

技術詞彙表

本技術詞彙表載有本文件所用與本公司及其業務有關的若干詞彙的釋義。該等詞彙及其涵義未必始終與該等詞彙的標準行業涵義或用法一致，且未必可與在本公司同行業經營的其他公司所採用的類似名稱詞彙直接比較。

「抽象化」	指	簡化複雜系統細節的過程，通過提供較高層級介面，使用戶或應用程序能夠與底層資源互動，而無須了解當中的技術複雜性
「加速器」	指	一種硬件組件，通過將特定任務由中央處理器轉移至專用處理器來提高算力性能
「自適應路由」	指	一種路由方法，根據實時網絡狀況動態選擇或調整網絡路徑，以提高傳輸效率及整體網絡性能
「智能體」	指	能在極少人工干預的情況下自主規劃、作出決策並執行多步任務以達成特定目標的AI系統
「AI」	指	人工智能，為計算機科學的一個分科，專注於利用機器模擬人類智能
「AI智能體」	指	一種能夠代表用戶自主執行特定任務以達成既定目標的智能系統
「AI芯片」	指	專門設計或優化以執行AI相關計算（如機器學習訓練及推理）的半導體芯片
「AI算力」	指	用於開發、訓練、部署及運行人工智能模型及應用的計算資源及處理能力。
「AI算力集群」	指	一個由相互連接的計算資源組成的集群，旨在協同工作，為人工智能工作負載（包括模型訓練、微調、推理及部署）提供算力
「AI基礎設施」	指	支持開發、訓練、部署及運行AI模型與應用的硬件、軟件及系統資源。其通常包括算力、網絡、儲存以及相關平台軟件與系統能力
「AI工作負載」	指	與人工智能模型及應用相關的計算任務，包括數據處理、模型訓練、微調、推理及部署

技術詞彙表

「AllReduce」	指	並行計算中的一種集合通信操作，在該操作中，來自所有參與進程的數據將透過指定操作（例如求和）進行合併，並將結果分發回所有進程，通常用於在分佈式AI訓練期間同步模型參數
「誤碼率」	指	通信信道上數據流中接收到的位元數，因噪聲、干擾、失真或位元同步錯誤而發生每單位時間的改變
「緩存」	指	將經常存取的數據暫存於高速存儲器，以縮短檢索時間並提高處理效率
「檢查點」	指	在特定時間點保存的模型訓練狀態快照，以便在發生系統故障或中斷時，能夠從該時間點恢復訓練
「CNP」	指	擁塞通知報文，一種發送至數據源的網絡控制信息，用於表示沿傳輸路徑偵測到擁塞，從而提示數據源降低其發送速率
「變異系數」	指	一種標準化的離散程度度量，以標準差與均值的比值計算得出，通常以百分比表示，用於評估數據集的相對變異性
「集合通信庫」	指	一種軟件庫，提供標準化例程，用於協調分佈式計算系統中多個處理器或節點之間的數據交換與同步
「編譯」	指	將高階編程代碼轉化為可由硬件或運行環境執行的低階指令的過程
「計算節點」	指	算力集群內提供處理能力、記憶體及儲存資源以執行計算任務的單一服務器或機器
「算力資源利用率」	指	可用算力資源（如CPU、內存、存儲及網絡帶寬）相對於其總容量的實際使用程度，作為評估系統效率、容量規劃及成本優化的關鍵指標
「算力利用率線性度」	指	一項衡量指標，用以量化當算力資源增加時，系統效能的擴展程度與理想比例增長的接近程度
「容器化」	指	一種將軟件應用程序及其依賴項打包成稱為容器的獨立、可移植單元的技術，從而實現跨不同計算環境的一致部署及執行

