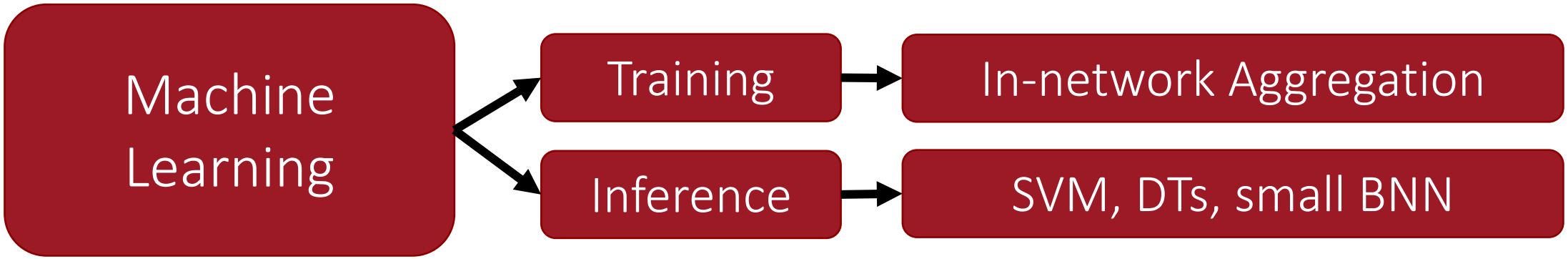
# Distributed DNN Serving in the Network Data Plane

