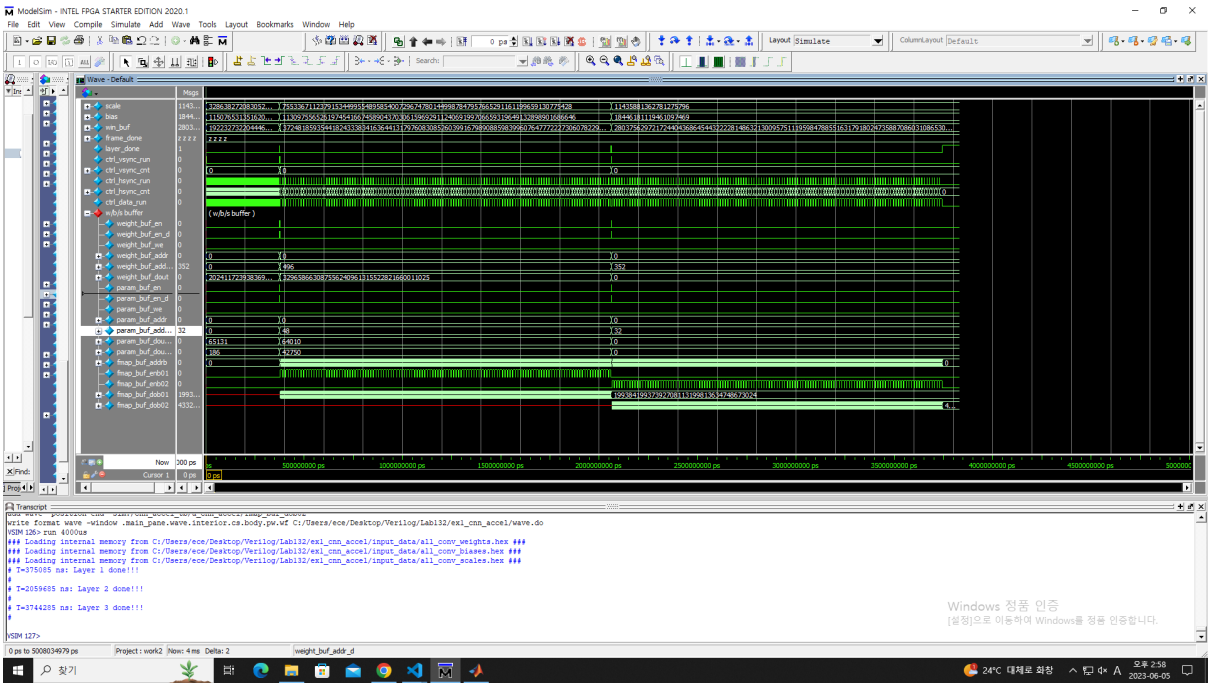
Homework 12: CNN Accelerator and DMA

SATYAM / 학생 / 조선해양공학과 **(2023-81784)**

****

Problem 1 (30p): CNN Accelerator

1. Weight/bias/scale buffers (10p)



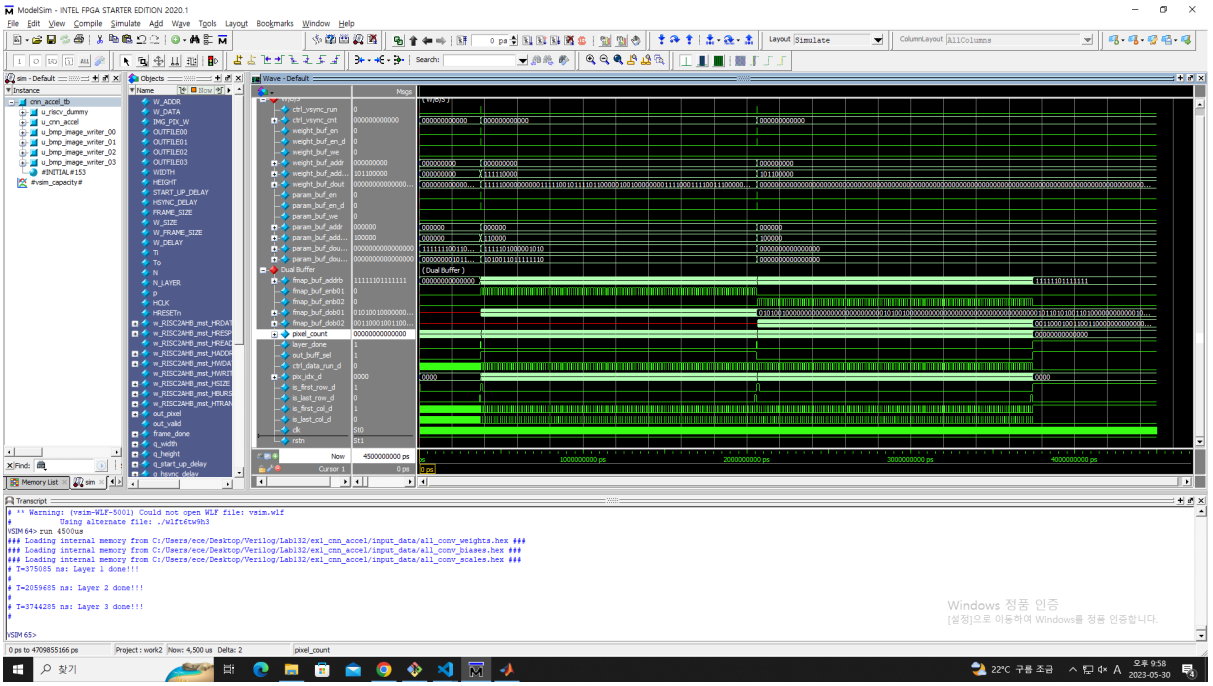
a. (3p) Determine the buffer sizes and the minimum number of cycles