技術詞彙表

「融合式」	指	在單一AI算力集群或其底層架構內，整合並統一多種運算元件（例如運算、網路及儲存），從而實現協調運作、提升效率並簡化系統設計
「融合」	指	將不同的元件、系統或資源整合並協調的過程或能力
「CPU」	指	中央處理器，一套複雜的電子電路，負責運行機器的操作系統和應用程序
「解碼」	指	人工智能推理的後續階段，模型會根據處理後的輸入情境及先前生成的詞元，按順序生成輸出詞元，通常一次生成一個詞元
「稠密」	指	一種人工智能模型架構，其中所有或幾乎所有模型參數參與均處理每個輸入或詞元
「分佈式系統」	指	一種計算架構，當中多台互連機器協同工作，作為一個統一系統執行任務、共享資源及協調進程
「DRAM」	指	動態隨機存取存儲器，一種將數據儲存於電容器中的易失性存儲器，需要定期刷新，並通常用作計算系統中的主存儲器
「ECN」	指	顯式擁塞通知，一種網絡機制，其以擁塞指示標記數據包而非將其丟棄，使端點能夠在無數據包遺失的情況下偵測並應對網絡擁塞
「EP」	指	專家並行，一種分佈式計算技術，將MoE模型的不同專家子網絡分配至多個設備，從而實現於訓練及推理期間的並行處理與通信
「微調」	指	在較小的特定任務數據集上對預訓練AI模型進行進一步訓練的過程，以使其能力適應特定應用或領域
「基於流切片的動態流量均衡」	指	一種網絡流量管理方法，將數據流分割成若干個由短暫空間間隔分隔的小流，並根據實時網絡狀況將該等小流動態地分配到多條網絡路徑上，從而改善負載均衡、減少網絡擁塞並提高傳輸效率
「基礎模型」	指	一種基於廣泛且多元化數據集開發的大規模預訓練模型，旨在作為通用模型以用於解決各類任務

技術詞彙表

「框架」	指	一種結構化軟件平台，提供預建工具、庫及介面，以促進AI模型及應用程序的開發、訓練及部署
「Gbps」	指	每秒千兆比特
「GDR」	指	GPUDirect RDMA，一種數據傳輸技術，允許GPU內存與兼容的網絡或外部設備之間直接通信，無需通過主機內存進行不必要的數據複製，亦無需大量CPU參與
「GDS」	指	GPUDirect Storage，一種能夠實現存儲設備與GPU顯存之間直接數據傳輸的技術，可繞過CPU及系統內存，從而減少延遲並提高GPU加速工作負載的數據吞吐量
「GPU」	指	圖形處理單元，一種專門設計用於操作及變更記憶體的電子電路，以加速在擬輸出至顯示裝置的幀緩衝器中生成圖像
「高帶寬內存」	指	一種高性能內存，採用垂直堆疊的DRAM和寬數據接口，可提供顯著更高的數據傳輸帶寬和能效，通常用於支持AI計算等數據密集型工作負載
「HDD」	指	硬盤驅動器，一種使用旋轉磁盤讀取和寫入數字數據的數據存儲設備
「羊群效應」	指	大量請求或進程同時集中於同一資源的現象，會導致需求激增，進而造成擁塞、爭用或效能下降
「熱切換」	指	在系統保持運行的情況下更換或重新配置硬件或軟件組件，而無須停機或中斷正在進行的進程
「推理」	指	已訓練的人工智能模型利用其學習參數處理輸入數據並產生輸出的過程，與開發或優化該等參數的訓練階段不同
「推理實例」	指	已部署的AI模型的單一副本，獲分配專用計算資源以處理推理請求
「推理副本」	指	已部署AI模型的多個相同實例，同時運行以處理並行推理請求，從而提高吞吐量及可用性
「輸入／輸出」或「I/O」	指	輸入／輸出，處理計算系統與外部設備之間數據傳輸的介面程序

技術詞彙表

「抖動」	指	通過網絡傳輸的數據包在時延或延遲上的變化，其可能影響數據通信的一致性及效能
「關鍵客戶」	指	於往績記錄期間任何期間／年度的收益貢獻超過人民幣10百萬元的客戶
「內核」	指	在多個GPU線程之間並行執行計算的函數
「KV緩存」	指	鍵值緩存，一種於模型推理期間使用的內存機制，用於儲存注意力層先前計算的鍵值對，從而減少冗餘計算並提高推理速度與效率
「KV儲存」	指	鍵值儲存，一種以獨特鍵及其相關值對形式組織及檢索數據的數據儲存系統，從而實現快速高效的數據存取
「長上下文」	指	於單次推理操作中處理及保留大量輸入數據或文本的能力
「LoRA」	指	低秩適應，一種具參數效率的微調技術，通過引入低秩矩陣僅更新模型的一小部分參數，從而實現對特定任務更快、資源消耗較低的適應
「無損」	指	指並無丟失或降低任何數據的數據傳輸或壓縮，確保原始信息的完全保真度
「基於MEMS的OCS」	指	基於微機電系統的光路交換機，一種光路交換機，其利用通過微機電系統技術製造的微小機械反射鏡或可移動組件，在連接埠之間對光信號進行物理重定向
「MFU」	指	模型浮點運算利用率，一種指標，表示於運行AI模型時GPU浮點運算能力的有效利用率，反映硬件與軟件的優化程度
「MoE」	指	混合專家模型，一種人工智能模型架構，使用一組子網絡（或稱「專家」），針對每個輸入僅激活其中的一個子集。這使得模型能根據任務性質分配計算資源，從而實現高效擴展，在保持效率的同時提升性能
「MTTR」	指	平均修復時間
「多模態」	指	一種人工智能模型架構，旨在支援處理及整合兩種或多種資料模態（例如文字、圖像、音訊或視訊），使該模型能執行涉及多種類型輸入和輸出的任務

