技術詞彙表

本詞彙表載有本文件所用有關本公司及我們業務的若干詞彙的解釋。本詞彙表所載術語及其涵義可能與該等詞彙的業內標準涵義或用法不同。

詞彙		釋義
AI智能體	指	由AI模型驅動的自主軟件實體，可在複雜環境中進行交互、規劃和採取行動，通常用於企業和客戶服務情境。
人工智能生成內容(AIGC)	指	利用人工智能技術自動生成文本、圖像、音頻、視頻等內容的技術或方式。
應用程序接口(API)	指	一套預先定義的規則和工具，允許不同的軟件應用程序相互通信、交換數據並調用彼此的功能，而無需了解對方內部的複雜實現細節。
AR/VR	指	增強現實／虛擬實境，分別將數字內容疊加到真實世界或創造沉浸式模擬環境的技術。
ASIC	指	專用集成電路，一種專為特定任務而設計的芯片，通常為目標工作負載提供更高效率但缺乏靈活性。
BF16	指	16位腦浮點數格式，一種用於AI訓練和推理的數值格式，可平衡範圍和精確度，對於大模型尤其有效。
CU	指	計算單元，本公司GPU+ (Minsky™)架構的專屬組件，支持計算任務，兼容CUDA/OpenCL編程模型。
統一計算架構(CUDA)	指	公司N推出的一種並行計算平台和編程模型，旨在讓開發者能夠利用GPU（圖形處理器）中數以千計的計算核心，來執行遠超CPU處理能力的大規模並行計算任務。
登臨深度學習神經網絡引擎(DLNNE)	指	本公司專屬的一款高性能網絡推理優化器和運行時引擎。它充分利用了本公司的GPU+架構芯片產品的特性，針對網絡推理應用場景實現了極致的推理速度、低延遲和高吞吐量。
電子製造服務(EMS)	指	根據客戶設計規範組裝、測試及封裝電子組件或電路板的提供商。
FP32/TF32/BF16/FP16/INT8/INT4	指	AI模型計算中使用的數值精度格式。這些格式分別指32位浮點數、TensorFloat-32、bfloat16、16位浮點數，以及8位和4位整數表示。低位格式通常可提供更快的運算速度，並降低推理任務的耗電量。

技術詞彙表

詞彙		釋義
FP4/FP8	指	4位和8位浮點數精度格式，用於低功耗、高密度的AI推理工作負載。
GPGPU	指	通用圖形處理器，一種不僅用於圖形渲染，還用於加速廣泛的通用運算工作負載的GPU，特別是AI和科學運算。
GPU	指	圖形處理器，一種專為處理並行任務(例如圖像處理和AI工作負載)而設計的處理器。
GPU+ (Minsky™)	指	我們的專有GPU芯片架構，以片內異構的方式將GPGPU核心、AI加速單元及高帶寬片上內存子系統集成於單顆芯片。
GS/GF/XL系列	指	專為高性能雲計算、邊緣AI推理和訓練以及大模型運算所設計的GPGPU加速卡產品線。
瀚銘(Hamming®)軟件棧	指	本公司的集成軟件平台和工具包，提供各類編譯程序、驅動程序棧、運行時環境和開發工具，專為實現GPU+(Minsky™)架構系列芯片的最佳效能而設計。
HBM	指	高帶寬內存，一種用於高性能GPU的內存，與傳統內存相比，可提供更快的數據傳輸速率和更低的功耗。
集成電路(IC)	指	一種將多個電子組件與電路集成到單一芯片上的半導體器件。
智能監考	指	用於教育測試場景的AI監考系統，可實時監控和評估學生的行為和表現。
KS系列	指	專為機器人、人工智能個人計算機和嵌入式等邊緣和終端運算場景所設計的新一代AI推理卡產品線，同時還能夠支持訓推一體及大模型運算。
LLM	指	大語言模型，一種在大量文字語料庫上訓練的基礎模型，用於回答問題、總結和自然語言生成等任務。
MES	指	製造執行系統，可在工業製造環境中實時管理、監控和追蹤生產流程的軟件系統。
MindSpore	指	由中國一家技術公司開發的開源深度學習框架，針對AI應用中的訓練和推理進行優化。

技術詞彙表

詞彙		釋義
NoC	指	片上網絡，一種通訊子系統，可讓芯片中的處理單元與內存之間進行高速數據交換。
原設備製造商	指	原設備製造商，通常指將第三方組件整合為以自有品牌銷售的完整系統的公司。
OS	指	操作系統，一組通過分配資源(包括內存、CPU、輸入／輸出設備及文件存儲)管理計算機硬件及應用程序的軟件。
OSAT	指	外包半導體組裝和測試提供商，執行半導體芯片封裝和最終測試的第三方服務提供商。
PCB	指	印刷電路板，一種用於支撐及連接電路組件的機械基板。
印刷電路板組裝(PCBA)	指	將半導體芯片與電子組件安裝到印刷電路板上，以形成功能電子模塊的過程。
RAG	指	檢索增強生成，LLM應用中的一種技術，實時檢索外部數據來源以增強模型的輸出，常用於企業問答系統。
RISC-V	指	基於已建立的精簡指令集計算(RISC)原理的免費開放標準指令集架構。
SDK	指	軟件開發工具包，一組軟件工具、庫和文檔的集合，可讓開發人員為特定平台建立應用程序。
SIMT	指	單指令多線程，一種GPU執行模型，多線程在不同的數據元素上並行執行相同的指令。
SLAM	指	即時定位與地圖構建，用於機器人和AR的運算程序，在構建未知環境地圖的同時追蹤設備在該環境中的位置。
Stable Diffusion	指	廣泛使用的開源文本到圖像生成AI模型。常用於AIGC應用程序。

技術詞彙表

詞彙		釋義
流片	指	將完成的集成電路設計轉化為可商業化部署的成品芯片的過程。該過程主要包括(i)將經驗證的芯片設計數據轉換為GDSII（圖形數據庫系統II）格式數據；(ii)根據GDSII數據生成光罩數據；(iii)在晶圓代工廠使用光罩，通過光刻、蝕刻、沉積等晶圓製程步驟製造半成品晶圓；及(iv)對晶圓進行切割、封裝及測試，以生產成品芯片。流片的主要成本包括光罩成本、晶圓製造成本及封裝測試成本，不同製程節點之間的成本差異主要來自光罩及晶圓成本，並隨相關製程節點的成熟度而變化。
張量引擎	指	針對優化張量運算（例如矩陣乘法）的計算單元，常用於AI模型訓練和推理。
TPU	指	張量處理單元，一種AI加速單元，專門用於神經網絡中以張量為基礎的運算。
UL/L系列	指	專為邊緣和終端推理及訓推一體場景、智能裝置和嵌入式系統等所設計的低功耗GPGPU加速卡產品線。
vGPU	指	虛擬GPU，一種讓多台虛擬機共享單個物理GPU的技術，可提高數據中心的資源利用率。
vLLM	指	專為大語言模型服務而優化的高吞吐量和內存高效的推理引擎。