required for weight/bias/scale preloading by completing the following table:

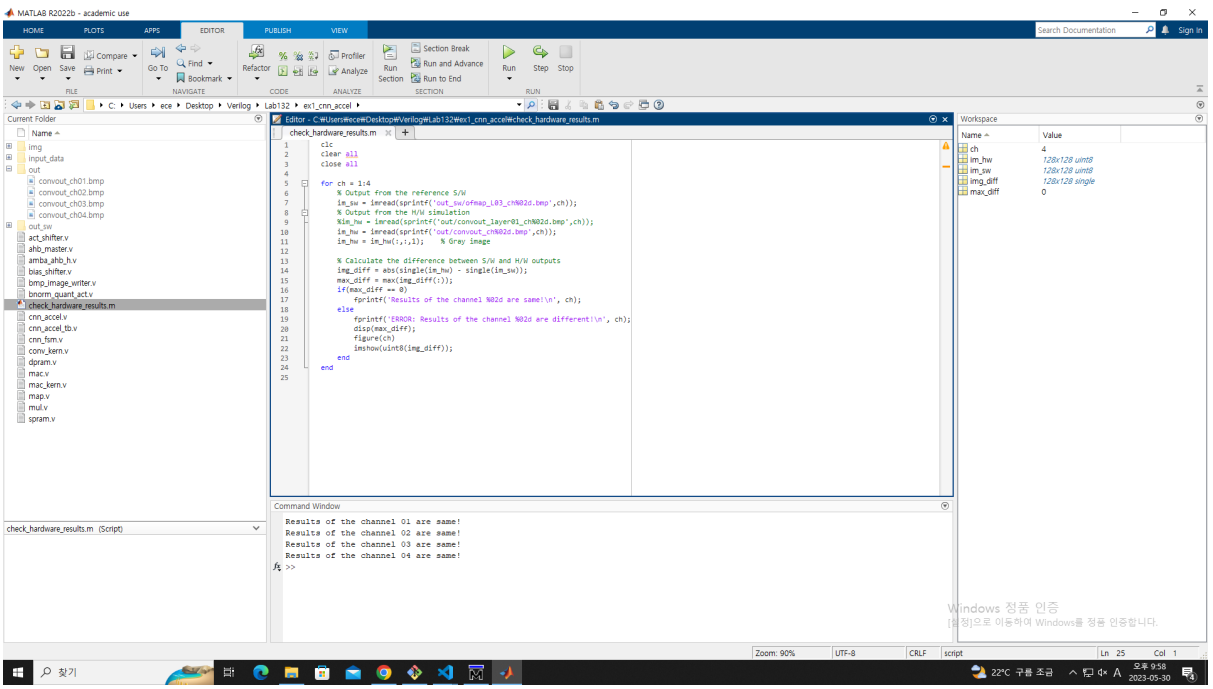
Buffer	The number of cycles			The size of buffers		
	Layer 1	Layer 2	Layer 3	Word (bit)	No. of words	Size (bit)
Weight	16	144	144	128	304	38912
Bias	16	16	16	16	48	768
Scale	16	16	16	16	48	768

