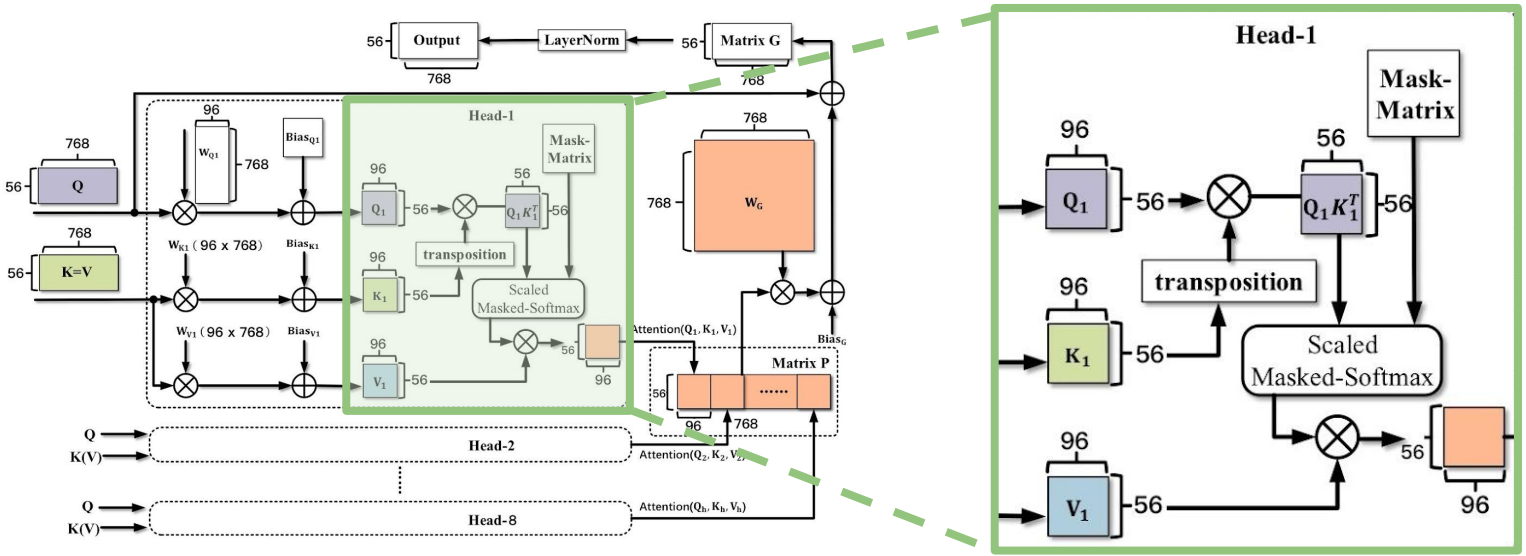
# ASoC final project report

hardware accelerator of attention mechanism using HLS

Group 4 蔡東翰

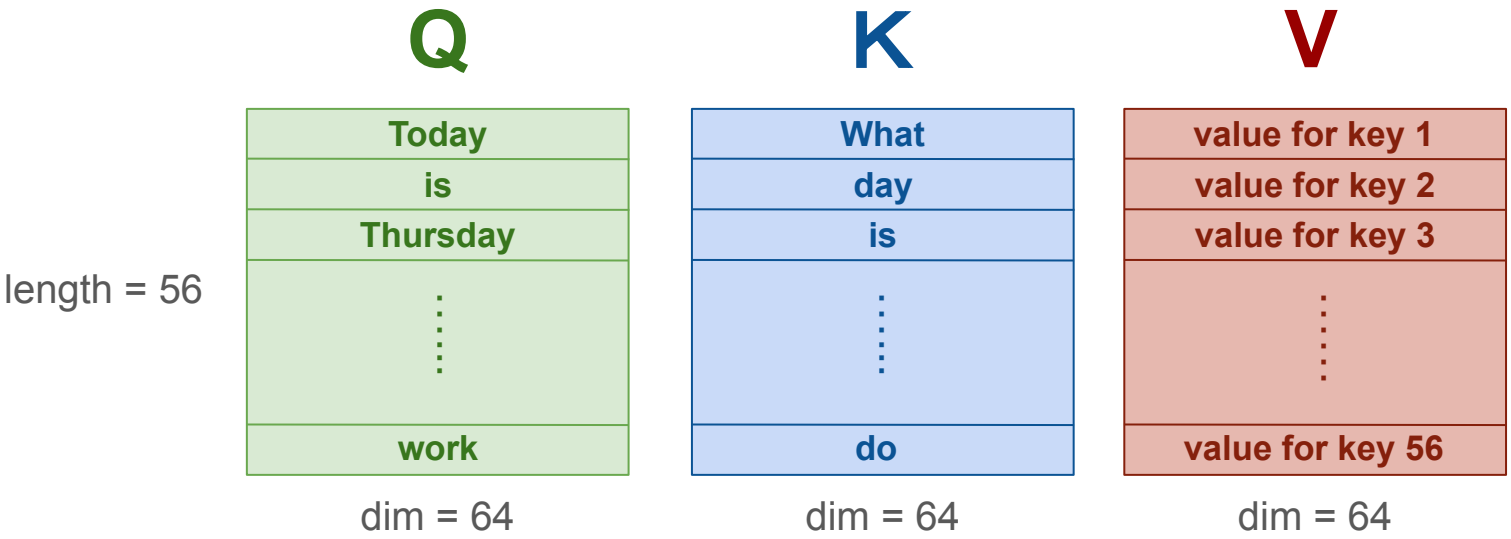
# 功能調整

- 刪減了前後需要額外權重(weight)的矩陣乘法
- 原因:其運算簡單但耗費大量時間與空間，若包含進去則無法體現真正核心部分的優化進步(在大多論文的晶片功能亦只包含核心，或許這就是原因)