**Kamran Razavi**†, George Karlos‡, Vinod Nigade‡, Max Mühlhäuser†, Lin Wang†‡

†University of Darmstadt

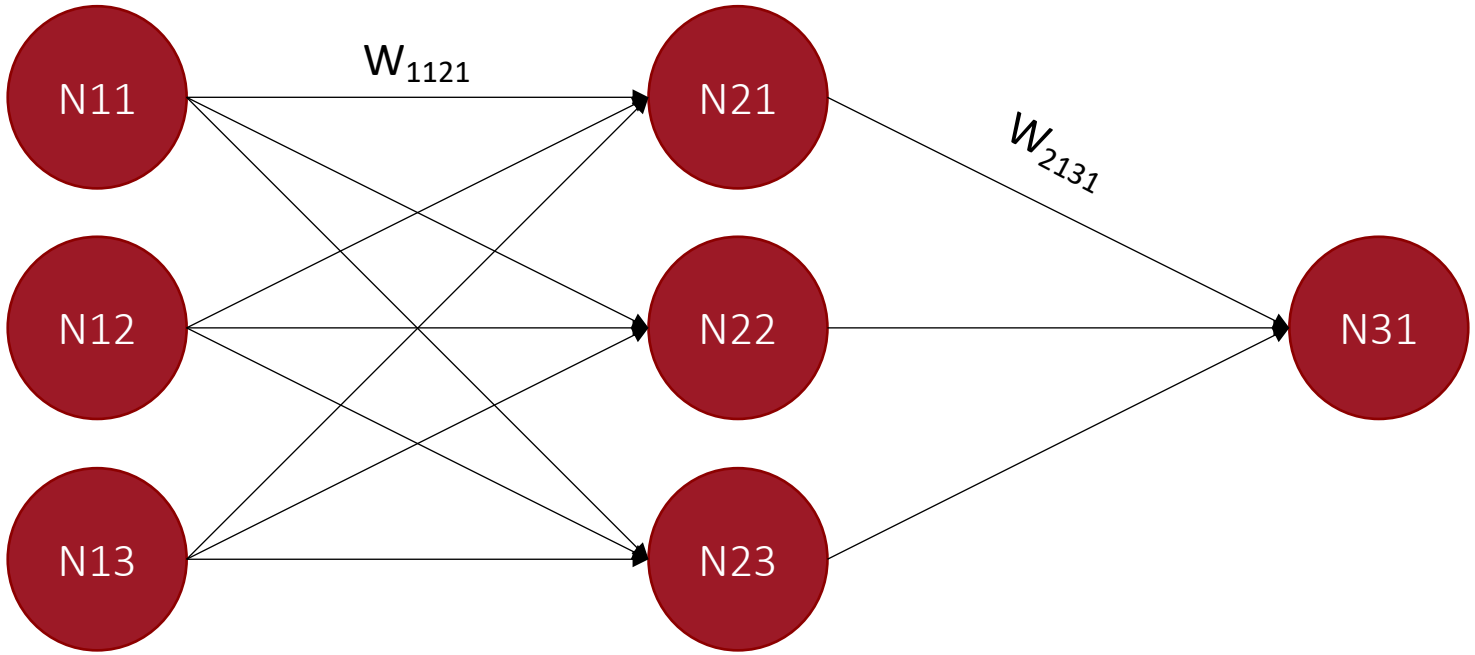
‡ Vrije Universiteit Amsterdam

# Programmable Network

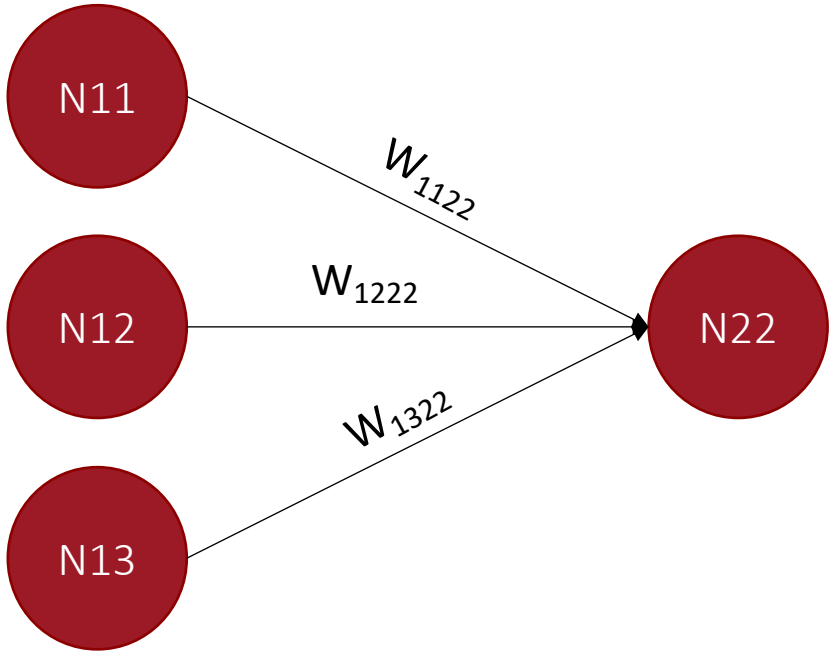



*Can we leverage a programmable network to perform DNN serving?*

# Deep Neural Network Inference

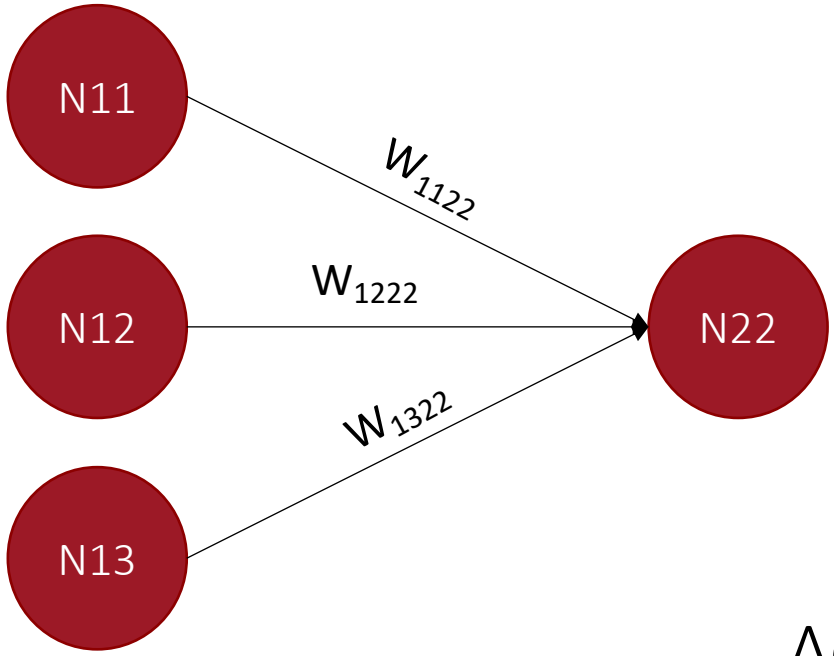


# Deep Neural Network Inference




$$N_{22} = N_{11} \times W_{1122} + N_{12} \times W_{1222} + N_{13} \times W_{1322}$$

