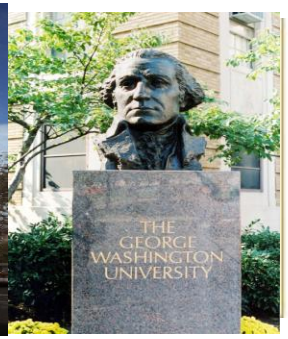


Photonic Processor for Fully Discretized Neural Networks

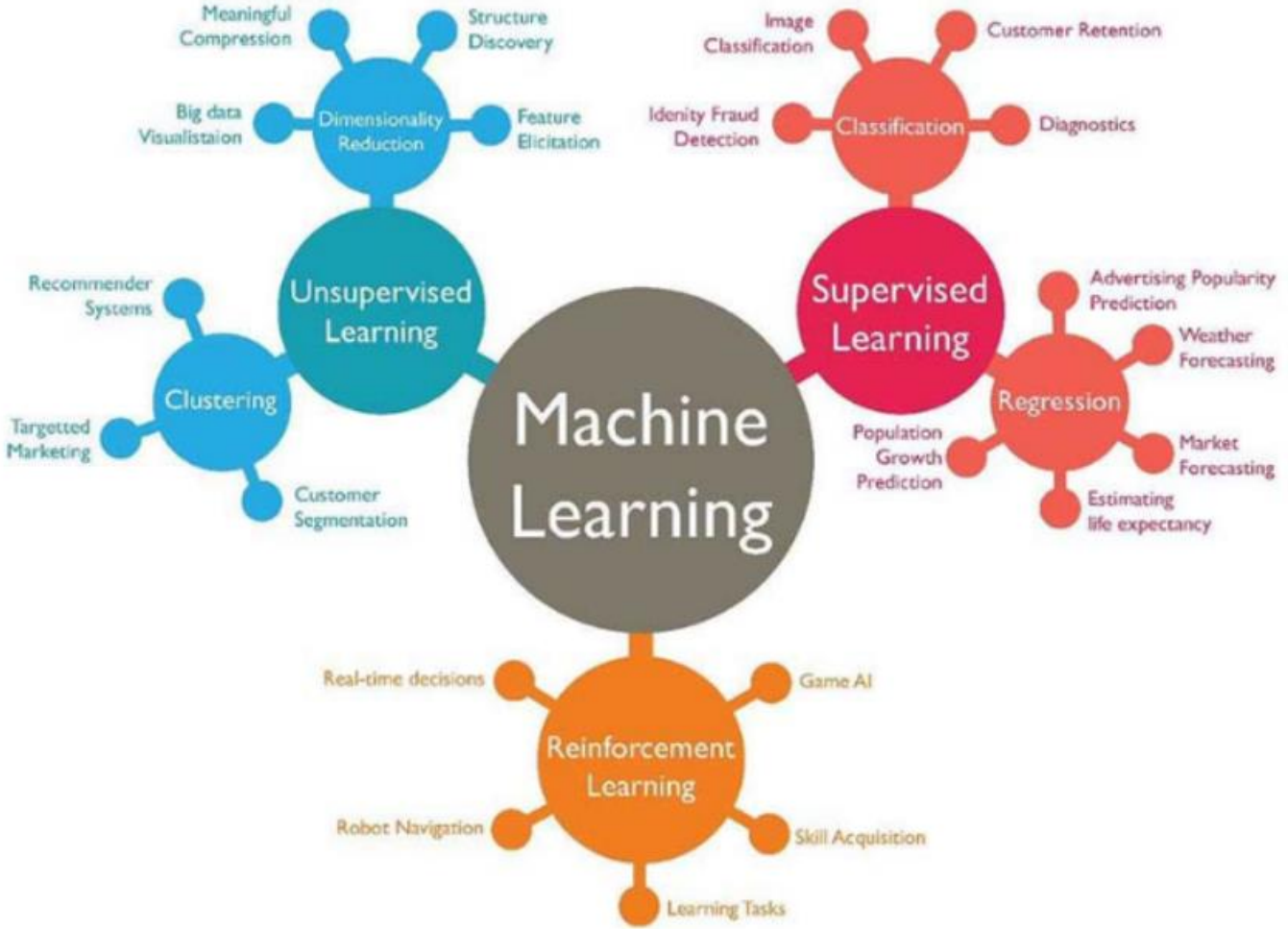
Jeff Anderson, Shuai Sun, Yousra Alkabani, Volker Sorger, Tarek El-Ghazawi

The George Washington University

July 2019

Introduction to ML

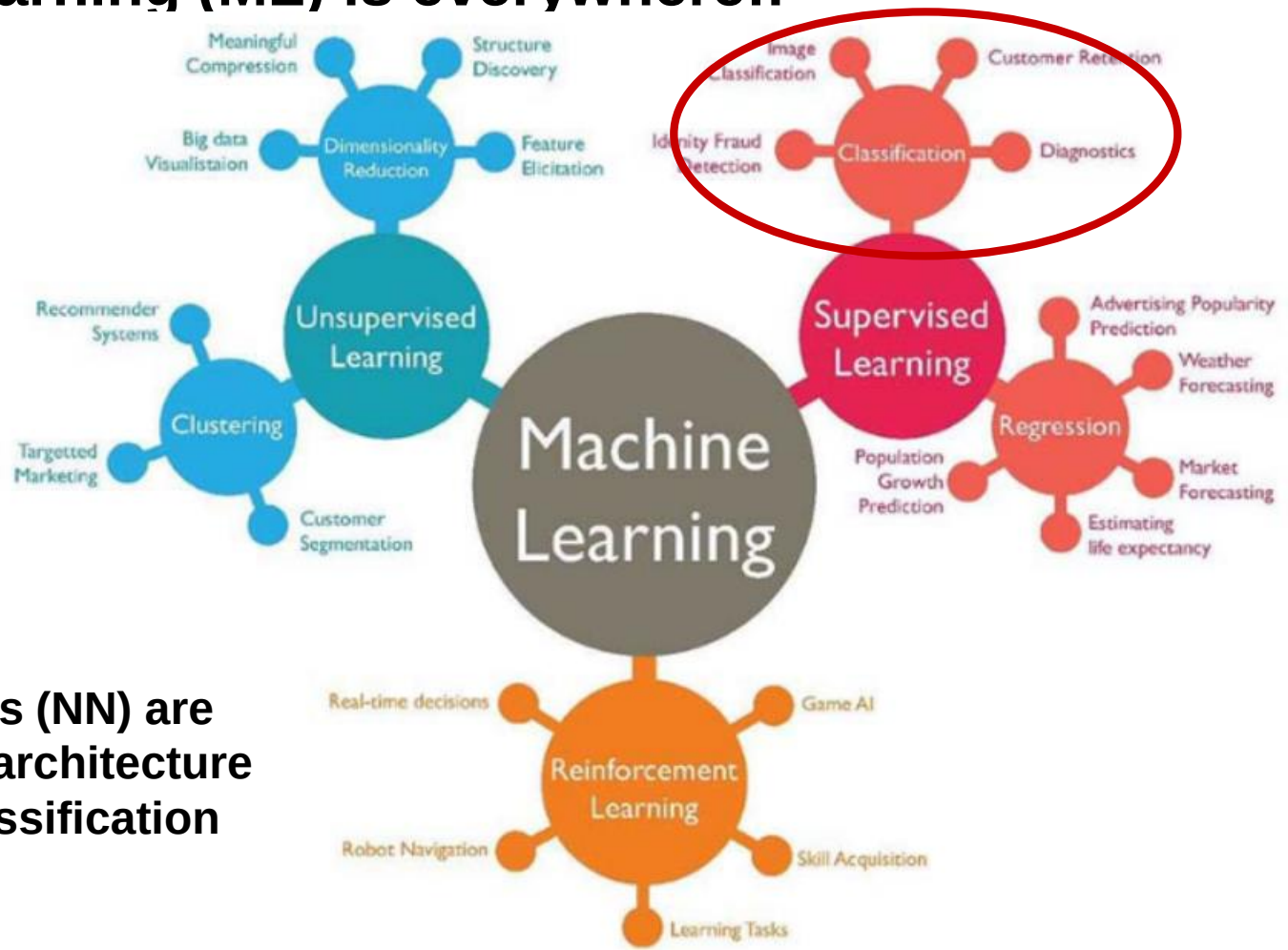
- **Machine Learning (ML) is everywhere!!**



https://www.datasciencecentral.com/profiles/blog/show?id=6448529%3ABlogPost%3A598753&commentId=6448529%3AComment%3A599182&xg_source=activity