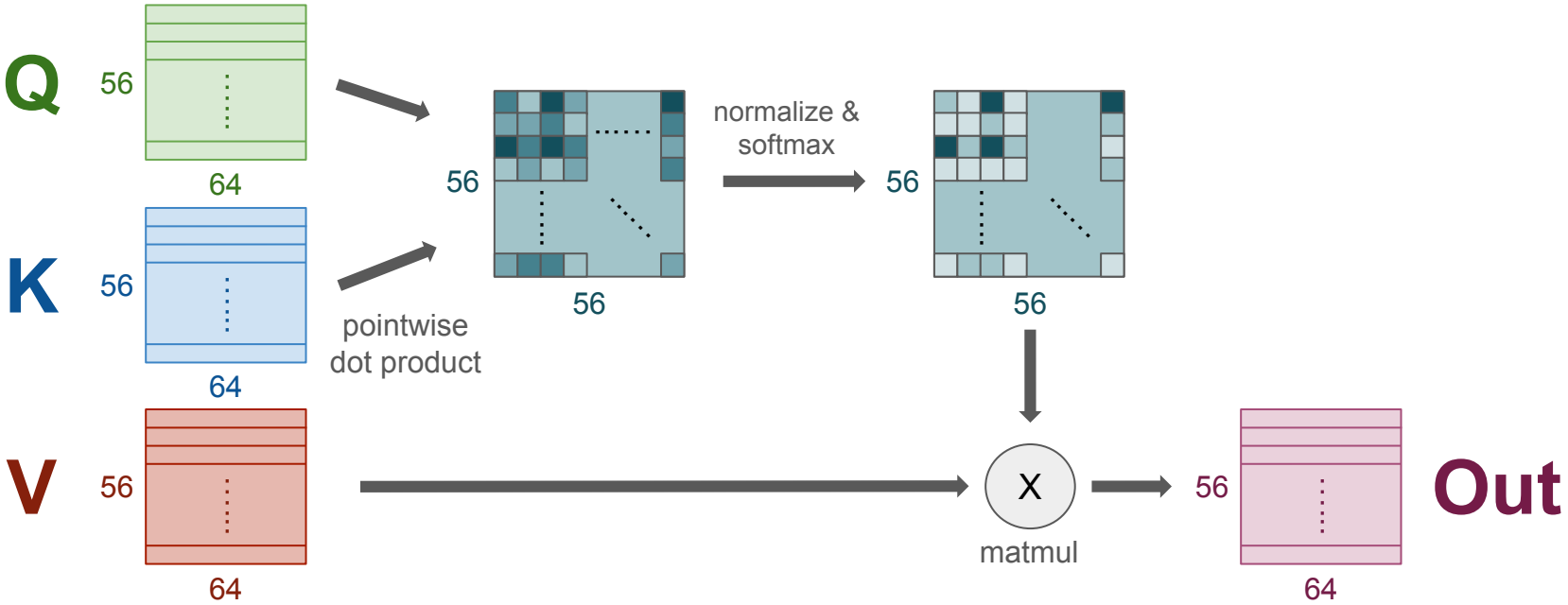
# 功能重新講述 (1)

- 輸入：會依序進行8個head的運算，其中每一個head有一組Q, K, V
- Q, K, V之大小相同，最大為56 * 64 * 16bit，Q與K可視為兩文本，V可視為權重



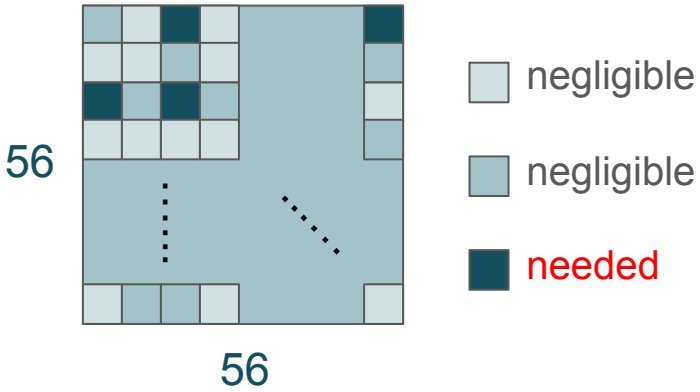
# 功能重新講述 (2)

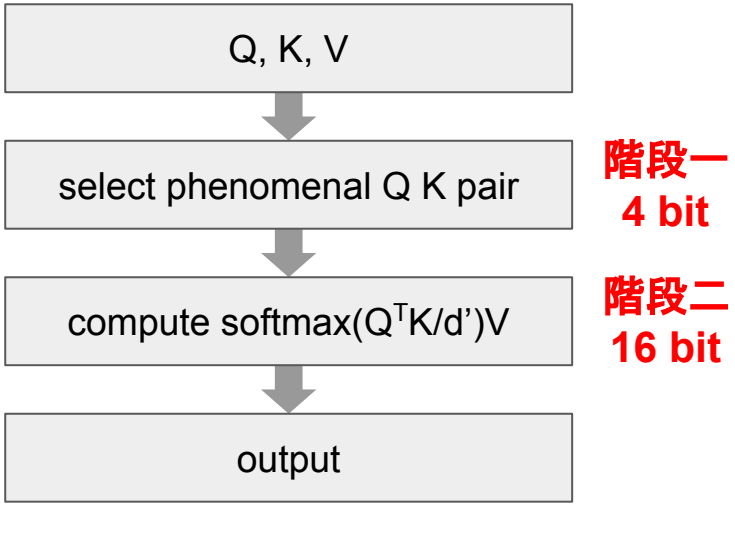
- 對於Q的每個位置，皆與K的每個位置用內積做比對，內積大代表關係強
- 比對結果經過縮放與softmax校正後，分別乘上相對應的V即為輸出



# 功能重新講述 (3)

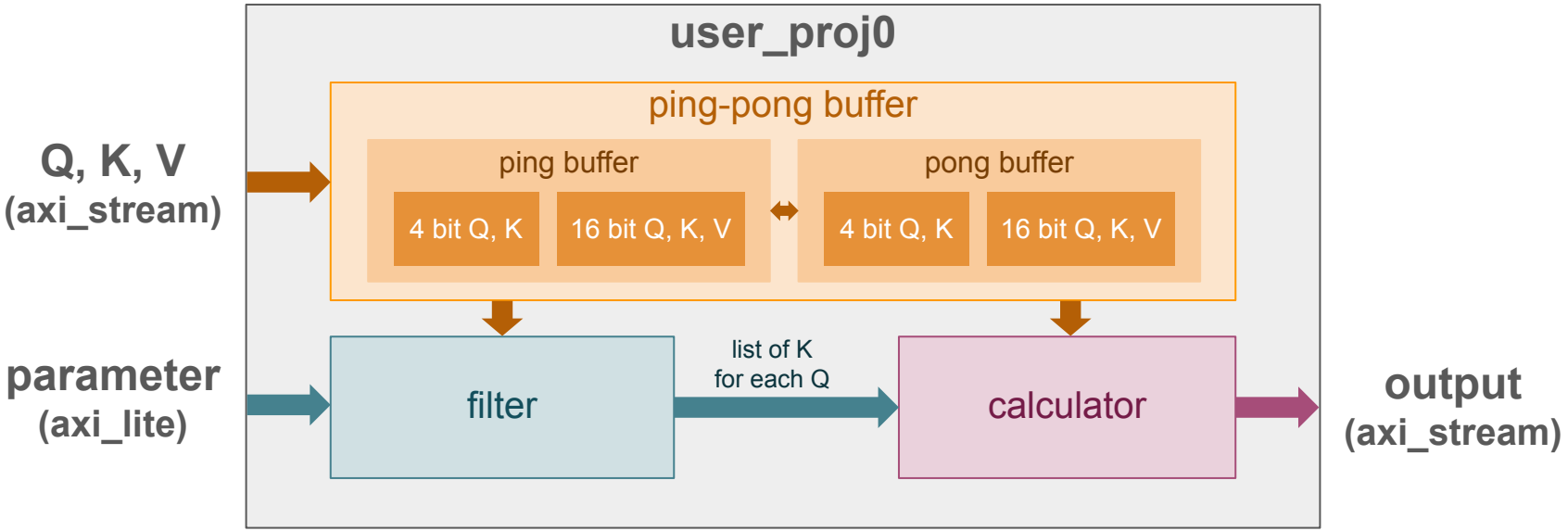
- 經過softmax會使原本較小的值變成更趨近0，等於其實不需要考慮此組Q,K pair
- 先用少bit數去尋找不考慮的Q, K pair，再對剩下來的進行full bit運算



