1. Systolic architectures

(a) Given an input vector $\mathbf{A}$:$[a_1, a_2, a_3, a_4, a_5, a_6, a_7]$ and weight vector $\mathbf{W}$:$[w_1, w_2, w_3]$, design systolic arrays of processing engines (PEs) for each of the design specifications listed below to perform 1D-convolution of input and weight vectors resulting in the output vector $\mathbf{Y}$:$[y_1, y_2, ...]$, where $y_n = w_1x_n + w_2x_{n+1} + w_3x_{n+2}$: [15 pts]

(1) Weights - stationary in PEs; Inputs - broadcasted; Outputs - move systolically.

(2) Weights - stationary in PEs; Inputs - move systolically; Outputs - evaluated through fan-in of partial sums from PEs.

(3) Weights - stationary in PEs; Inputs and outputs move systolically in opposite direction.

Clearly indicate all the signals, appropriately timed, and consider PE as a black box that computes MAC.

(b) Design a 2-D mesh of systolic PEs to carry out matrix-matrix multiplication between $\mathbf{A}_{3\times3}$ and $\mathbf{W}_{3\times3}$ matrices with output stationary dataflow. Clearly illustrate all the signals and timed dataflow and consider PE as a black box that computes MAC in one cycle. [20 pts]

(1) Calculate total cycles for the above operation.

(2) Consider mixed-precision operation where weights are quantized to 8 bits (Int), inputs quantized to 16 bits (half-float) and outputs are single precision floating-point. What is the suitable dataflow (input or weight or output stationarity) for area efficiency?

(3) Calculate the PE utilization to process above matrices when the dataflow is weight stationary and size of the PE array is 32 ×32.

(c) Consider the following parameters of a CNN model (FastRCNN): [25 pts]

Ow: The width of output feature map
Oh: The height of output feature map
kw: The width of filter
kh: The height of filter
C: The number of channels
F: The number of filters

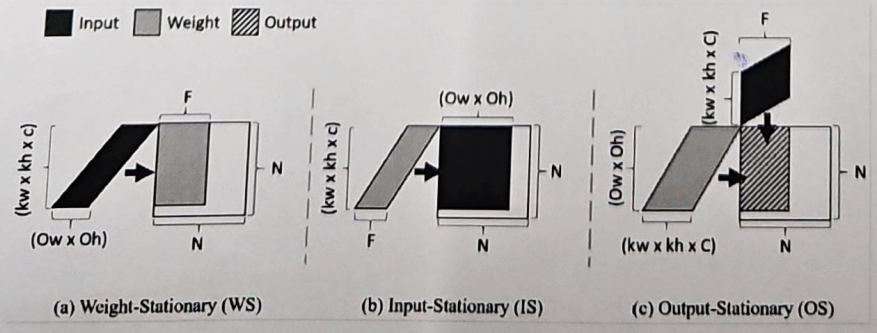

Figure 1: Execution workflow of three dataflows

Each layer of the model is parameterized as (Ow, Oh, kw, kh, C, F): layer L1(56, 56, 1, 1, 64, 256), layer L2(26, 26, 3, 3, 128, 128), layer L3(14, 14, 1, 1, 512, 1024). The size of systolic array is 128 ×128.

(1) Calculate the execution time of all three dataflows for all three layers on the systolic array (refer Fig. 1).

(2) Can we opt for a single dataflow for all three layers on the systolic array?

(3) Which of the dataflow policies (input or weight or output stationary) is the best for each of the three layers.

(d) Design a multi-dataflow processing engine (PE) and datapath controller micro- architectures for the port mapping given below. [10 pts]

DO THIS.

	Weight-stationary	Input-stationary	Output-stationary
preload	*Weight*	*Input*	X
Input_0	X	X	*Weight*
Input_1	*PartialSum*	*PartialSum*	X
Input_2	*Input*	*Weight*	*Input*
Output_0	X	X	*Weight*
Output_1	*PartialSum'*	*PartialSum'*	X
Output_2	*Input*	*Weight*	*Input*

Figure 2: Port mapping in each dataflow

2. Model Compression

Consider the following Convolutional Neural Network (CNN) designed for CIFAR-10 classification (input image size: 32×32×3):

- Conv1: 3×3 kernel, stride = 1, padding = 1, 32 output channels
- Conv2: 3×3 kernel, stride = 1, padding = 1, 64 output channels, followed by 2×2 max-pooling
- FC1: Fully connected layer with 512 hidden units
- FC2: Fully connected layer with 10 output units (for classification)

(a) Basics of CNNs [10 pts]

1. Compute the number of trainable parameters and MACs (Multiply-Accumulate operations) in each layer and total parameters in the network.
2. Suppose the Conv2 is replaced by Conv2 (depthwise separable) — depthwise 3×3 (stride = 1, padding =1) on the incoming channels, followed by a pointwise 1×1 to 64 output channels, then a 2×2 max-pool (stride 2). Compute the number of trainable parameters and MACs again.

(b) Pruning [20 pts]

1. Suppose we do unstructured pruning to remove 50% of weights in Conv2. What is the MACs and number of parameters that need to be stored? Explain your answer.

2. Suppose we do 3x3 filter pruning of 50% of the filters on Conv2. What is the MACs and number of parameters that need to be stored? Explain your answer with appropriate examples of the Matrix multiplication that happens.

(c) Discuss the pros and cons of Uniform/Linear Quantization vs Nonlinear/K-Means based quantization? [20 pts]