

2026 智慧創新大賞

# 基於 FPGA 之無 DSP 的 BitMamba-2 推論加速器

A DSP-Free FPGA Accelerator for BitMamba-2 Inference

DE10-Nano | Cyclone V | 零 DSP | 純 C 推論引擎

# 研究動機與背景

1. Memory Wall : 1B 參數 FP16 模型需 6 GB, DE10-Nano 僅 1 GB DDR3
2. 運算資源 : 低成本 FPGA 僅 112 DSP, 傳統 MAC 不可行
3. KV Cache : Transformer  $O(N^2)$  記憶體, 2048 tokens 即佔 157.29 MB

三元權重  $\{-1, 0, +1\}$  的突破 :

- 乘法退化為加減法, 硬體複雜度大幅降低
- Microsoft BitNet b1.58 證明三元權重可匹配 FP16 性能
- 模型大小從 2 GB 壓縮至 250 MB (1B 參數)

**目標 : 在完全不使用 DSP 的低成本 FPGA 上運行 1B 參數大語言模型**

# 為什麼選 Mamba (SSM)

Transformer KV cache =  $O(N^2)$ , 碰到 Memory Wall

Mamba-2 以固定隱狀態壓縮歷史 :  $O(1)$  記憶體

BitMamba-2 混合精度 : 90% 三元權重 + SSM 變數 FP32

雙軌策略 :

