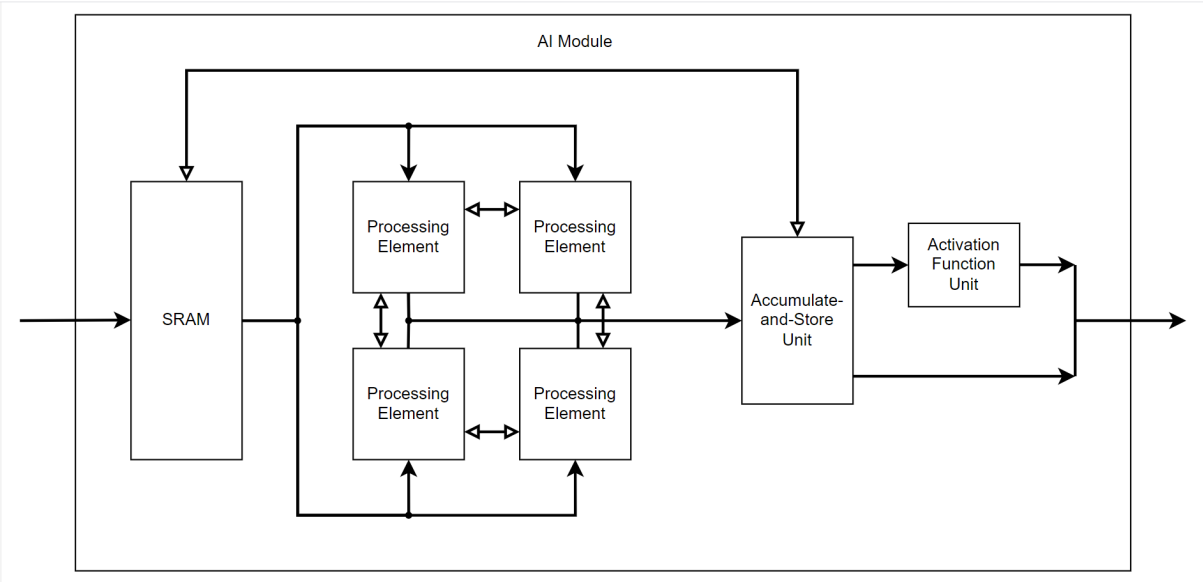
AI Module

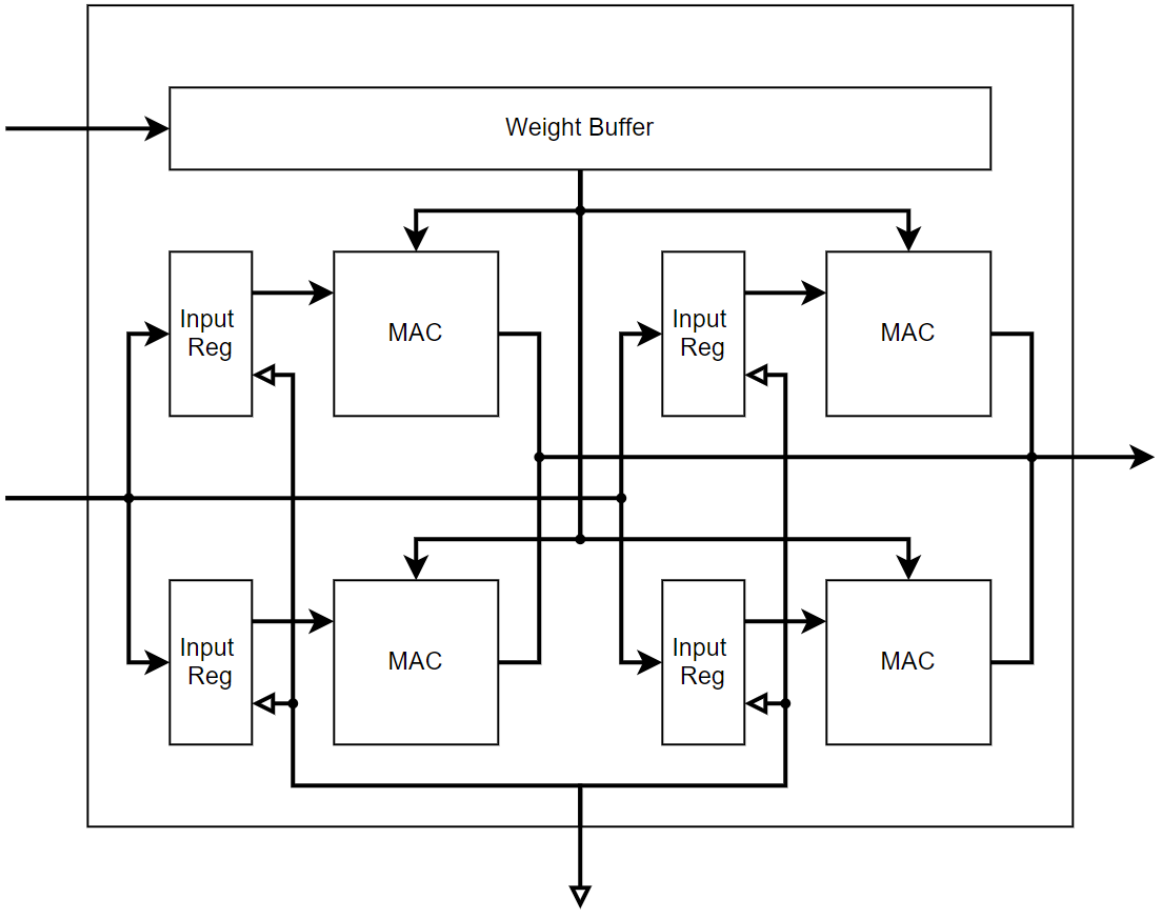
1. Kiến trúc hệ thống



Hệ thống bao gồm 4 thành phần và kết nối như trên hình:

- **SRAM** đóng vai trò lưu trữ các giá trị *weight*, *input feature map (input neurons)* và *bias*. **SRAM** phân phối các giá trị này tới cho dãy các **Processing Unit** và **Accumulate-and-Store Unit**.
- **Processing Unit** đóng vai trò thực hiện tính toán các phép toán cộng và nhân trong mạng *Deep Learning*. Đầu vào của **Processing Unit** là các *weight* và *input feature map (input neurons)* từ SRAM. Các giá trị sau khi tính toán của **Processing Unit** được gửi tới cho **Accumulate-and-Store Unit** để thực hiện cộng dồn các giá trị**.