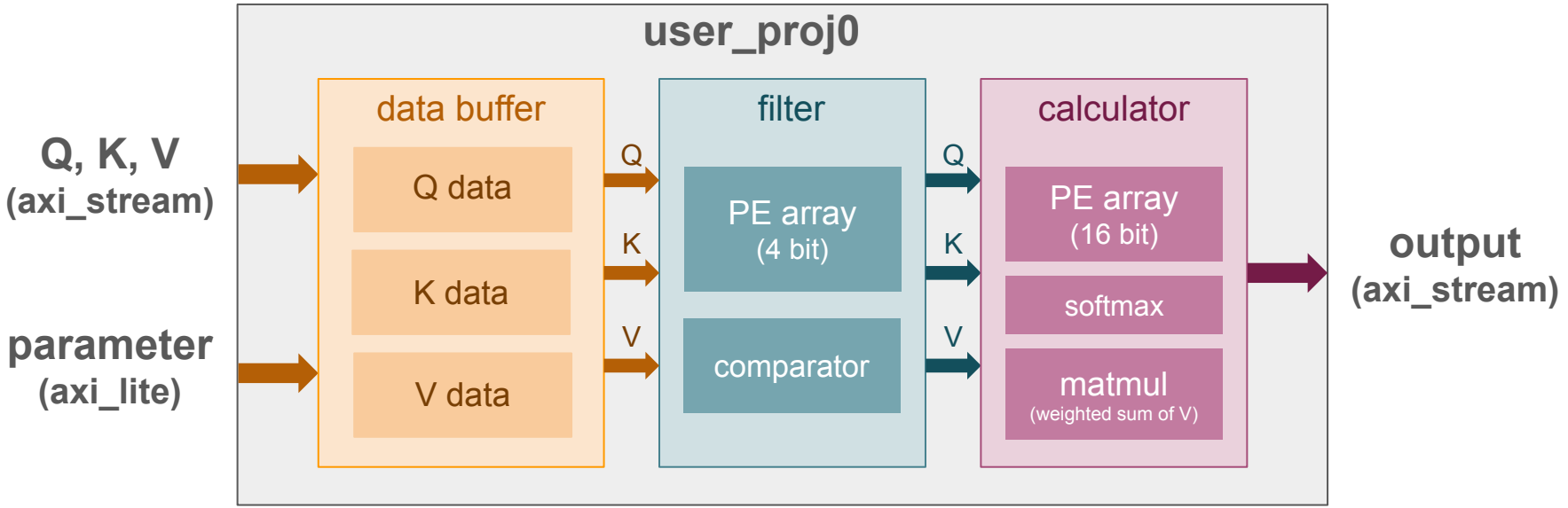
# 設計架構圖 (原本構想)

