

技術詞彙表

於本文件中，除文義另有所指外，本文件所用有關本公司及我們業務的若干詞彙的解釋及定義具有下文所載含義。這些詞彙及其含義可能並非總是符合標準行業含義或這些詞彙的用法。

「智能體 AI」	指	一類AI系統，可在極少的人為干預下自主感知環境、做出決策並執行多步驟任務，通常會協調多個AI智能體來實現複雜目標
「AgentiCTRL」	指	我們專為先進AI智能體訓練打造的強化學習平台，深耕基於Transformer的大模型訓練場景，為智能體AI的研究與部署提供全面的工程基礎設施
「AI」	指	人工智能，計算機科學分支之一，開發能夠執行感知、學習、推理和決策制定等典型人類智能型任務的系統
「AI智能體」	指	一種使用AI技術代表用戶或其他系統自主執行任務並達到目標的系統或程序
「AIDC」	指	智算中心，是一種專為承載AI業務的海量算力需求而設計和優化的專用數據中心。AIDC高度依賴於服務器，以支持AI模型的訓練和AI推理任務的運行
「AIFS」	指	AI基礎軟件，即定制化的AI軟件解決方案，提供工具和平台，協助客戶開發、訓練和部署AI模型，從而優化內部工作流程
「AI基礎設施」	指	開發、部署和管理AI應用程序及解決方案所需的硬件、軟件、工具和流程的組合，構成支持AI開發與運維的基礎層
「Alaya Lab」	指	我們專為AI研究人員和工程師打造的一站式AI開發平台，提供了一個貫穿AI項目完整生命週期的統一開發環境，涵蓋從數據準備、模型構建到調試和部署的各個環節

技術詞彙表

「API」	指	應用程序編程接口，一套用於實現不同軟件應用程序之間交互通信的規範協議、程序及工具。就AI而言，API為外部應用程序提供標準化入口，可直接調用AI服務或模型，使用者毋須了解底層算法或架構
「CaaS」	指	容器即服務，是一種雲服務模式，允許用戶通過基於容器的虛擬化技術來管理和部署容器、應用程序及集群，通常構建在Kubernetes等容器編排平台上
「算力池」	指	我們透過Alaya NeW OS調度及營運以提供智算雲服務的總算力，包括(i)租賃自我們交付予智算基礎設施解決方案客戶的AIDC的算力、(ii)租賃自我們根據合作安排管理的第三方AIDC的算力，及(iii)由我們自有的服務器產生的算力
「可控算力」	指	我們直接控制的算力，包括我們自有的服務器算力及根據合約安排取得的算力，並由我們負責其架構規劃、系統部署及持續調度
「集群」	指	一組相互連接的服務器，透過共享計算資源共同作為一個統一系統運作
「CPU」	指	中央處理器，計算機中進行大部分處理的主要元件
「DCU」	指	算力計量單位，是我們自研的算力計費和計量標準，界定為相當於持續一小時312 TFLOPS的FP16算力。該標準消除了底層硬件差異，實現了算力的標準化計量、交易和計費
「已交付的算力」	指	我們通過智算基礎設施解決方案向客戶交付的AIDC算力
「DingoFS」	指	一個完全開源混合存儲系統，專為滿足AI工作負載的大規模並行讀寫需求而設計，支持分鐘級檢查點優化，並能高效存儲和訪問艾字節級數據
「DingoStack」	指	我們基於OpenStack的專有基礎設施平台，可為大規模智算環境優化第三層轉發、路由性能及網絡管理能力