** Ngoài ra, giữa các **Processing Unit** còn có thể giao tiếp với nhau để chia sẻ giá trị *input feature map* phục vụ *Convolutional layer*.
- **Accumulate-and-Store Unit** đóng vai trò cộng dồn giá trị một phần khi tính toán cho *output feature map (output neuron)* và giá trị *bias* tương ứng.
- **Activation Function Unit: Activation Function Unit** đóng vai trò thực hiện tính toán các hàm kích hoạt (*activation function*) của neuron như *ReLU, sigmoid, tanh*, …

2. Processing Unit



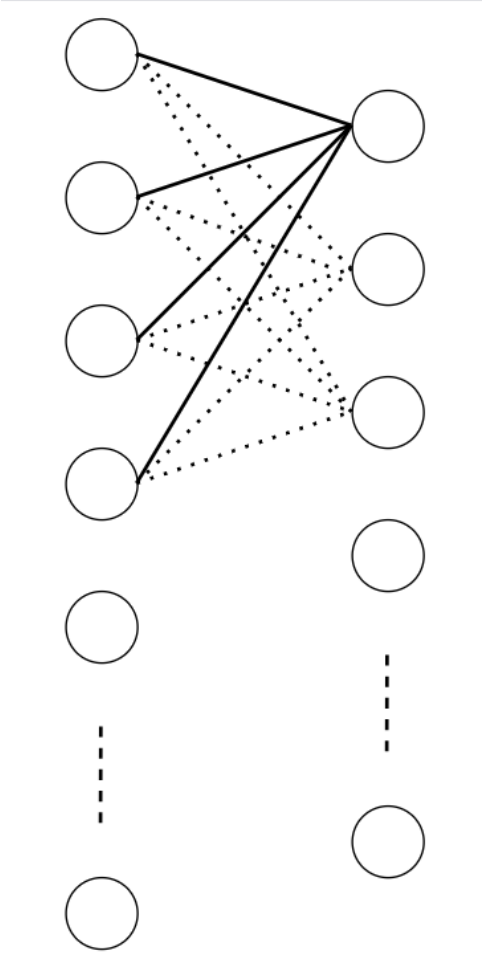
Các thành phần chính của **Processing Unit** là:

