

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。

「智能體基礎設施」	指	智能體基礎設施包括一套基礎設施，旨在構建一個AI代理可以在其中自主運行的雲環境，反映未來AI代理（而不是人類）將成為雲服務主要用戶的未來
「AI」	指	人工智能
「AI集群」	指	一組共同工作以執行人工智能相關任務的計算機或服務器
「AI雲計算」	指	通過雲端交付計算資源、儲存和AI開發工具，以支持人工智能模型的訓練、推理和部署。其將雲基礎設施的可擴展性與AI工作負載的高性能要求相結合
「AI原生」	指	在設計之初即以人工智能作為核心組成部分而構建的產品、服務或系統，而非在既有產品、服務或系統中額外整合人工智能功能
「AIDC」	指	AI數據中心，是一種專門設計和優化的數據中心，旨在處理AI工作負載的巨大計算需求。AIDC高度依賴於GPU，為支持AI模型的訓練和AI推理的運行而設計
「AIGC」	指	AI生成的內容，即利用AI技術生成的內容
「AI微調」	指	將預訓練模型調整為執行特定任務或使用案例的流程
「AI基礎設施」	指	開發、部署和管理AI應用程序及解決方案的必要硬件、軟件和框架的組合
「AIOps」	指	運用人工智能及機器學習技術以管理及維護基礎設施的先進方法
「API」	指	一組允許不同軟件應用程序相互通信的既定協定、程序及工具。就AI而言，API提供一種標準方式，讓外部應用程序能夠使用AI驅動的服務或模型，讓使用者毋須了解底層算法或架構

技術詞彙表

「帶寬」	指	數據透過網絡在設備、邊緣節點及中央數據中心之間傳輸的最大速率
「CDN」	指	內容分發網絡，一個由接近最終用戶的地理分佈式節點組成的網絡，旨在更快地將內容（如網頁、圖像及視頻）分發給用戶。通過將內容緩存在離用戶較近的服務器上，CDN縮短了數據傳輸距離，提高了加載速度和穩定性
「CPU」	指	中央處理單元，計算機的主要組件，負責執行計算機內部的大部分信息處理
「CPU核心」	指	中央處理器(CPU)內能夠獨立執行其自身計算任務的處理單元
「計算節點」	指	提供計算資源（如CPU核心、GPU、記憶體及存儲資源）的實體或虛擬服務器
「跨雲擴展」	指	在多個雲端環境中動態分配資源以處理不同工作負載並確保高可用性的能力
「併發能力」	指	雲基礎設施同時處理大量人工智能請求、會話或智能體運行的能力
「邊緣裸金屬」	指	部署在邊緣為計算提供專用硬件資源的物理獨立服務器。與多個用戶或進程共享同一物理機器的虛擬化環境不同，邊緣裸金屬提供基礎設施的直接和非共享途徑，提供高性能、低資源開銷和精確控制
「邊緣CDN」	指	一種地理分佈式服務器網絡連同軟件及應用，通過優化服務器的分佈，使服務器更接近最終用戶和數據源（與傳統的CDN相比），從而加快數據傳輸和縮短緩衝時間
「邊緣雲」	指	一種雲計算模式，其中數據存儲和處理能力等雲資源被移至更靠近數據生成位置及最終用戶所在位置
「邊緣計算」	指	一種分佈式計算模式，使計算和存儲更接近數據源，與雲計算相比，其具有改進的響應時間和更好的帶寬利用率等優點

技術詞彙表

「邊緣容器」	指	容器是虛擬化單元，使開發者可將其應用程序打包成單一可移植單元。邊緣容器的功能與傳統容器類似，但部署在邊緣設備或節點上，讓應用程序在數據生成位置或附近的線上邊緣計算節點中高效運行，從而降低時延並提高即時應用的性能。
「邊緣節點」	指	計算資源，由一個或多個靠近網絡邊緣數據源的設備組成，旨在促進數據源附近的數據處理和分析
「生成式AI」	指	AI的一個分支領域，專注於基於從現有數據中學習到的模式生成文本、圖像、視頻或音頻等新內容，例如多模態內容生成、聲音克隆及創建虛擬主机
「GPU」	指	最初為渲染圖形而設計的專用處理器，但現在廣泛用於加速人工智能等領域的並行計算，特別是用於深度學習模型的訓練和推理
「IDC」	指	互聯網數據中心，是一種專門的物理設施，旨在容納大量的計算機系統、網絡設備以及相關的基礎設施（如電源、冷卻系統和安全系統）。它主要依賴於CPU，為支持各種通用型IT工作負載而設計。現時IDC的定義更廣泛，包括傳統基於CPU的IDC及基於GPU的AIDC
「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應用的基礎
「延遲」	指	用戶設備或AI智能體向雲端發送數據或請求，至接收到模型輸出結果或回應之間所產生的時間延遲
「MB」	指	兆字節，衡量計算機存儲空間的標準。1MB等於約100萬字節
「模型API」	指	該接口允許開發人員通過標準化的API調用訪問和使用預訓練的AI模型，從而無需管理相關基礎設施即可完成文本生成、圖像識別或語音處理等任務

技術詞彙表

「模型訓練」	指	使用數據優化機器學習模型參數的過程，使其能夠學習模式並執行特定任務。其通常涉及前向傳遞、損失計算和反向傳播
「模型推理」	指	將經過訓練的機器學習模型應用於新輸入數據以生成預測、決策或結構化輸出的過程。其表示模型學習參數的部署時執行，通常涉及實際場景中的複雜推理、模式識別或語言理解
「多雲平台」	指	在單一架構中使用由多個雲提供商提供的雲計算服務
「多模態模型」	指	一種經過訓練的AI模型，可處理和理解多種類型的輸入數據（如文本、圖像、音頻或視頻），使其能夠以統一的方式跨不同模態執行推理和生成
「開源模型」	指	可供免費使用的可修改及重新發放源代碼的模型
「PUE」	指	電力使用效率，用來衡量數據中心能源效率的指標
「QoS」	指	服務質量
「SDK」	指	軟件開發工具包，一組可用於創建和開發應用程序的軟件開發工具
「SLA」	指	服務水平協議，服務供應商與客戶之間的合約，概述所提供的服務及預期的性能水平
「Token」	指	AI模型處理的基本數據單位
「VPC」	指	虛擬私有雲為公共雲中的一個安全、隔離的區域，允許客戶像在私有雲中一樣運行代碼、儲存數據、託管網站以及執行其他任務，但基礎設施由公共雲提供商託管