採購規範書

壹、採購目的

隨著大型語言模型（LLM）於邊緣裝置之應用需求增加，單一邊緣 AI 運算平台於模型容量與推論效能方面逐漸面臨限制。本案擬針對 MediaTek Genio 720 開發板建置多晶片大型語言模型推論測試環境，驗證 Tensor Parallel（TP）架構下之模型切分、跨板資料傳遞、Collective Operation 及模型推論效能。

本案將以 Llama 3.1 8B W4A16 模型為主要測試對象，於 Dual-EVK TP2 環境進行完整模型執行、AllReduce／AllGather 功能驗證、Continuous Decode，以及 Prefill 至 Decode KV 狀態銜接等功能與效能測試，並量測 TTFT、Decode TPS、Full-model TPS、Latency 與跨晶片資料傳輸量等指標，作為後續多晶片邊緣 AI 大型語言模型推論架構優化與擴充之技術依據。

貳、工作時程

委託生效日起至 20 工作日。

參、驗收項目

(一) Dual-EVK TP2 多晶片推論環境建置

以 MediaTek Genio 720 Dual-EVK 建置 Tensor Parallel 2（TP2）大型語言模型推論測試環境，採 Llama 3.1 8B W4A16 模型進行驗證，完成 32 Layers 模型之多晶片切分與執行流程。

(二) 模型平行化功能正確性驗證

完成 Full32 Host 32 Layers 模型執行、NCC Compile、Layer Handoff 及跨晶片資料傳遞驗證，並確認完整模型於 Dual-EVK TP2 架構下可正確執行。

(三) Collective Operation 功能驗證

完成 Attention 及 MLP 階段之 AllReduce 功能驗證。

(四) Decode 效能測試

進行 Dual-EVK TP2 Continuous Decode 效能測試，量測 Throughput（TPS）、Latency（ms/token）、TX/RX Bytes 及 AllReduce Calls 等指標。

(五) 交付「MediaTek Genio 720 多晶片大型語言模型推論效能測試報告」電子檔 1 份。

肆、智財權權利歸屬

(一) 本案執行所產出之成果，其智慧財產權歸屬於經濟部與工業技術研究院所有。

(二) 受委託單位須確保提供驗收之檔案內所使用任何圖片、音樂、內容均不得侵犯其他第三者之智財權。同時，獲得相關口頭、書面資訊有保密之責，經委託單位同意之外，一律不得對外露出。

伍、經費支付

(一) 完成本案各項測試工作，並依據委託單位意見調整，且完成『MediaTek Genio 720 多晶片大型語言模型推論效能測試報告』驗收後，支付簽約總經費。