- 有一個ping-pong buffer，可以支援兩階段的運算
- 難處：ping-pong的轉換，多來源的讀取與寫入難以用hls有效率的實踐

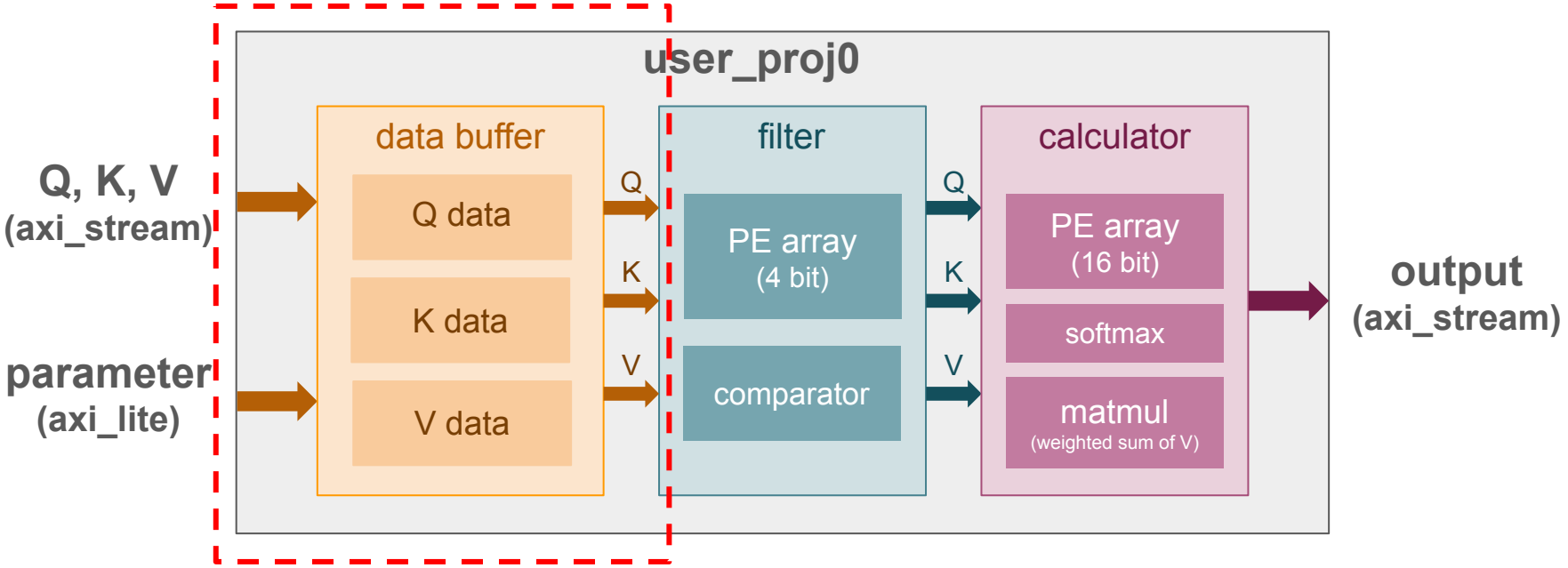


# 設計架構圖 (最終)

- 讓每個步驟變成data driven (類似lab-catapult-edge)
- 難處:中間不確定資料長度處需處理資料傳遞分界

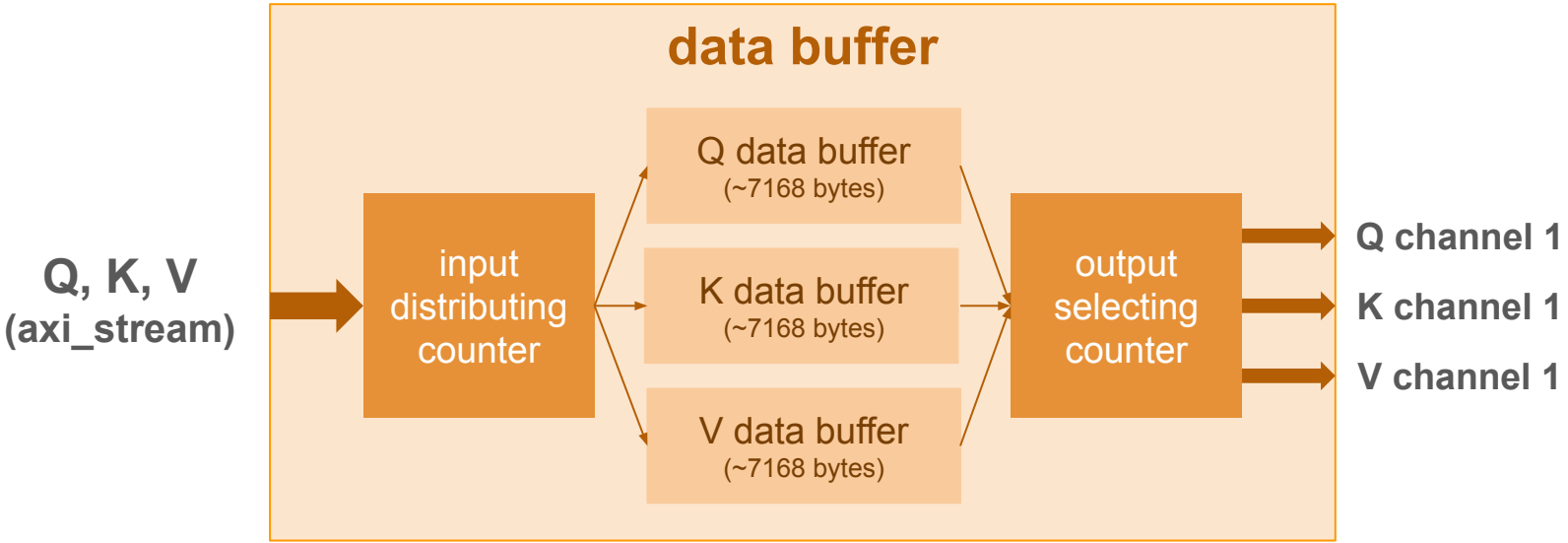


# 設計架構細部 — data buffer (1)



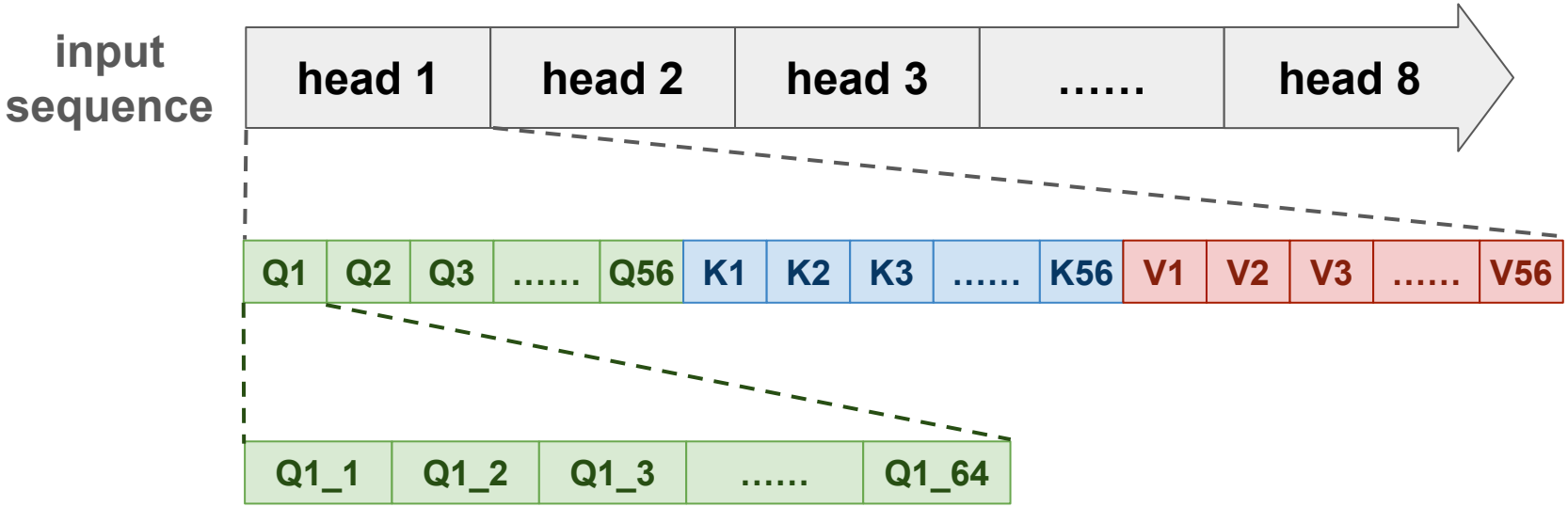
# 設計架構細部 — data buffer (2)

- 接收來自axi-stream的Q, K, V data, 將其存在內部buffer內
- 依照一定的順序將Q, K, V分三個通道傳入filter中 (Q, K, V channel 1)



