

TRƯỜNG ĐẠI HỌC BÁCH KHOA TP.HCM
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH



ĐỒ ÁN THIẾT KẾ
KỸ THUẬT MÁY TÍNH

Thiết kế SoC RISC-V tích hợp EdgeAI
cho ứng dụng IoT

Học kỳ 251

GVHD: PGS. TS. Trần Ngọc Thịnh
ThS. Huỳnh Phúc Nghị

STT	Họ và tên	MSSV	Ghi chú
1	Lâm Nữ Uyển Nhi	2212429	
2	Vũ Đức Lâm	2211824	

TP. Hồ Chí Minh, Tháng 12/2025

Mục lục

Danh mục Ký hiệu và Chữ viết tắt	viii
1 Giới thiệu đề tài	1
1.1 Đặt vấn đề	1
1.2 Mục tiêu đề tài	1
1.3 Phạm vi đề tài	1
1.4 Phân chia công việc	1
1.5 Cấu trúc báo cáo	1
2 Cơ sở lý thuyết	2
2.1 Tổng quan về Mạng nơ-ron tích chập (CNN)	2
2.2 Kỹ thuật thiết kế bộ tăng tốc phần cứng (AI Accelerator)	2
2.3 Kiến trúc System-on-Chip trên nền tảng FPGA	2
3 Phân tích yêu cầu và Kiến trúc tổng quan	3
3.1 Phân tích yêu cầu thiết kế	3
3.2 Kiến trúc hệ thống tổng quan	3
3.3 Đặc tả giao diện kết nối	3
4 Thiết kế kiến trúc Accelerator và Chiến lược dòng dữ liệu	4
4.1 Phân tích toán học của phép tính tích chập	4
4.1.1 Standard Convolution (Tích chập tiêu chuẩn) . .	5
4.1.1.1 Công thức toán học tổng quát	5

4.1.1.2	Cấu trúc vòng lặp (Loop Nest)	5
4.1.2	Depthwise Separable Convolution	6
4.1.2.1	Depthwise Convolution (DW)	7
4.1.2.2	Pointwise Convolution (PW)	8
4.1.3	Yêu cầu đối với Kiến trúc thống nhất (Unified Architecture)	8
4.1.4	Kỹ thuật Gập Batch Normalization (BN Folding)	8
4.1.4.1	Công thức biến đổi trọng số	9
4.2	Chiến lược phân mảnh và Dòng dữ liệu để xuất	10
4.2.1	Định nghĩa khái niệm "Tile" (Mảnh dữ liệu)	10
4.2.2	Phương pháp Phân mảnh không gian dữ liệu (Space Partitioning)	11
4.2.2.1	Công thức chia khối (Block Calculation)	11
4.2.3	Mô hình hóa toán học và Tham số thiết kế	12
4.2.3.1	Công thức tính kích thước không gian toàn cục	12
4.2.3.2	Tính toán kích thước Output Tile theo cơ chế Ping-Pong	13
4.2.3.3	Phân tích số lượng Pass và Dữ liệu biên	13
4.2.3.4	Trường hợp Depthwise Convolution	14
4.2.4	Thuật toán Điều phối Pass (Pass Scheduling)	14
4.2.4.1	Thuật toán cho Standard Convolution	15
4.2.4.2	Thuật toán cho Depthwise Convolution	15
4.2.5	Phân tích vấn đề tại biên và Dữ liệu dội ra	16
4.2.5.1	Cơ sở hình thành Dữ liệu dội ra	17
4.2.6	Cơ chế Ping-Pong Buffer và Logic xử lý hàng hợp lệ	18
4.2.7	Thuật toán Điều phối và Xoay vòng bộ nhớ	19
4.3	Thiết kế kiến trúc vi mô (Micro-architecture)	23
4.3.1	Sơ đồ khối tổng quát hệ thống	23

4.3.2	Tổ chức Mảng tính toán (Processing Hierarchy)	25
4.3.2.1	Mảng xử lý (Process Array - PA)	25
4.3.2.2	Đơn vị xử lý (Process Unit - PU)	25
4.3.2.3	Phần tử xử lý (Process Element - PE)	26
4.3.3	Đánh giá thời gian thực thi (Performance Estimation)	27
4.3.3.1	Thời gian xử lý một Pass cơ sở (T_{pass})	27
4.3.3.2	Tổng thời gian thực thi (T_{total})	28
4.3.4	Chiến lược Chia giấu độ trễ và Mô hình hiệu năng toàn hệ thống	28
4.3.4.1	Cơ chế hoạt động	29
4.3.4.2	Các kịch bản hiệu năng (Performance Scenarios)	29
4.3.4.3	Tổng thời gian toàn mạng (Model Latency)	32
4.4	Tối ưu hóa Tham số và Cơ chế Sinh mã cấu hình	32
4.4.1	Bài toán Tối ưu hóa Không gian thiết kế	32
4.4.1.1	Các ràng buộc phần cứng (Hardware Constraints)	33
4.4.2	Chiến lược Tìm kiếm và Sinh mã (Search & Generate)	33
4.4.3	Cấu trúc Lệnh cấu hình (Layer Descriptor)	34
5	Hiện thực nền tảng SoC	35
5.1	Môi trường và Công cụ hiện thực	35
5.2	Cấu hình hệ thống xử lý (Processing System)	35
5.3	Thiết kế hệ thống kết nối (Interconnect Subsystem)	35
5.4	Tích hợp và Kiểm thử nền tảng cơ sở	35
6	Đánh giá hiệu năng lý thuyết	36
6.1	Phương pháp đánh giá: Mô hình Roofline	36

6.2	Ước lượng độ trễ và Tài nguyên	36
6.3	So sánh với các nghiên cứu liên quan	36
7	Kế hoạch phát triển	37
7.1	Đánh giá mức độ hoàn thành Giai đoạn 1	37
7.2	Kế hoạch thực hiện Giai đoạn 2	37
7.3	Tiến độ dự kiến	37

Danh sách hình vẽ

Danh sách bảng biểu

Table 4.1 Bảng tham số thiết kế và ánh xạ ký hiệu 12

Table 4.2 Cấu trúc dữ liệu cấu hình cho một Layer 34

Danh mục Ký hiệu và Chữ viết tắt

Ký hiệu	Ý nghĩa
H, W	Chiều cao và chiều rộng của đặc trưng đầu vào (Input Feature Map)
C	Số lượng kênh đầu vào (Input Channels)
N_f	Số lượng bộ lọc / Số kênh đầu ra (Number of Filters / Output Channels)
H_{out}, W_{out}	Chiều cao và chiều rộng của đặc trưng đầu ra (Output Feature Map)
R, S	Chiều cao và chiều rộng của bộ lọc (Kernel Height, Kernel Width)
P	Kích thước vùng đệm (Padding)
Str (hoặc U)	Bước trượt (Stride)
T_h	Chiều cao của mảnh dữ liệu đầu vào trong một Pass (Tile Height)
T_c	Số kênh đầu vào được xử lý song song trong một Pass (Tile Input Channels)
T_m	Số bộ lọc được tính toán song song trong một Pass (Tile Output Channels)
T_{ho}	Chiều cao hợp lệ của mảnh dữ liệu đầu ra trong một Pass

Ký hiệu	Ý nghĩa
b	Số chu kỳ đồng hồ để truyền một giá trị dữ liệu (Cycles per Data Transfer)
T_{comp}	Thời gian tính toán (Computation time)
T_{load}	Thời gian nạp dữ liệu (Load time)
T_{store}	Thời gian ghi dữ liệu (Store time)
T_{pass}	Thời gian hoàn thành một Pass
I	Tensor dữ liệu đầu vào
O	Tensor dữ liệu đầu ra
W	Tensor trọng số (Weights)
B	Vector hệ số chêch (Bias)
μ, σ	Giá trị trung bình (Mean) và Phương sai (Variance) trong Batch Norm
γ, β	Tham số tỉ lệ (Scale) và dịch chuyển (Shift) trong Batch Norm

Viết tắt	Ý nghĩa
AI	Trí tuệ nhân tạo (Artificial Intelligence)
CNN	Mạng nơ-ron tích chập (Convolutional Neural Network)
DNN	Mạng nơ-ron sâu (Deep Neural Network)
FPGA	Mảng cổng lập trình được dạng trường (Field-Programmable Gate Array)
SoC	Hệ thống trên chip (System-on-Chip)
RTL	Mức chuyển giao thanh ghi (Register Transfer Level)
IFM	Đặc trưng đầu vào (Input Feature Map)
OFM	Đặc trưng đầu ra (Output Feature Map)
PE	Phần tử xử lý (Processing Element)
PU	Đơn vị xử lý (Processing Unit - Chứa nhiều PE)
PA	Mảng xử lý (Process Array - Chứa nhiều PU)
MAC	Phép tính Nhân-Cộng tích lũy (Multiply-Accumulate)

Ký hiệu	Ý nghĩa
BRAM	Block RAM (Bộ nhớ nội trên FPGA)
DMA	Truy cập bộ nhớ trực tiếp (Direct Memory Access)
AXI	Giao diện mở rộng nâng cao (Advanced eXtensible Interface)
DSP	Digital Signal Processing (Khối xử lý tín hiệu số trên FPGA)
LUT	Bảng tra (Look-Up Table)
FF	Flip-Flop
DRAM	Bộ nhớ truy cập ngẫu nhiên động (Dynamic Random Access Memory - Bộ nhớ ngoài)
FSM	Máy trạng thái hữu hạn (Finite State Machine)

Chapter 1

Giới thiệu đề tài

Chương này trình bày tổng quan về bối cảnh nghiên cứu, xác định mục tiêu cụ thể, phạm vi thực hiện và phân công nhiệm vụ giữa các thành viên trong nhóm.