Introduction to ML

- **Machine Learning (ML) is everywhere!!**

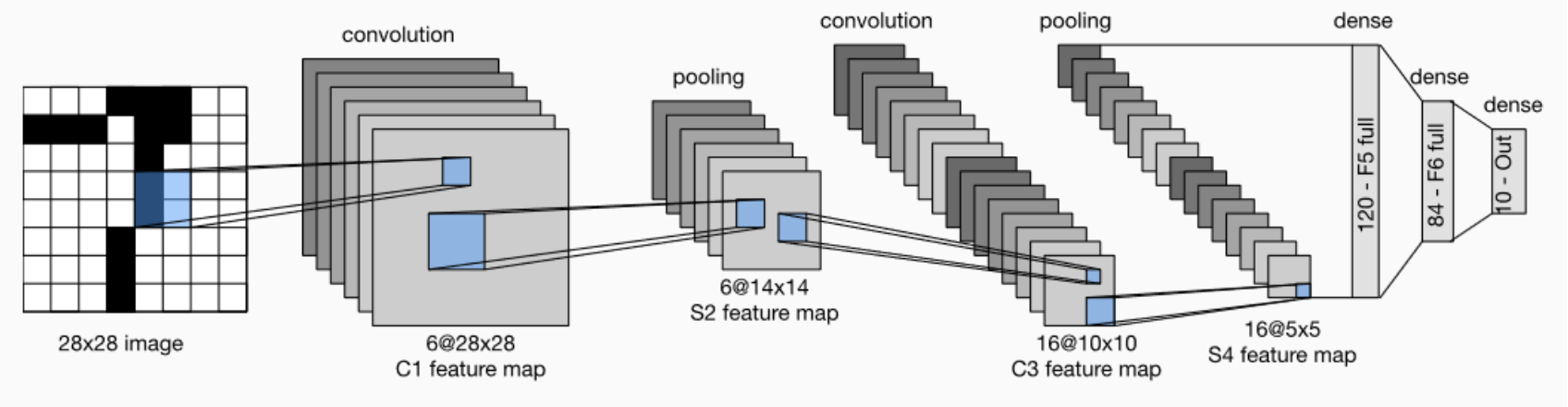


- **Neural Networks (NN) are the underlying architecture of many ML classification systems.**

https://www.datasciencecentral.com/profiles/blog/show?id=6448529%3ABlogPost%3A598753&commentId=6448529%3AComment%3A599182&xg_source=activity

Your First Convolutional Neural Network

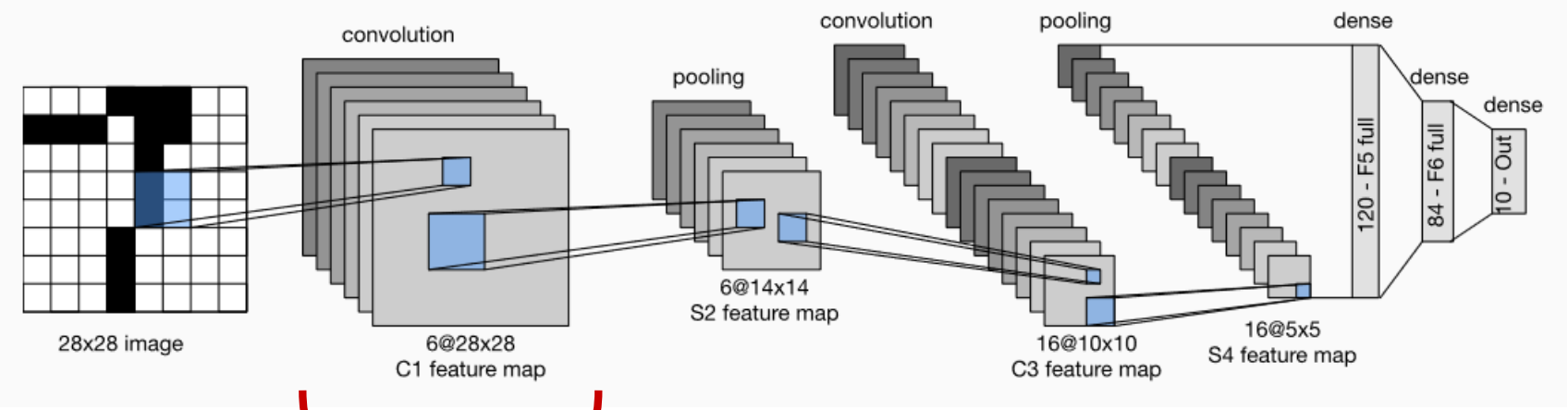
- **Lenet-5 is used for handwriting recognition**



