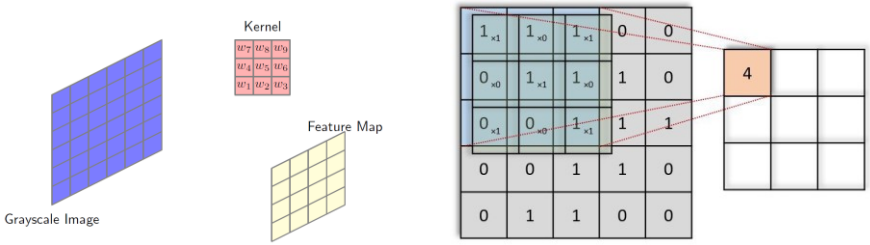
HDL HW5:
CNN Accelerator for VGG-16

Outlines

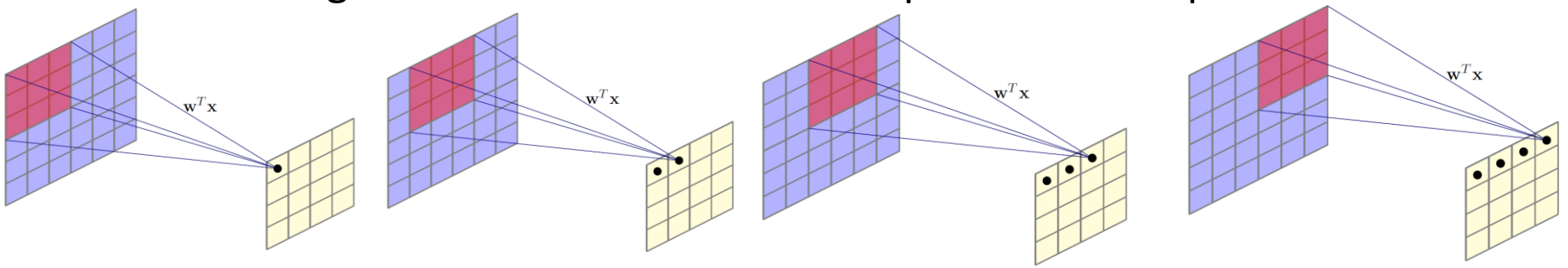
- Deep Neural Network (DNN) hardware accelerator
  - ICP=x, KWP=9, OCP=y
  - each PE computes partial sums for x input channels
    - ✓ multiply-adder-tree (MAT) structure for each PE
  - y PEs in parallel, each for an output channel
  - execution of first two layers of VGG-16 Model
  - calculate total execution cycles and memory accesses
- could reduce ICP=x->1, and/or OCP=y->1
  - fewer number of multipliers => smaller area/power => less simulation time => more execution cycles
  - determined ICP, OCP by yourself
- memory is in testbench, not in hardware (hardware memory in next bonus homework)

2D Convolution

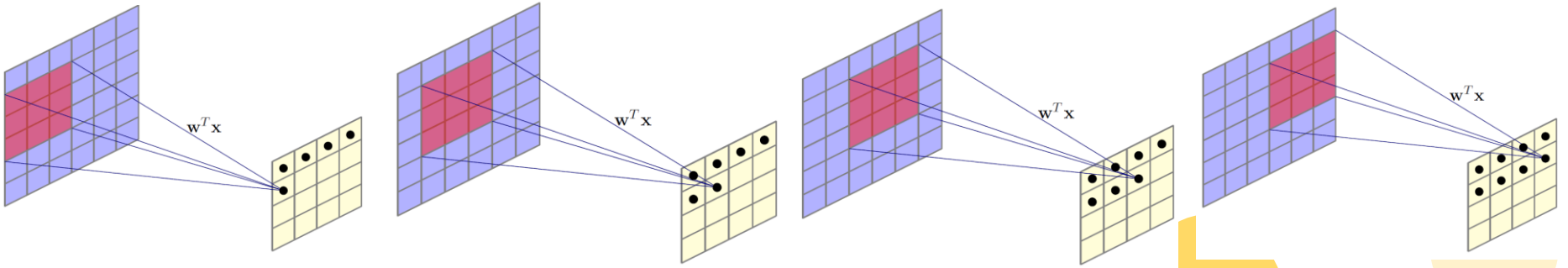
https://miro.medium.com/max/875/1*wpbLgTW_lopZ6JtDqVByuA.gif