1.1 Đặt vấn đề

1.2 Mục tiêu đề tài

1.3 Phạm vi đề tài

1.4 Phân chia công việc

1.5 Cấu trúc báo cáo

Chapter 2

Cơ sở lý thuyết

Chương này cung cấp các kiến thức nền tảng về Mạng nơ-ron tích chập (CNN), các kỹ thuật thiết kế phần cứng cho AI và kiến trúc System-on-Chip trên FPGA.

2.1 Tổng quan về Mạng nơ-ron tích chập (CNN)

2.2 Kỹ thuật thiết kế bộ tăng tốc phần cứng (AI Accelerator)

2.3 Kiến trúc System-on-Chip trên nền tảng FPGA

Chapter 3

Phân tích yêu cầu và Kiến trúc tổng quan

Chương này phân tích các ràng buộc thiết kế từ đó để xuất kiến trúc tổng thể của hệ thống SoC tích hợp AI Accelerator.

3.1 Phân tích yêu cầu thiết kế

3.2 Kiến trúc hệ thống tổng quan

3.3 Đặc tả giao diện kết nối

Chapter 4

Thiết kế kiến trúc Accelerator và Chiến lược dòng dữ liệu

Chương này trình bày chi tiết thiết kế của lõi IP Accelerator, bao gồm phân tích toán học, chiến lược tối ưu dòng dữ liệu và kiến trúc vi mô.

4.1 Phân tích toán học của phép tính tích chập

Để đảm bảo tính linh hoạt cho kiến trúc phần cứng, giúp hệ thống có khả năng hỗ trợ đa dạng các mô hình mạng nơ-ron từ kinh điển (như VGG16) đến các mô hình tối ưu cho thiết bị biên (như MobileNet), nhóm thực hiện đề tài đã tập trung phân tích đặc tả toán học của hai loại phép tính cốt lõi: **Standard Convolution** và **Depthwise Separable Convolution**.

Việc hiểu rõ bản chất toán học và cấu trúc dữ liệu của các phép tính này (bao gồm cả cơ chế xử lý biên - Padding) là cơ sở quan trọng để chúng tôi thiết kế nên một kiến trúc thống nhất (Unified Architecture).

4.1.1 Standard Convolution (Tích chập tiêu chuẩn)

Đây là phép tính nền tảng trong hầu hết các mạng CNN truyền thống. Về mặt toán học, tích chập tiêu chuẩn thực hiện việc trượt bộ lọc (filter) trên không gian đầu vào (H, W) , đồng thời tích lũy giá trị qua toàn bộ chiều sâu của kênh (Channels).

4.1.1.1 Công thức toán học tổng quát

Xét một lớp tích chập với đầu vào I có kích thước $C \times H_{in} \times W_{in}$ và bộ trọng số W có kích thước $M \times C \times R \times S$. Tham số Padding (P) được sử dụng để giữ nguyên kích thước không gian hoặc kiểm soát việc giảm kích thước. Giá trị đầu ra O tại kênh m , vị trí (h, w) được xác định bởi:

$$O[m][h][w] = B[m] + \sum_{c=0}^{C-1} \sum_{r=0}^{R-1} \sum_{s=0}^{S-1} I[c][h \cdot U + r - P][w \cdot U + s - P] \times W[m][c][r][s] \quad (4.1)$$

Trong đó:

- U : Bước trượt (Stride).
- P : Số lượng điểm ảnh đệm thêm vào mỗi cạnh (Padding).
- Điều kiện biên: Nếu chỉ số truy cập I nằm ngoài phạm vi $[0, H_{in} - 1]$ hoặc $[0, W_{in} - 1]$, giá trị trả về là 0 (Zero-padding).

4.1.1.2 Cấu trúc vòng lặp (Loop Nest)

Với giả thiết kích thước batch $N = 1$, chúng tôi mô hình hóa phép tính này dưới dạng 6 vòng lặp lồng nhau. Việc xử lý Padding thường được thực hiện bằng phần cứng chuyên dụng (Padding Logic) để tránh truy cập bộ nhớ ngoài vùng cho phép.

Algorithm 1: Standard Convolution (Standard Conv2D)

Input: $I[C][H_{in}][W_{in}]$, $W[M][C][R][S]$, Padding P , Stride U

Output: $O[M][H_{out}][W_{out}]$

for $m = 0$ **to** $M - 1$ **do**

for $c = 0$ **to** $C - 1$ **do**

for $h = 0$ **to** $H_{out} - 1$ **do**

for $w = 0$ **to** $W_{out} - 1$ **do**

for $r = 0$ **to** $R - 1$ **do**

for $s = 0$ **to** $S - 1$ **do**

$h_{in} = h \cdot U + r - P$

$w_{in} = w \cdot U + s - P$

if $h_{in} \geq 0 \wedge h_{in} < H_{in} \wedge w_{in} \geq 0 \wedge w_{in} < W_{in}$ **then**

$val = I[c][h_{in}][w_{in}]$

else

$val = 0$ /* Zero Padding */

end

$O[m][h][w] \leftarrow O[m][h][w] + val \times W[m][c][r][s]$

end

end

end

end

end

4.1.2 Depthwise Separable Convolution

Để giảm chi phí tính toán cho các thiết bị biên, các mô hình như MobileNet sử dụng kỹ thuật **Depthwise Separable Convolution**, tách phép chập chuẩn thành hai bước: **Depthwise (DW)** và **Pointwise (PW)**.

4.1.2.1 Depthwise Convolution (DW)

Phép tính này áp dụng bộ lọc riêng cho từng kênh đầu vào. Công thức tính toán bao gồm tham số Padding như sau:

$$O_{dw}[c][h][w] = \sum_{r=0}^{R-1} \sum_{s=0}^{S-1} I[c][h \cdot U + r - P][w \cdot U + s - P] \times W_{dw}[c][r][s] \quad (4.2)$$

Nhận xét: Việc xử lý Padding trong Depthwise cũng tương tự như Standard Conv, tuy nhiên do tính độc lập giữa các kênh, bộ điều khiển (Controller) cần đảm bảo logic Padding hoạt động chính xác cho từng luồng tính toán song song.

Algorithm 2: Depthwise Convolution (với Padding)

```

for  $c = 0$  to  $C - 1$  /* Parallel Channels */
  do
    for  $h = 0$  to  $H_{out} - 1$  do
      for  $w = 0$  to  $W_{out} - 1$  do
        for  $r = 0$  to  $R - 1$  do
          for  $s = 0$  to  $S - 1$  do
             $h_{in} = h \cdot U + r - P$ 
             $w_{in} = w \cdot U + s - P$ 
            if  $h_{in} \in [0, H_{in}) \wedge w_{in} \in [0, W_{in})$  then
               $O_{dw}[c][h][w] += I[c][h_{in}][w_{in}] \times W_{dw}[c][r][s]$ 
            end
          end
        end
      end
    end
  end

```

4.1.2.2 Pointwise Convolution (PW)

Pointwise Convolution là tích chập chuẩn với kernel 1×1 . Do kích thước kernel là 1×1 , tham số Padding thường được đặt bằng 0 ($P = 0$) và Stride $U = 1$ để giữ nguyên kích thước khung gian (H, W) , chỉ thay đổi số kênh từ C sang M .

$$O_{pw}[m][h][w] = \sum_{c=0}^{C-1} I[c][h][w] \times W_{pw}[m][c] \quad (4.3)$$

4.1.3 Yêu cầu đối với Kiến trúc thống nhất (Unified Architecture)

Từ các phân tích trên, nhóm nhận thấy rằng để bộ tăng tốc hoạt động hiệu quả cho cả hai trường hợp, kiến trúc phần cứng cần giải quyết được bài toán "kép":

- Cơ chế xử lý Padding động:** Phần cứng cần có khối logic để tự động chèn giá trị 0 khi chỉ số tính toán $(h \cdot U + r - P)$ bị âm hoặc vượt quá kích thước ảnh, thay vì phải tốn tài nguyên bộ nhớ để lưu trữ các viền số 0 thực tế.
- Tính linh hoạt của Mảng PE:** Các đơn vị tính toán cần có khả năng chuyển đổi chế độ giữa tích lũy theo khung gian (Standard/Pointwise) và tính toán độc lập theo kênh (Depthwise).