BitMamba-2 (SSM)  $\rightarrow$  KV Cache 消滅, 固定隱狀態  $O(1)$

BitNet (Transformer)  $\rightarrow$  TurboQuant 4x 壓縮

兩種模型均使用 T-MAC 加速器核心

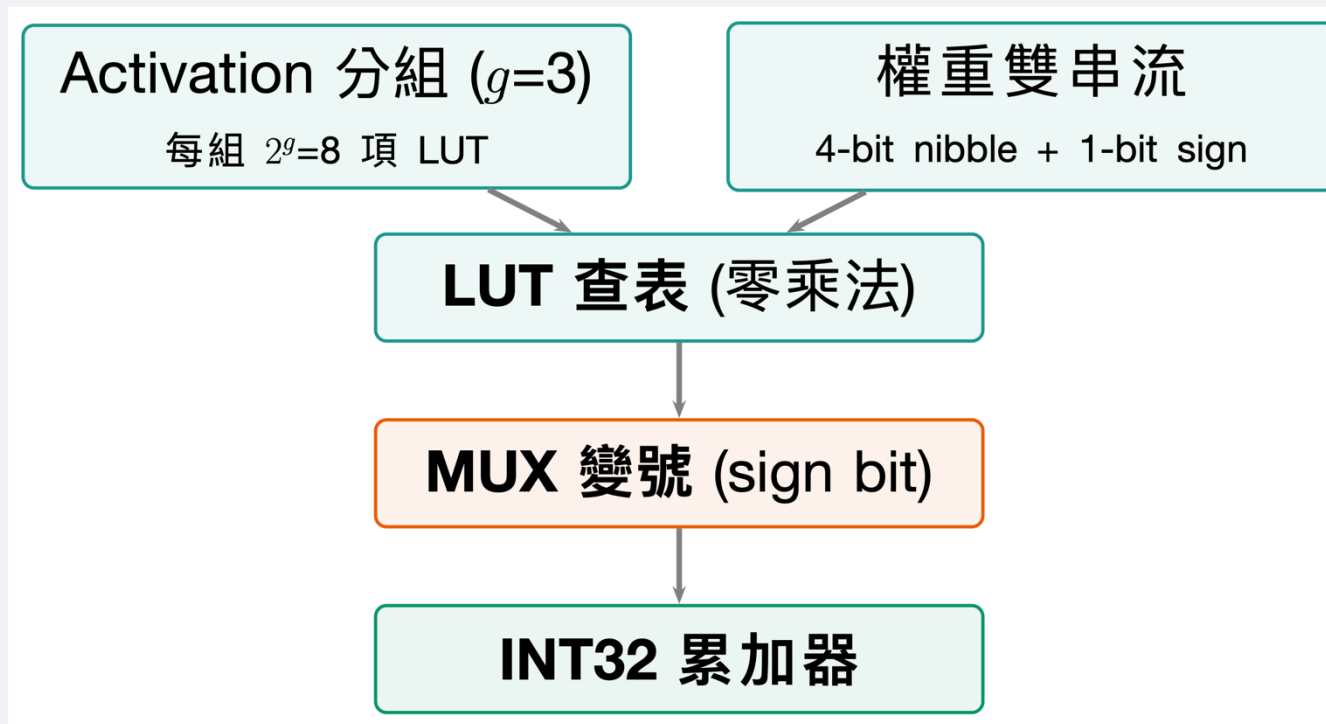