- Convolution to generate the 1st row of an output feature map



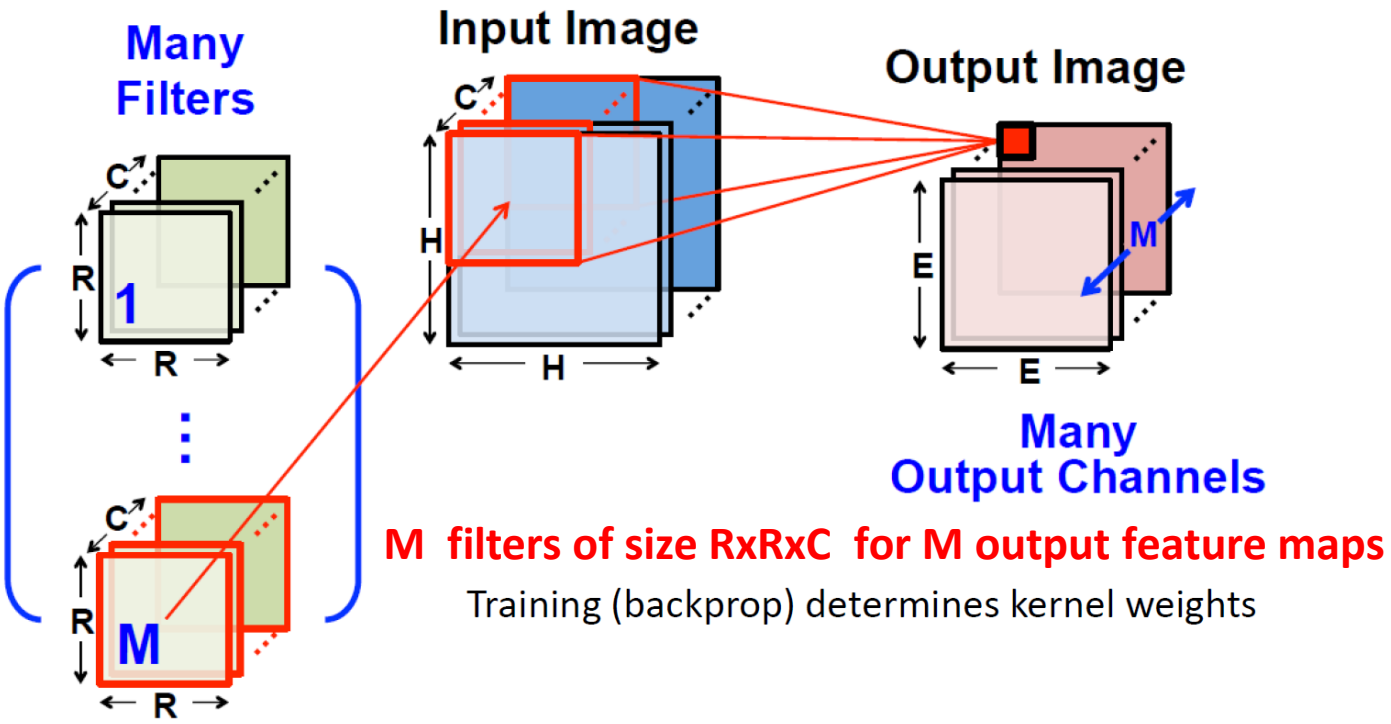
- Convolution to generate the 2nd row of an output feature map



...

3D Convolution (With illustration)