4.1.4 Kỹ thuật Gộp Batch Normalization (BN Folding)

Trong các mạng CNN hiện đại như MobileNet, lớp Batch Normalization (BN) thường được đặt ngay sau lớp Convolution để chuẩn hóa phân phối dữ liệu, giúp mạng hội tụ nhanh hơn. Công thức tính toán của lớp BN trong quá trình suy luận (Inference) cho một kênh m là:

$$y = \frac{x - \mu_m}{\sqrt{\sigma_m^2 + \epsilon}} \cdot \gamma_m + \beta_m \quad (4.4)$$

Trong đó:

- x : Giá trị đầu ra từ lớp Convolution (trước khi qua hàm kích hoạt).
- μ_m, σ_m : Giá trị trung bình (mean) và phương sai (variance) động (running statistics) của kênh m .
- γ_m, β_m : Tham số tỉ lệ (scale) và dịch chuyển (shift) được học trong quá trình huấn luyện.
- ϵ : Hằng số nhỏ để tránh chia cho 0.

Việc thực hiện trực tiếp công thức này trên phần cứng rất tốn kém do yêu cầu các phép toán phức tạp như căn bậc hai và phép chia. Tuy nhiên, do tại thời điểm suy luận, các tham số $\mu, \sigma, \gamma, \beta$ đều là hằng số, chúng tôi áp dụng kỹ thuật **BN Folding** để gộp toàn bộ phép tính BN vào trọng số (W) và bias (B) của lớp Convolution đi trước nó.

4.1.4.1 Công thức biến đổi trọng số

Giả sử đầu ra của lớp Convolution là $x = W_{orig} \cdot Input + B_{orig}$. Khi thay vào công thức BN, ta có:

$$y = \frac{(W_{orig} \cdot Input + B_{orig}) - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta \quad (4.5)$$

Phương trình trên có thể được viết lại dưới dạng một phép Convolution mới với trọng số W' và bias B' :

$$y = W' \cdot Input + B' \quad (4.6)$$

Trong đó, các tham số mới được tính toán trước (offline) bởi phần mềm

(driver) trước khi nạp xuống phần cứng:

$$W'[m][c][r][s] = W_{orig}[m][c][r][s] \cdot \frac{\gamma_m}{\sqrt{\sigma_m^2 + \epsilon}} \quad (4.7)$$

$$B'[m] = (B_{orig}[m] - \mu_m) \cdot \frac{\gamma_m}{\sqrt{\sigma_m^2 + \epsilon}} + \beta_m \quad (4.8)$$

Kết luận thiết kế: Nhờ kỹ thuật BN Folding, kiến trúc phần cứng của chúng tôi **không cần** thiết kế khối chức năng riêng cho Batch Normalization. Accelerator chỉ cần thực hiện phép tính Convolution bình thường với bộ trọng số (W', B') đã được tinh chỉnh, giúp tiết kiệm đáng kể tài nguyên DSP và giảm độ trễ xử lý.

4.2 Chiến lược phân mảnh và Dòng dữ liệu đè xuất

Để xử lý các Feature Map kích thước lớn trên tài nguyên phần cứng giới hạn, chúng tôi áp dụng chiến lược phân mảnh dữ liệu (Tiling) không chồng lấn. Mục này sẽ trình bày chi tiết cách thức chia nhỏ không gian dữ liệu và thuật toán điều phối các bước tính toán (Passes).

4.2.1 Định nghĩa khái niệm "Tile" (Mảnh dữ liệu)

Trong kiến trúc này, một "Tile" được định nghĩa là một phần nhỏ của khối dữ liệu gốc với kích thước được tối ưu hóa cho dung lượng bộ nhớ on-chip. Hệ thống xử lý ba loại Tile chính tương ứng với ba luồng dữ liệu. Đầu tiên, Input Tile (Mảnh đầu vào) là khối dữ liệu được cắt ra từ Input Feature Map gốc với kích thước $T_c \times T_h \times W$. Việc giới hạn số kênh nạp vào là T_c và chiều cao là T_h giúp dữ liệu vừa vặn với bộ nhớ đệm, trong khi chiều rộng W được giữ nguyên để tận dụng tính liên tục của dữ liệu trong bộ nhớ (Burst Read). Tiếp theo, Weight Tile (Mảnh trọng số) tập hợp các bộ lọc

cần thiết để xử lý cho Input Tile hiện tại, có kích thước $T_m \times T_c \times R \times S$. Cuối cùng, kết quả tính toán tương ứng tạo ra Output Tile (Mảnh đầu ra) với kích thước $T_m \times T_h \times W$.

4.2.2 Phương pháp Phân mảnh không gian dữ liệu (Space Partitioning)

Không gian tính toán của một lớp tích chập được định nghĩa bởi ba chiều chính: Chiều cao không gian (H), Chiều sâu kênh đầu vào (C), và Số lượng bộ lọc/kênh đầu ra (M). Chúng tôi chia nhỏ không gian này thành các khối (Block/Tile) độc lập dựa trên tham số phần cứng.

4.2.2.1 Công thức chia khối (Block Calculation)

Giả sử phần cứng có khả năng xử lý song song một khối dữ liệu kích thước $T_c \times T_h \times W$ và tạo ra T_m kênh đầu ra. Số lượng khối (Blocks) trên mỗi chiều được tính như sau:

- **Số khối theo chiều dọc (N_h):** Ảnh đầu vào chiều cao H được cắt thành N_h phần không chồng lấn.

$$N_h = \lceil \frac{H}{T_h} \rceil \quad (4.9)$$

Ví dụ: Với $H = 21, T_h = 11$, ta có $N_h = 2$ khối (Khối 0: hàng 0-10; Khối 1: hàng 11-20).

- **Số khối kênh đầu vào (N_c):** Tổng C kênh được chia thành N_c nhóm.

$$N_c = \lceil \frac{C}{T_c} \rceil \quad (4.10)$$

- **Số khối kênh đầu ra (N_m):** Tổng M bộ lọc được chia thành N_m nhóm.

$$N_m = \lceil \frac{M}{T_m} \rceil \quad (4.11)$$

Một đơn vị xử lý cơ sở, gọi là **1 Pass**, chính là quá trình hệ thống xử lý hoàn tất cho một cặp Input Tile và Weight Tile để cập nhật giá trị cho một Output Tile.

4.2.3 Mô hình hóa toán học và Tham số thiết kế

Để hiện thực hóa chiến lược phân mảnh, chúng tôi xây dựng mô hình toán học cho lớp tích chập thứ i . Các ký hiệu và tham số thiết kế được chuẩn hóa trong Bảng 4.1.

Bảng 4.1: Bảng tham số thiết kế và ánh xạ ký hiệu

Nhóm tham số	Ký hiệu	Mô tả
Filter	S, R	Độ rộng (w_f) và Độ dài (h_f) của bộ lọc
	N_f	Tổng số bộ lọc (Filters)
Feature Map	W, H, C	Kích thước Input Feature Map (Rộng, Dài, Số kênh)
	W_{out}, H_{out}, N_f	Kích thước Output Feature Map
Tiling (Pass)	T_h	Chiều cao IFM nạp trong 1 pass (h)
	T_c	Số kênh IFM tính toán song song (k)
	T_m	Số bộ lọc tính toán song song (m)
Output Tile	T_{ho}	Chiều cao OFM hợp lệ tạo ra trong 1 pass (h_o)
Khác	P, Str	Padding và Stride

4.2.3.1 Công thức tính kích thước không gian toàn cục

Dựa trên nguyên lý tích chập, một layer cần xử lý khôi dữ liệu vào $W \times H \times C$. Kích thước không gian của Output Feature Map (OFM) toàn cục được tính như sau:

$$W_{out} = \left\lfloor \frac{W - S + 2P}{Str} \right\rfloor + 1; \quad H_{out} = \left\lfloor \frac{H - R + 2P}{Str} \right\rfloor + 1 \quad (4.12)$$

4.2.3.2 Tính toán kích thước Output Tile theo cơ chế Ping-Pong

Trong chiến lược phân mảnh đề xuất, chiều cao đầu ra hợp lệ (T_{ho}) không cố định mà phụ thuộc vào việc hệ thống có tận dụng được dữ liệu dôi ra (Residual Data) từ pass trước đó hay không.

Giả sử bước trượt $Str = 1$ và Padding được xử lý tại biên ảnh gốc, số lượng hàng Output hợp lệ ghi xuống DRAM trong mỗi Pass được xác định bởi công thức:

$$T_{ho} = \begin{cases} T_h - R + 1 & \text{nếu là Tile đầu tiên } (Tile_idx = 0) \\ T_h & \text{nếu là các Tile tiếp theo } (Tile_idx > 0) \end{cases} \quad (4.13)$$

Giải thích:

- **Tile đầu tiên ($T_{ho} = T_h - R + 1$):** Do không có dữ liệu tích lũy từ phía trên, $R - 1$ hàng cuối cùng không đủ dữ liệu lân cận để hoàn thành phép tính tích chập, trở thành dữ liệu dôi ra (Residual) được lưu vào Buffer.
- **Các Tile tiếp theo ($T_{ho} = T_h$):** Hệ thống nạp T_h hàng input mới, kết hợp với $R - 1$ hàng residual từ tile trước đó. Điều này cho phép hoàn thiện $R - 1$ hàng biên cũ và tính trọng ven phần thân mới, tạo ra đủ T_h hàng output hợp lệ.

