UCLA Samueli Computer Science

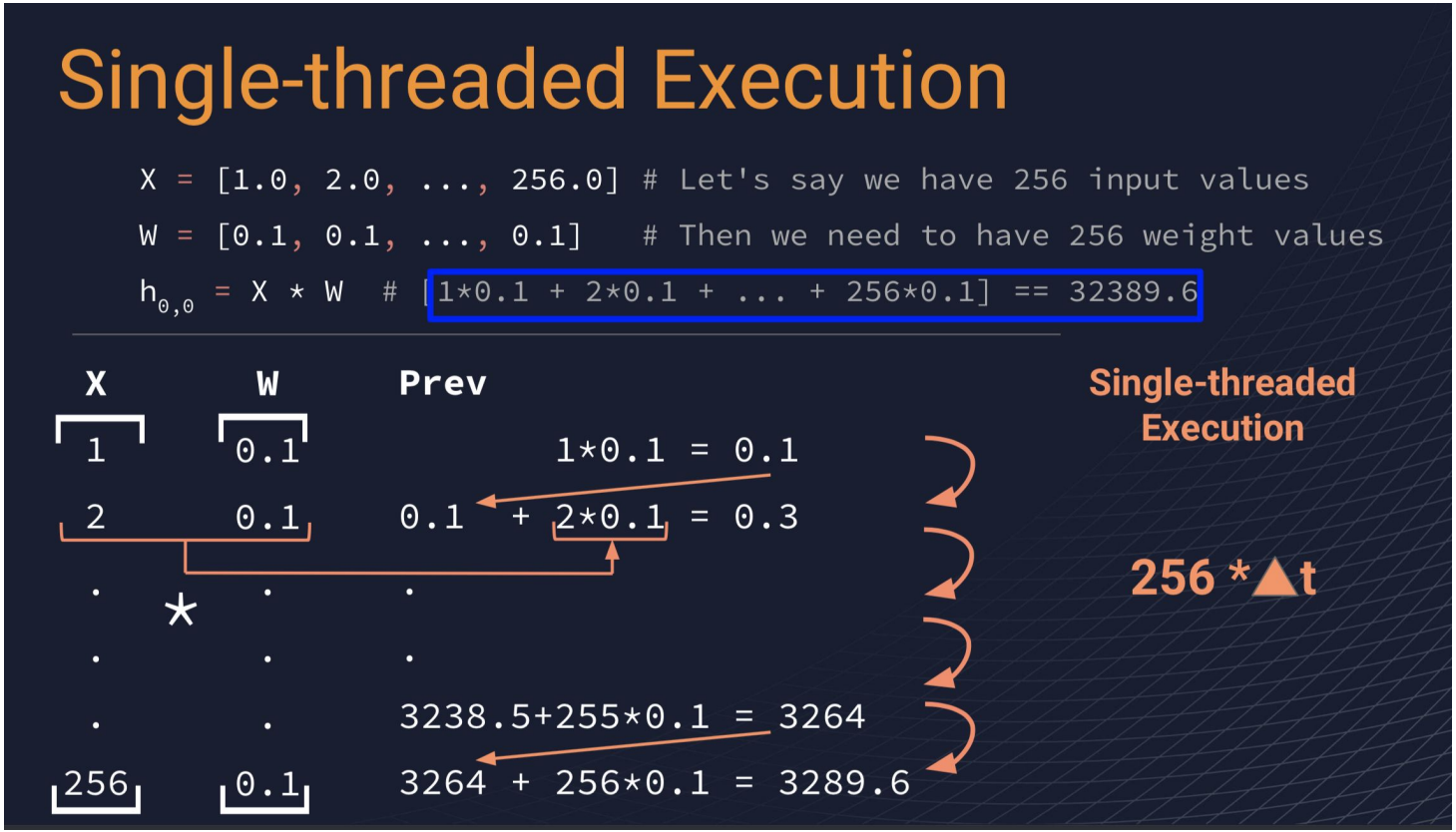
CS145 Discussion: Week 6 (Add-On)