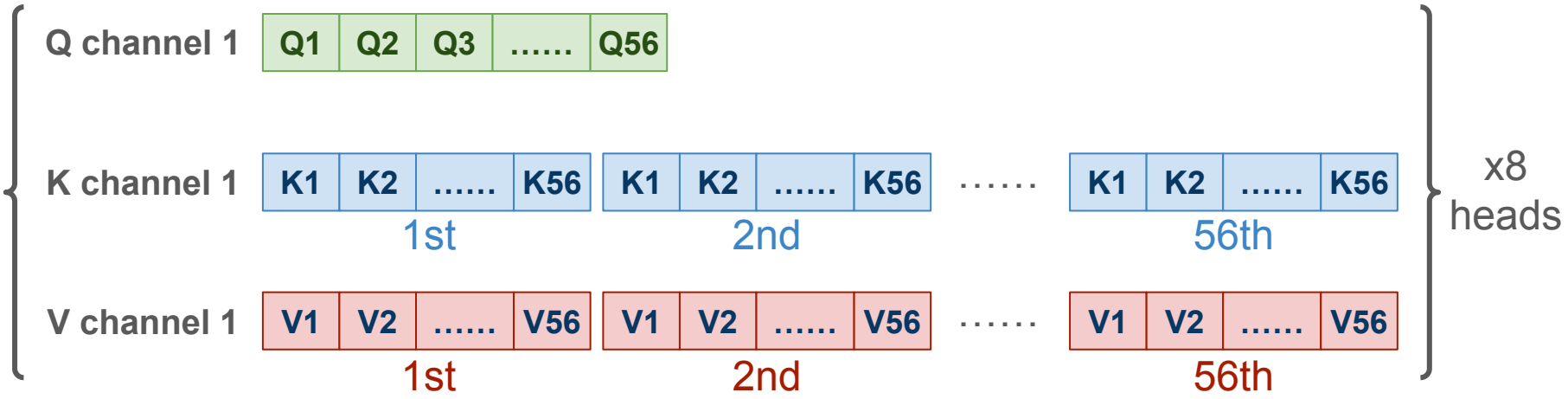
# 設計架構細部 — data buffer (3)

- 因為一次只能stream in一筆資料，故input由tb依順序傳入
- 順序：head 1 (Q □ K □ V) □ head 2 (Q □ K □ V) □ … head 8 (Q □ K □ V)

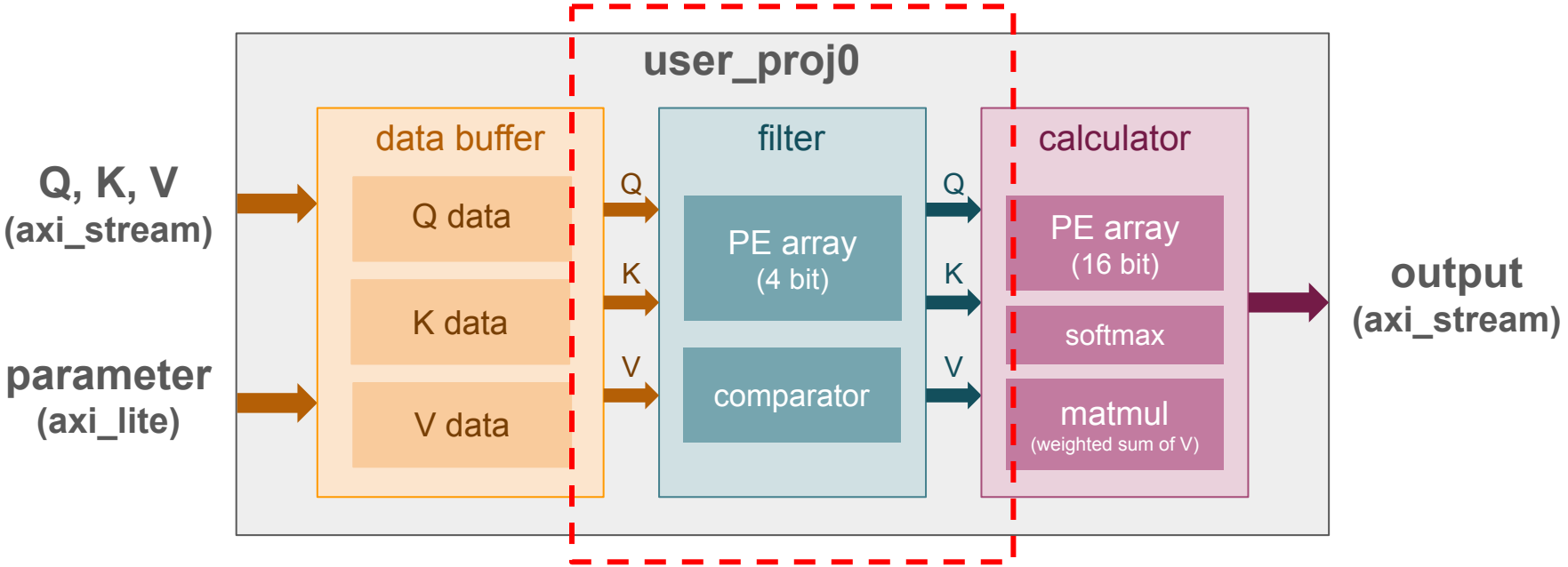


# 設計架構細部 — data buffer (4)

- 在資料存入內部buffer中後，分成三個通道將資料往下傳
- 每個head： Q通道傳入Q1~Q56, K, V通道傳入重複56次K1~K56及V1~V56
- 其中Q1代表(Q1_1, Q1_2, Q1_3, … , Q1_64)依順序之集合