4.2.3.3 Phân tích số lượng Pass và Dữ liệu biên

Tổng số pass cần thiết được xác định bởi tích số các phân mảnh trên 3 chiều:

$$Total_Pass = \underbrace{\left\lceil \frac{C}{T_c} \right\rceil}_{N_c} \times \underbrace{\left\lceil \frac{H}{T_h} \right\rceil}_{N_h} \times \underbrace{\left\lceil \frac{N_f}{T_m} \right\rceil}_{N_m} \quad (4.14)$$

Do kích thước layer thường không chia hết cho kích thước Tile, lượng dữ liệu xử lý trong các pass cuối (Boundary Passes) sẽ khác biệt:

- **Phân mảnh theo kênh (C):** Để hoàn thiện một phần output, cần thực hiện $\lceil C/T_c \rceil$ pass tích lũy. Pass cuối cùng xử lý phần dư: $T_h \times W \times (C \pmod{T_c})$.
- **Phân mảnh theo chiều dọc (H):** Cần $\lceil H/T_h \rceil$ bước trượt dọc.
 - Pass cuối cùng (Last Tile) xử lý chiều cao input dư: $H_{rem} = H \pmod{T_h}$.
 - Số hàng Output hợp lệ của Pass cuối cũng tuân theo logic Ping-Pong: Nếu H_{rem} đủ lớn, nó sẽ tạo ra H_{rem} hàng output (do thừa hưởng biên trên).
- **Phân mảnh theo số bộ lọc (N_f):** Pass cuối xử lý số bộ lọc dư: $N_f \pmod{T_m}$.

4.2.3.4 Trường hợp Depthwise Convolution

Đối với Depthwise Convolution ($N_f = C$), quy trình được đơn giản hóa do không có vòng lặp tích lũy kênh ($N_c = 1$):

- Tổng số Pass: $\lceil H/T_h \rceil \times \lceil N_f/T_m \rceil$.
- Kích thước Output Tile T_{ho} vẫn tuân theo quy tắc (3) nêu trên.

4.2.4 Thuật toán Điều phối Pass (Pass Scheduling)

Trình tự thực hiện các Pass phụ thuộc vào loại tích chập (Standard hay Depthwise) để tối ưu hóa việc tái sử dụng dữ liệu biên (như đã phân tích ở mục Dữ liệu dội ra).

4.2.4.1 Thuật toán cho Standard Convolution

Trong Standard Convolution, một điểm ảnh đầu ra cần tổng hợp dữ liệu từ **tất cả** các khối kênh đầu vào (N_c). Do đó, ta cần vòng lặp tích lũy (Reduction Loop) chạy qua N_c trước khi chuyển sang khối chiều cao khác.

Algorithm 3: Lịch trình Pass cho Standard Convolution

Input: N_m (Output Blocks), N_h (Height Blocks), N_c (Input Blocks)

for $m = 0$ **to** $N_m - 1$ **do**

1. Load Weights for Output Block m (Weight Stationary)

for $h = 0$ **to** $N_h - 1$ **do**

for $c = 0$ **to** $N_c - 1$ **do**

Pass (m, h, c):

- Nạp Input Block (c, h) kích thước $T_c \times T_h$

- Tính toán với Weight Block (m, c)

- Cộng dồn kết quả vào Buffer hiện tại (A hoặc B)

end

2. Xử lý biên & Ghi Output:

- Sau khi cộng đủ N_c passes: Output Block (m, h) đã hoàn tất (Valid).

- Ghi phần Valid xuống DRAM.

- Hoán đổi Ping-Pong Buffer (để dùng phần Dôi ra cho $h + 1$).

end

end

4.2.4.2 Thuật toán cho Depthwise Convolution

Trong Depthwise Convolution, kênh Input thứ i chỉ tính toán với kênh Filter thứ i . Do đó $N_c = N_m$ (số nhóm kênh Input bằng số nhóm kênh Output) và không có sự cộng dồn chéo giữa các nhóm. Vòng lặp tích lũy biến mất.

Algorithm 4: Lịch trình Pass cho Depthwise Convolution

Input: N_m (Channel Groups), N_h (Height Blocks)

for $g = 0$ **to** $N_m - 1$ **do**

1. Load Weights for Group g

for $h = 0$ **to** $N_h - 1$ **do**

Pass (g, h) :

- Nạp Input Block (g, h) kích thước $T_c \times T_h$
- Tính toán với Weight Block g
- Tạo ra ngay kết quả Output Block (g, h) (không cần cộng dồn)

2. Xử lý biên & Ghi Output:

- Ghi ngay phần Valid xuống DRAM.
- Hoán đổi Ping-Pong Buffer (lưu phần Dôi ra cho $h + 1$).

end

end

Pass 0	Pass 1	Pass 2	Pass 3
row 0-10, channel 0-10, filter 0	row 0-10, channel 11-20, filter 0	row 11-20, channel 0-10, filter 0	row 11-20, channel 11-20, filter 0
Pass 4	Pass 5	Pass 6	Pass 7
row 0-10, channel 0-10, filter 1	row 0-10, channel 11-20, filter 1	row 11-20, channel 0-10, filter 1	row 11-20, channel 11-20, filter 1

(a) Standard Convolution ($H = 21, M = 2$)

Pass 0	Pass 1	Pass 2	Pass 3
row 0-10, channel 0-10, filter 0-10	row 11-20, channel 0-10, filter 0-10	row 0-10, channel 11-20, filter 11-20	row 11-20, channel 11-20, filter 11-20

(b) Depthwise Convolution ($H = 21, M = 21$)

Hình 4.1: Minh họa chiến lược phân chia Pass cho hai loại tích chập với $T_h = 11$. (a) Standard Convolution chia thành 2 phần theo chiều dọc và tích lũy theo chiều sâu. (b) Depthwise Convolution xử lý độc lập từng nhóm kênh và chia 2 phần theo chiều dọc.

4.2.5 Phân tích vấn đề tại biên và Dữ liệu dội ra

Mặc dù chiến lược phân mảnh dữ liệu được áp dụng trên cả chiều kênh và chiều không gian, tác động của chúng lên luồng dữ liệu là khác nhau.

- Việc chia nhỏ chiều kênh (C, M) dẫn đến bài toán tích lũy tổng riêng

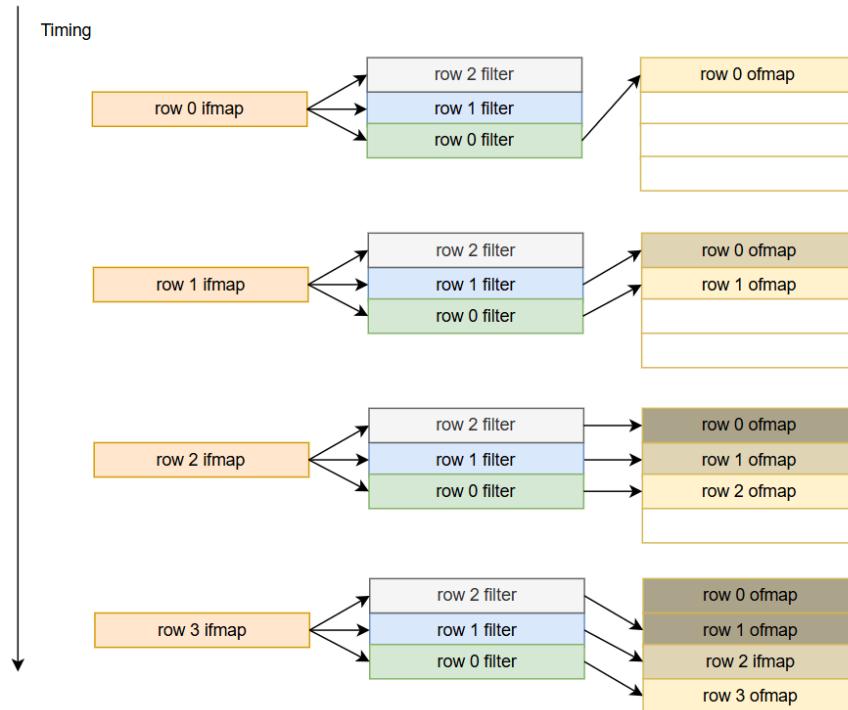
(Partial Sum Accumulation).

- Việc chia nhỏ chiều không gian (H) dẫn đến bài toán thiếu dữ liệu vùng lân cận cho cửa sổ trượt (Sliding Window Boundary).

Mục này tập trung phân tích vấn đề tại biên không gian, nguyên nhân chính dẫn đến sự cần thiết của cơ chế Ping-Pong Buffer đặc thù.

4.2.5.1 Cơ sở hình thành Dữ liệu dôi ra

Khi bộ lọc trượt theo chiều dọc, tại các hàng cuối cùng của một Tile không gian (gọi là Tile H_k), bộ lọc cần dữ liệu của các hàng tiếp theo (thuộc Tile H_{k+1}) để hoàn thành phép tính.