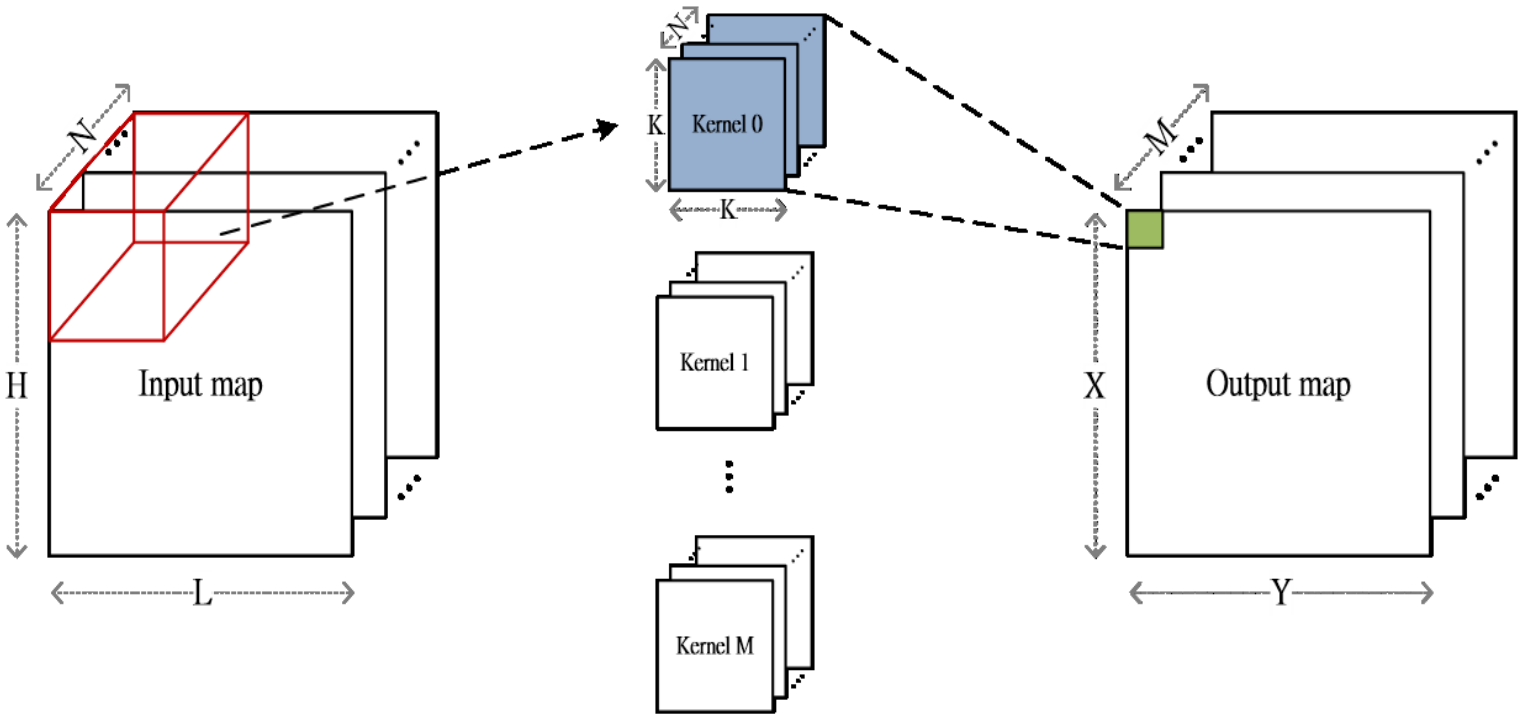
- **C** input feature maps => **M** output feature maps
- Need **CxM** filter kernels, **C** filter kernels for each of the **M** output channels
- What are the total numbers of parameters and Multiply-Add (MAD) operations?



3D Convolution (Mathematical Expression)