https://www.d2l.ai/chapter_convolutional-neural-networks/lenet.html

Your First Convolutional Neural Network

- **Lenet-5 is used for handwriting recognition**

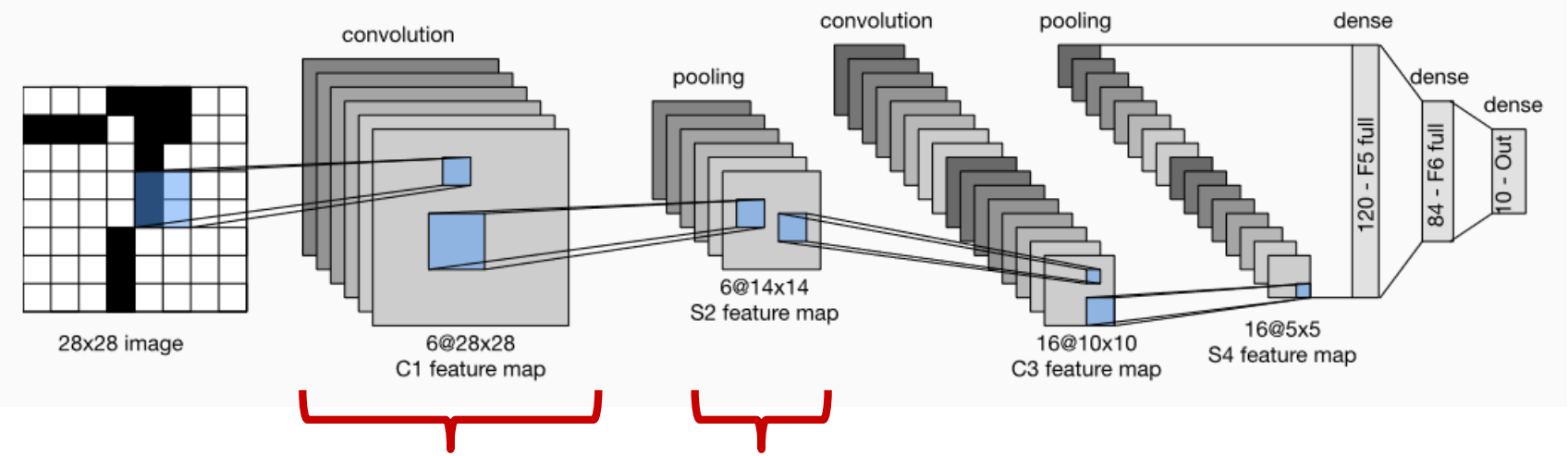


  - **Feature extraction identifies interesting features**

https://www.d2l.ai/chapter_convolutional-neural-networks/lenet.html

Your First Convolutional Neural Network

- **Lenet-5 is used for handwriting recognition**

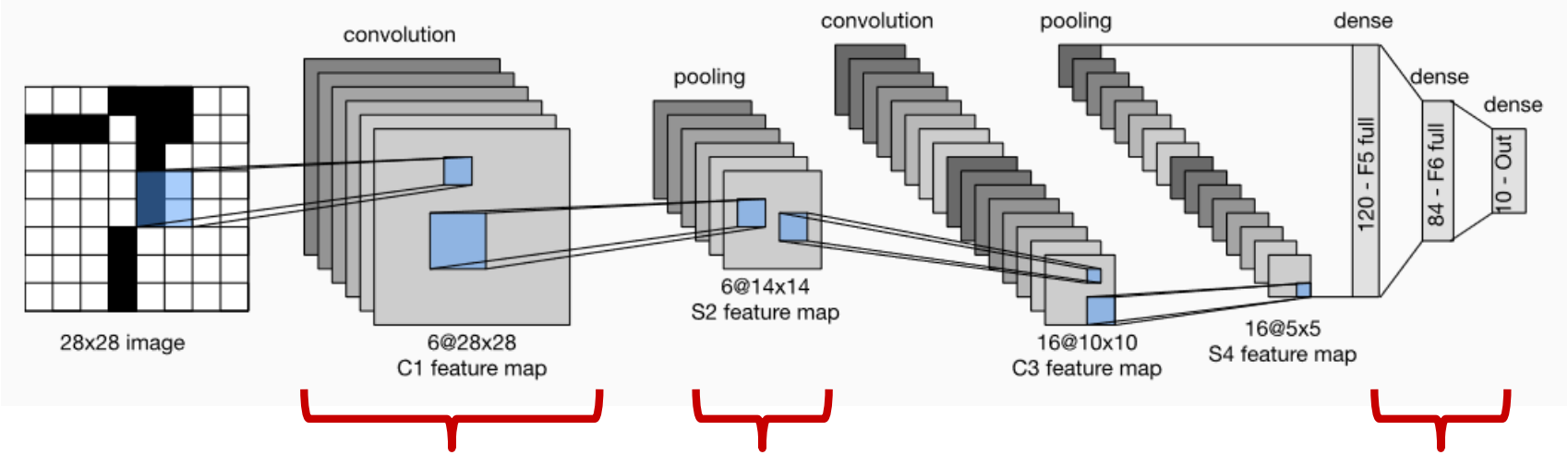


  - **Feature extraction identifies interesting features**

https://www.d2l.ai/chapter_convolutional-neural-networks/lenet.html

Your First Convolutional Neural Network

- **Lenet-5 is used for handwriting recognition**

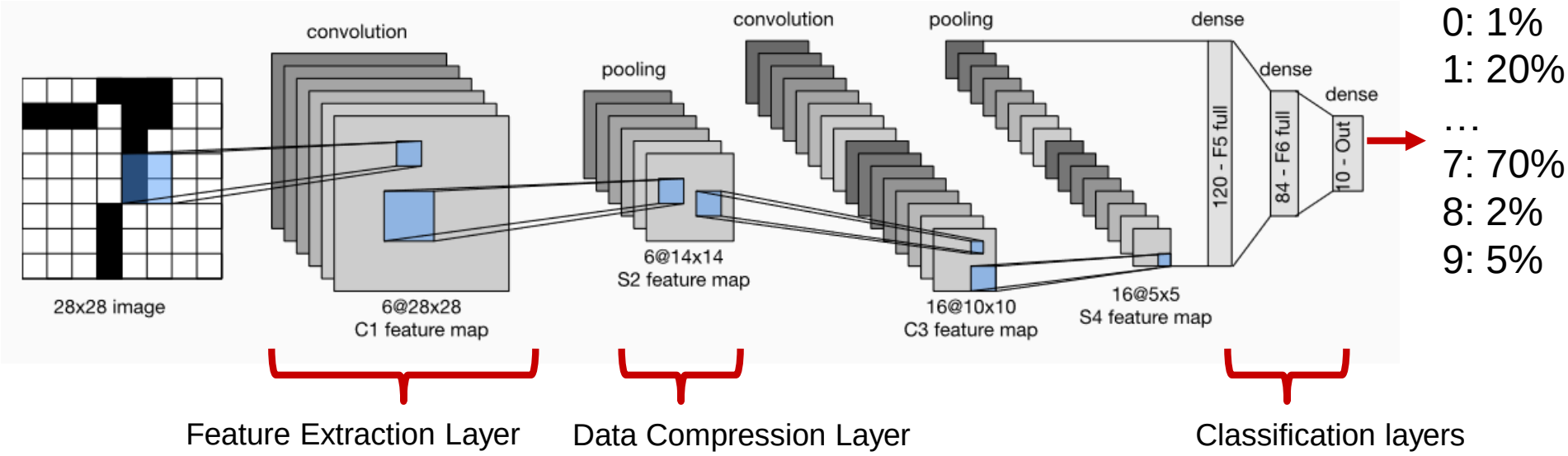


  - **Feature extraction identifies interesting features**
  - **Classification uses features to identify digit**

https://www.d2l.ai/chapter_convolutional-neural-networks/lenet.html

Your First Convolutional Neural Network

- **Lenet-5 is used for handwriting recognition**

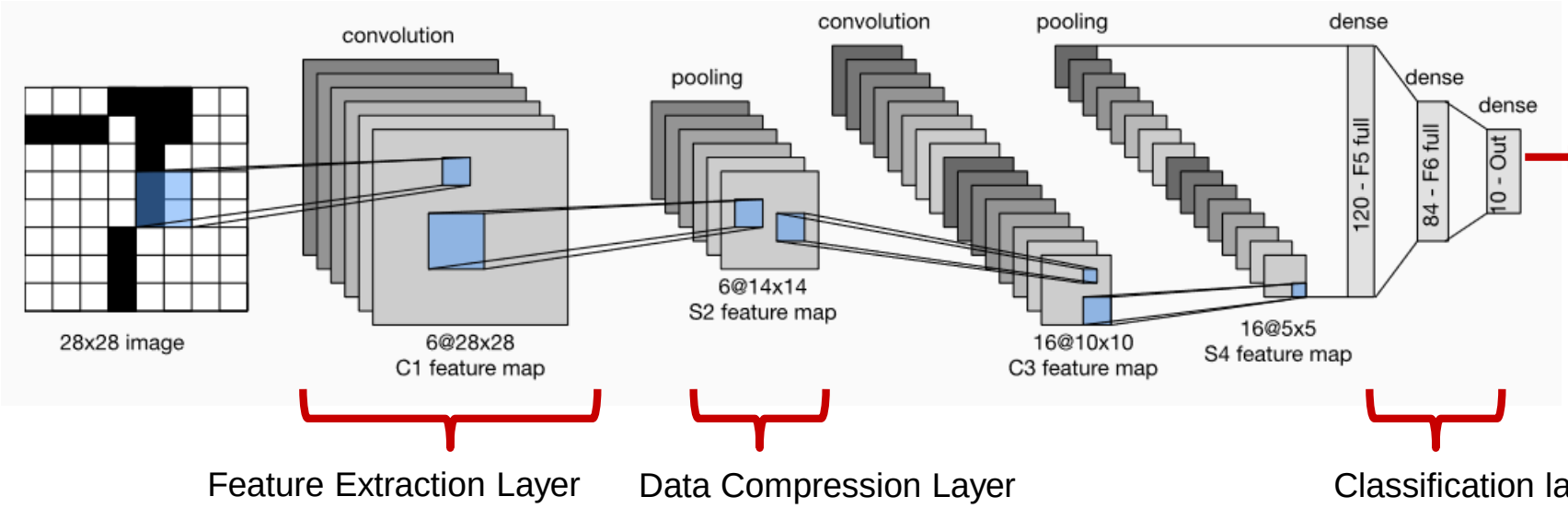


  - **Feature extraction identifies interesting features**
  - **Classification uses features to identify digit**

https://www.d2l.ai/chapter_convolutional-neural-networks/lenet.html

Your First Convolutional Neural Network

- **Lenet-5 is used for handwriting recognition**



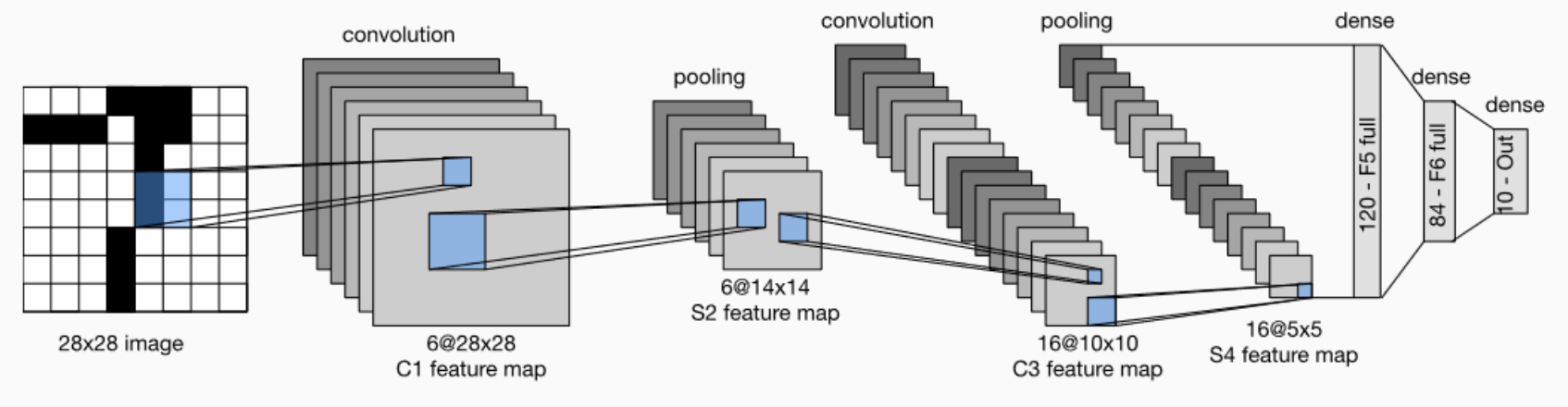
  - **Feature extraction identifies interesting features**
  - **Classification uses features to identify digit**

- **NNs are comprised of layers of neurons**
  - **Neurons (Yj) execute multiply-accumulate :** $Y_j = bias + \sum_{i=1}^{s} in_{i,j-1} \times w_{i,j}$