Hình 4.2: Sơ đồ minh họa quá trình tính toán tích chập và sự hình thành dữ liệu dôi ra (Residual Data) trong một pass với tile đầu vào $T_h = 4$ và bộ lọc kích thước 3×3 trong trường hợp số kênh của ifmap feature là 1.

Xét ví dụ cụ thể với Tile đầu vào có kích thước chiều cao $T_h = 4$ (các hàng 0, 1, 2, 3) và bộ lọc kích thước 3×3 . Khi thực hiện tích chập:

- Hàng 0, 1:** Có đầy đủ dữ liệu lân cận (trong phạm vi Tile hiện tại)

→ Tạo ra kết quả hoàn chỉnh (*Valid Output*).

- **Hàng 2:** Cần dữ liệu hàng [2, 3, 4]. Thiếu hàng 4 (thuộc Tile H_{k+1})
→ Kết quả chưa hoàn thiện.
- **Hàng 3:** Cần dữ liệu hàng [3, 4, 5]. Thiếu hàng 4, 5 (thuộc Tile H_{k+1}) → Kết quả chưa hoàn thiện.

Các kết quả tại hàng 2 và 3 được gọi là **Dữ liệu dôi ra (Residual Data)**. Số lượng hàng dôi ra luôn là $R - 1$. Để đảm bảo tính đúng đắn mà không cần nạp lại phần dữ liệu Input [2, 3] khi xử lý Tile H_{k+1} , hệ thống cần lưu trữ các giá trị dôi ra này và cộng dồn chúng với kết quả tính toán từ Tile tiếp theo.

4.2.6 Cơ chế Ping-Pong Buffer và Logic xử lý hàng hợp lệ

Để xử lý dữ liệu dôi ra (Residual Data) tại biên dưới của mỗi tile mà không cần nạp lại Input, hệ thống sử dụng hai bộ đệm đầu ra $Buffer_A$ và $Buffer_B$ hoạt động luân phiên.

Điểm quan trọng trong chiến lược này là số lượng hàng đầu ra hợp lệ (Valid Rows) sẽ khác nhau giữa Tile đầu tiên và các Tile tiếp theo:

- **Tile đầu tiên ($h = 0$):** Do không có dữ liệu tích lũy từ phía trên, bộ lọc trượt qua T_h hàng đầu vào chỉ tạo ra được $T_h - R + 1$ hàng đầu ra hoàn chỉnh. $R - 1$ hàng cuối cùng là dữ liệu dôi ra.
- **Các Tile tiếp theo ($h > 0$):** Nhờ tận dụng $R - 1$ hàng dôi ra từ bước trước (đã lưu trong Buffer), hệ thống sẽ hoàn thiện được các hàng này. Tổng số hàng hoàn chỉnh được ghi xuống DRAM trong bước này là đủ T_h hàng.

4.2.7 Thuật toán Điều phối và Xoay vòng bộ nhớ

Thuật toán 5, 6 và 7 mô tả chi tiết quy trình quản lý bộ nhớ và luồng dữ liệu, minh họa rõ sự khác biệt khi xử lý Tile đầu tiên và các Tile sau.

Algorithm 5: Lịch trình Pass cho Standard Conv (Phần 1: Tích lũy)

Input: N_m, N_h, N_c

for $m = 0$ **to** $N_m - 1$ **do**

1. Load Weights...

for $h = 0$ **to** $N_h - 1$ **do**

for $c = 0$ **to** $N_c - 1$ **do**

Pass (m, h, c) : ...

... (Code phần tích lũy) ...

end

(Xem tiếp xử lý biên ở Thuật toán 6)

end

end

Algorithm 6: Lịch trình Pass cho Standard Conv (Phần 2: Xử lý biên)

... Tiếp tục từ vòng lặp h của Thuật toán 5

foreach Tile h đã hoàn tất tích lũy **do**

2. Xử lý biên & Ghi Output:

- Kiểm tra điều kiện biên...

- Ghi phần Valid xuống DRAM.

- Hoán đổi Ping-Pong Buffer.

end

Algorithm 7: Lịch trình Pass cho Depthwise Convolution

Input: N_m (Số nhóm kênh), N_h (Số khối đọc)

Output: DRAM (Valid Output Feature Map)

Initialize pointers: $Buf_{curr} \leftarrow A$, $Buf_{next} \leftarrow B$

for $m = 0$ **to** $N_m - 1$ **do**

1. Load Weights for Group m

for $h = 0$ **to** $N_h - 1$ **do**

Pass (m, h):

- Nạp Input Tile (m, h) kích thước $T_c \times T_h$

- Tính toán Depthwise (1-to-1 mapping)

2. Xử lý biên & Quản lý Ping-Pong:

if $h == 0$ **then**

// Trường hợp Tile đầu tiên

- Lưu $T_h - R + 1$ hàng Output hợp lệ vào Buf_{curr}

- Lưu phần dư ($R - 1$ hàng cuối) vào Buf_{next}

- **Drain:** Ghi Buf_{curr} xuống DRAM

else

// Các Tile tiếp theo (Tận dụng Residual)

- Hoàn thiện $R - 1$ hàng biên (từ Buf_{curr} cũ)

- Tạo thêm các hàng thân mới (đủ T_h hàng)

- Lưu phần dư mới vào Buf_{next}

- **Drain:** Ghi toàn bộ T_h hàng hợp lệ xuống DRAM

end

3. Chuẩn bị cho tile tiếp theo:

- Clear Buf_{curr}

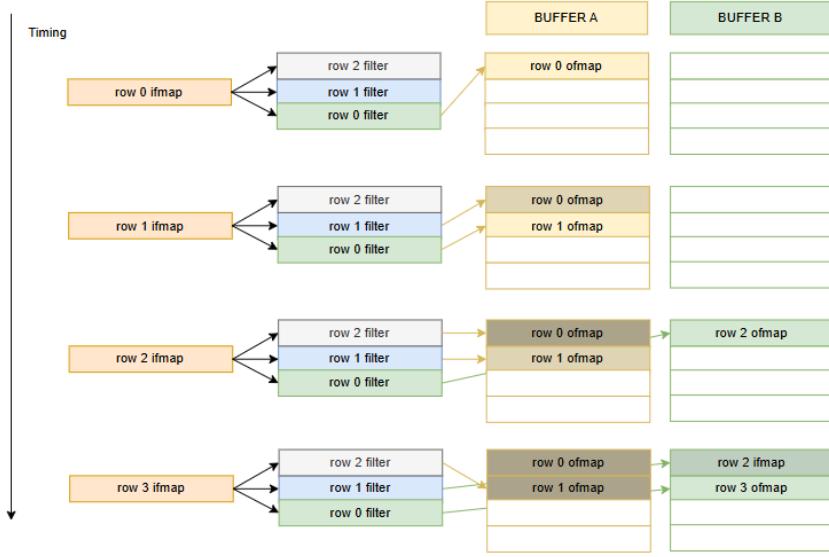
- **Swap pointers:** $Buf_{curr} \leftrightarrow Buf_{next}$

end

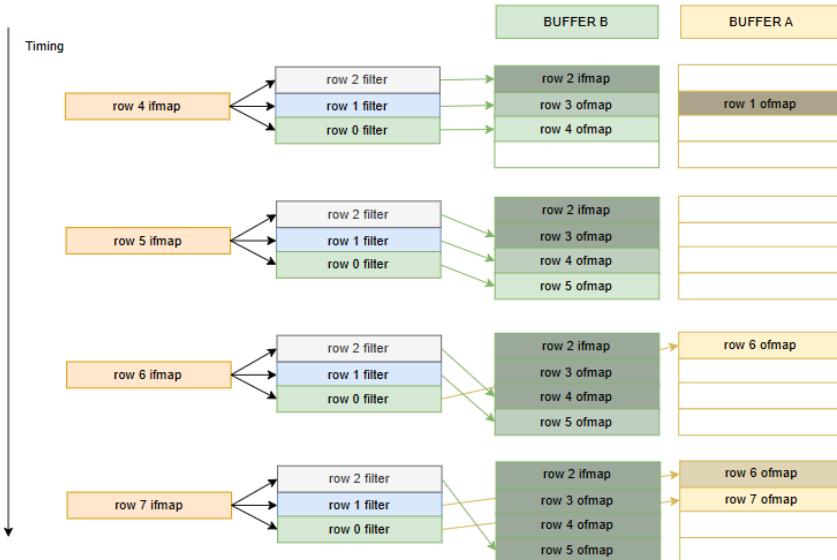
end

Giải thích cơ chế:

1. Tại vòng lặp h , Buf_{curr} đóng vai trò tích lũy kết quả chính, còn Buf_{next} đóng vai trò hứng các giá trị dội ra (Residual) cho tương lai.
2. Khi $h = 0$: Chúng ta ghi các hàng dội ra vào đầu Buf_{next} .
3. Khi chuyển sang $h = 1$: Ta thực hiện Swap. Lúc này Buf_{curr} (vốn là Buf_{next} cũ) đã có sẵn dữ liệu dội ra ở các hàng đầu. Việc tính toán tiếp tục cộng dồn vào đó, biến chúng thành kết quả hoàn chỉnh (Valid).
4. Quá trình ghi xuống DRAM (Drain) ở $h > 0$ sẽ ghi toàn bộ T_h hàng, bao gồm cả những hàng vừa được hoàn thiện từ dữ liệu dội ra.