技術詞彙表

「節點」	指	網絡或系統內的個別計算單元或設備（例如服務器、處理器或工作站），能夠處理、發送或接收數據
「NPO」或「近封裝光學」	指	一種光互連架構，其中光模塊或光引擎在物理上緊鄰交換芯片、處理器或其他高性能計算組件，從而縮短電傳輸距離，提高帶寬和功率效率，並支持高速數據通信
「NVMe SSD」	指	非易失性內存表達固態驅動器，一種使用閃存並通過NVMe協議連接的高效能儲存設備，與傳統儲存介面相比，其可提供更快的數據讀寫速度
「OCS」	指	光路交換機，一種在無須將信號轉換為電信號的情況下於連接埠之間建立專用光信號路徑的網絡交換機，從而實現高頻寬、低延遲的數據傳輸
「光模塊」	指	一種將電信號轉換為光信號（反之亦然）的設備，用於通過光纖網絡傳輸數據
「正交背板」	指	一種電路板架構，當中線路卡與交換卡相互垂直連接，從而實現網絡交換機或計算機箱內的高密度、高頻寬互連
「數據包」	指	格式化以通過網絡傳輸的數據單元
「數據包轉發率」	指	網絡設備（例如交換機）處理及轉發數據包的速率，通常以每秒數據包數來衡量
「並行計算」	指	一種計算形式，當中多個計算或進程同時執行（通常跨越多個處理器或核心），以提高處理速度與效率
「參數」	指	基礎模型於其訓練過程中學習及調整的內部數值變數或配置。該等參數有效蘊含模型從其訓練所用的廣泛數據集中提取的知識、模式及關係
「劃分」	指	將單一物理硬件資源（如GPU）劃分為多個隔離區段，該等區段可獨立分配予不同任務或用戶
「PCB」	指	印刷電路板，一種由絕緣材料製成的平板，其上蝕刻或印有導電路徑，用於對電子組件進行機械支撐及電氣連接

技術詞彙表

「PFLOPS」	指	PetaFLOPS，一種計算效能單位，等於每秒一千萬億次浮點運算，通常用於衡量高性能計算系統的處理能力
「後訓練」	指	預訓練之後的階段，當中AI模型通過微調、對齊或強化學習等技術得到進一步完善，以提升其於特定任務或預期行為上的表現
「預填充」	指	人工智能推理的初始階段，模型會處理輸入提示或語境，計算相應的內部表示和緩存數據，並為後續的輸出生成做好準備
「預訓練」	指	在對AI模型進行特定任務微調之前，以大規模數據集對其進行訓練以學習一般模式及表示的初始階段
「Python」	指	一種廣泛使用的通用編程語言，以其可讀性及多功能性著稱，常用於人工智能、數據科學及軟件開發
「PyTorch」	指	一個廣泛用於構建、訓練及部署AI模型的開源機器學習框架，以其靈活性及動態計算能力著稱
「RDMA」	指	遠端直接記憶體存取
「強化學習」	指	一種機器學習方法，當中AI模型基於其行動結果接收以獎勵或懲罰形式的反饋，從而學習作出決策
「關鍵客戶回購率」	指	截至最後實際可行日期曾向本集團購買最少兩次任何產品或服務的關鍵客戶的百分比，按(i)該等關鍵客戶數目除以(ii)往績記錄期間的關鍵客戶總數計算
「RLHF」	指	基於人類反饋的強化學習，一種訓練技術，當中利用人類評估反饋作為獎勵信號來完善AI模型，引導模型產生更符合人類偏好及期望的輸出
「RoCE」	指	基於融合以太網的RDMA，一種允許通過以太網進行RDMA的網絡協議