https://www.d2l.ai/chapter_convolutional-neural-networks/lenet.html

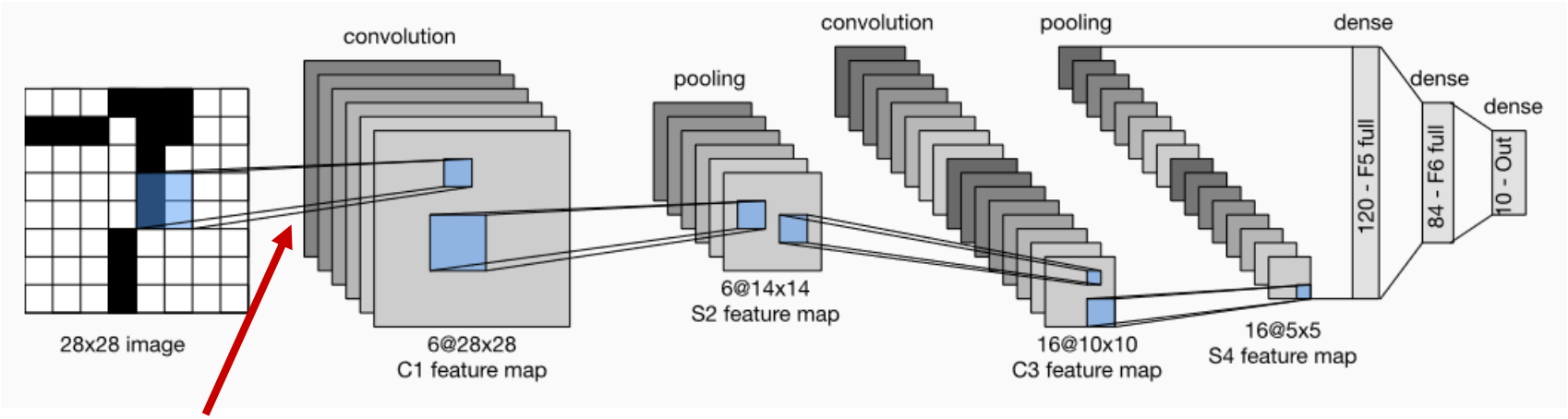
NNs Require Many Operations

- **How many MAC operations are needed?**



NNs Require Many Operations

- **How many MAC operations are needed?**

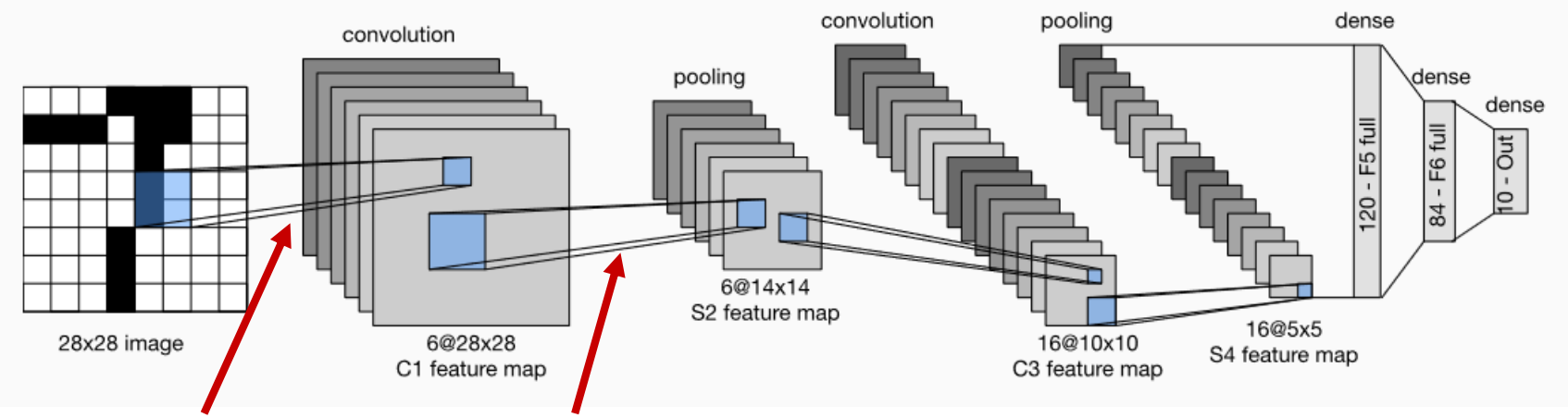


5x5 kernel =
25-element MAC
Convolution

C1 = 28x28x6x(25) = 117600

NNs Require Many Operations

- **How many MAC operations are needed?**



5x5 kernel = 25-element MAC

2x2 kernel = 4-element MAC

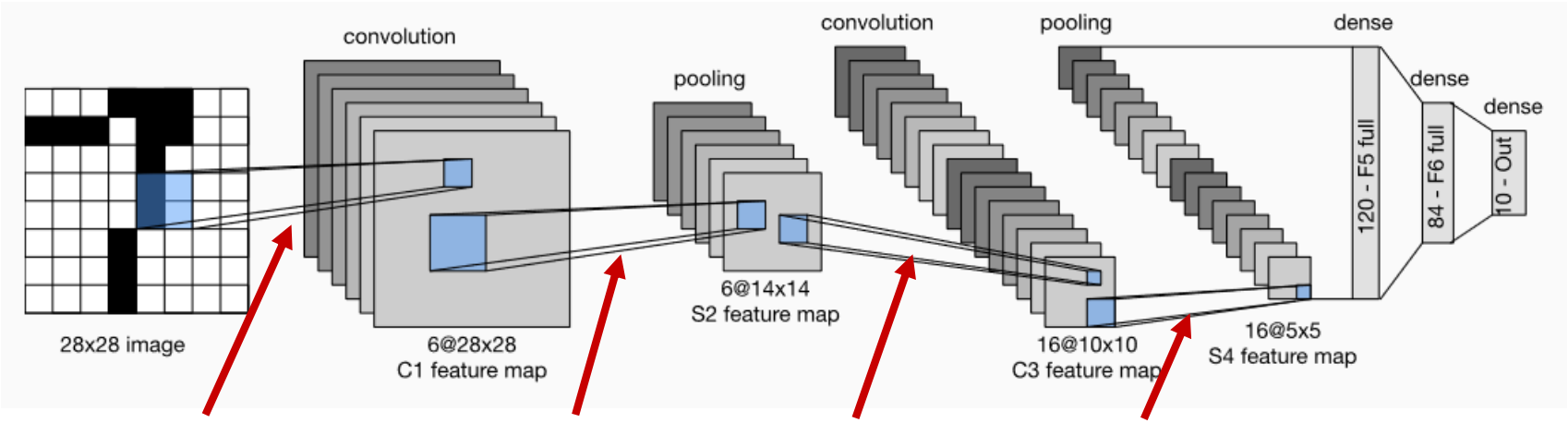
Convolution

C1 = 28x28x6x(25) = 117600

S2 = 14x14x6x(4) = 4704

NNs Require Many Operations

- **How many MAC operations are needed?**



5x5 kernel = 25-element MAC

2x2 kernel = 4-element MAC

5x5 kernel

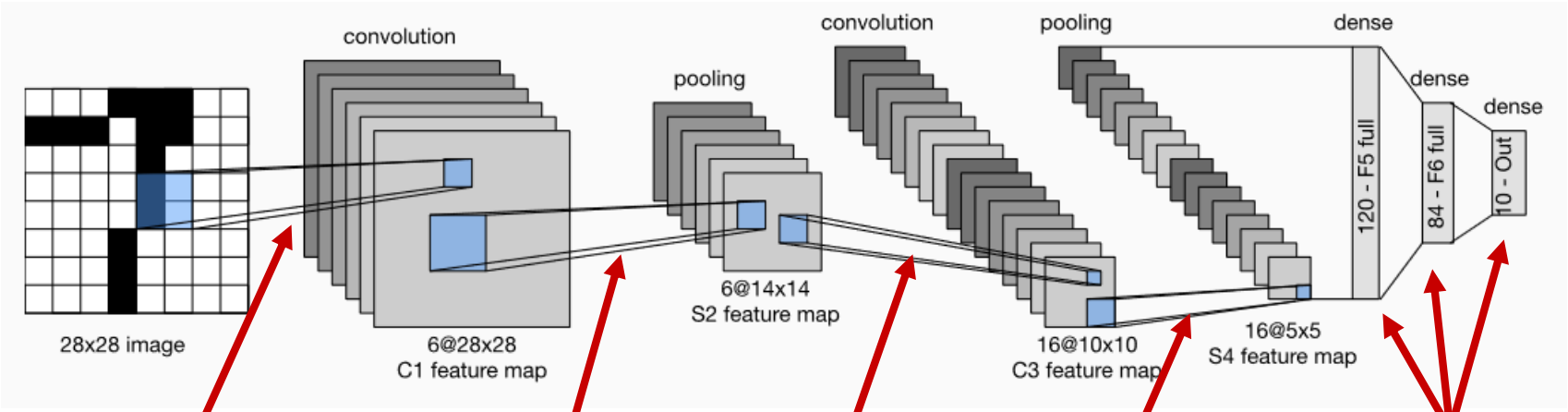
2x2 kernel

Convolution

C1 = 28x28x6x(25) = 117600
S2 = 14x14x6x(4) = 4704
C3 = 10x10x16x(25) = 40000
S4 = 1x1x16x(25) = 400

NNs Require Many Operations

- **How many MAC operations are needed?**



5x5 kernel = 25-element MAC Convolution

2x2 kernel = 4-element MAC

5x5 kernel

2x2 kernel

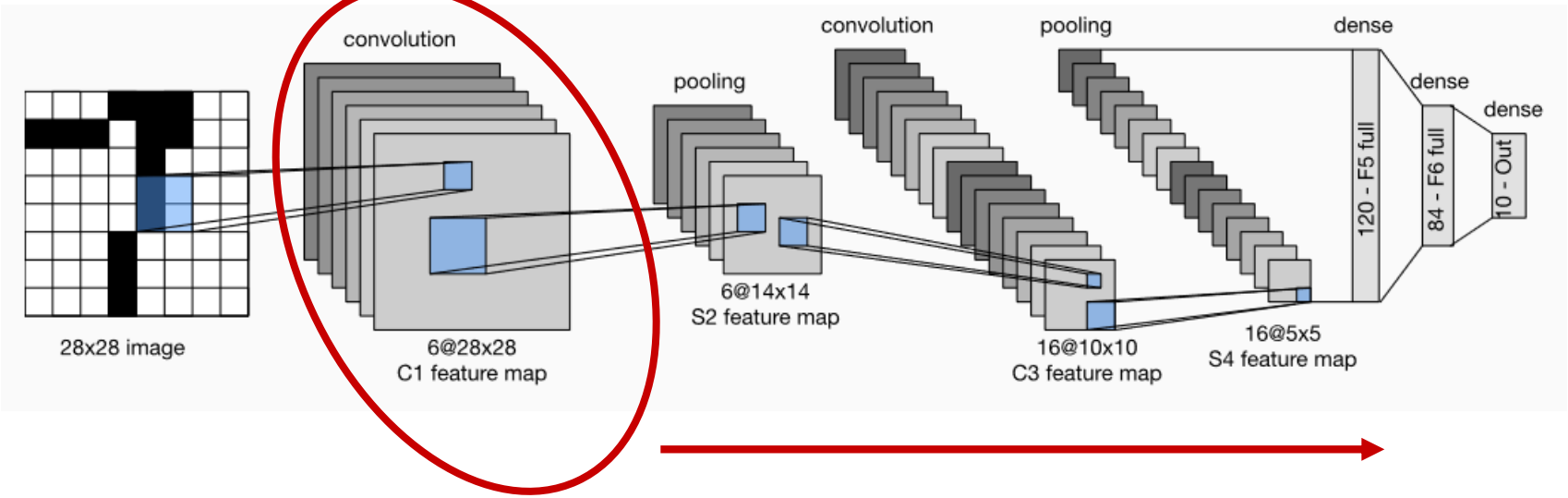
Input x Neurons Fully-Connected

C1 = 28x28x6x(25) = 117600
S2 = 14x14x6x(4) = 4704
C3 = 10x10x16x(25) = 40000
S4 = 1x1x16x(25) = 400
F5 = 120x(400) = 48000
F6 = 84x(120) = 10080
Out = 10x(84) = 840