$\rightarrow$  證明零 DSP FPGA 的通用性

# T-MAC 查表取代乘法器

核心思想：同一批 activation 被重複乘上不同權重

T-MAC 將重複計算預先算好放進 LUT

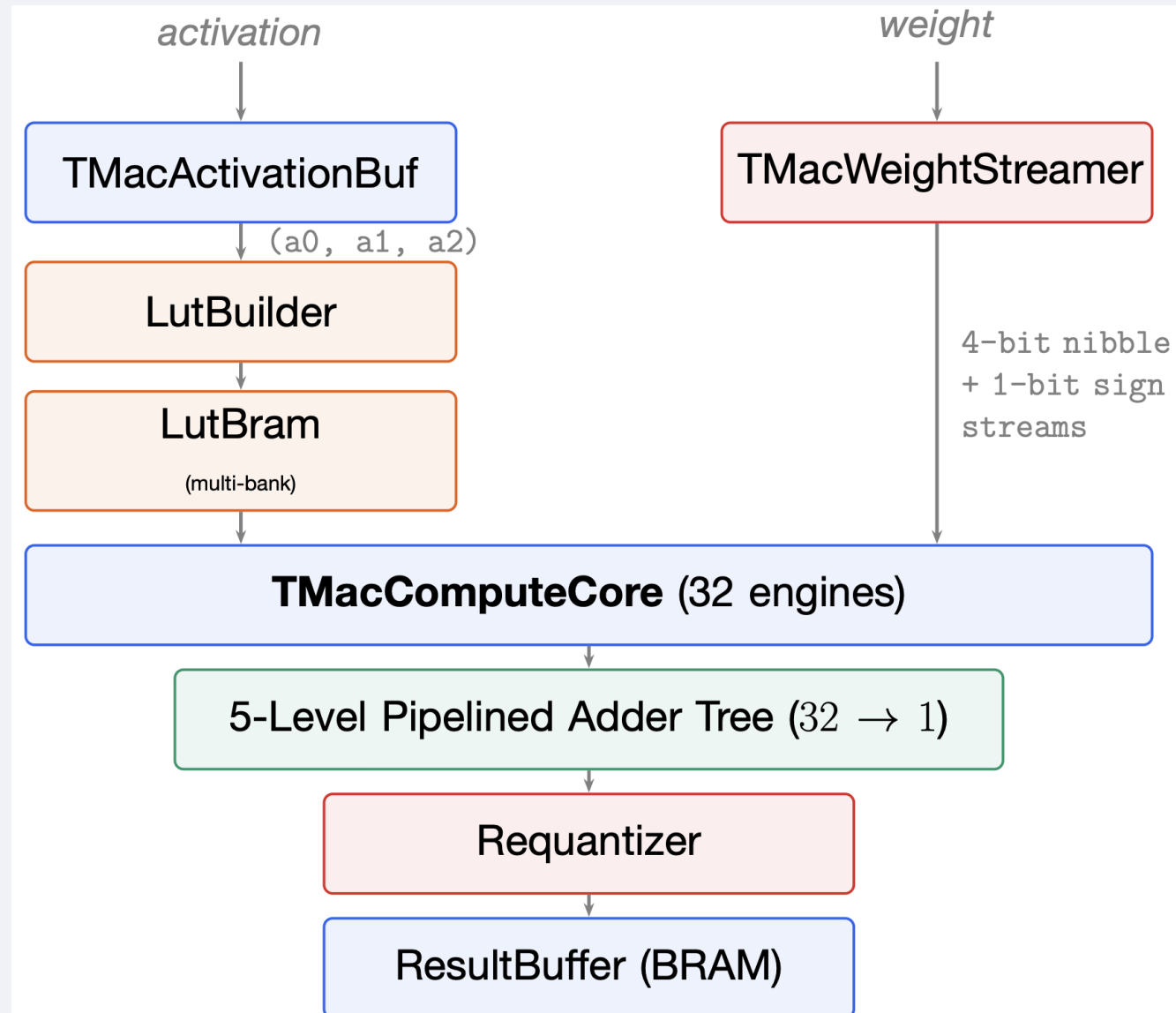