- **Weight Buffer**: lưu trữ các giá trị *weight* từ **SRAM** và phân phối tới dãy các **MAC**.
- **Input Reg**: lưu trữ các giá trị của *input feature map (input neuron)* và phân phối tới **MAC** tương ứng. Mỗi **Input Reg** được liên kết với một **MAC**. Ngoài ra, các **Input Reg** này có liên kết với các **Input Reg** ở **Processing Unit** phục vụ mục đích chia sẻ dữ liệu *input feature map (input neuron)* đối với *Convolutional layer*.
- **MAC**: thực hiện tính toán phép nhân và cộng dồn ($a_0 * b_0 + a_1 * b_1 + a_2 * b_2 + \ldots$) từ các giá trị *weight* và *input feature map (input neuron)*.

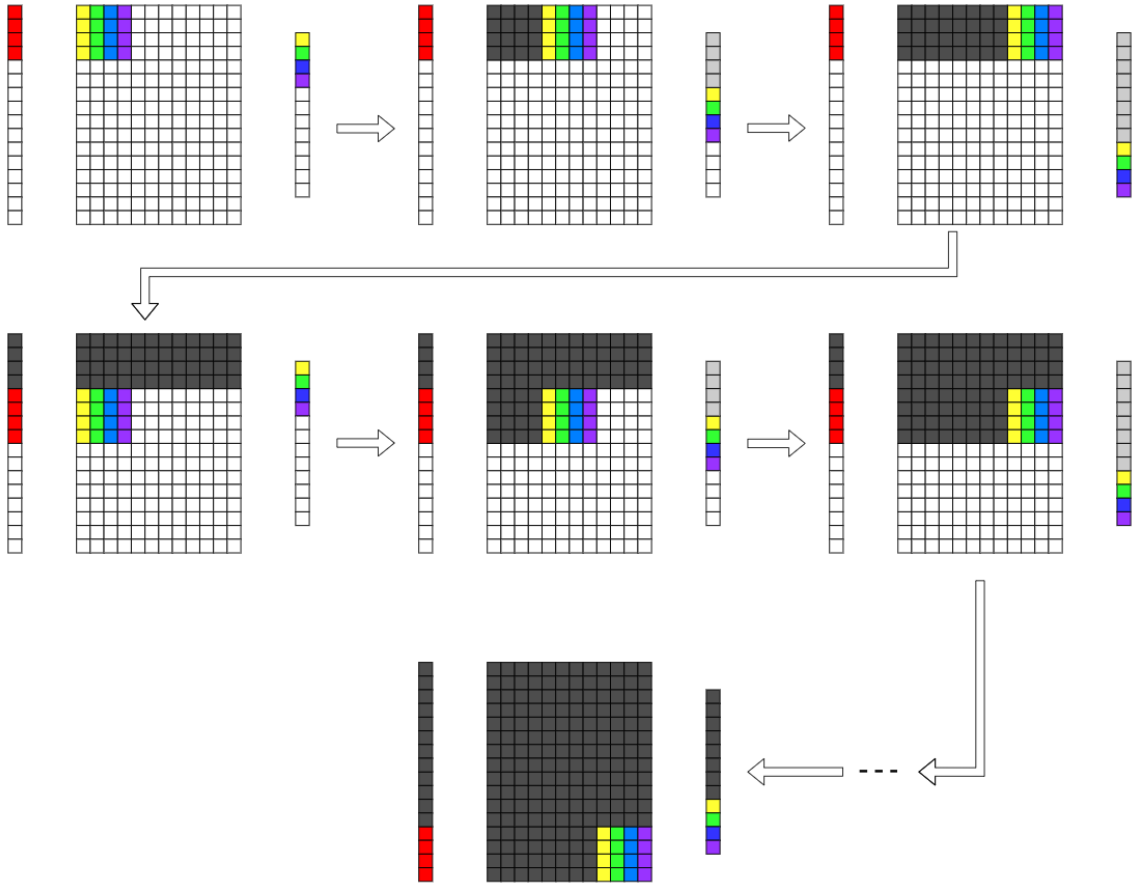
3. Data flow

3.1 Fully-Connected Layer

Mô hình của *Fullly-Connected Layer* mô hình như bên dưới:



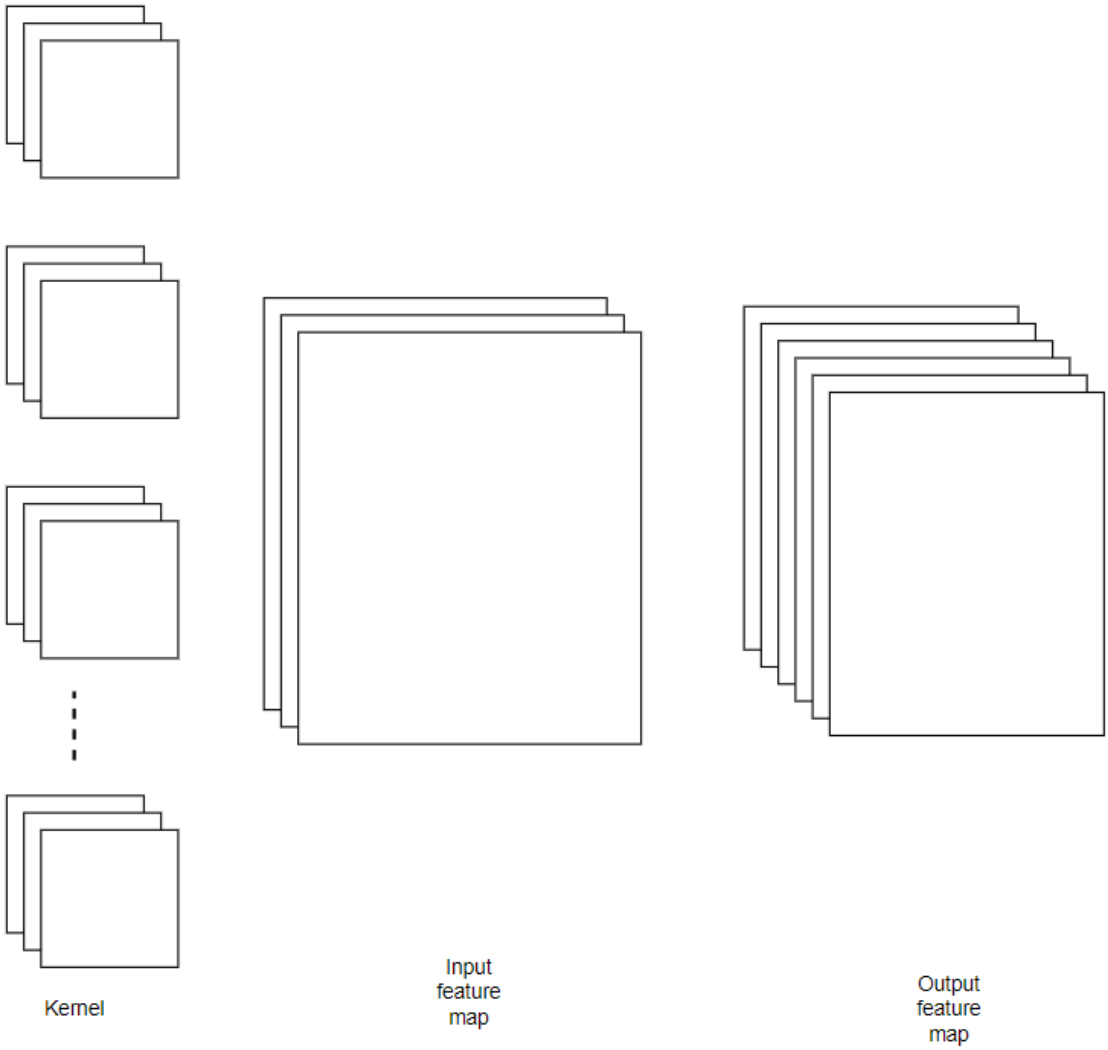
Ở layer này, các tác vụ bộ nhớ áp đảo hoàn toàn so với các tác vụ tính toán. Nguyên nhân của việc này là do số lượng *weight* cực kỳ lớn. Mô hình tính toán được đề xuất như sau:



