法繼續租賃該等物業。

截至2025年12月31日，由我們租賃或擁有的物業均無佔我們綜合總資產15%或以上的賬面值。因此，根據《香港上市規則》第5章及《公司（豁免公司及招股章程遵從條文）公告》（香港法例第32L章）第6(2)條，本文件獲豁免遵守《公司（清盤及雜項條文）條例》第342(1)(b)條有關《公司（清盤及雜項條文）條例》附表三第34(2)段（其要求就本集團在土地或建築物中的所有權益出具估值報告）的規定。

法律程序及合規

於往績記錄期間及直至最後實際可行日期，我們或我們的任何董事並無涉及任何我們認為會對我們的業務、經營業績、財務狀況或聲譽及合規產生重大不利影響的實際或未決法律、仲裁或行政程序（包括任何破產或接管程序）。

於往績記錄期間及截至最後實際可行日期，我們並無任何違反法律及法規的重大不合規行為，且我們並無經歷任何系統性不合規事件。我們的董事認為，於往績記錄期間及直至最後實際可行日期，我們在所有重大方面均已遵守我們營運所在司法管轄區的所有相關法律法規。

業 務

牌照及許可

我們的中國法律顧問已告知，我們已自中國相關政府部門取得在所有重大方面進行我們業務所需的所有牌照、許可及證書，且該等牌照、許可及批准截至本文件日期仍然有效。我們亦預期未來在重續該等牌照、許可及證書時不會遇到任何重大困難。下表載列目前由我們持有的重大牌照、許可及批准：

持有人	牌照／許可	頒發機構	簽發日期	屆滿日期
本公司	電信設備進網許可	中華人民共和國工業和信息化部	2025年4月3日	2028年4月3日
本公司	電信設備進網許可	中華人民共和國工業和信息化部	2026年4月14日	2029年4月14日
上海基流光聯 . .	電信設備進網許可	中華人民共和國工業和信息化部	2025年10月31日	2028年10月31日
上海基流光聯 . .	增值電信業務經營許可證	中華人民共和國工業和信息化部	2026年4月16日	2031年4月16日

風險管理及內部控制

我們的董事會負責確保我們維持健全且有效的內部控制及風險管理系統，以保障我們股東的投資及我們的資產。我們的風險管理系統旨在：(i)識別可能影響我們的潛在事件，確保相關風險控制在與我們目標相符的可接受且適當的水平內；(ii)促進對內部及外部利益相關者準確可靠的信息溝通；(iii)確保我們的業務運營合規；(iv)提升我們業務活動的效率及效力，從而減少實現運營目標的不確定性；及(v)為重大風險建立危機管理計劃，保護本公司免受因災難性風險或人為錯誤造成的巨大損失。

我們的審計委員會負責審查與風險管理相關的規章及主要目標，向董事會提交年度綜合風險管理報告，審查重大風險的風險管理戰略及解決方案，並處理董事會授權的與綜合風險管理相關的其他事項。我們的內部審計部門的任務是實施我們的風險管理政策及系統。其他部門及業務單位在進行風險管理工作時由內部審計部門進行監督。

我們風險管理工作的一般操作程序包括：(i)收集與風險管理相關的信息，包括歷史數據及未來預測。具體而言，在戰略風險方面，我們廣泛收集國內及國際公司因未能控制戰略風險而遭受損失的案例；在財務風險方面，我們收集由失控財務風險引發的財務危機案例；在市場風險方

業 務

面，我們收集公司因忽視市場風險及缺乏應對措施而遭受損失的案例；在運營風險方面，我們收集與我們及行業相關的信息；而在法律風險方面，我們收集公司因忽視法律及監管風險且缺乏應對措施而遭受損失的案例；(ii)評估風險的嚴重性及可能性（或頻率及概率）；(iii)高級管理層協助董事會根據業務特徵確定統一的風險偏好及容忍度，並建立風險警戒線及相應的應對措施；(iv)高級管理層根據風險管理戰略為重大風險制定風險管理解決方案；及(v)高級管理層監督風險管理的實施，並根據變動及任何不足之處及時作出改善。我們的方法亦涉及持續完善內部控制機制，並定期審查我們的風險管理政策及內部控制措施以維持其有效性及充分性，確保我們的營運符合行業標準及監管要求。

獎項及表彰

下表載列截至最後實際可行日期我們獲得的主要獎項及表彰。

年份	獎項／表彰	頒發機構／部門
2025年	未來網絡領先創新科技成果「自主可控智算通信系統」	中國通信學會未來網絡領先創新科技成果專家委員會
2024年	大規模智算中心AI雲平台示範案例	中國通信標準化協會算網融合產業及標準推進委員會
2024年	中國算力卓越企業獎	中國智能計算產業聯盟
2024年	創新引擎獎	DPU與AI網絡大會委員會
2023年	中國算力先進技術獎	中國智能計算產業聯盟