流程：



資源平衡：利用 BRAM 作為 LUT，平衡 ALM 與 BRAM 的資源用量

結果：完全不使用乘法器，DSP 使用量 = 0

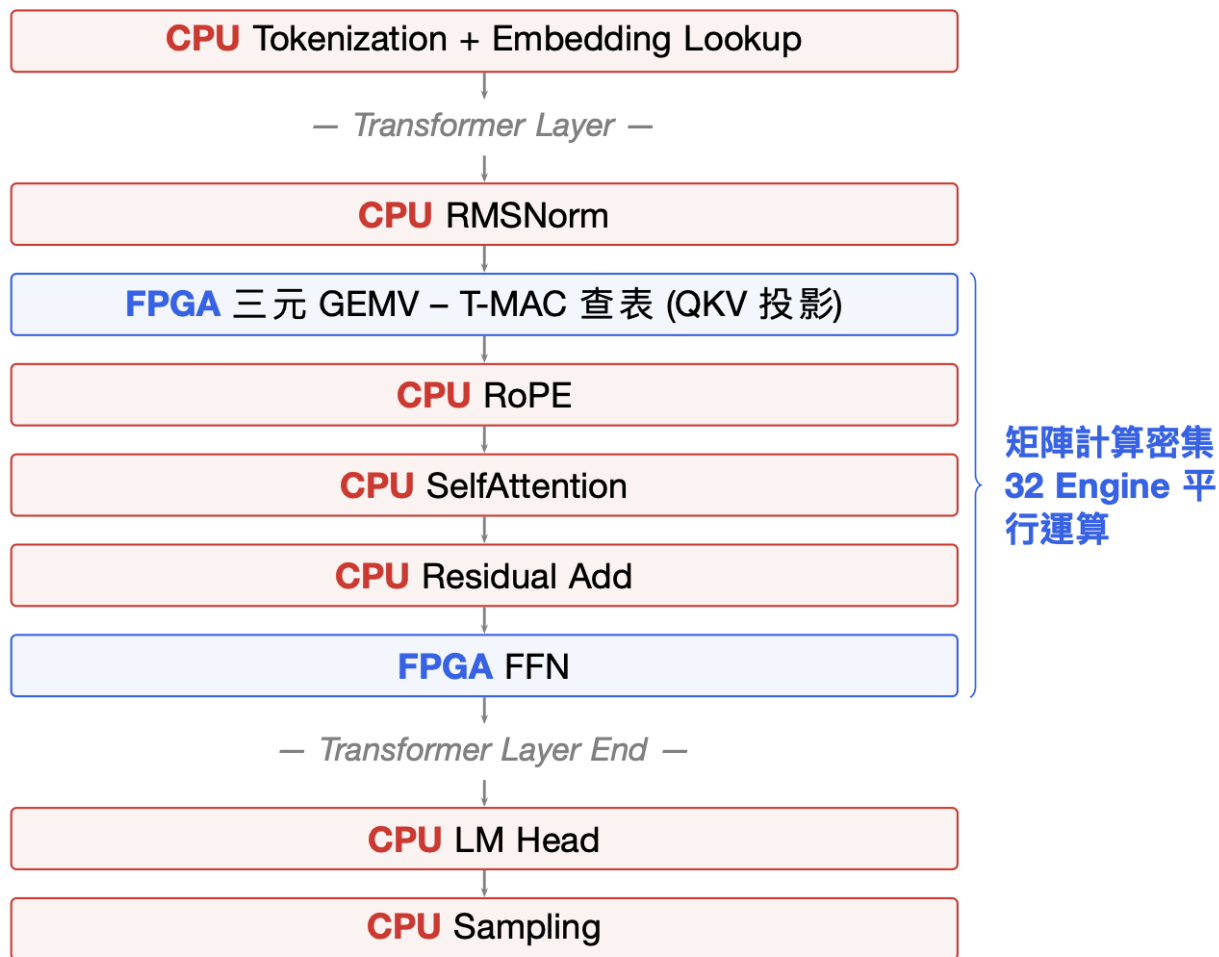
# FPGA 內部硬體架構 (TMacAccelerator)