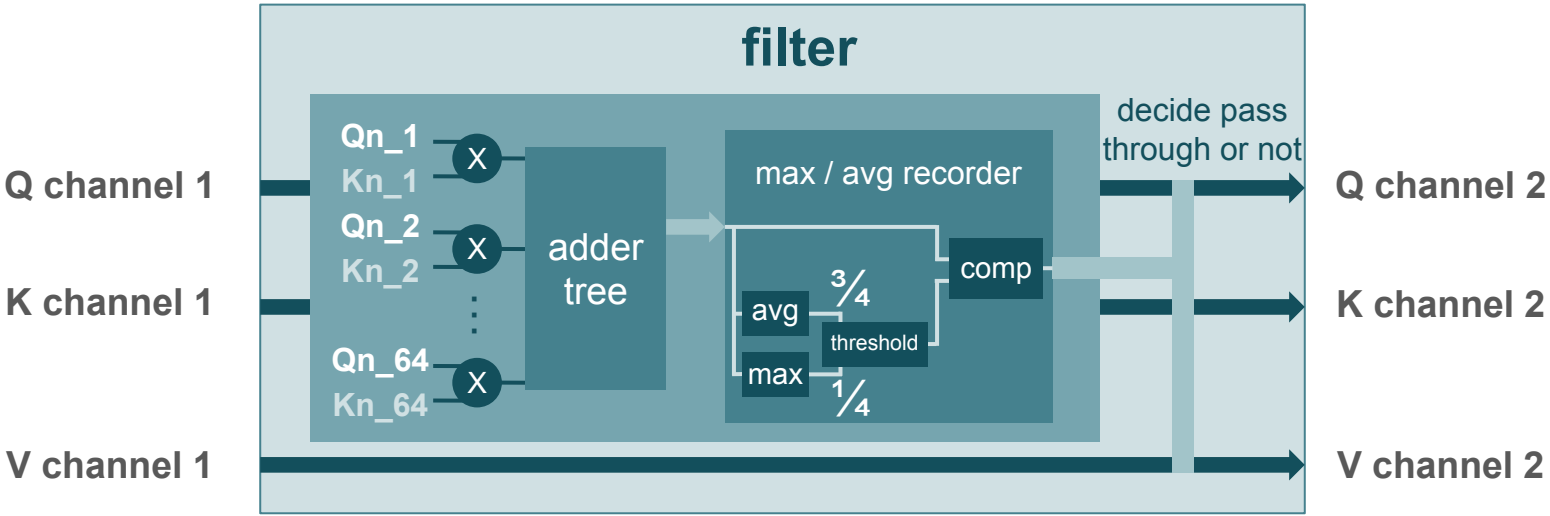


# 設計架構細部 — filter (1)



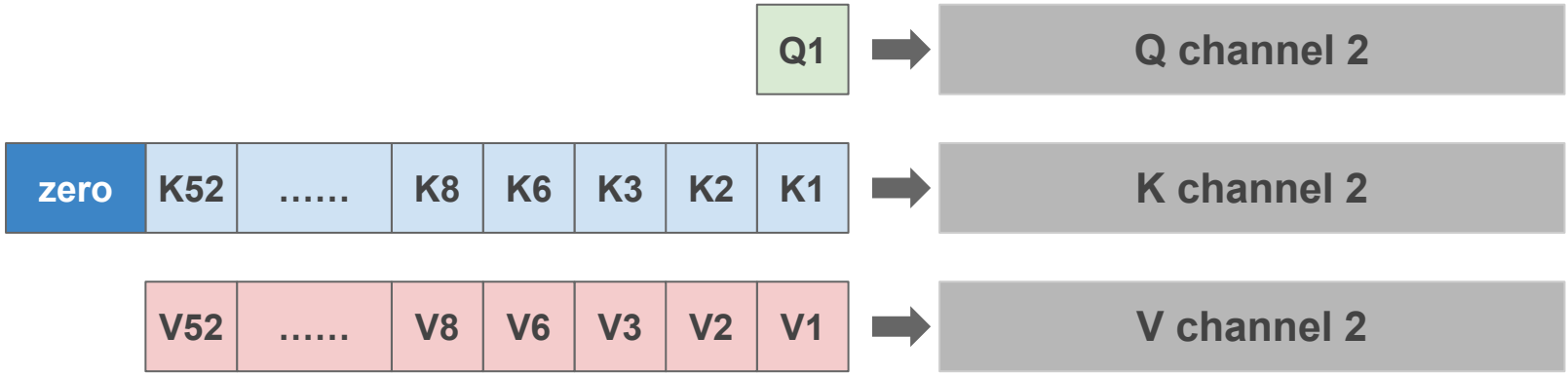
# 設計架構細部 — filter (2)

- 對進來的Q，使用每一個K與其進行4bit的內積
- 紀錄該Q目前內積的最大值與平均值，並依此判斷要不要留 (¼ max + ¾ avg)
- 若不留則直接捨棄此K, V，若留則將此K, V傳入K, V channel 2
- 有效減少每個Q需要做full bit運算的K數量 (56 □ 不到10個)

