針對 PIM 執行的 Benchmark 特性分析：辨識適合 DPU 的工作負載特徵

Characterizing Benchmarks for PIM Execution: Identifying Features of DPU-Suitable Workloads

指導教授：何建忠

專題成員：王葆凱、李宜頲

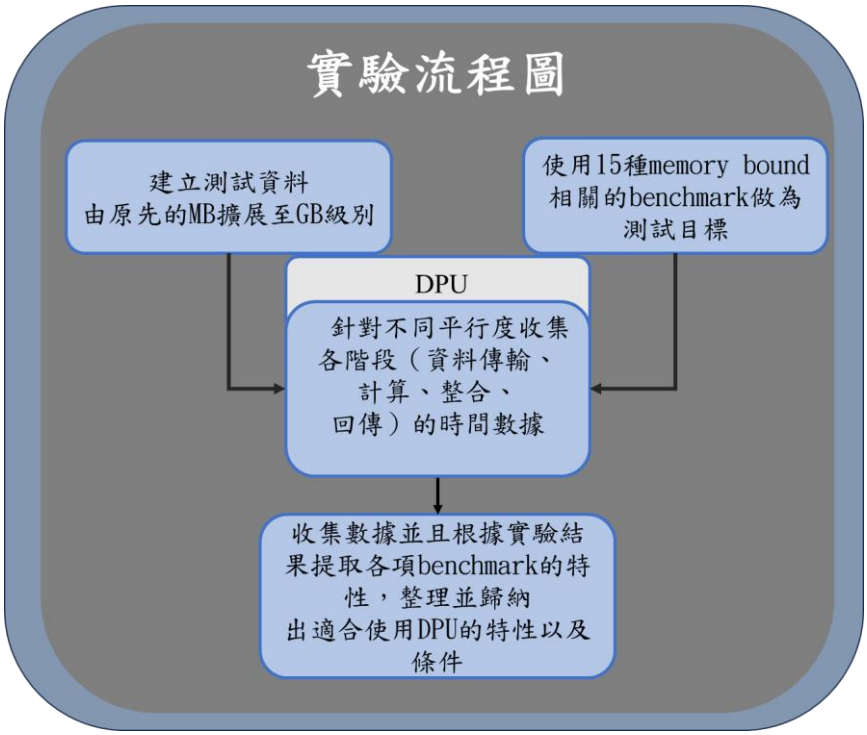
開發工具：Python、C

測試環境：UPMEM cloud

1、 簡介：

在現代運算應用（如資料庫與圖像處理）中，記憶體頻寬與延遲已成為影響效能的關鍵瓶頸。為了減少數據移動所帶來的開銷，記憶體內運算（Processing-in-Memory, PIM） 將運算單元直接整合至記憶體中，使計算能夠在數據儲存的地方直接執行，以提升計算效率並降低傳輸成本。

本研究聚焦於 PIM 架構，特別是其 DPU（DRAM Processing Unit）設計，並針對一系列基準測試（benchmarks）進行分析。並且擴大資料量規模至 GB 等級資料量，探索當資料量增加時，DPU 在執行效能與適用性上的變化。

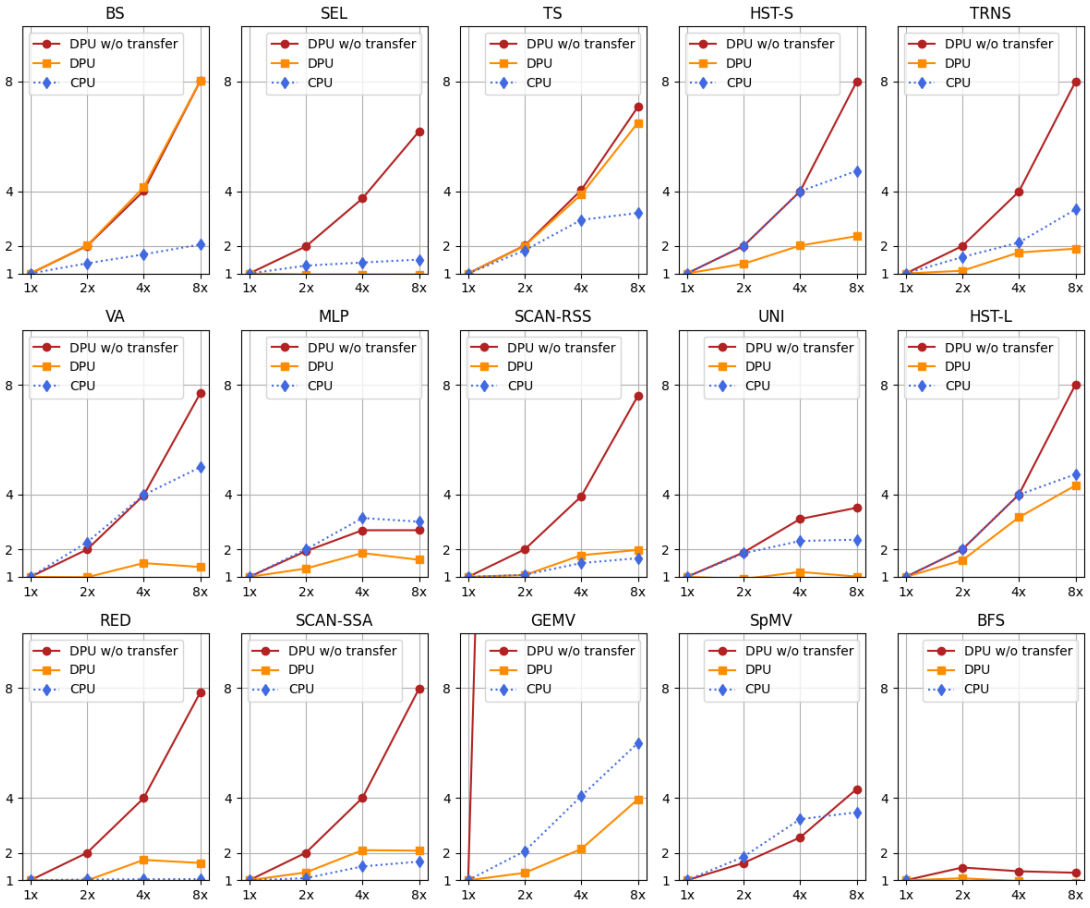


圖一為實驗流程圖

二、測試結果：

根據實驗結果，所有 benchmark 可歸納為三類：

1. 適合 DPU 的操作（如 TS、BS、RED、SCAN-RSS、SCAN-SSA）：
   當回傳資料量極小（近 O(1)），且傳輸階段可平行加速時，整體效能可隨 DPU 數量近線性提升，最適合部署於 DPU 架構。對於 SCAN-SSA，SCAN-RSS 這兩個操作充分發揮了 DPU 在 MRAM 存取上的優勢 —— 其 MRAM 存取次數分別為 3N 與 4N。即使 DPU 到 CPU 的資料傳輸時間複雜度不是 O(1)，仍能展現 DPU 的平行處理效能。
2. 特定條件下適合 DPU 的操作（如 SEL、UNI、BFS）：
   若資料具高重複性或圖結構簡單，可顯著降低回傳資料量與通訊開銷，在特定情境下仍可受益於 DPU 的平行傳輸與運算能力。
3. 不適合 DPU 的操作（如 VA、GEMV、SPMV、MLP）：
   當資料傳輸或同步開銷成為效能瓶頸，且難以隨 DPU 數量縮減，將導致平行化效益受限，不宜使用 DPU 架構執行。



圖二為測試結果的加速折線圖