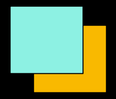
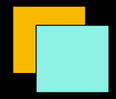
# Introduction to Systolic Array

12/21/2021



NTHU Logic Design Laboratory (Fall. 2021)



By Prof. Chun-Yi Lee

# Agenda

Introduction

Architecture

Example

# Agenda

Introduction

Architecture

Example

# Introduction (1/2)

- Many of your final projects involves DNN inferencing
- As you may know the inference of a linear layer in a DNN can be expressed by matrix multiplication.
- How?
- How about convolutional layers?

# Introduction (2/2)

- Serial computation of matrix multiplication has O(n^3) complexity!
- How about breaking it down to many vector inner product and parallelize it?
- The degree of parallization?
- Bottleneck is the memory bandwidth
- Useful data are thrown away!

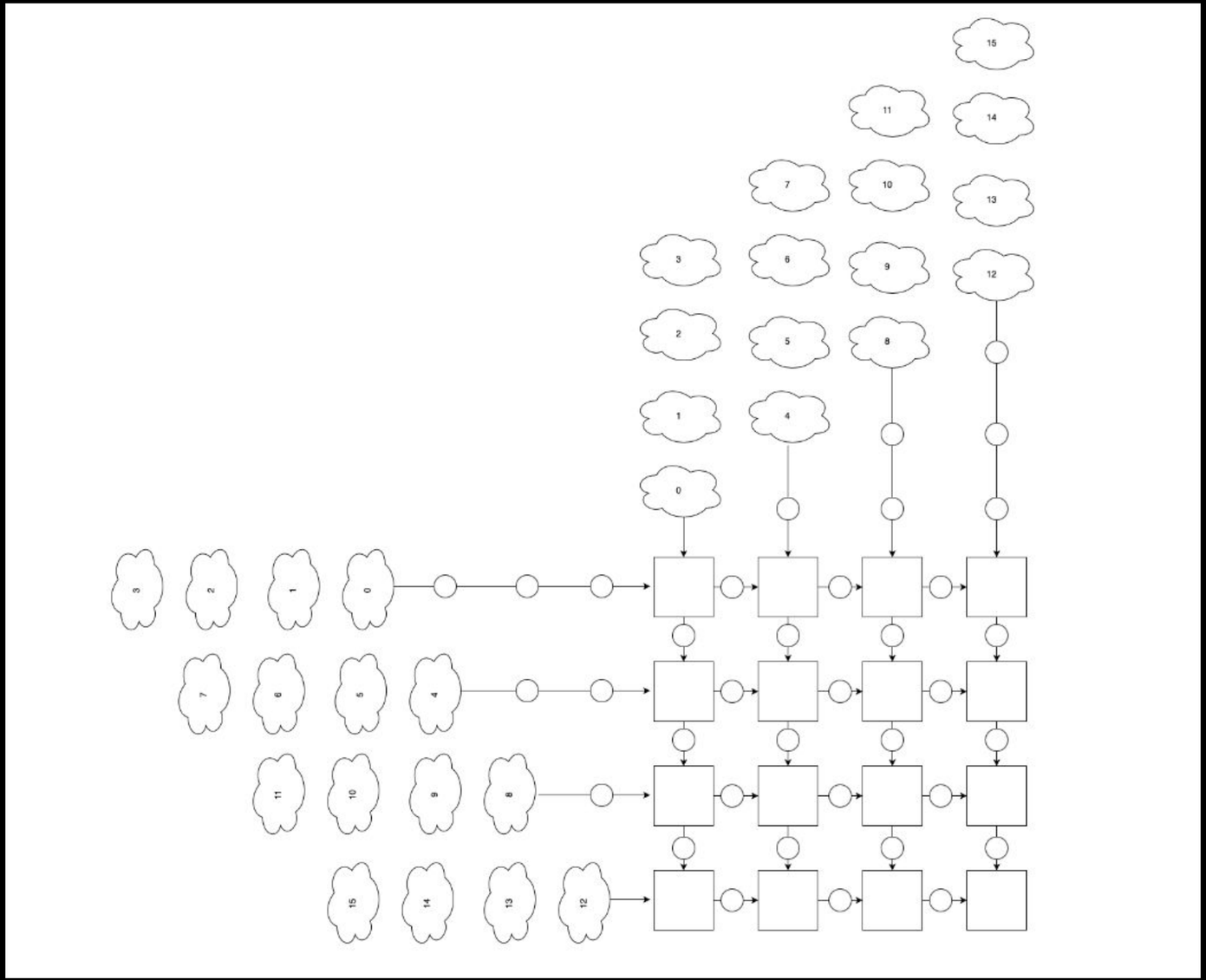
# Agenda

Introduction

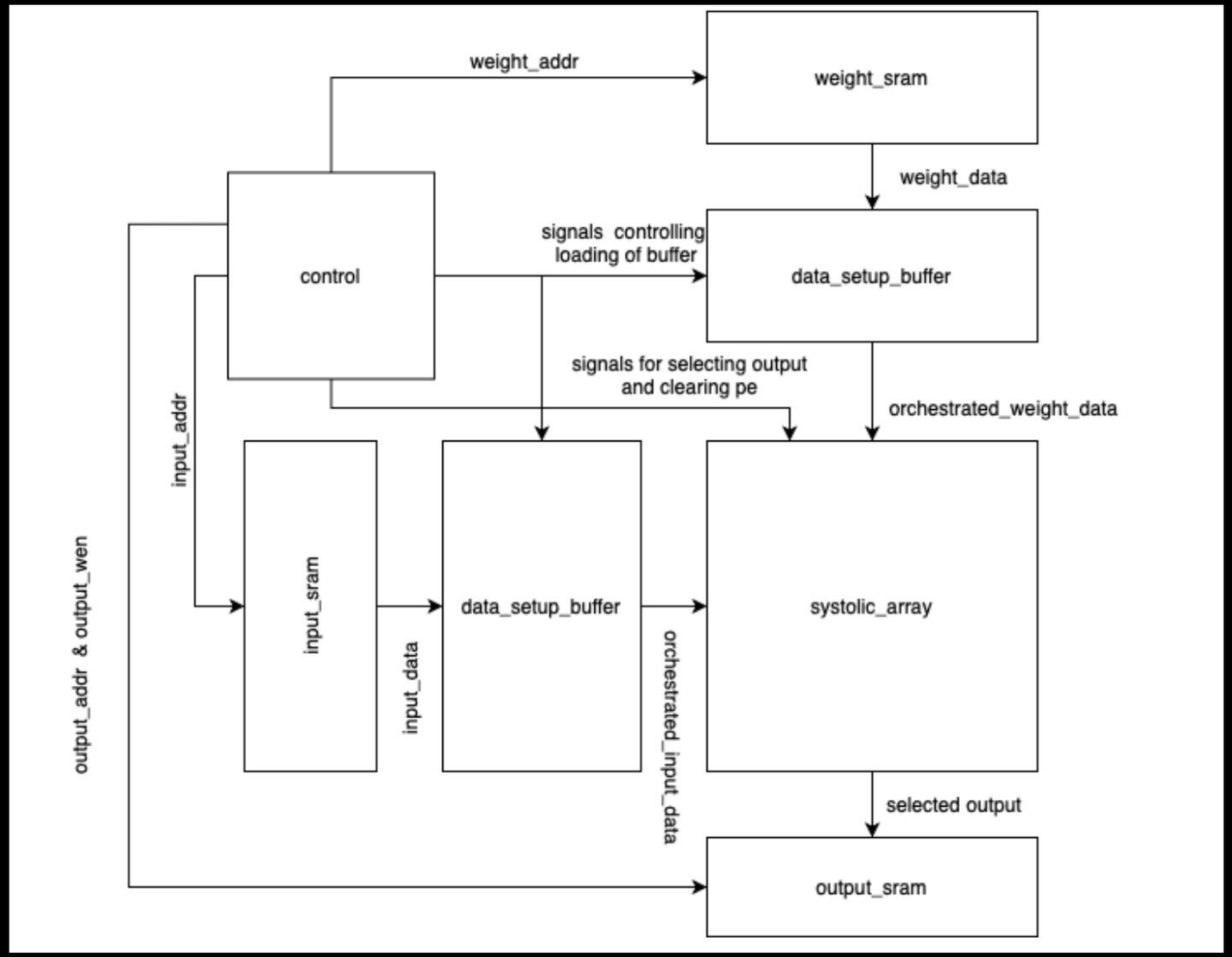
Architecture

Example

# Architecture (1/2)



# Architecture (2/2)



# Agenda

Introduction

Architecture

Example

# Example (1/1)

- Google TPU
- NVIDIA Tensor Core
- And many more …

# Reference

- Prof. Chih-Tsun Huang
- Prof. Yong-Long Lin
  https://www.youtube.com/watch?v=6PFgOMHAUo0&t=83s
- And myself