b. (2p) Explain how the base addresses q_base_addr_weight and q_base_addr_param are used.

This q_base_addr_weight and q_base_addr_param are base addresses used to tell which address to start reading the data from for weight and param (biases/scales) in each layer.

2. Dual buffers for feature maps (20p)





a. (2p) Explain how the flags q_is_first_layer and out_buff_sel are used in this code.

Using the q_is_first_layer and the out_buff_sel we select which buffer we want to put data to, first or second buffer. So if it's the second layer then q_is_first_layer will be false and out_buff_sel will be true to select the first buffer. And if it's the third layer then q_is_first_layer will be false and out_buff_sel will be false to select the second buffer.

b. (3p) Determine the buffer sizes by completing the following table:

Buffer	Buffer size		
	Word (bit)	Number of words	Size (bit)
Input buffer (in_img)	8	16384	131072
Buffer 1	128	16384	2097152
Buffer 2	128	16384	2097152

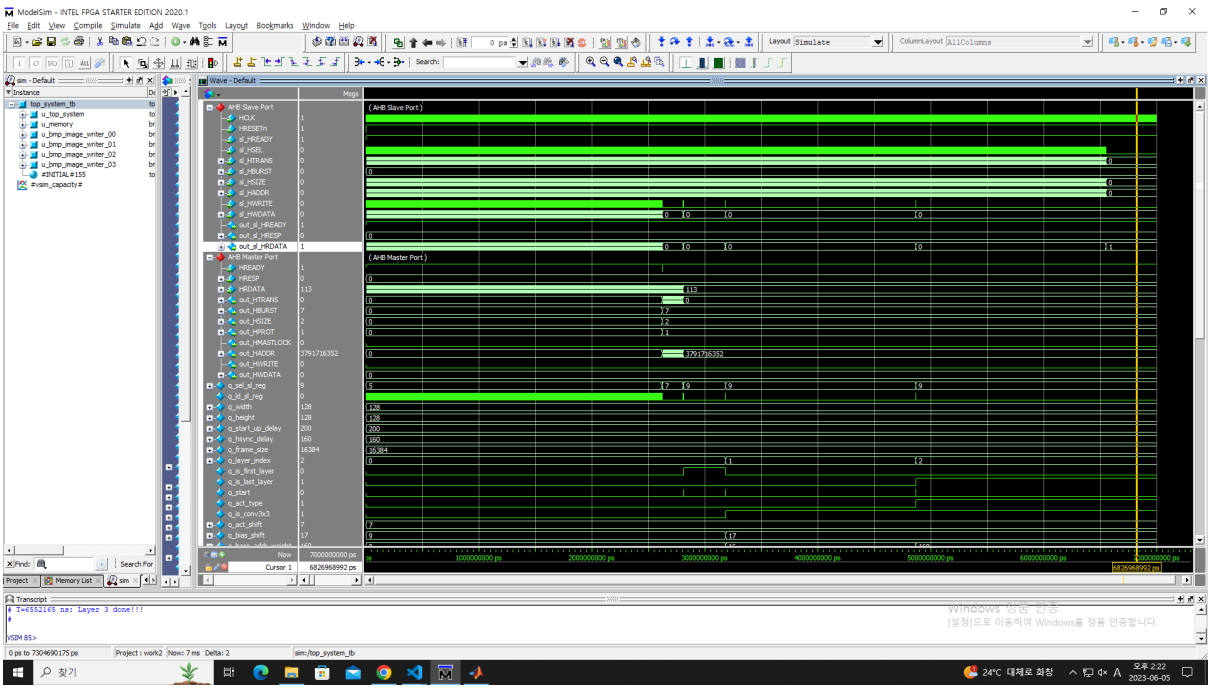
c. (2p) As explained in the class, in our CNN accelerator (cnn_accel.v), each convolution kernel (conv_kern.v) is used to generate an output feature map. In each convolution kernel, Ti multipliers are used to do convolution on the pixels from multiple input feature maps. Fig. 1-5 shows the captured waveform of convolution kernels 0 and 8 when doing the simulation with time = 375us. Explain why the outputs of multipliers 9~15 are zero.