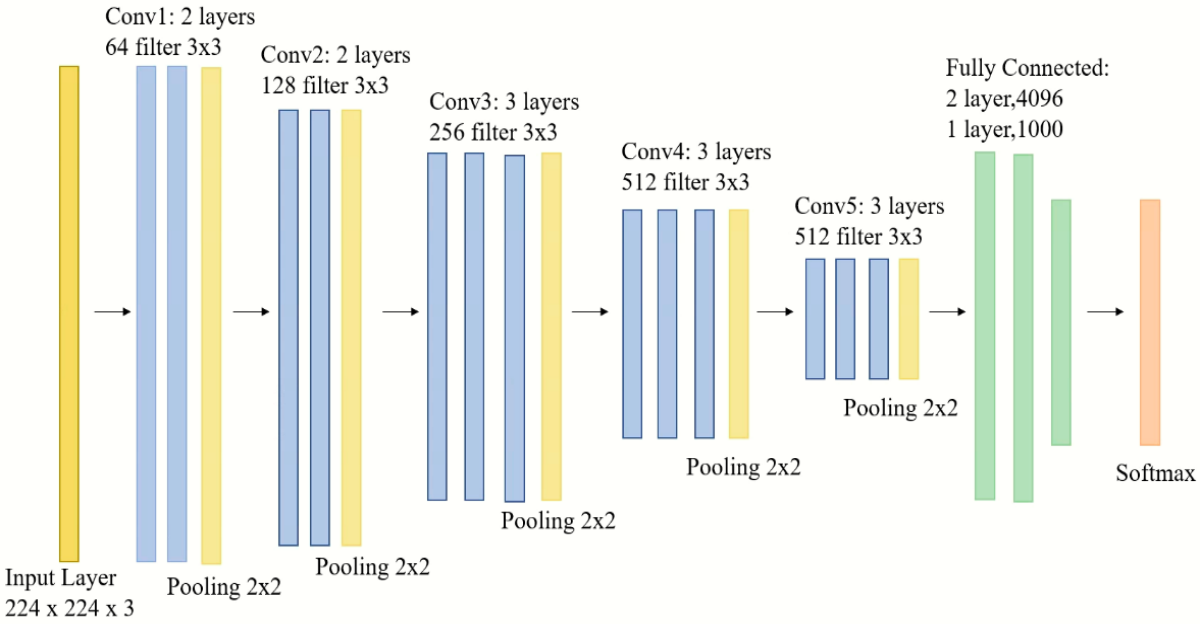
- *N* input feature maps (input channels) *I* of size H*W
- *M* output feature maps (output channels*) O* of size X*Y
- Filter kernel *W* of size K*K

$$O[m][x][y] \mathrel{+}= \text{bias}[m] + \sum_{i=0}^{K-1} \sum_{j=0}^{K-1} W[m][n][i][j] * I[n][x+i][y+j]$$