(a) Giai đoạn 1 (Xử lý Tile H_k): Buffer A tích lũy kết quả Valid, Buffer B lưu trữ dữ liệu Dôi ra (Residual).



(b) Giai đoạn 2 (Xử lý Tile H_{k+1}): Buffer B hoàn thiện kết quả biên (từ Residual cũ), Buffer A lưu trữ dữ liệu Dôi ra mới.

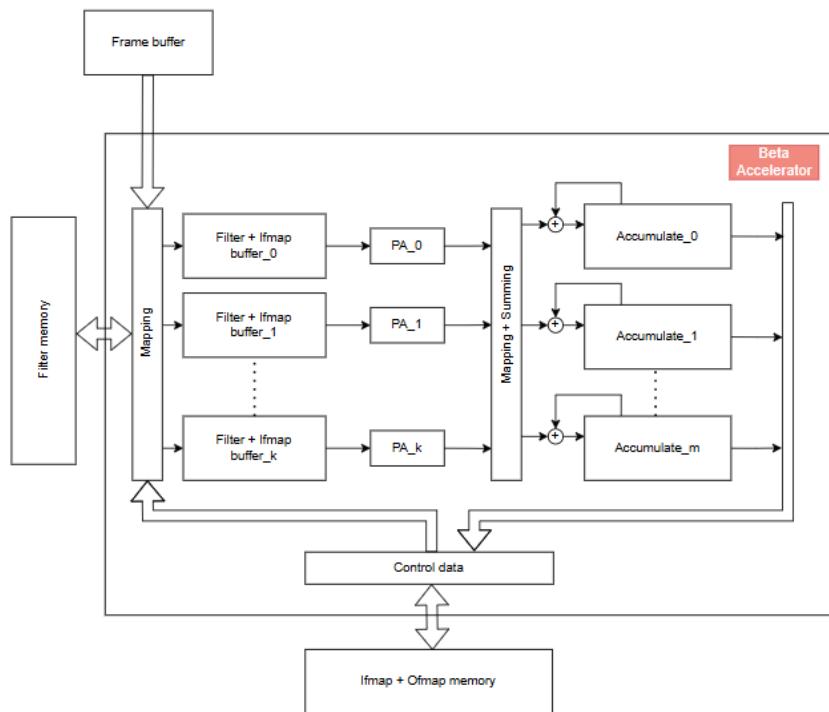
Hình 4.3: Sơ đồ luồng dữ liệu minh họa cơ chế Ping-Pong Buffer dùng để quản lý vùng dữ liệu dôi ra (Residual Data). Hệ thống luân phiên vai trò của Buffer A và B để đảm bảo tính liên tục của phép tính biên mà không cần nạp lại dữ liệu đầu vào.

4.3 Thiết kế kiến trúc vi mô (Micro-architecture)

Dựa trên chiến lược dòng dữ liệu đã phân tích, nhóm đề xuất kiến trúc phần cứng chuyên dụng mang tên **Beta Accelerator**. Kiến trúc này được thiết kế để tối ưu hóa khả năng tính toán song song ở mức bộ lọc (Filter parallelism) và mức kênh (Channel parallelism), đồng thời hỗ trợ cơ chế quản lý bộ nhớ Ping-Pong để che giấu độ trễ truy cập.

4.3.1 Sơ đồ khái quát hệ thống

Sơ đồ tổng thể của Beta Accelerator được trình bày trong Hình 4.4. Hệ thống bao gồm các khái niệm chính sau:



Hình 4.4: Sơ đồ khái quát kiến trúc Beta Accelerator

- **Khái Controller (Controller):** Đóng vai trò bộ điều khiển trung tâm và quản lý giao tiếp với bộ nhớ ngoài. Do hệ thống chỉ sử dụng một bus dữ liệu chung (shared data bus) cho cả luồng vào và luồng ra, khái này chịu trách nhiệm điều phối (arbitration) tài nguyên bus,

quyết định thời điểm thực hiện nạp dữ liệu đầu vào (Load IFM) hoặc ghi kết quả đầu ra (Store OFM) để tránh xung đột dữ liệu.

- **Khối Mapping (Dispatcher):** Chịu trách nhiệm phân phối dữ liệu IFM và trọng số (Weights) từ các bus dữ liệu chính tới các bộ nhớ đệm cục bộ của từng đơn vị tính toán, đảm bảo băng thông và tính đồng bộ.
- **Hệ thống Bộ đệm (Filter + Ifmap Buffer):** Được tổ chức theo cơ chế **Ping-Pong Buffer** (Double Buffering) để cho phép nạp dữ liệu cho Pass $k+1$ trong khi Pass k đang được tính toán.
 - Mỗi khối buffer lưu trữ một tile IFM kích thước $T_h \times W$ của 1 kênh.
 - Đồng thời lưu trữ bộ trọng số kích thước $S \times R \times T_m$ (trong đó S, R là kích thước filter, T_m là số filter tính song song).
- **Mảng xử lý (Process Array - PA):** Là trái tim tính toán của hệ thống, bao gồm T_c khối PA hoạt động song song. Mỗi khối PA phụ trách xử lý 1 kênh đầu vào (Input Channel) và T_m bộ lọc tương ứng.
- **Khối Tổng hợp (Reduction Unit - Mapping + Summary):** Thực hiện chức năng cộng dồn (Reduction) kết quả từ T_c khối PA. Do tích chập là phép tổng trọng số qua các kênh, khối này sẽ cộng giá trị Partial Sum từ các kênh IFM khác nhau để tạo ra T_m giá trị OFM bán hoàn chỉnh.
- **Khối Tích lũy (Accumulator):** Sử dụng bộ nhớ Ping-Pong để lưu trữ và cộng dồn kết quả qua các Pass (theo chiều sâu kênh C). Khi một điểm ảnh OFM đã được tích lũy đủ số kênh cần thiết, nó sẽ được gửi đi thông qua bus dữ liệu chung và vị trí nhớ đó sẽ được reset về 0 để chuẩn bị cho lượt tính mới.

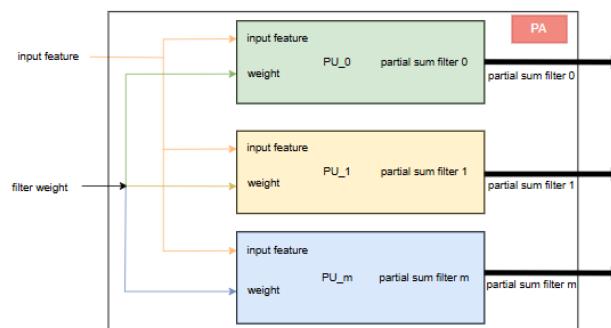
4.3.2 Tổ chức Mảng tính toán (Processing Hierarchy)

Kiến trúc tính toán được tổ chức theo mô hình phân cấp gồm 3 tầng: Process Array (PA), Process Unit (PU) và Process Element (PE).

4.3.2.1 Mảng xử lý (Process Array - PA)

Khối PA (Hình 4.5) được thiết kế để khai thác tính song song mức kênh đầu ra (Output Channel Parallelism).

- Mỗi PA chịu trách nhiệm tính toán cho 1 kênh đầu vào (Input Channel) duy nhất nhưng tạo ra kết quả cho T_m bộ lọc (Filters) khác nhau.
- Luồng dữ liệu:** Trọng số đầu vào (Input Filter Weights) được rẽ nhánh (demultiplex) tới các PU cụ thể (ví dụ: PU_0 nhận trọng số của Filter 0). Ngược lại, dữ liệu IFM được quảng bá (broadcast) dùng chung cho tất cả các PU trong cùng một PA, giúp tiết kiệm băng thông đọc IFM.

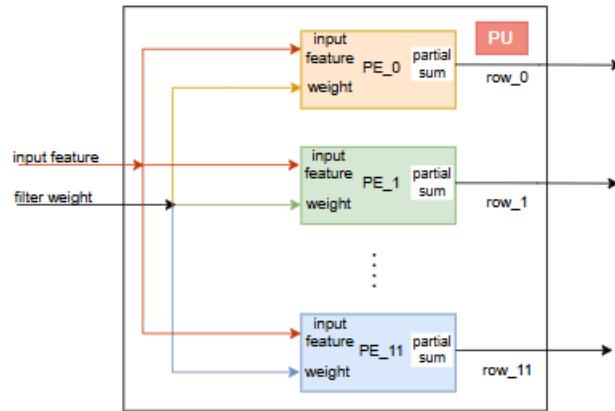


Hình 4.5: Kiến trúc bên trong khối Process Array (PA)

4.3.2.2 Đơn vị xử lý (Process Unit - PU)

Mỗi PU (Hình 4.6) bao gồm 11 phần tử xử lý (PE) hoạt động song song, tương ứng với khả năng hỗ trợ kích thước bộ lọc tối đa là 11×11 (chiều cao $R = 11$).