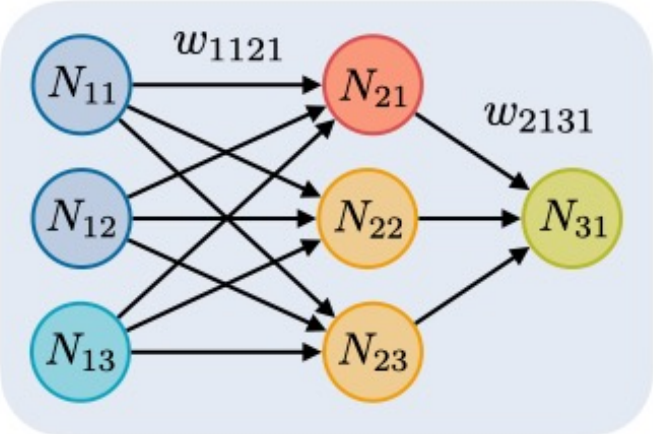
# Deep Neural Network Inference




$$N_{22} = F(N_{11} \times W_{1122} + N_{12} \times W_{1222} + N_{13} \times W_{1322})$$

Activation Function

# DNN Serving in the Network



# On-Switch Execution

## No FPU: 1.1*2 is time consuming!

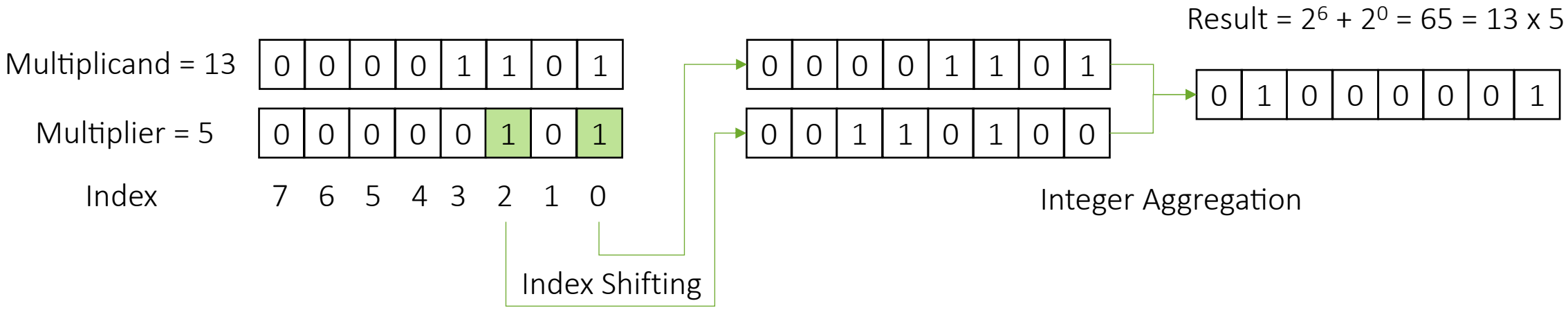
- Int-8 (first bit is the sign bit)
- Parallel multiplication

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Multiplicand = 13 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| Multiplier = 5 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |

# On-Switch Execution

## No FPU: 1.1*2 is time consuming!

- Int-8 (first bit is the sign bit)
- Parallel multiplication




8 ALUs per multiplication -> # of parallel multiplication ~= N/8

# On-Switch Execution

## No FPU: 1.1*2 is time consuming!

- Int-8 (first bit is the sign bit)
- Parallel multiplication

## Activation functions

For example, ReLU:

- If the sign bit == 1, put 0
- Else keep the number

## Layer specific requirements: e.g. MaxPooling

- State storage
- Order of packets

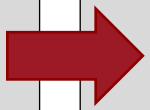
# Case Study: Mini-AlexNet

7 Layers

15M Parameters

91M Operations

16MB Memory



2ms Latency

| | mini-AlexNet (CIFAR-10) |
|---|---|
| **Layer 1** | Input: $32 \times 32 \times 3$ |
| **Layer 2** | Conv1: $3 \times 3 \times 3 \times 64$ |
| **Layer 3** | Conv2: $3 \times 3 \times 64 \times 192$ |
| **Layer 4** | Conv3: $3 \times 3 \times 192 \times 384$ |
| **Layer 5** | FC1: 4096 |
| **Layer 6** | FC2: 2048 |
| **Layer 7** | FC3 (Output): 10 |