VGG-16 CNN Model

- VGG-16 CNN Model
  - 13 convolutional layers
  - 3 fully connected layers

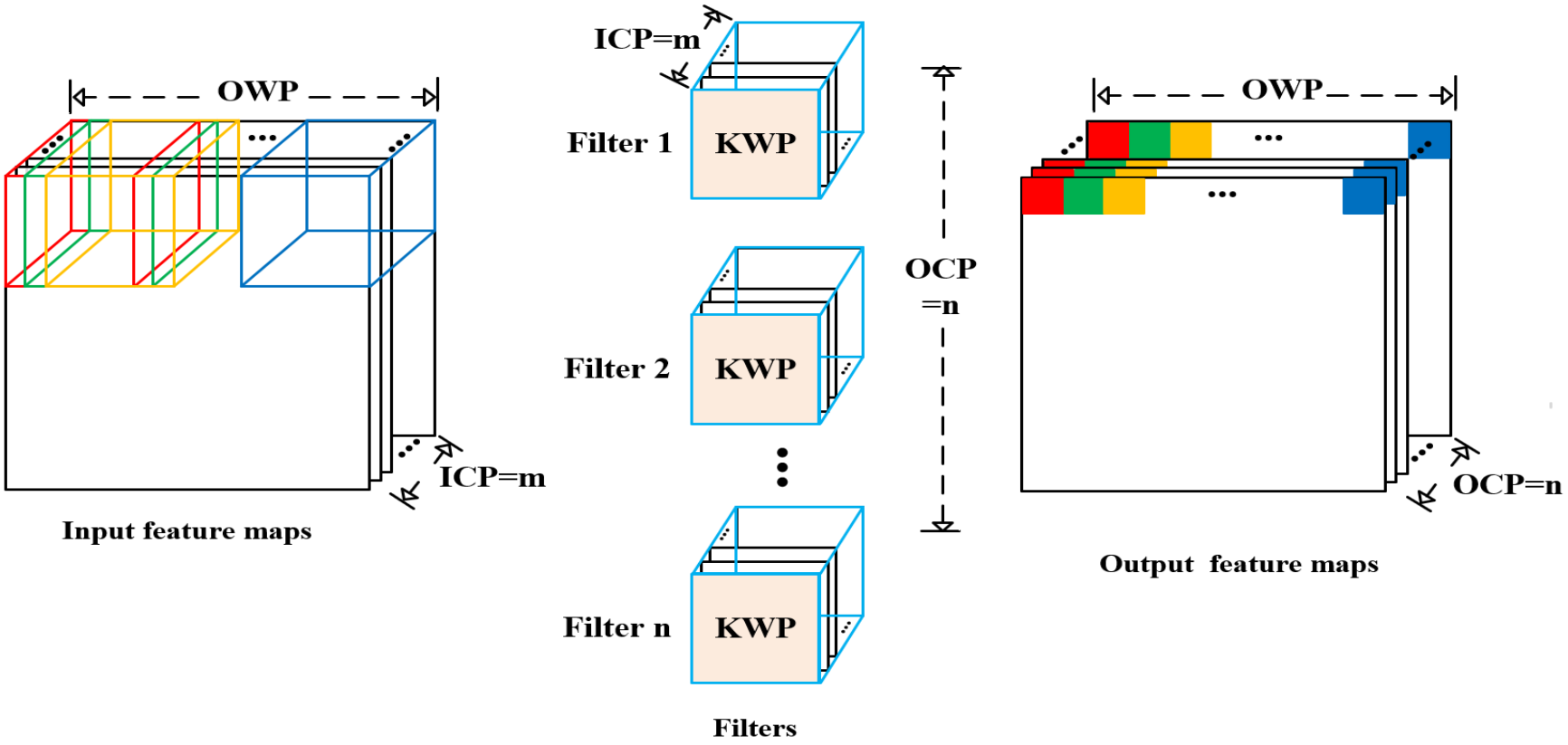


- First two layers of VGG-16

# of input channels	# of output channels	featuremap size	Filter Kernel size	Stride
3	64	224x224	3x3	1
64	64	224x224	3x3	1