- Mỗi PE trong PU chịu trách nhiệm tính toán tích chập cho **1 hàng** của bộ lọc (Filter Row).
- Các PE hoạt động đồng bộ. Sau mỗi khoảng thời gian ΔT chu kỳ, PU sẽ tạo ra một cột kết quả gồm R giá trị tương ứng với R hàng của bộ lọc.

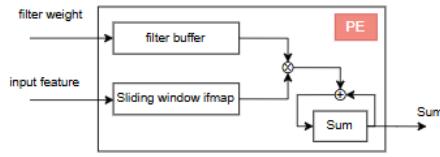


Hình 4.6: Kiến trúc khối Process Unit (PU) với các PE hoạt động song song

4.3.2.3 Phần tử xử lý (Process Element - PE)

PE là đơn vị tính toán cơ sở nhỏ nhất (Hình 4.7), thực hiện phép tính nhân-cộng (MAC).

- **Filter Buffer:** Lưu trữ S giá trị trọng số của một hàng filter ($1 \times S$). Buffer này hoạt động theo chế độ Weight Stationary, giữ giá trị không đổi trong suốt quá trình thực hiện 1 Pass.
- **Sliding Window Register:** Chứa S giá trị IFM ($1 \times S$). Đây là thanh ghi dịch, sau mỗi ΔT chu kỳ, dữ liệu sẽ dịch di 1 vị trí (stride = 1) để thực hiện phép trượt cửa sổ.
- Vì mỗi PE chứa 1 bộ nhân và 1 bộ cộng, để tính tích chập 1 hàng kích thước S , hệ thống cần $\Delta T = S$ chu kỳ.



Hình 4.7: Cấu trúc bên trong một Process Element (PE)

4.3.3 Đánh giá thời gian thực thi (Performance Estimation)

Thời gian thực thi của hệ thống phụ thuộc vào loại lớp tích chập (Standard hay Depthwise) do sự khác biệt trong chiến lược luồng dữ liệu.

4.3.3.1 Thời gian xử lý một Pass cơ sở (T_{pass})

Dựa trên kiến trúc Pipeline của các Process Element (PE), thời gian để hoàn thành tính toán cho một tile có chiều cao T_h và độ rộng OFM W_{out} được xác định bởi:

$$T_{pass} = [(W_{out} - 1) \times (S + U - 1) + S] \times T_h \quad (4.15)$$

Trong đó:

- S : Kích thước bộ lọc (Filter width).
- U : Bước trượt (Stride).
- W_{out} : Chiều rộng của OFM.
- $(S + U - 1)$: Số chu kỳ trung bình để tính một điểm ảnh tiếp theo nhờ tối ưu hóa Pipeline (khi $U = 1$, thời gian này là S).

4.3.3.2 Tổng thời gian thực thi (T_{total})

Trường hợp 1: Standard Convolution

Với tích chập tiêu chuẩn, mỗi điểm ảnh đầu ra là tổng hợp của tất cả C kênh đầu vào. Hệ thống phải thực hiện vòng lặp tích lũy qua các khối kênh T_c .

$$T_{total_std} = \underbrace{\left[\frac{N_f}{T_m} \right]}_{\text{Output Blocks}} \times \underbrace{\left[\frac{C}{T_c} \right]}_{\text{Input Blocks}} \times \underbrace{\left[\frac{H}{T_h} \right]}_{\text{Height Blocks}} \times T_{pass} \quad (4.16)$$

Trường hợp 2: Depthwise Convolution

Với tích chập chiều sâu, các kênh hoạt động độc lập ($N_f = C$). Hệ thống không cần thực hiện vòng lặp tích lũy kênh đầu vào ($\lceil C/T_c \rceil$ bị loại bỏ). Các nhóm kênh được xử lý song song dựa trên khả năng của phần cứng (T_m).

$$T_{total_dw} = \underbrace{\left[\frac{N_f}{T_m} \right]}_{\text{Channel Groups}} \times \underbrace{\left[\frac{H}{T_h} \right]}_{\text{Height Blocks}} \times T_{pass} \quad (4.17)$$

Nhận xét: So với Standard Convolution, Depthwise Convolution giảm được hệ số $\lceil C/T_c \rceil$ lần số lượng tính toán, giúp tăng tốc độ xử lý đáng kể đối với các mạng nhẹ (Lightweight CNNs) như MobileNet.

4.3.4 Chiến lược Che giấu độ trễ và Mô hình hiệu năng toàn hệ thống

Để tối ưu hóa hiệu năng, Beta Accelerator áp dụng kỹ thuật **Che giấu độ trễ (Latency Hiding)** thông qua cơ chế Ping-Pong Buffer. Mục tiêu là thực hiện song song quá trình tính toán (Computation) và quá trình truyền tải dữ liệu (Data Transfer) để ẩn đi thời gian giao tiếp với bộ nhớ ngoài.

4.3.4.1 Cơ chế hoạt động

Trước khi bắt đầu tính toán Pass đầu tiên, hệ thống cần nạp đầy đủ dữ liệu (IFM, Weights) vào buffer. Sau giai đoạn khởi tạo này, quy trình hoạt động theo nguyên lý "gối đầu":

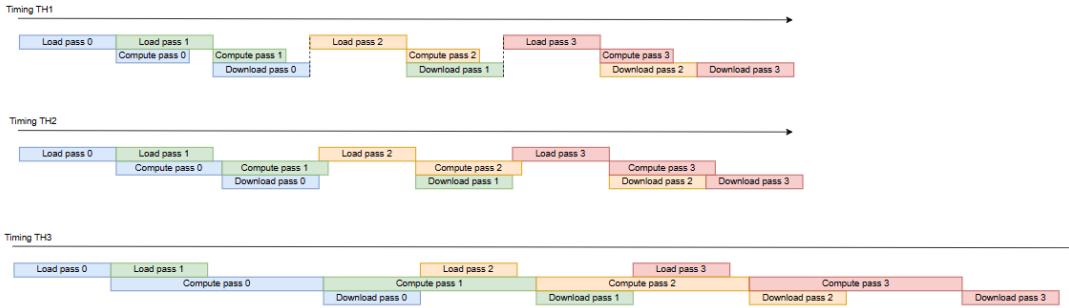
- Trong khi lõi tính toán đang xử lý Pass i , bộ điều khiển DMA đồng thời nạp dữ liệu cho Pass $i + 1$ vào nửa còn lại của Buffer.
- Đồng thời, kết quả của Pass $i - 1$ (nếu đã hoàn tất) được ghi trả về bộ nhớ ngoài.

Do hệ thống sử dụng bus dữ liệu dùng chung (Shared Data Bus) cho cả luồng nạp (Load) và ghi (Store), băng thông bộ nhớ phải được chia sẻ thời gian. Bộ điều khiển sẽ ưu tiên nạp Pass tiếp theo, sau đó mới đến ghi Pass trước đó (hoặc xen kẽ tùy theo chính sách trọng tài).

4.3.4.2 Các kịch bản hiệu năng (Performance Scenarios)

Gọi T_{load} là thời gian nạp 1 Input Pass, T_{store} là thời gian ghi 1 Output Pass, và T_{comp} là thời gian tính toán 1 Pass (T_{pass} đã tính ở mục 4.3.3). Ta định nghĩa tham số b là **số chu kỳ đồng hồ cần thiết để truyền 1 giá trị dữ liệu** (Cycles per Data Transfer).

Mô hình thời gian hoàn thành 1 layer được phân tích dựa trên sự chênh lệch giữa năng lực tính toán và băng thông bộ nhớ, được minh họa trong Hình 4.8.



Hình 4.8: Biểu đồ thời gian thực thi trong 3 trường hợp: (Trên cùng) Memory Bound 1, (Giữa) Memory Bound 2, (Dưới cùng) Compute Bound.

Trường hợp 1: Memory Bound 1 (Nghẽn băng thông nghiêm trọng)

Xảy ra khi thời gian nạp dữ liệu lớn hơn thời gian tính toán ($T_{load} \geq T_{comp}$). Lỗi tính toán phải chờ dữ liệu nạp xong mới có thể chạy. Tổng thời gian hoàn thành layer được quyết định chủ yếu bởi tổng lượng dữ liệu cần truyền tải (Input + Output).

- **Đối với Standard Convolution:** Do phải nạp lại Input Feature Map cho mỗi nhóm Filter khác nhau (nếu không đủ bộ nhớ on-chip), tổng thời gian là:

$$T_{total} \approx \left[\left(H \times W \times C \times \left\lceil \frac{N_f}{T_m} \right\rceil \right) + (H_{out} \times W_{out} \times N_f) \right] \times b \quad (4.18)$$

- **Đối với Depthwise Convolution:** Mỗi kênh Input chỉ tương tác với 1 kênh Filter tương ứng ($N_f = C$), nên Input Feature Map chỉ cần nạp 1 lần duy nhất:

$$T_{total} \approx [(H \times W \times C) + (H_{out} \times W_{out} \times C)] \times b \quad (4.19)$$

Trường hợp 2: Memory Bound 2 (Nghẽn băng thông trung bình)

Xảy ra khi thời gian tính toán nhanh hơn tổng thời gian nạp và ghi, nhưng

chậm hơn thời gian nạp ($T_{load} < T_{comp} < T_{load} + T_{store}$). Lúc này, thời gian thực thi bao gồm thời gian nạp, ghi và một phần chênh lệch thời gian tính toán.

$$T_{total} \approx T_{total_IO} + (T_{comp} - T_{load}) \quad (4.20)$$

Trong đó T_{total_IO} được tính theo công thức tại Trường hợp 1 tùy thuộc loại Convolution.