= > 200k MACs

NNs Require Many Operations

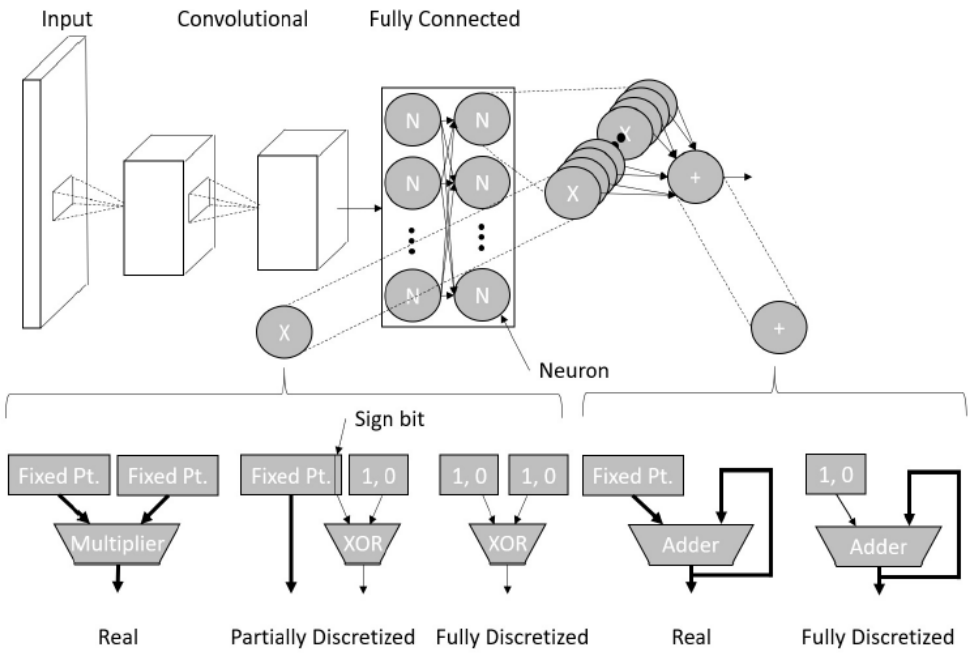
- **How many MAC operations are needed?**



- **Parallelize layers to reduce latency**
  - **Increase in hardware**

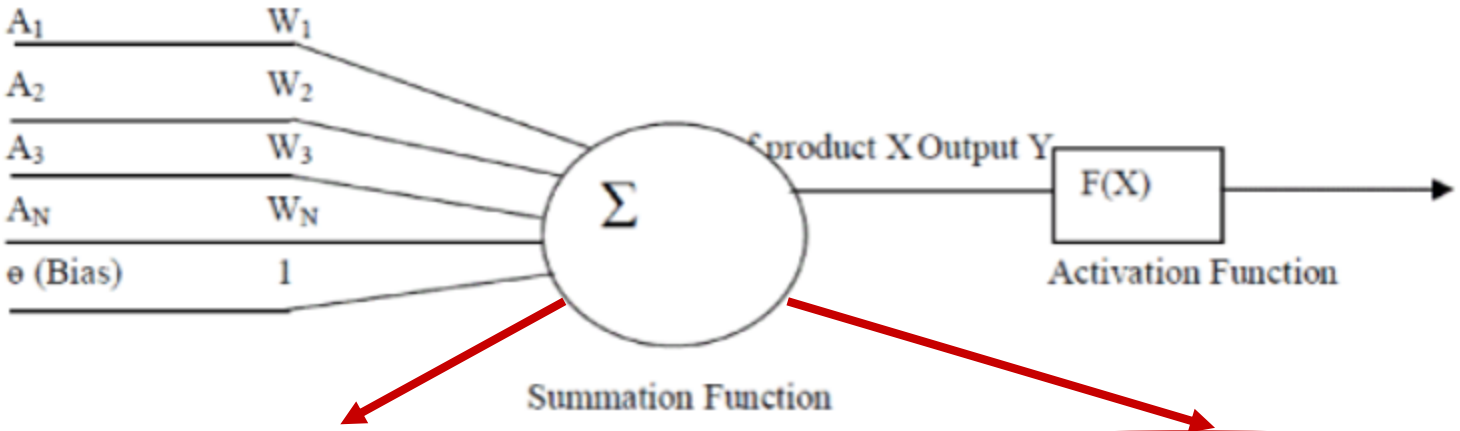
NN Architectural Optimizations

- **Discretization of the NN**
  - Partially discretized NN reduces weights
    - {-1,0,1} (Ternary Connect)
    - {-1,1} (Binary Connect)
  - Fully discretized NN reduces weights and I/O
    - {-1,0,1} (Ternarized NN)
    - {-1,1} (Binarized NN)
  - Reduces latency
    - Smaller HW footprint
    - Simple operations
      - » XOR, SUM
  - Sum called “popcount”
    - Population count
    - Smaller than accumulator

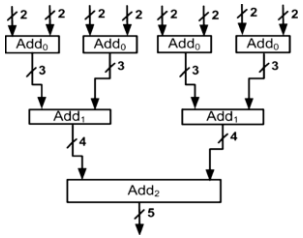


Accumulation Drives Latency

- **MAC consists of parallel multiplies and a summation**
  - **Latency of parallel multiplies = latency of one multiply**
  - **Latency of summation is a system bottleneck**

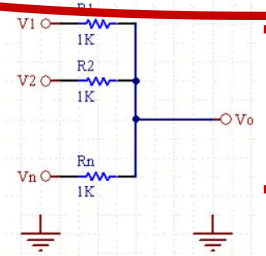


Digital VLSI



Latency = $(t_{add} \times (log_2(N_{inputs}) + 1))$

Analog Electronic (neuromorphic)

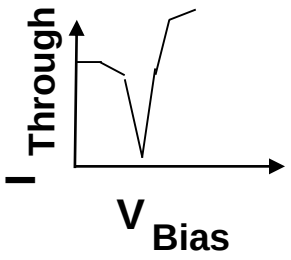


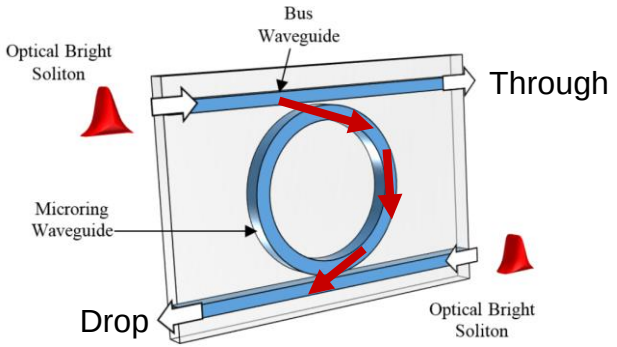
Resistors
Memristors
MOS transistors

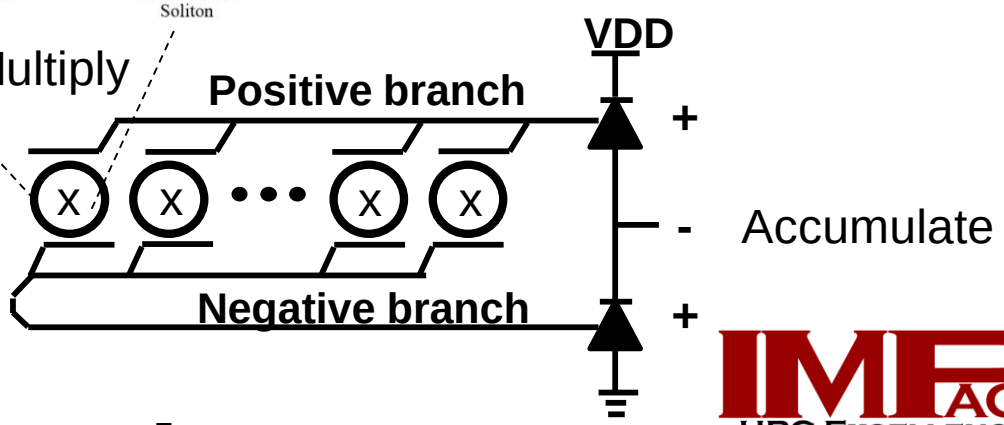
Latency = RC = τ

How to Improve on Analog Summation?