技術詞彙表

「DKS」	指	專屬容器計算服務是我們的一項雲服務，為企業客戶提供對計算集群的專屬訪問權限，確保性能隔離，並滿足大規模、高強度分佈式AI模型訓練任務的穩定性要求
「EFLOPS」	指	ExaFLOPS，一種計算效能單位，代表每秒完成一百京(10^{18})次浮點運算，指衡量大型智算中心及高性能計算系統整體處理能力時常用的最高數量級。其相當於1,000 PFLOPS
「微調」	指	利用專屬小數據集對預訓練模型進行微調，使其能適配特定任務或特定應用場景，與從頭開始訓練相比，這種方法算力成本相對更低，可快速實現模型的專業化適配
「FLOPS」	指	每秒浮點運算次數，為衡量處理器或計算系統運算效能的指標，表示每秒可執行的浮點算術運算次數
「FP16」	指	16位浮點格式，一種用於AI模型訓練與推理的數值表示標準，與通常用於加速深度學習工作負載的32位浮點(FP32)相比，此格式能減少內存使用量並提高計算吞吐量
「生成式AI」	指	AI的一個分支領域，專注於基於從現有數據習得的模式生成文本、圖像、視頻或音頻等新內容
「極核訓練」	指	我們基於Kubernetes的serverless分佈式訓練服務，旨在通過將基礎設施管理、資源調度及分佈式框架適配封裝到一個統一的核心任務引擎中，解決企業級AI訓練在運算及管理方面的複雜性
「IDC」	指	互聯網數據中心，一種專門的物理設施，用以容納大量的計算機系統、網絡設備及相關基礎設施，如電源、冷卻系統和安全系統。現時IDC的定義更廣泛，包括傳統基於CPU的IDC及AIDC
「IDE」	指	集成開發環境，是為開發者提供用於編寫、編輯、調試和測試代碼統一界面的綜合軟件應用程序

技術詞彙表

「Kubernetes」	指	一個開源的容器編排平台，可自動在服務器集群中部署、擴展和管理容器化應用程序，並提供資源調度、負載均衡和自愈功能
「大語言模型」	指	大語言模型，一種針對大規模文本數據進行訓練的深度學習模型，用於理解、處理和生成人類語言。其專注於自然語言任務，為多個基於語言的AI應用程序的基礎
「MaaS」	指	模型即服務，一種基於雲的服務，可通過API提供預訓練機器學習模型訪問權限，使企業無需具備模型開發、訓練或基礎設施管理方面的大量內部專業知識即可部署AI應用程序
「MFU」	指	模型FLOP利用率，這是一項衡量AI模型訓練或推理過程中計算系統效率的指標，通過對比硬件實際算力吞吐量與理論峰值算力得出，結果以百分比形式表示
「ModelOps」	指	模型運維，即一套貫穿人工智能模型管理全生命週期的執行規範、工具及流程，涵蓋模型開發、訓練編排、版本管理、部署上線、運行監控與迭代優化各環節，旨在提升人工智能模型落地生產環境的效率與可靠性
「PFLOPS」	指	每秒千萬億次浮點運算，一種計算效能單位，代表每秒完成一千萬億(10^{15})次浮點運算，通常用於衡量智算中心及計算集群的整體處理能力
「研發」	指	研究與開發
「RDMA」	指	遠程直接內存訪問，可實現兩台計算機內存之間的直接數據讀寫，無需經由雙方操作系統轉發，實現高吞吐量、低時延的數據傳輸，是開展大規模分佈式AI訓練的關鍵技術
「強化學習」或「RL」	指	一種機器學習範式。智能體透過與環境互動學習決策策略，以獲取最大化累計獎勵；該技術廣泛應用於大語言模型後訓練，以及研發具備複雜推理與決策能力的自主AI智能體

技術詞彙表

「SaaS」	指	軟件即服務，屬雲服務的一種類型。該模式下軟件採用訂閱制授權，並由服務商集中託管及提供服務
「serverless」	指	一種雲計算執行模式。由雲服務商代用戶動態管理計算資源的分配及配置，應用程序及工作負載可自動彈性擴容，用戶無需自行部署、管理或維護服務器
「TFLOPS」	指	teraFLOPS，一種計算效能單位，代表每秒完成一萬億(10^{12})次浮點運算，多用於標示計算服務器的處理吞吐量
「詞元」	指	AI模型(尤其是大語言模型)處理的文本或數據的基本單元；詞元可表示詞語、子詞片段、標點符號或其他離散輸入單位，並作為輸入處理及輸出生成的基本元素。在特定互動中消耗的詞元數量通常被用作衡量及就AI模型使用情況計費的基礎
「Transformer」	指	一種深度學習神經網絡架構。其運用注意力機制判斷輸入內容的權重，可高效處理文本等序列數據，是目前絕大多數主流大語言模型的底層核心架構
「VKS」	指	虛擬容器計算服務是我們的雲服務層產品，專為AI企業、研究機構及開發者提供靈活的serverless計算資源；依託serverless架構，可根據實時工作負載需求，在數秒內完成計算資源彈性擴展
「VLA」	指	視覺－語言－動作，為多模態AI模型架構。整合視覺感知、自然語言理解與實體動作控制能力，賦予機器人系統及具身AI智能體解讀視覺畫面、語言指令，並轉化為實際行動的能力