

## 課程名稱：17828\_CUDA New Features and beyond 課程摘要

本摘要由 Gemini AI 協助生成，僅供參考

### 1. 適合對象與能力門檻

- **適合族群：**中階主管、資料分析師、AI/HPC 軟體工程師、系統架構師與 GPU 算力優化專家。此課程深入探討並剖析 GPU 底層硬體與軟體架構，適合需要大幅提升運算效能並控制資料中心能耗的專業人員。
- **能力門檻：**普通至進階。建議具備 C/C++ 或 Python 基礎程式設計能力，並對計算機結構、平行運算（Parallel Computing）及基礎矩陣運算有一定程度的了解。

### 2. 課程核心大綱與架構圖

- **章節架構：**
  - **章節一：能耗與平行計算新浪潮** — 解析為何運算效能受限於能耗（Performance per Watt），以及從單純追求速度轉向能效比的趨勢。
  - **章節二：張量核心（Tensor Cores）與混合精度算力突破** — 探討 FP64、FP16 到混合精度（如 LU 分解、GMRES 迭代演練）在維持精度的同時節省數倍能耗與大幅提升效能。
  - **章節三：核心融合（Kernel Fusion）與 JIT 即時編譯** — 介紹如何透過將多個運算步驟融合以減少資料搬移，並利用 JIT 編譯器優化動態工作流程。
  - **章節四：Python 與 CUDA 生態系的深度結合** — 說明如何透過 CuBLASdx、Cutlass、Warp 及 Legate 將 C++ 底層極致效能轉化為 Python 開發者的完整加速體驗。
  - **章節五：分散式系統加速與 GPU Direct 技術進化** — 剖析 GPUDirect Async 到 Kernel Initiated 如何突破 CPU 瓶頸，實現低延遲與高頻寬通訊。
  - **章節六：Grace Hopper 架構與動態任務圖（Conditional CUDA Graphs）** — 介紹統一記憶體架構（NVLink C2C）與 CUDA 12.4

具備控制流（If / While）的動態圖運算。

### 3. 必備前置知識清單

- **C/C++ 或 Python 程式語言**：具備基礎或中階語法撰寫能力，了解變數、迴圈與函式庫調用。
- **平行運算與矩陣運算概念**：理解矩陣乘法、向量運算及基礎的線性代數觀念。
- **計算機組織與記憶體層級**：對 CPU 與 GPU 的基本架構差異（如 Cache、Bandwidth、Latency）有初步概念。

### 4. 專有名詞與概念辭典

- **Tensor Cores（張量核心）**：NVIDIA GPU 中專門用於加速矩陣乘法與累加運算（GEMM）的硬體單元，能在極低能耗下提供爆發性的算力。
- **Kernel Fusion（核心融合）**：將多個原本獨立執行的 GPU 運算核心合併為一個，避免中間資料反覆寫回主記憶體，進而大幅減少資料搬移的功耗與延遲。
- **GPUDirect Kernel Initiated**：一種讓 GPU 能夠直接發起並控制網路卡（NIC）進行資料傳輸的技術，完全將 CPU 從傳輸路徑中剔除，將通訊步數降至極致。
- **Conditional CUDA Graphs（條件式 CUDA 任務圖）**：CUDA 12.4 推出的新功能，支援在 GPU 內部直接進行 if 與 while 等邏輯判斷，無需每次透過 CPU 判定是否進行下一輪迭代。

### 5. 課程學習職能對照

- **對應職能名稱**：AI/HPC 軟體工程師、平行運算架構師、高效能資料分析師
- **主要工作任務/能力指標**：
  - **分析與決策能力**：能衡量演算法的能耗與效能（Performance per Watt），選擇適當的混合精度與優化模型。
  - **工具操作與執行能力**：能掌握 CUDA 平台、Tensor Cores、Cutlass、Warp、Legate 以及動態 CUDA Graphs，進行複雜運算

管線的優化與極速部署。