- **Micro-Ring Resonator (MRR) enables optical equivalent to analog electric summation**
  - **Latency not influenced by RC constant**
  - **Wavelength division multiplexing (WDM) enables parallel operation with no increase in hardware**

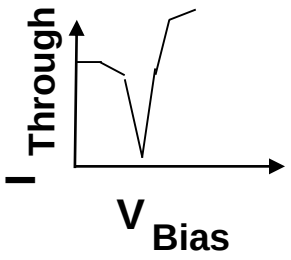


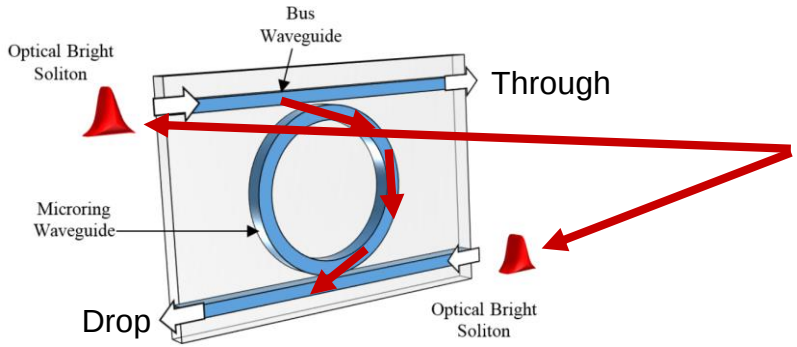




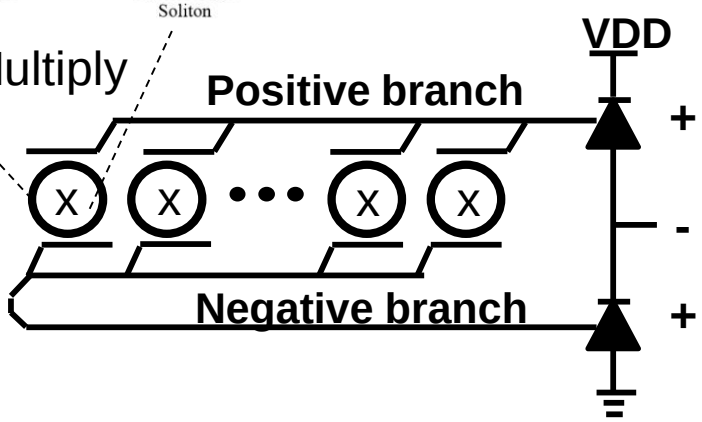
How to Improve on Analog Summation?

- **Micro-Ring Resonator (MRR) enables optical equivalent to analog electric summation**
  - **Latency not influenced by RC constant**
  - **Wavelength division multiplexing (WDM) enables parallel operation with no increase in hardware**



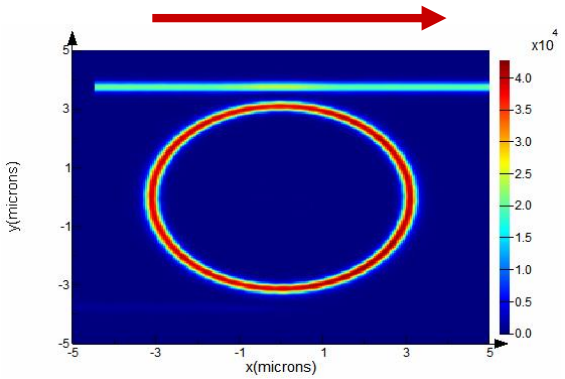


Single bit encoded to 2-bit MRR input

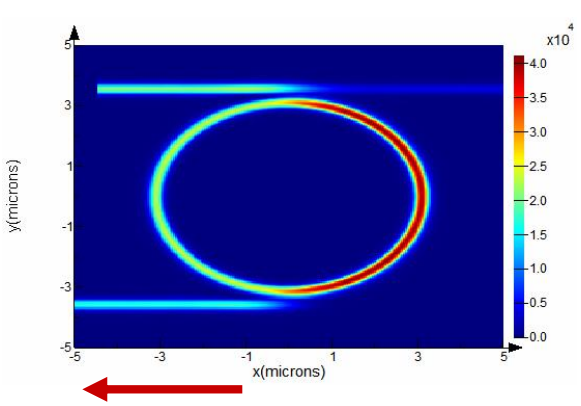


MRRs As Discretized Multiplier

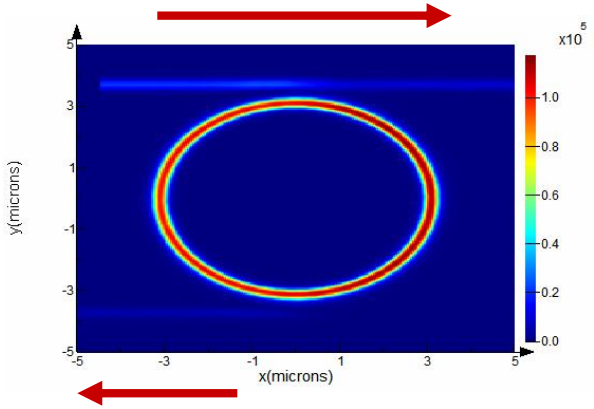
x1

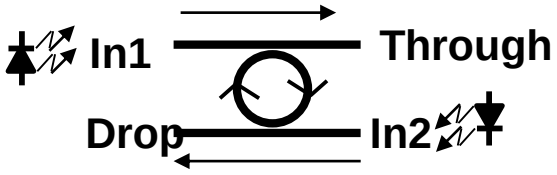


x -1



x0





Digital

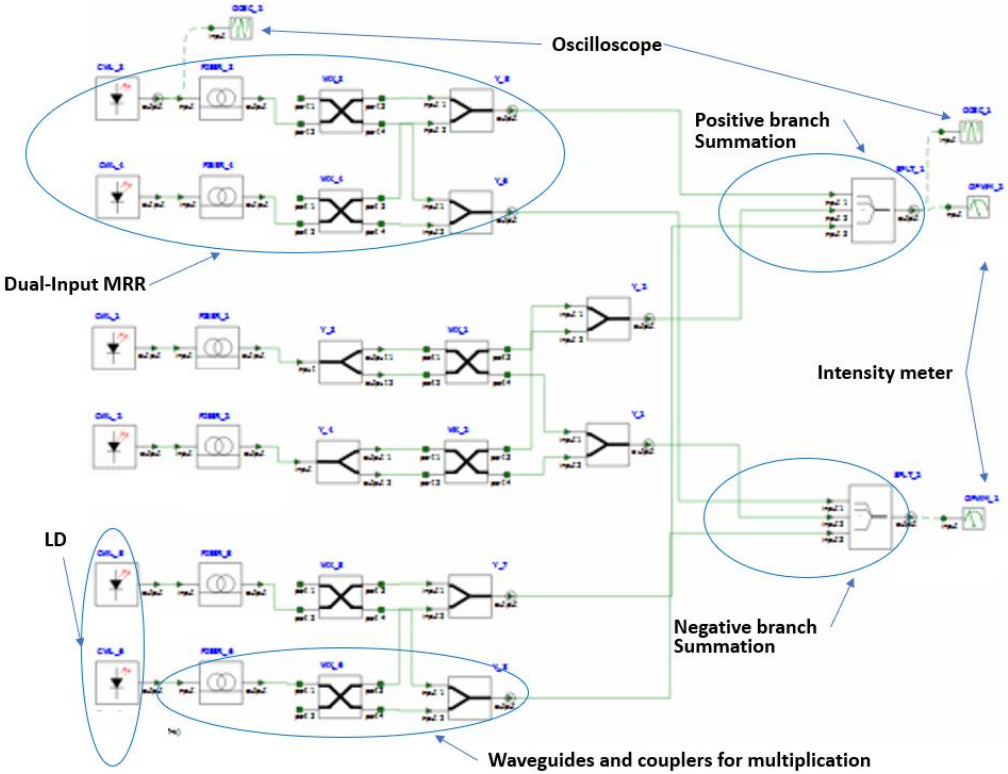
A	B	Y
0	0	0
0	1	0
0	-1	0
1	0	0
1	1	1
1	-1	-1
-1	0	0
-1	1	-1
-1	-1	1

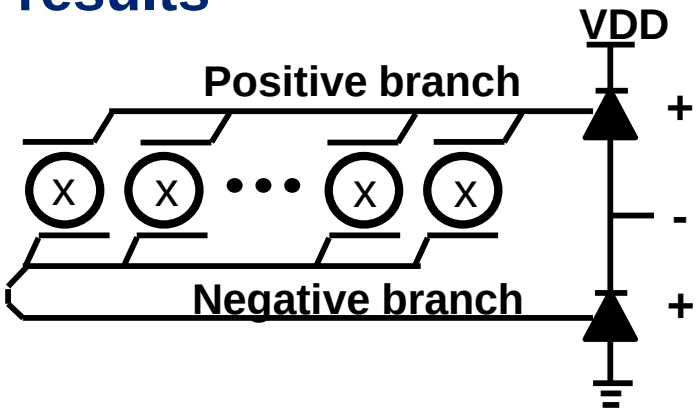
Photonic Encoded

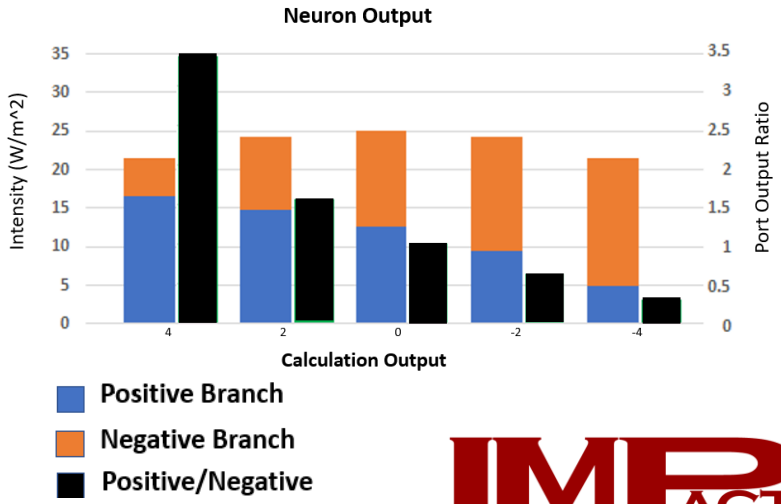
A		B			Y
in1 (%)	in2 (%)	bias	through (%)	drop (%)	through/drop
0	0	0	0	0	1
0	0	1	0	0	1
0	0	-1	0	0	1
100	0	0	50	50	1
100	0	1	100	0	> 1
100	0	-1	0	100	< 1
0	100	0	50	50	1
0	100	1	0	100	< 1
0	100	-1	100	0	> 1