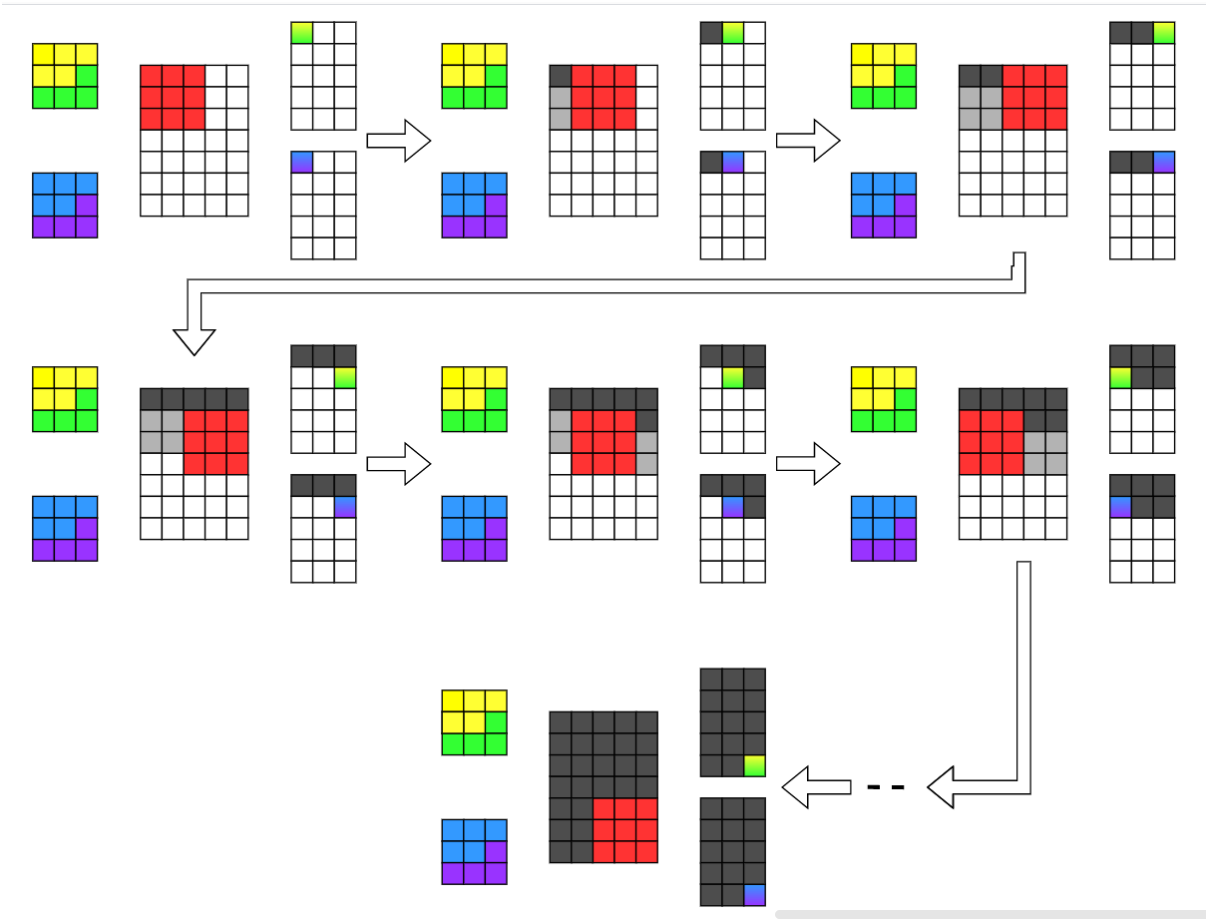
Mô hình này khai thác tối đa việc tái sử dụng *input feature map*.

3.2 Convolutional Layer

Mô hình của *Connvolutional Layer* mô hình như bên dưới:



Ở layer này, các tác vụ tính toán áp đảo hoàn toàn so với các tác vụ bộ nhớ. Mô hình tính toán được đề xuất như sau:



Mô hình này khai thác tối đa việc tái sử dụng *input feature map*.