技術詞彙表

「運行環境」	指	提供必要資源、庫及服務以於運行期間執行及管理AI工作負載的底層軟件基礎設施
「沙箱」	指	一個隔離且受控的環境，以旨在限制對其他環境、系統、資源或數據造成意外影響的方式，開展活動
「SerDes」	指	串行器／解串器，一種在串行與並行格式之間轉換數據的硬件組件，從而實現通過通信鏈路的高速數據傳輸
「傾斜」	指	計算資源之間工作負載或數據的分佈不均，這可能導致效能失衡，使得部分資源被過度利用，而其他資源仍未獲充分利用
「SLA」	指	服務級別協議，為服務提供商與客戶之間簽訂的正式合同，定義了預期的服務級別，包括正常運行時間、響應時間及支持質量等性能指標，以及在未達到標準時的補救措施
「SOTA」	指	最先進的，模型、算法或系統於特定時間點在公認基準或評估指標上所達到的最高性能水平或技術成就
「超級計算機」	指	一種高效能計算系統，能夠處理極大量數據並執行複雜計算，其速度顯著高於傳統計算機
「SuperPod」	指	一個大規模AI算力單元，由多個互連的計算服務器、GPU或其他加速器、高性能網絡設備、存儲系統及相關軟件組成，旨在作為一體化集群運行，以支持大規模AI訓練及推理工作負載
「交換機」	指	一種網絡設備，用於互連服務器、加速器、存儲系統及其他計算基礎設施，並在AI算力集群內引導它們之間的數據流量
「Tbps」	指	每秒太比特
「遙測」	指	自動收集、傳輸及報告來自網絡設備及系統的運行及性能數據，用於監控、分析及管理目的
「張量」	指	張量是用於表示及處理機器學習模型中數據的多維數值數組
「詞元」	指	文本單元，為模型輸入處理及輸出生成的基本元素
「工具」	指	AI智能體可調用以執行特定任務（如檢索數據、執行代碼或與外部系統互動）的外部函數、服務或資源

技術詞彙表

「工具調用」	指	AI模型於其推理過程中識別並調用外部工具或函數以完成既定任務的能力
「Top-K路由」	指	MoE模型中的一種選擇機制，將每個輸入詞元引導至得分最高的K個專家子網絡進行處理，從而平衡計算效率與模型效能
「拓撲」	指	網絡或計算系統內節點、設備或組件的佈置或互連模式
「TP」	指	張量並行，一種分佈式計算技術，將單個模型張量劃分到多個設備上，實現並行計算和通信，以訓練或運行超過單個設備內存容量的大規模AI模型
「訓練」	指	利用數據及計算資源開發人工智能模型的過程，通過優化模型參數，使其能夠根據給定的輸入執行指定任務或生成輸出
「Triton」	指	一種開源程式語言及編譯器，使開發人員能夠為深度學習工作負載編寫高效的GPU內核代碼，而無需具備低階硬件編程專業知識
「驗證」	指	通過一組指定參考數據對模型輸出進行評估，以評核其準確度及效能的過程
「虛擬化」	指	一種將物理硬件資源（如GPU）抽象化為多個虛擬實例的技術，允許於不同用戶或工作負載之間對其進行獨立共享、分配及管理
「VLA模型」	指	視覺－語言－動作模型，一種整合視覺感知、語言理解及動作規劃能力的AI模型，使其能夠解釋視覺及文本輸入並產生相應動作
「線程束」	指	一種硬件執行單元（常見於GPU），包含一組以鎖步方式並行執行相同指令的線程，其可針對特定工作負載或操作進行優化或專門設計
「世界模型」	指	一種AI模型，其學習並維持物理或虛擬環境的內部表徵，使其能夠對世界的運作方式進行模擬、預測及推理
「白盒」	指	基於開放和可定制設計的硬件或系統架構，其規格透明，允許用戶靈活選擇、修改和優化軟件組件，而無需依賴專有解決方案
「零拷貝數據傳輸」	指	一種數據傳輸技術，可消除通信期間記憶體緩衝區之間不必要的數據中間複製，使數據能夠直接於來源與目的地之間移動，從而減少延遲及CPU開銷