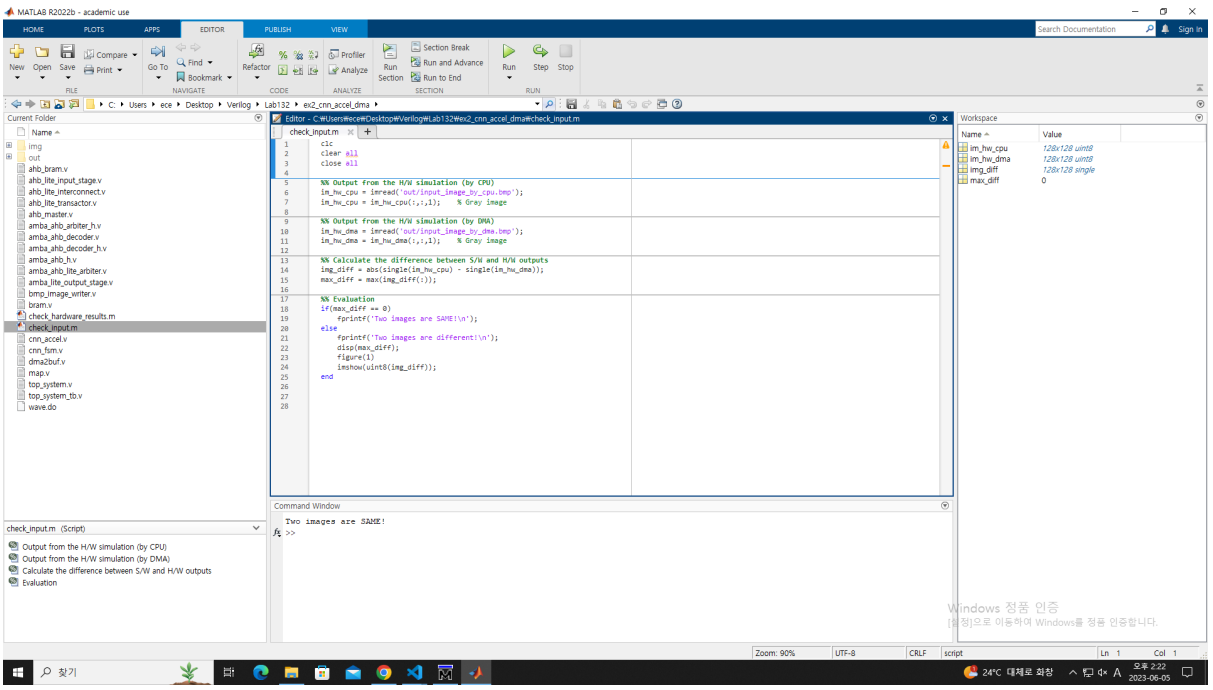
Because we need only 9 convolution kernels to calculate the pixel as we are doing 3x3 convolution, which uses a 3 by 3 matrix to calculate the output from a pixel.

d. (2p) After the simulation, the output of the last layer is stored in conv_output_L03.txt, as shown in Fig. 6. Explain the file's data format.

Only the last 8 characters are used for 4 images which are produced in the last layer's output. So 2 characters (4*2=8 bit) defines one pixel of an image out of the four.

Problem 2 (10p): DMA





a. (2p) By analyzing the data patterns of the AHB master port and the AHB slave port of cnn_accel in Problem 2, explain how DMA can speed up the data reading process compared to the CPU.

DMA is used to speed up the reading process compared to CPU by requesting and getting larger Blocks of data to the CNN accelerator. And meanwhile we are freeing the CPU from outsourcing its work to DMA, which is directly accessing CPU’s memory to output the data for CNN accelerator.