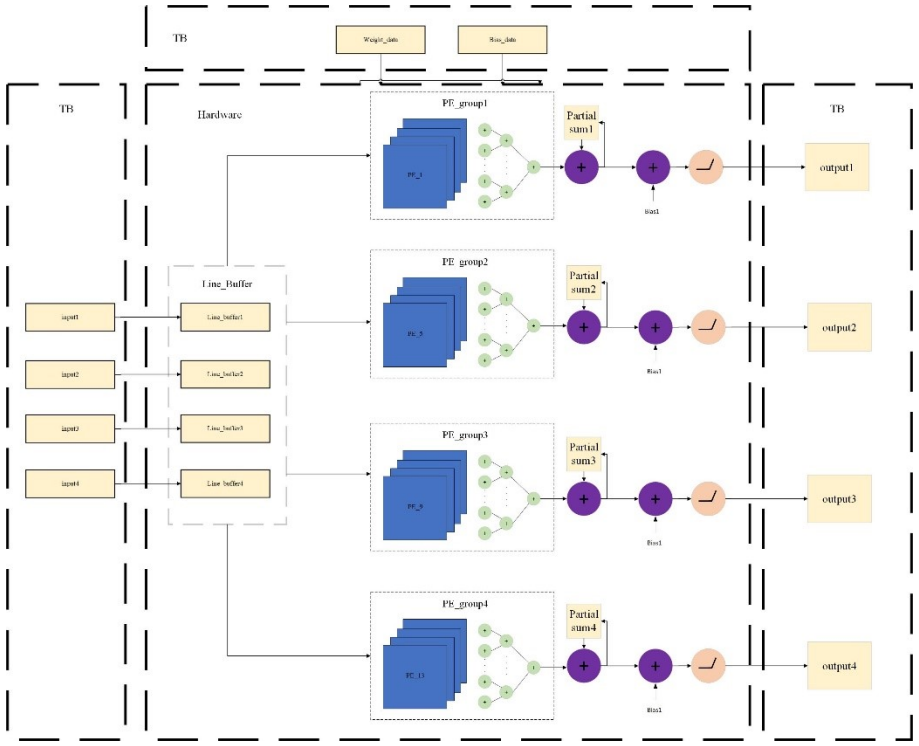
Hardware Parallelism

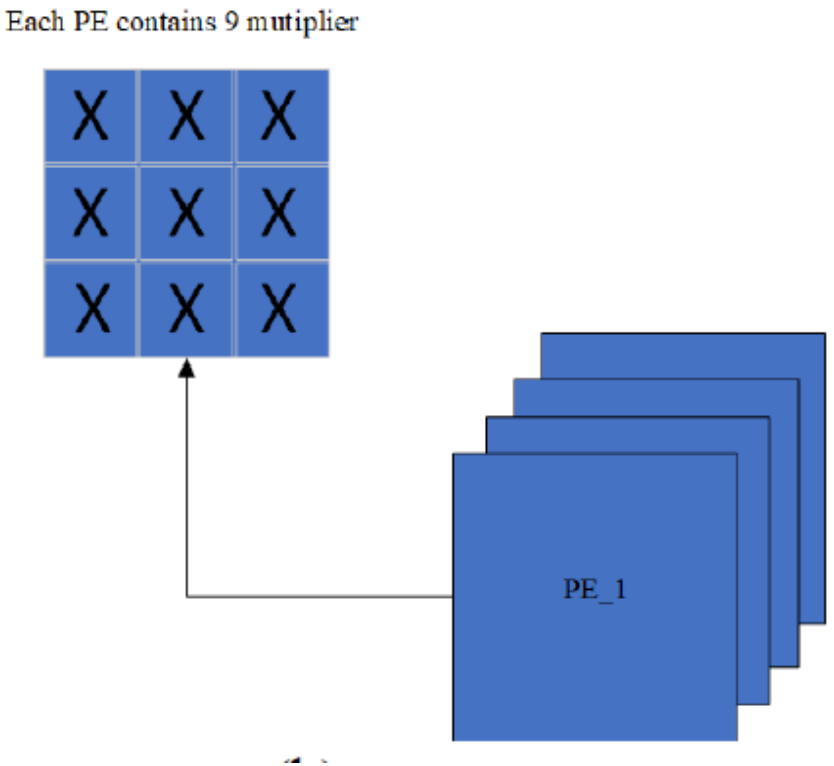
- Input Channel Parallelism (ICP)
  - ICP=m, m inpout channels are processed in parallel
- Output Channel Parallelsim (OCP)
  - OCP=n, n output channels are computed in parallel
- Kernel Window Parallelism (KWP)
  - KWP=k, k filter kernel coefficients are processed in parallel



Overall Architecture (ICP=4, KWP=9, OCP=4)

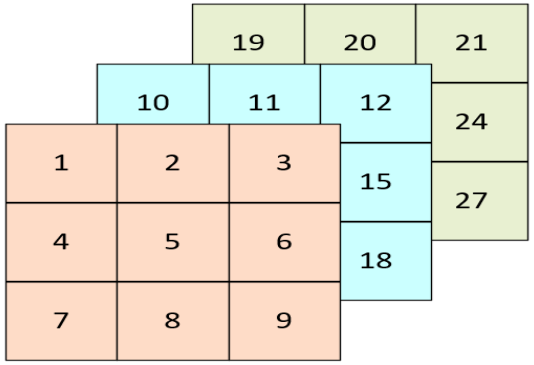
- PE contains multiply-adder-tree (MAT) for ICP=4 and KWP=9
- 4 PEs in parallel (OCP=4), each for an output channel
- Input/Weight/Output memory
- line buffers support concurrent data access