# BiTurbo 推理引擎流程拆解

## CPU NEON + FPGA 混合加速

不適合 FPGA 的運算交由 HPS 處理，適合的交由 FPGA 加速



為什麼分工：

FPGA：三元 GEMV + FFN

佔推理 85%+ 時間

查表適合 FPGA LUT

CPU：RMSNorm, RoPE,

SelfAttention, Residual Add

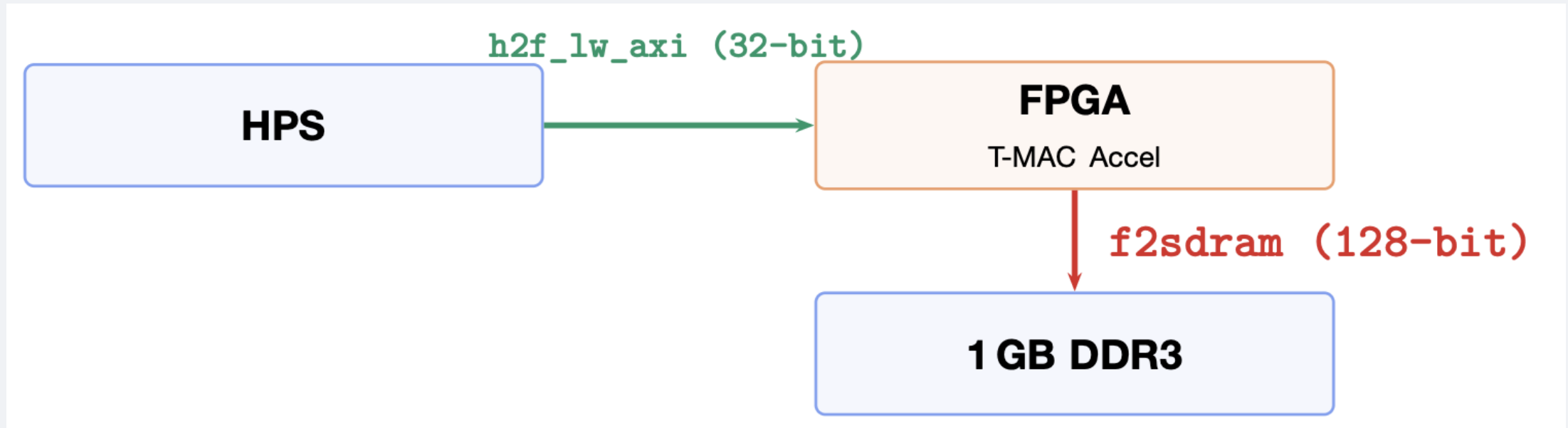
需浮點精度或複雜控制流

矩陣計算密集

32 Engine 平行運算

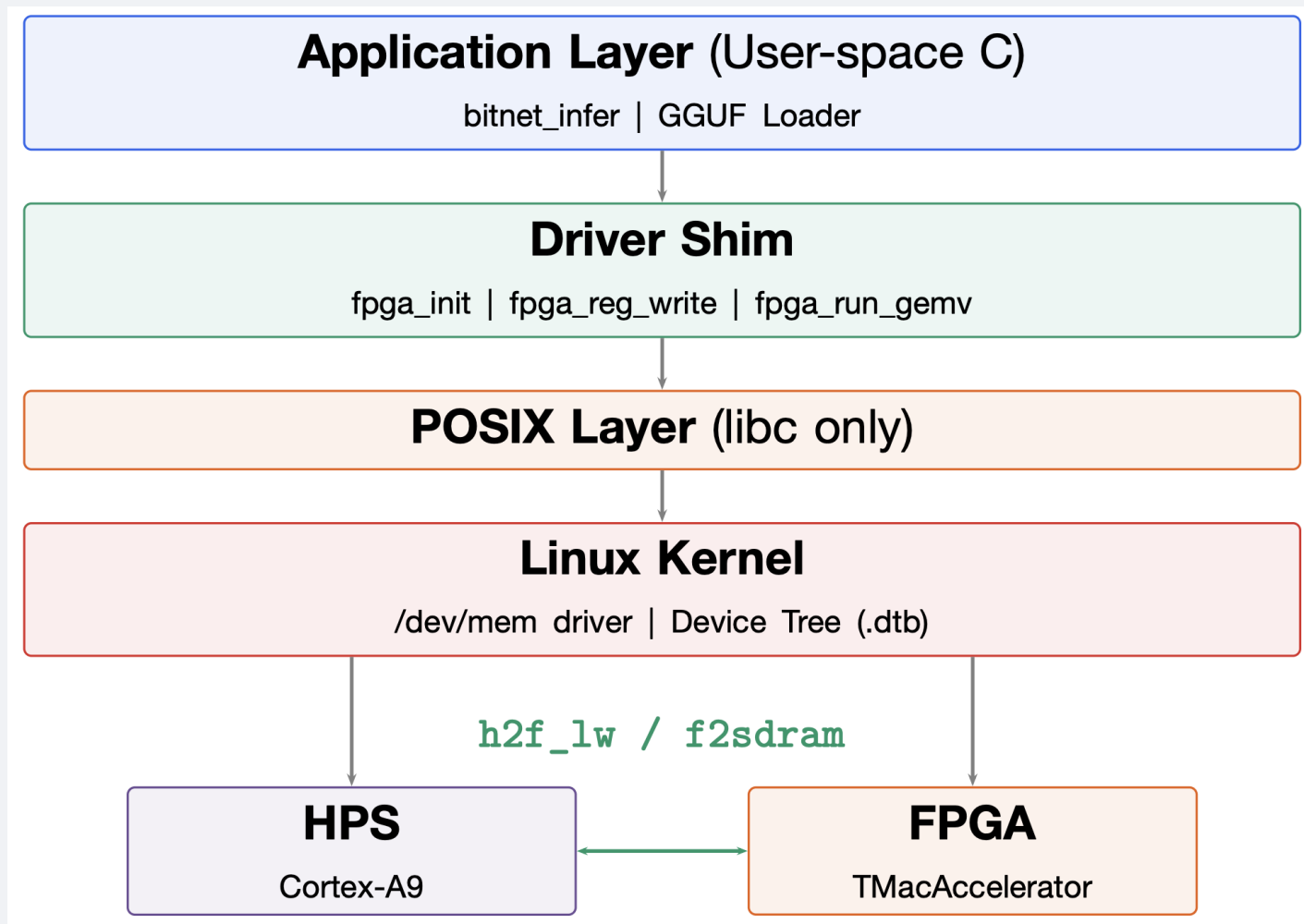
# 系統架構：Chisel + HPS Bridge + 軟體堆疊

Chisel 硬體設計與 HPS Bridge：



# 系統架構：Chisel + HPS Bridge + 軟體堆疊

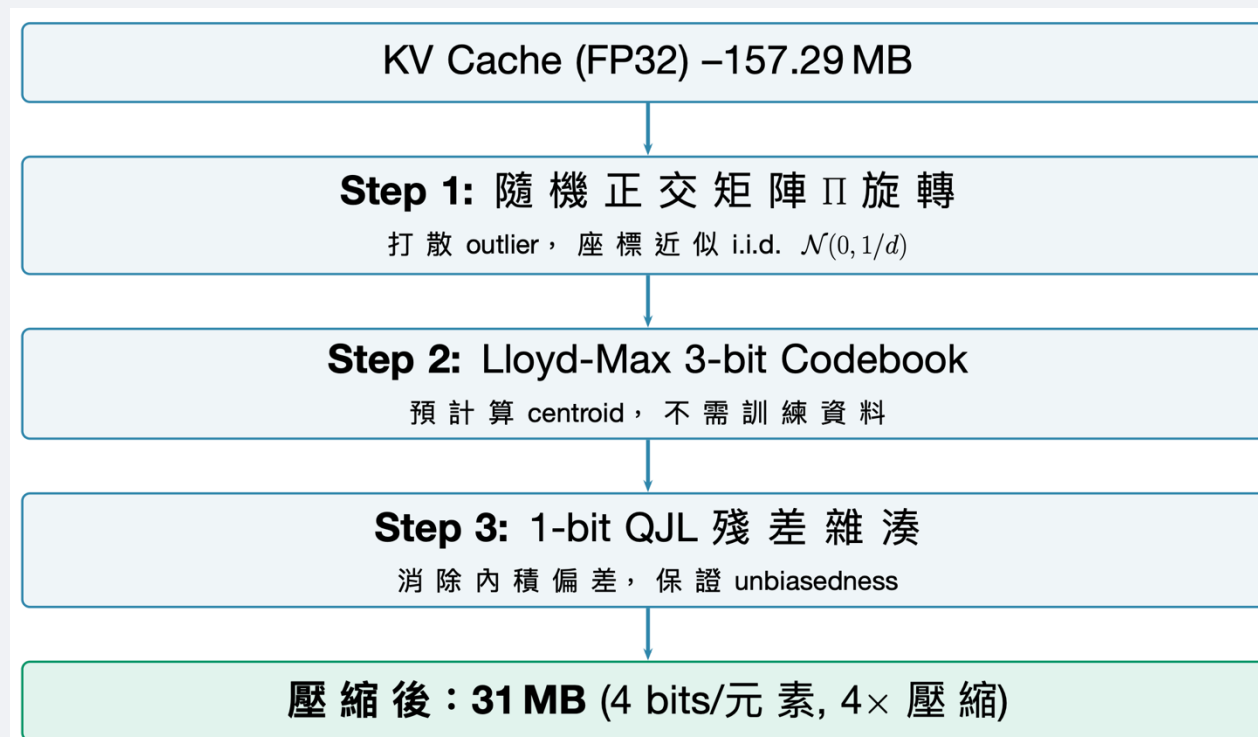
HPS 軟體堆疊：



# TurboQuant KV Cache 壓縮

問題：BitNet-b1.58-2B-4T 的 KV Cache 在 2048 tokens 時達 157.29 MB  
無法與模型權重共存於 1 GB DDR3

TurboQuant 三步壓縮：



# 雙軌記憶體策略：消滅 vs 壓縮

|          | BitMamba-2 (SSM)      | BitNet (Transformer)  |
|----------|-----------------------|-----------------------|
| KV Cache | 消滅 — 固定隱狀態 $O(1)$     | 壓縮 — TurboQuant 4x    |
| 記憶體      | 常數，與序列長度無關            | 線性，但壓縮後可容納            |
| 權重格式     | 三元 $\{-1,0,+1\}$ 混合精度 | 三元 $\{-1,0,+1\}$ 混合精度 |
| 推論核心     | T-MAC 查表              | T-MAC 查表              |
| 參數量      | 255M / 1B             | 2B-4T                 |
| 模型大小     | 250 MB / 620 MB       | 690 MB                |