Photonic Neuron

- **System Simulations using INTERCONNECT.**
  - **S-parameters derived from MODE results**

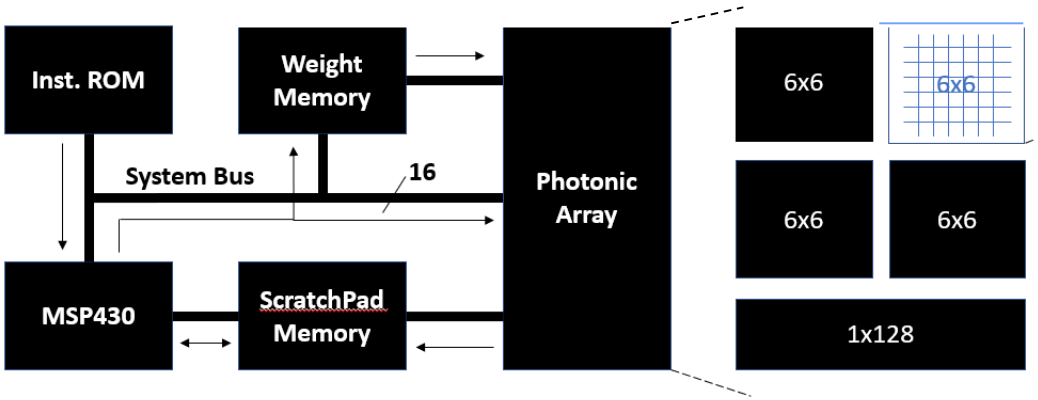






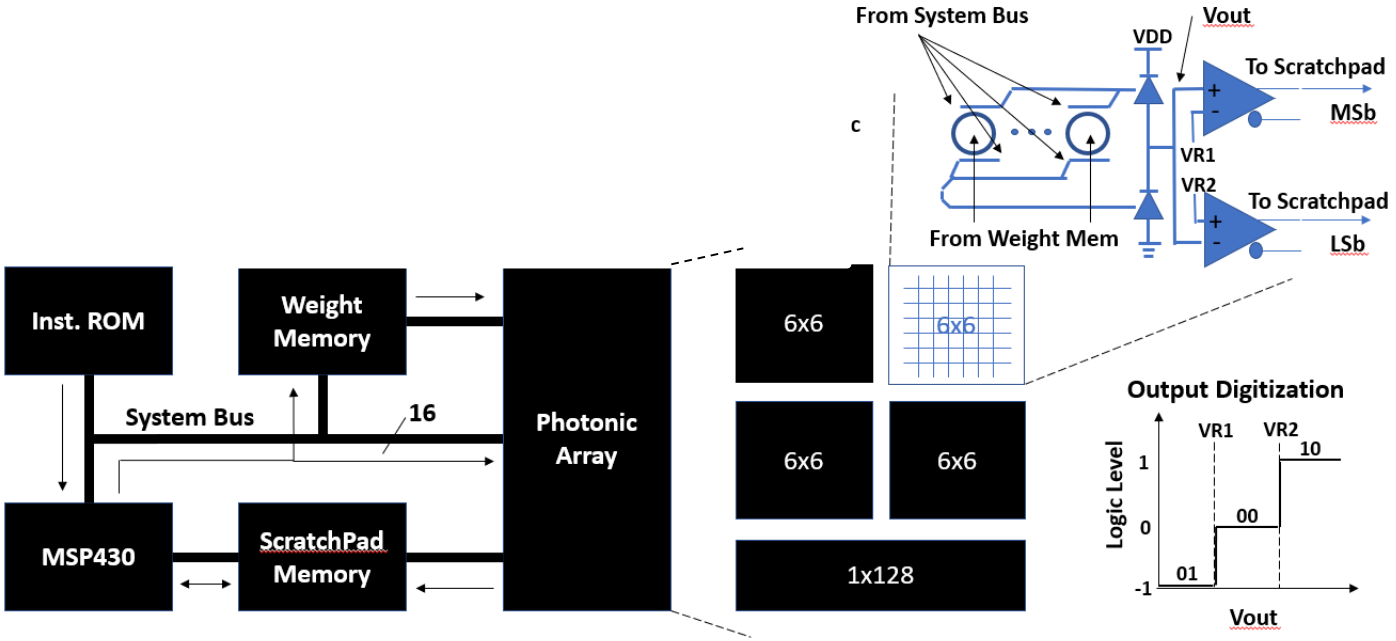
Photonic NN Processor Architecture

- **Loosely-coupled architecture controlled by MSP430**



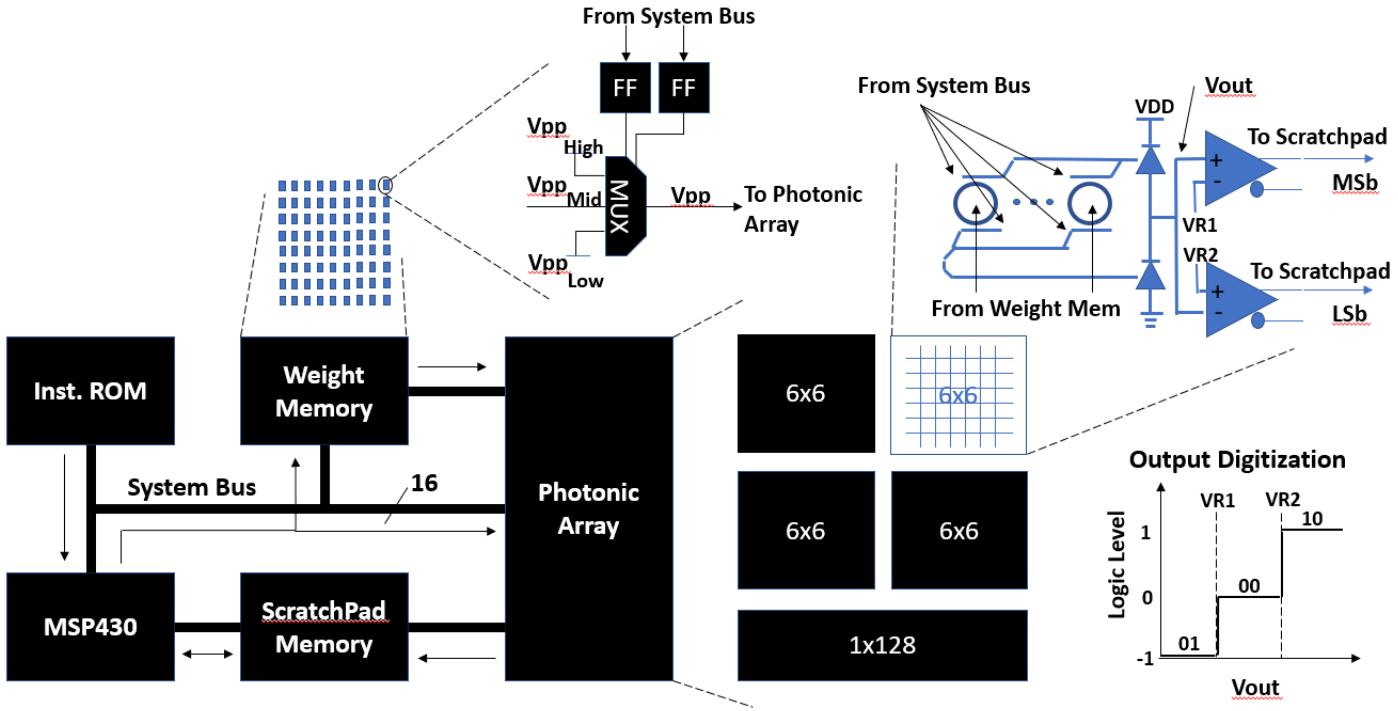
Photonic NN Processor Architecture

- **Loosely-coupled architecture controlled by MSP430**



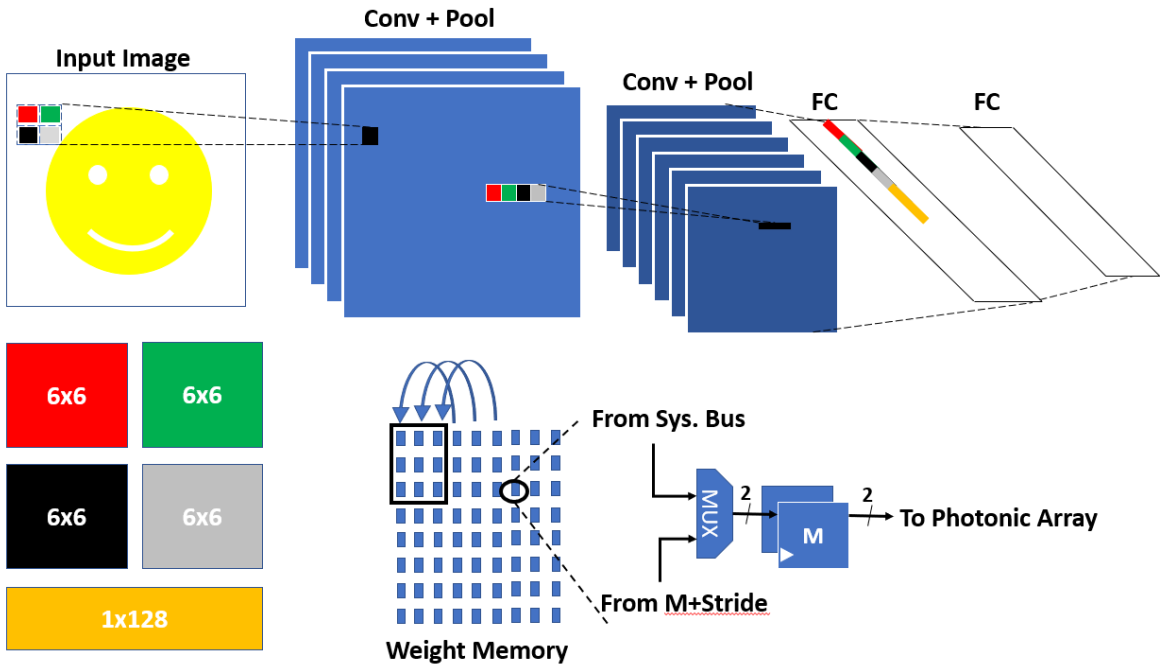
Photonic NN Processor Architecture

- **Loosely-coupled architecture controlled by MSP430**
  - **Weight memory selects Vref to bias MRR**
  - **Discretized values enable HW reduction**