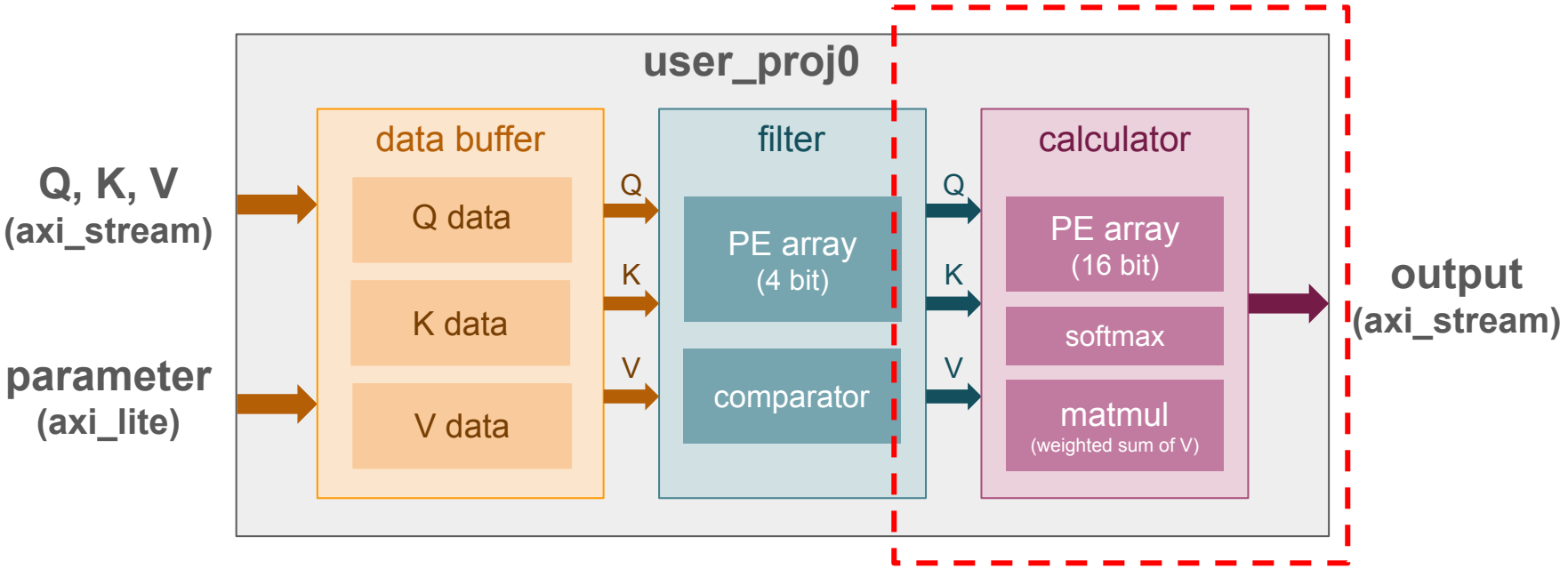


# 設計架構細部 — filter (2)

- 由於我們不能事先預知64個K中會有幾個被篩選出來，因此需要一個flag指示對於每個Q的最後一個K的位置
- 我們用全0向量作為判斷

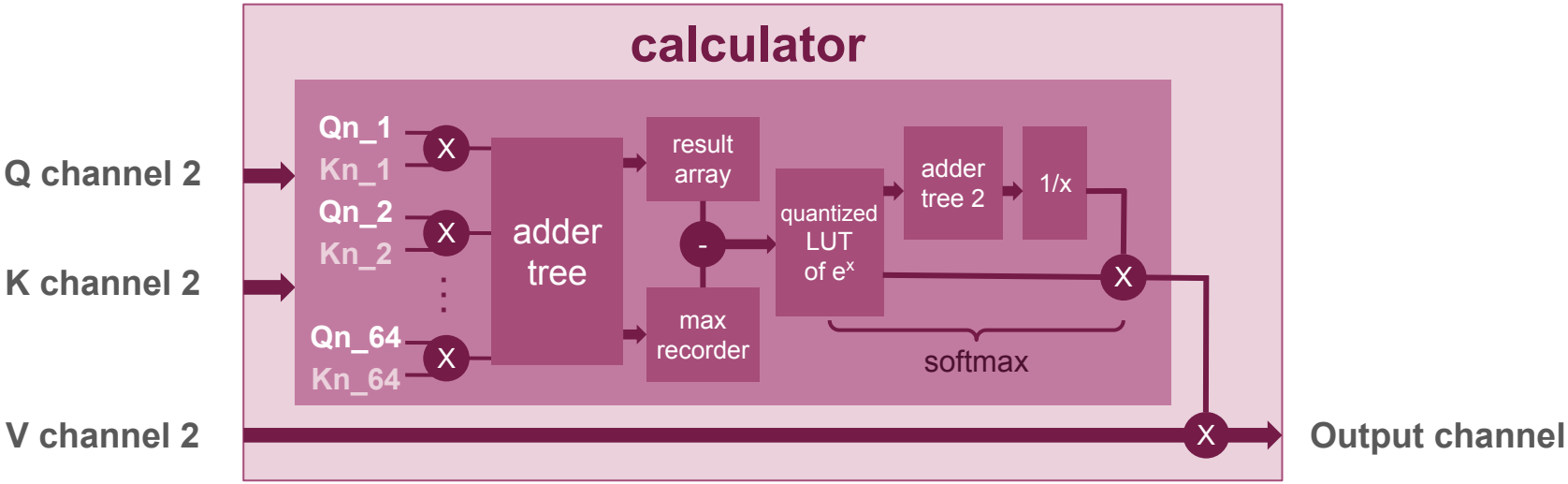


# 設計架構細部 — calculator (1)



# 設計架構細部 — calculator (2)

- 對進來的Q，使用每一個被篩選過的K與其進行16bit的內積
- 存下對每個Q的所有內積結果，將其做對dim做縮放後再softmax
- softmax結果乘上相對應的V並作累加就得到輸出



# CDesignChecker

| FATAL | Violated | Waived |
|---|---|---|
| | | |
| **ERROR** | **Violated** | **Waived** |
| AOB - Arithmetic Operator with Boolean | 0 | 0 |
| RRT - Reset referenced in thread | 0 | 0 |
| | | |
| **WARNING** | **Violated** | **Waived** |
| ACC - Accumulator of native C type | 0 | 0 |
| ACS - Accumulator of saturated type | 0 | 0 |
| AIC - Assignment used Instead of Comparison | 0 | 0 |
| ALS - Ac_int Left Shift check | 0 | 0 |
| AWE - Assignments Without Effect | 0 | 0 |
| CBU - Conditional break in Unrolled Loop | 0 | 0 |
| CCC - Static constant comparison | 0 | 0 |
| CGR - Conditional Guard in Rolled Loop | 1 | 0 |
| CIA - Comparison Instead of Assignment | 0 | 0 |
| CNS - Constant condition of if/switch | 0 | 0 |
| CWB - Case Without Break | 0 | 0 |
| DIU - Dynamic Index in Unrolled Loop | 0 | 0 |
| FVI - For Loop with Variable Iterations | 19 | 0 |
| FXD - Mixed fixed and non-fixed datatypes | 0 | 0 |
| MDB - Missing Default Branch | 0 | 0 |
| NCO - No Contribution to Output | 0 | 0 |
| OSA - Optimal Size Accumulator | 3 | 0 |
| PDD - Platform dependent datatype (long) | 0 | 0 |
| RIU - Rolled loop Inside Unrolled loop | 0 | 0 |
| SAT - Sub-optimal Adder Tree | 0 | 0 |
| SUD - Suboptimal Use of Divide and Modulus Operator | 1 | 0 |

# HLS 合成策略

- **迴圈**

  乘除法迴圈:pipeline (ii=1) (增加乘除法器的使用效率)
  賦值迴圈:unroll (簡單的操作可以一次做完)
  使用channel迴圈:no change

- **通道**

  使用ccs_in/out_wait (input 是wait coupled)

- **儲存單元**

  buffer中儲存一個head的資料:RAM 1R/1W (減少面積)
  其餘:register (方便讀存資料)

# QuestaSIM result

```
# Checking results
# 'q_chan1_data'
#    capture count        = 448
#    comparison count     = 448
#    ignore count         = 0
#    error count          = 0
#    stuck in dut fifo    = 0
#    stuck in golden fifo = 0
# 'k_chan1_data'
#    capture count        = 25088
#    comparison count     = 25088
#    ignore count         = 0
#    error count          = 0
#    stuck in dut fifo    = 0
#    stuck in golden fifo = 0
# 'v_chan1_data'
#    capture count        = 25088
#    comparison count     = 25088
#    ignore count         = 0
#    error count          = 0
#    stuck in dut fifo    = 0
#    stuck in golden fifo = 0
#
# Info: scverify_top/user_tb: Simulation PASSED @ 35150166 ns
# ** Note: (vsim-6574) SystemC simulation stopped by user.
```

buffer

```
# Checking results
# 'q_chan2_data'
#    capture count        = 448
#    comparison count     = 448
#    ignore count         = 0
#    error count          = 0
#    stuck in dut fifo    = 0
#    stuck in golden fifo = 0
# 'k_chan2_data'
#    capture count        = 9004
#    comparison count     = 9004
#    ignore count         = 0
#    error count          = 0
#    stuck in dut fifo    = 0
#    stuck in golden fifo = 0
# 'v_chan2_data'
#    capture count        = 8556
#    comparison count     = 8556
#    ignore count         = 0
#    error count          = 0
#    stuck in dut fifo    = 0
#    stuck in golden fifo = 0
#
# Info: scverify_top/user_tb: Simulation PASSED @ 17709766 ns
# ** Note: (vsim-6574) SystemC simulation stopped by user.
```

filter

```
# Info: Collecting data completed
#    captured 1 values of head
#    captured 1 values of length
#    captured 1 values of dim
#    captured 448 values of q_chan2_data
#    captured 9004 values of k_chan2_data
#    captured 8556 values of v_chan2_data
#    captured 28672 values of dout_chan
# Info: scverify_top/user_tb: Simulation completed
#
# Checking results
# 'dout_chan'
#    capture count        = 28672
#    comparison count     = 28672
#    ignore count         = 0
#    error count          = 0
#    stuck in dut fifo    = 0
#    stuck in golden fifo = 0
#
# Info: scverify_top/user_tb: Simulation PASSED @ 12405486 ns
# ** Note: (vsim-6574) SystemC simulation stopped by user.
```