兩種模型均使用 T-MAC 加速器核心 → 證明零 DSP FPGA 的通用性

# FPGA 資源使用率 + DE10-Nano 規格

Quartus (Cyclone V, 100 MHz)

| 配置     | ALMs        | M10K      | DSP |
|--------|-------------|-----------|-----|
| 8 Eng  | 3,655 (9%)  | 41 (7%)   | 0   |
| 16 Eng | 5,354 (13%) | 76 (14%)  | 0   |
| 32 Eng | 8,863 (21%) | 143 (26%) | 0   |

DE10-Nano 平台

| 規格   | 數值                |
|------|-------------------|
| SoC  | Cyclone V 5CSEBA6 |
| CPU  | ARM A9 800 MHz    |
| FPGA | 41,910 ALMs       |
| DSP  | 112 (未使用)         |
| RAM  | 1 GB DDR3 共享      |
| 功耗   | 4~5 W             |

# 推理速度 Benchmark

| 模型                   | 配置       | tok/s | RAM    |
|----------------------|----------|-------|--------|
| BitMamba-2 (255M)    | ARM-only | 3.42  | 250 MB |
|                      | ARM+FPGA | 5.85  |        |
| BitMamba-2 (1B)      | ARM-only | 0.60  | 620 MB |
|                      | ARM+FPGA | 3.0   |        |
| BitNet-b1.58 (2B-4T) | ARM-only | 0.02  | 690 MB |
|                      | ARM+FPGA | 0.75  |        |

DE10-Nano 實測數據

# 與現有邊緣 LLM 加速器比較

| 系統                | FPGA 平台    | 模型           | 效能        | DSP 使用 |
|-------------------|------------|--------------|-----------|--------|
| TeLLMe (2025)     | AMD KV260  | BitNet 0.73B | 25 tok/s  | 使用 DSP |
| LightMamba (2025) | Alveo U280 | Mamba2-2.7B  | 93 tok/s  | 使用 DSP |
| BiTurbo (ours)    | DE10-Nano  | BitMamba 1B  | 3.0 tok/s | 零 DSP  |

## 關鍵差異：

KV260 與 Alveo U280 皆為中高階 FPGA，資源量遠超 DE10-Nano

DE10-Nano 僅 41K ALMs vs KV260 的 117K LUTs vs U280 的 1.3M LUTs

本作品特色在零 DSP 條件下成功運行 1B LLM 的系統

在資源受限環境下的可行性驗證，而非追求絕對效能

# 應用場景

## 場景一：智慧農業邊緣 AI

農場環境監測 → 本地 LLM 即時分析 → 自然語言建議

隱私保護：裝置端推論，不經過雲端

低功耗：功耗 <10W，太陽能供電，長期無人部署

人機互動：Telegram Bot，語音/文字 → LLM 回覆

## 場景二：年長者居家智慧監測

語音輸入 → 即時轉譯 → LLM 辨識情境 → 判斷危險

即時警報：發送警報給長照中心或家人

隱私保護：本地離線推論

全天候：24 小時不間斷監測

共同特色：全鏈路在單一邊緣裝置完成，無需雲端、無需 GPU

# 結論與貢獻

1. 零 DSP 架構：T-MAC 查表完全取代乘法器，32 Engine 僅佔 21% ALMs
2. 雙模型支援：BitMamba-2 (SSM) + BitNet (Transformer) 均使用同一加速核心
3. TurboQuant：4x KV Cache 壓縮
4. 零廠商鎖定：完全開源可重現
5. 實測加速：LLM 推理 3.0 tok/s (1B)
6. 純 C 推論引擎：零函式庫依賴，ARM NEON 加速非矩陣運算

# 未來展望

1. BitEmbedding 壓縮：將 1B 模型 embedding 層從 614 MB 壓縮至 80-120 MB
2. 多模態擴展：整合語音辨識前處理 (Mel spectrogram)，實現語音輸入邊緣 AI
3. 推理優化：進一步優化推理過程的瓶頸，目標為可加速至記憶體頻寬的理論極限
4. 開源社群：完整開源硬體設計 + 推論引擎 + 模型轉換工具鏈

# 謝謝聆聽

基於 FPGA 之無 DSP 的 BitMamba-2 推論加速器