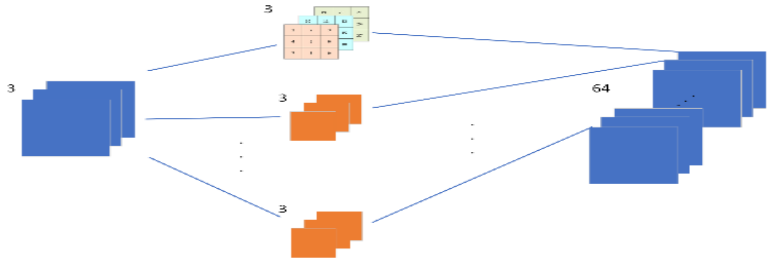




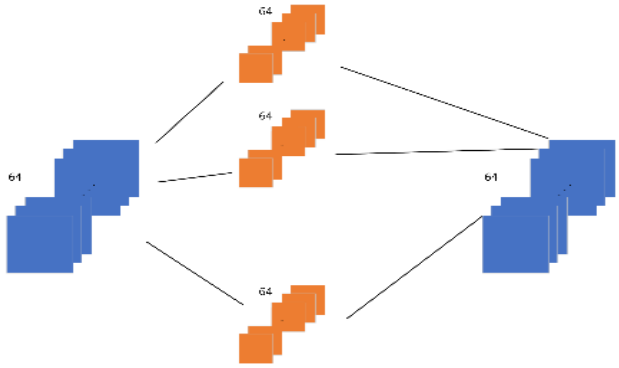
VGG Layers 1 and 2

- 1st layer (ICH=3, OCH=64)
  - use only 3 line buffers and MAT
  - use 4 PEs
  - execution cycles?
  - amount of memory access?
- 2nd layer (ICH=64, OCH=64)
  - use 4 line buffers and MAT
  - use 4 PEs
  - execution cycles?
  - amount of memory accesses?
- select ICP, OCP for your design
  - affect EDA tool simulation time and synthesis time
  - affect hardware execution cycles and amount of memory accesses





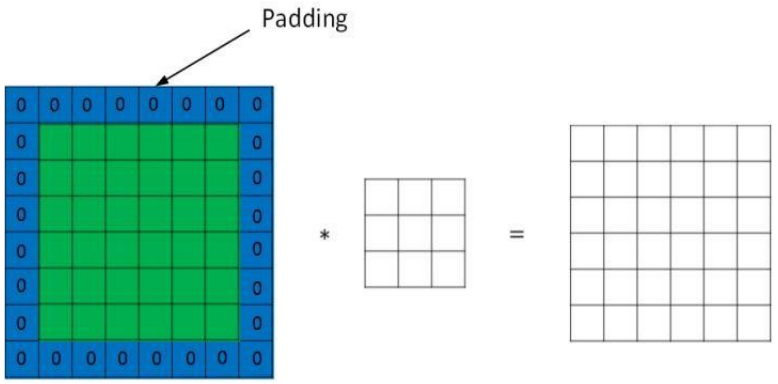
First layer's parameter counts:
(kernal_width*kernal_height*input_channels+1)*filters = (3*3*3+1)*64 = 1728 + 64

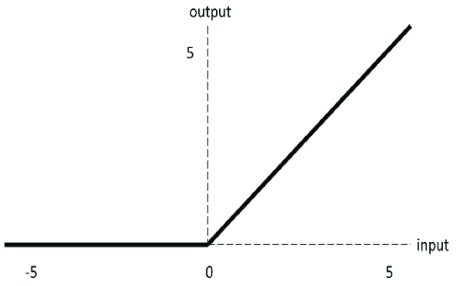


Second layer's parameter counts:
(kernal_width*kernal_height*input_channels+1)*filters = (3*3*64+1)*64 = 36864 + 64

Implementations

- Padding
  - make the output feature map size after convolution the same as that of input feature map \
  - ciould be done beforehand in testbench
- Line Buffers
  - ICP line buffers
    - ✓ i.e., one for each input channel
- 3x3 convolution units (PEs)
  - OCP PEs
    - ✓ i.e., one PE for each output channel
- ReLU
  - Force to zero for the negative convolution results
  - Non-linear activation function after each CNN layer
- ICP and OCP affects numbers of line buffers and PEs

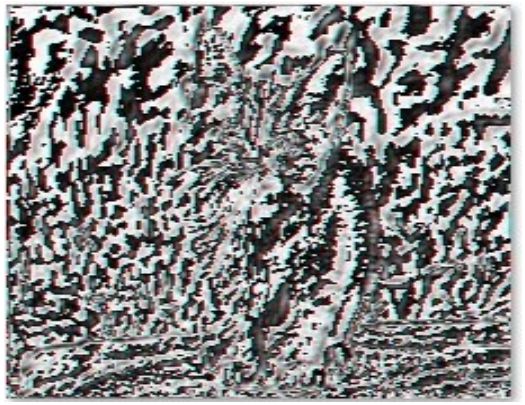




Some Feature Maps

- Original image

- 1st output feature map of the 1st layer
  - `3*3*3*64 filter weights
    - ✓3*3*3 filters for accumulation of three 3*3 convolution, each corresponding to an input channel
    - ✓first of the 64 output channels



- 1st output feature map of the 2nd layer
  - 3*3*64*64 filter weights
    - ✓3*3*64 filters for accumulation of 64 3*3 convolution, each corresponding to an input channel
    - ✓first of the 64 output channels