    - **Reduces latency since analog MUX selects voltage**



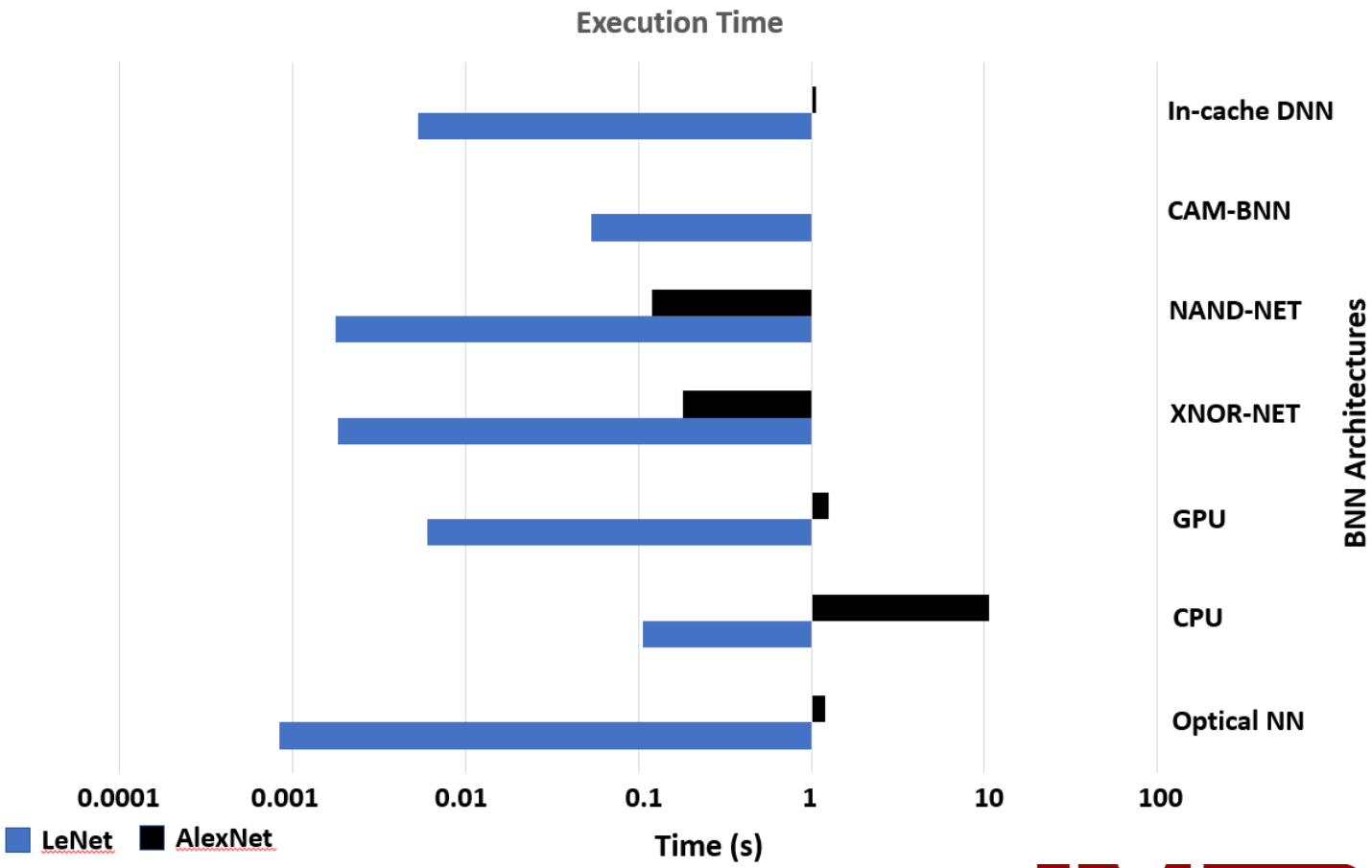
Photonic Computation of a CNN

- **Partial unrolling of convolutional layer reduces hardware requirements with minimal performance impact**
  - **Shift register enables emulation of dragging window**



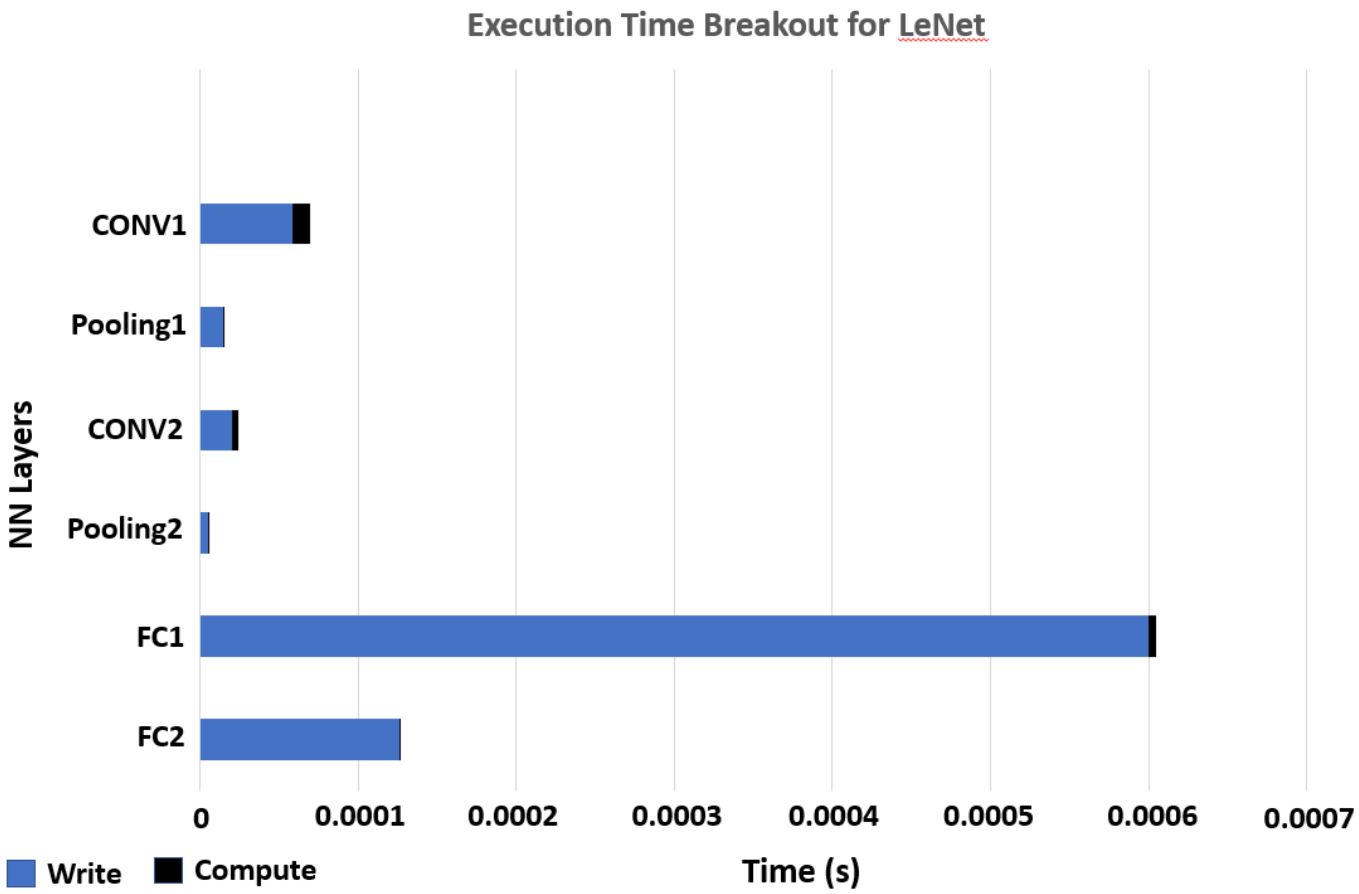
LeNet and AlexNet Simulation Results

- AlexNet performance lagged due to network size
- Larger NN processor reduces latency



LeNet Layer Analysis

- **Execution time is write-dominated**



Questions?

Efficiency Analysis

- **Due to WDM, large optical components rival performance per area of CMOS counterparts**