Trường hợp 3: Compute Bound (Nghẽn tính toán)

Xảy ra khi thời gian tính toán lớn hơn tổng thời gian nạp và ghi ($T_{comp} > T_{load} + T_{store}$). Lúc này, toàn bộ thời gian truyền tải dữ liệu (trừ pass đầu và cuối) được che giấu hoàn toàn bên dưới thời gian tính toán.

Công thức tổng quát:

$$T_{total} = T_{load_first_pass} + \sum_{all_passes} T_{comp} + T_{store_residual} \quad (4.21)$$

- **Đối với Standard Convolution:**

$$T_{total} \approx (T_h W T_c b) + \left(\left\lceil \frac{H}{T_h} \right\rceil \left\lceil \frac{C}{T_c} \right\rceil \left\lceil \frac{N_f}{T_m} \right\rceil \times T_{comp} \right) + T_{res_std} \quad (4.22)$$

Với T_{res_std} là thời gian ghi phần dư cuối cùng phụ thuộc số filter dư ($N_f \% T_m$):

$$T_{res_std} = \left(\left\lfloor \frac{H \% T_h}{Str} \right\rfloor + 1 \right) \times W_{out} \times (N_f \% T_m) \times b \quad (4.23)$$

- **Đối với Depthwise Convolution:** Do không có vòng lặp tích lũy kênh đầu vào (C/T_c), tổng số pass giảm đi đáng kể:

$$T_{total} \approx (T_h W T_c b) + \left(\left\lceil \frac{H}{T_h} \right\rceil \left\lceil \frac{C}{T_m} \right\rceil \times T_{comp} \right) + T_{res_dw} \quad (4.24)$$

Với T_{res_dw} là thời gian ghi phần dư cuối cùng phụ thuộc số kênh dư

$(C\%T_m)$:

$$T_{res_dw} = \left(\left\lfloor \frac{H\%T_h}{Str} \right\rfloor + 1 \right) \times W_{out} \times (C\%T_m) \times b \quad (4.25)$$

4.3.4.3 Tổng thời gian toàn mạng (Model Latency)

Thời gian thực thi của toàn bộ mô hình (Model) bao gồm N lớp tích chập là tổng thời gian của từng lớp, do sự phụ thuộc dữ liệu tuần tự giữa các lớp (Layer $i + 1$ cần OFM của Layer i làm IFM):

$$T_{model} = \sum_{i=1}^N T_{total}^{(i)} \quad (4.26)$$

Mục tiêu của bài toán tối ưu hóa thiết kế là tìm bộ tham số cấu hình (T_h, T_m, T_c) cho từng layer sao cho $T_{total}^{(i)}$ là nhỏ nhất, cân bằng giữa tài nguyên tính toán và băng thông bộ nhớ.

4.4 Tối ưu hóa Tham số và Cơ chế Sinh mã cấu hình

Sau khi xây dựng mô hình ước lượng hiệu năng, bước tiếp theo là xác định bộ tham số cấu hình tối ưu cho từng lớp mạng và chuyển đổi chúng thành chuỗi lệnh điều khiển (Instruction Stream) mà Controller có thể thực thi.

4.4.1 Bài toán Tối ưu hóa Không gian thiết kế

Với mỗi lớp tích chập thứ i , mục tiêu là tìm ra bộ ba tham số phân mảnh $\mathbf{S}_i = \{T_{h,i}, T_{c,i}, T_{m,i}\}$ sao cho thời gian thực thi tổng thể (T_{total}) là nhỏ nhất.

4.4.1.1 Các ràng buộc phần cứng (Hardware Constraints)

Bộ tham số được chọn bắt buộc phải thỏa mãn các giới hạn vật lý của FPGA. Các ràng buộc chính bao gồm:

1. **Dung lượng bộ nhớ on-chip (BRAM):** Tổng kích thước của các Tile (bao gồm cả cơ chế Ping-Pong nhân hệ số 2) không được vượt quá dung lượng BRAM dành riêng cho từng loại dữ liệu.

$$2 \times (T_c \times T_h \times W) \times DW \leq BRAM_{IFM_MAX} \quad (4.27)$$

$$2 \times (T_m \times T_c \times R \times S) \times DW \leq BRAM_{WGT_MAX} \quad (4.28)$$

2. **Tài nguyên tính toán:** Số lượng bộ lọc tính toán song song (T_m) không được vượt quá số lượng mảng xử lý (PA) vật lý có trên chip.

$$1 \leq T_m \leq N_{PA_MAX} \quad (4.29)$$

3. **Tính hợp lệ:** Kích thước Tile không được lớn hơn kích thước gốc của Feature Map.

$$1 \leq T_h \leq H; \quad 1 \leq T_c \leq C \quad (4.30)$$

4.4.2 Chiến lược Tìm kiếm và Sinh mã (Search & Generate)

Do không gian tìm kiếm tham số cho một lớp là hữu hạn, chúng tôi phát triển một công cụ phần mềm (Software Tool) chạy trên máy tính chủ (Host PC) để thực hiện quy trình sau:

- **Bước 1 - Quét tham số (Exhaustive Search):** Thuật toán duyệt qua toàn bộ các tổ hợp (T_h, T_c, T_m) khả dĩ.

- **Bước 2 - Kiểm tra ràng buộc:** Loại bỏ ngay các tổ hợp vi phạm ràng buộc BRAM hoặc tài nguyên tính toán nêu trên.
- **Bước 3 - Đánh giá hiệu năng:** Với các tổ hợp hợp lệ, phần mềm áp dụng các công thức mô hình hiệu năng (Mục 4.3.4) để tính T_{total} dự kiến. Cấu hình cho T_{total} nhỏ nhất sẽ được chọn.
- **Bước 4 - Đóng gói (Packing):** Các tham số tối ưu được đóng gói thành một chuỗi bit nhị phân (Binary Descriptor) để gửi xuống phần cứng.

4.4.3 Cấu trúc Lệnh cấu hình (Layer Descriptor)

Để Controller phần cứng hiểu và vận hành theo tham số đã tìm được, chúng tôi định nghĩa một cấu trúc dữ liệu điều khiển (Descriptor). Mỗi lớp mạng tương ứng với một Descriptor được lưu trong bộ nhớ lệnh.

Bảng 4.2: Cấu trúc dữ liệu cấu hình cho một Layer

Offset	Trường thông tin	Mô tả chức năng
0x00	Layer Info	Chứa H, W, C, N_f (Kích thước gốc)
0x04	Kernel Info	Chứa R, S, P, Str (Bộ lọc, Padding,Stride)
0x08	Tiling Config	Chứa T_h, T_c, T_m (Tham số tối ưu)
0x0C	IFM Base Addr	Địa chỉ bắt đầu của Input Feature Map
0x10	WGT Base Addr	Địa chỉ bắt đầu của Weights
0x14	OFM Base Addr	Địa chỉ bắt đầu của Output Feature Map
0x18	Control Flags	Loại lớp (Std/Depthwise), Activation (ReLU)...

Khối **Control Data** (Mục 4.3) sẽ đọc Descriptor này, giải mã và cài đặt các giá trị T_h, T_c, T_m vào các thanh ghi đếm của máy trạng thái (FSM). Nhờ đó, phần cứng có thể linh hoạt xử lý nhiều kích thước mạng khác nhau mà không cần thiết kế lại mạch RTL.

Chapter 5

Hiện thực nền tảng SoC

Chương này trình bày quá trình xây dựng hệ thống SoC cơ sở trên FPGA, bao gồm cấu hình vi xử lý, hệ thống bus và tích hợp các ngoại vi.

5.1 Môi trường và Công cụ hiện thực

5.2 Cấu hình hệ thống xử lý (Processing System)

5.3 Thiết kế hệ thống kết nối (Interconnect Subsystem)

5.4 Tích hợp và Kiểm thử nền tảng cơ sở

Chapter 6

Đánh giá hiệu năng lý thuyết

Chương này sử dụng các mô hình giải tích để ước lượng hiệu năng, độ trễ và tài nguyên tiêu thụ của kiến trúc đề xuất.

6.1 Phương pháp đánh giá: Mô hình Roofline

6.2 Ước lượng độ trễ và Tài nguyên

6.3 So sánh với các nghiên cứu liên quan

Chapter 7

Kế hoạch phát triển

Chương này tổng kết các kết quả đạt được trong Giai đoạn 1 và đề ra kế hoạch chi tiết cho việc hiện thực và kiểm thử trong Giai đoạn 2.

7.1 Đánh giá mức độ hoàn thành Giai đoạn

1

7.2 Kế hoạch thực hiện Giai đoạn 2

7.3 Tiến độ dự kiến