calculator

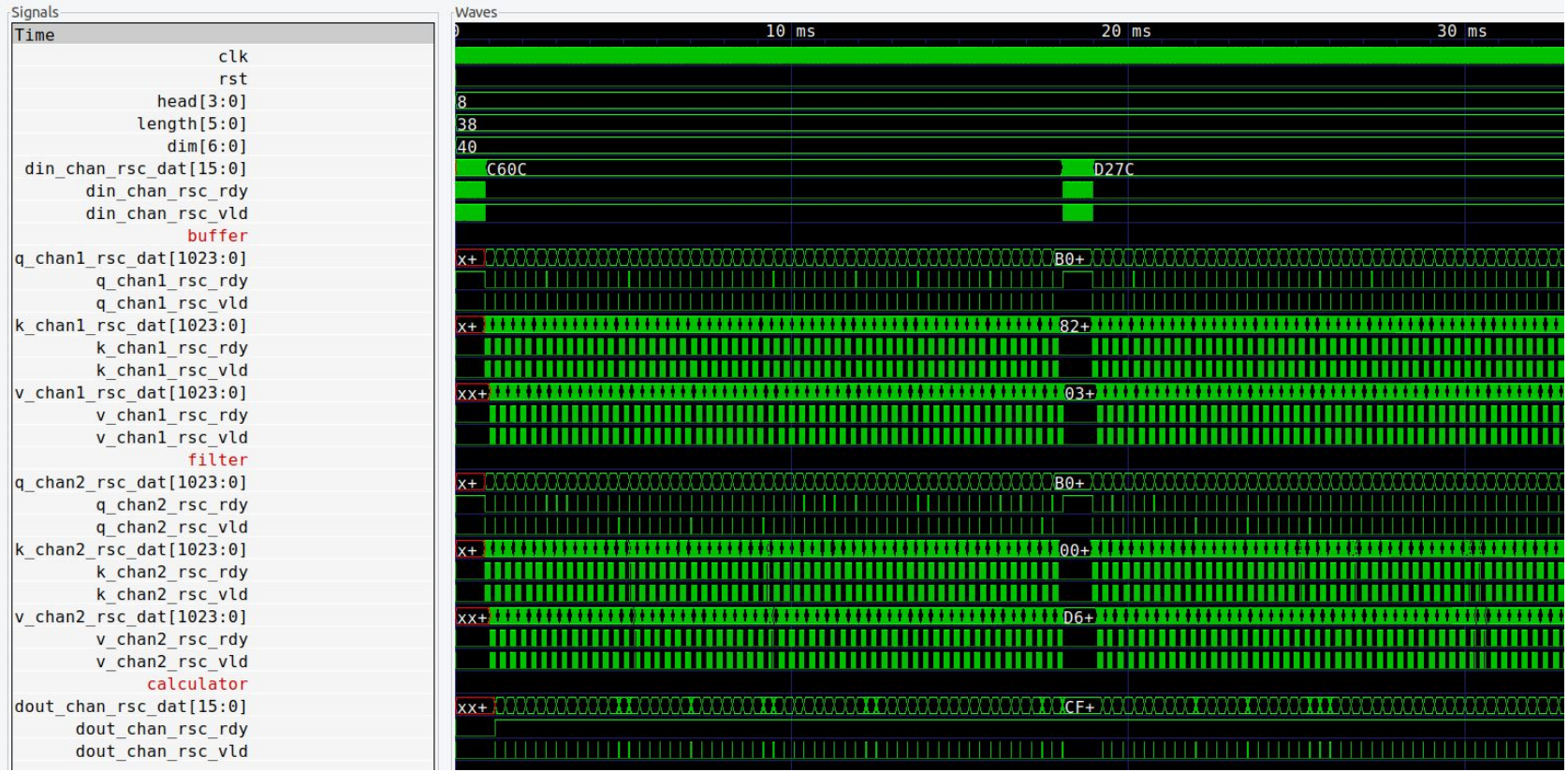
# 合成結果

Start Page | Table | Constraint Editor

Report: General

| Solution | Latency... | Latency... | Throug... | Throug... | Slack | Total Ar... |
|---|---|---|---|---|---|---|
| ATTENTION_IP::Attention_Buffer.v2 (extract) | 263 | 2630.00 | 268 | 2680.00 | 7.89 | 26409.68 |
| ATTENTION_IP::Attention_Filter.v2 (extract) | 12 | 120.00 | 17 | 170.00 | 7.54 | 28937.94 |
| ATTENTION_IP::Attention_Calculator.v4 (extract) | 12 | 120.00 | 17 | 170.00 | 3.30 | 110989.22 |
| **ATTENTION_IP::Attention_Top.v8** (extract) | **0** | **0.00** | **0** | **0.00** | **3.30** | **166336.84** |

# 執行結果 (FSIC)


Signals
Time
clk
rst
head[3:0]
length[5:0]
dim[6:0]
din_chan_rsc_dat[15:0]
din_chan_rsc_rdy
din_chan_rsc_vld
buffer
q_chan1_rsc_dat[1023:0]
q_chan1_rsc_rdy
q_chan1_rsc_vld
k_chan1_rsc_dat[1023:0]
k_chan1_rsc_rdy
k_chan1_rsc_vld
v_chan1_rsc_dat[1023:0]
v_chan1_rsc_rdy
v_chan1_rsc_vld
filter
q_chan2_rsc_dat[1023:0]
q_chan2_rsc_rdy
q_chan2_rsc_vld
k_chan2_rsc_dat[1023:0]
k_chan2_rsc_rdy
k_chan2_rsc_vld
v_chan2_rsc_dat[1023:0]
v_chan2_rsc_rdy
v_chan2_rsc_vld
calculator
dout_chan_rsc_dat[15:0]
dout_chan_rsc_rdy
dout_chan_rsc_vld
Waves
10 ms
20 ms
30 ms
8
38
40
C60C
D27C
B0+
82+
03+
B0+
00+
D6+
CF+

# 遇到問題與解決

- **問題:** HLS合成出來的module一直沒有拉高接收input data的ready

  **解決:** 後來發現是read channel寫在迴圈內，而迴圈沒有把結束條件拉出for
- **問題:** HLS合成出來兩個module中間傳遞卡住

  **解決:** data dependency沒有處理好，前一個module : (傳送K1, 傳送V1, 傳送K2, 傳送V2)，後一個module : (接收K1, 接收K2, ……)
- **問題:** 放到FPGA上無法成功讀取到輸出值

  **解決:** TBD……

# Github link

- https://github.com/DonghanTsai/asoc_final

# 參考資料

- https://arxiv.org/pdf/2110.09310
- https://github.com/bol-edu/caravel-soc_fpga-lab/tree/main/catapult_hls