A Story of Computing

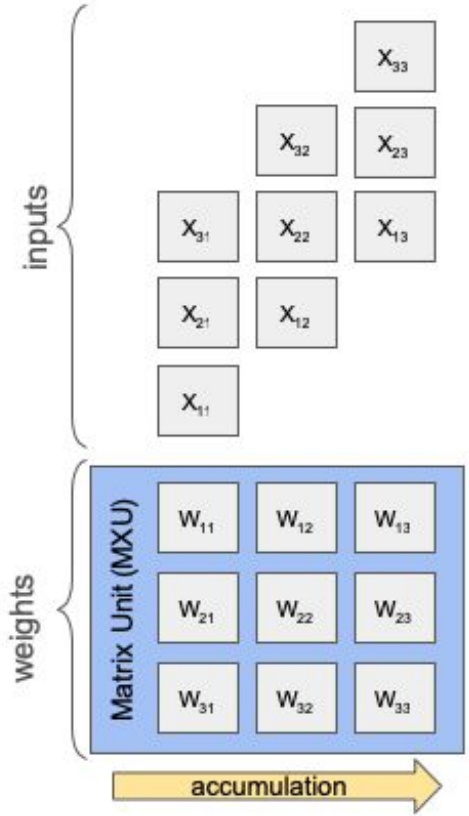
Junheng Hao

Friday, 11/13/2020

Single-thread Computing of *X*W*



UCLA Neural Networks: Computation Example

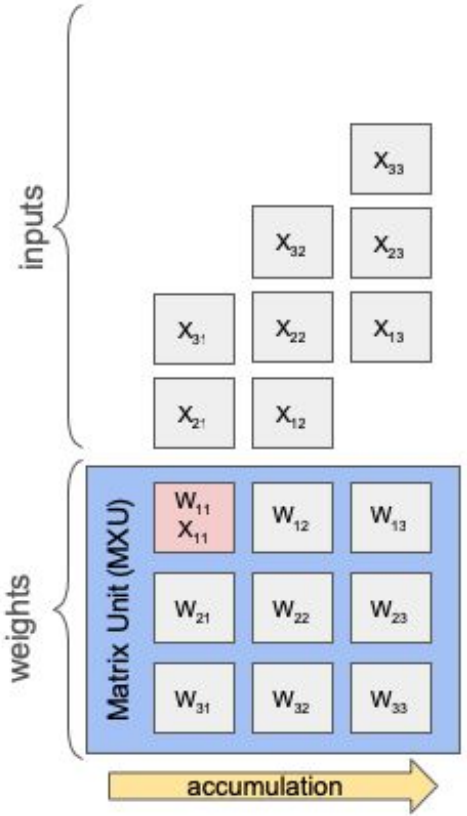


Matrix Unit Systolic Array

Computing $y = Wx$

3x3 systolic array
W = 3x3 matrix
Batch-size(x) = 3

UCLA Neural Networks: Computation Example

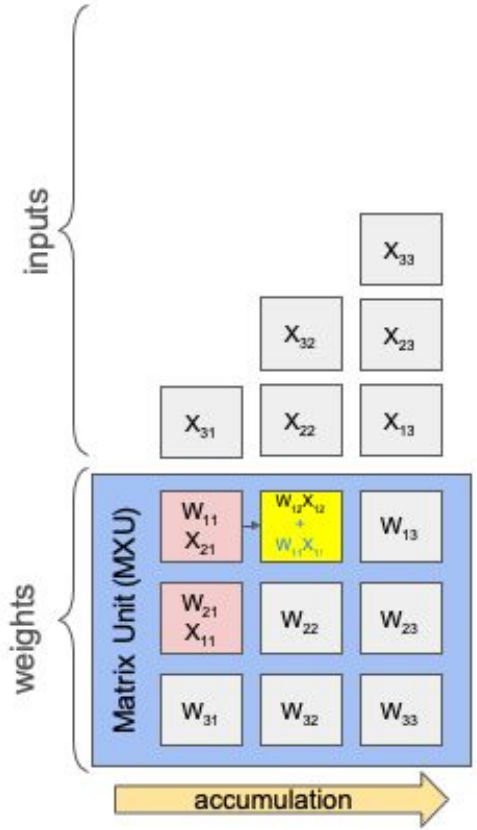


Matrix Unit Systolic Array

Computing $y = Wx$

with $W = 3x3$, batch-size$(x) = 3$

UCLA Neural Networks: Computation Example

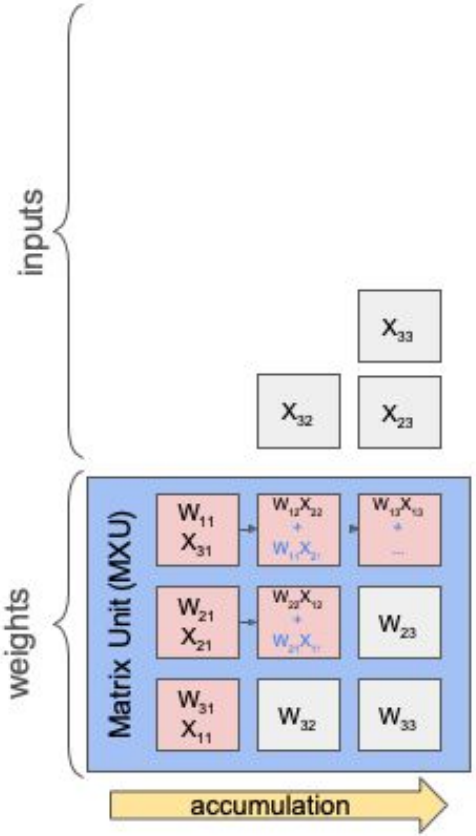


Matrix Unit Systolic Array

Computing $y = Wx$

with $W = 3x3$, batch-size(x) = 3

Neural Networks: Computation Example

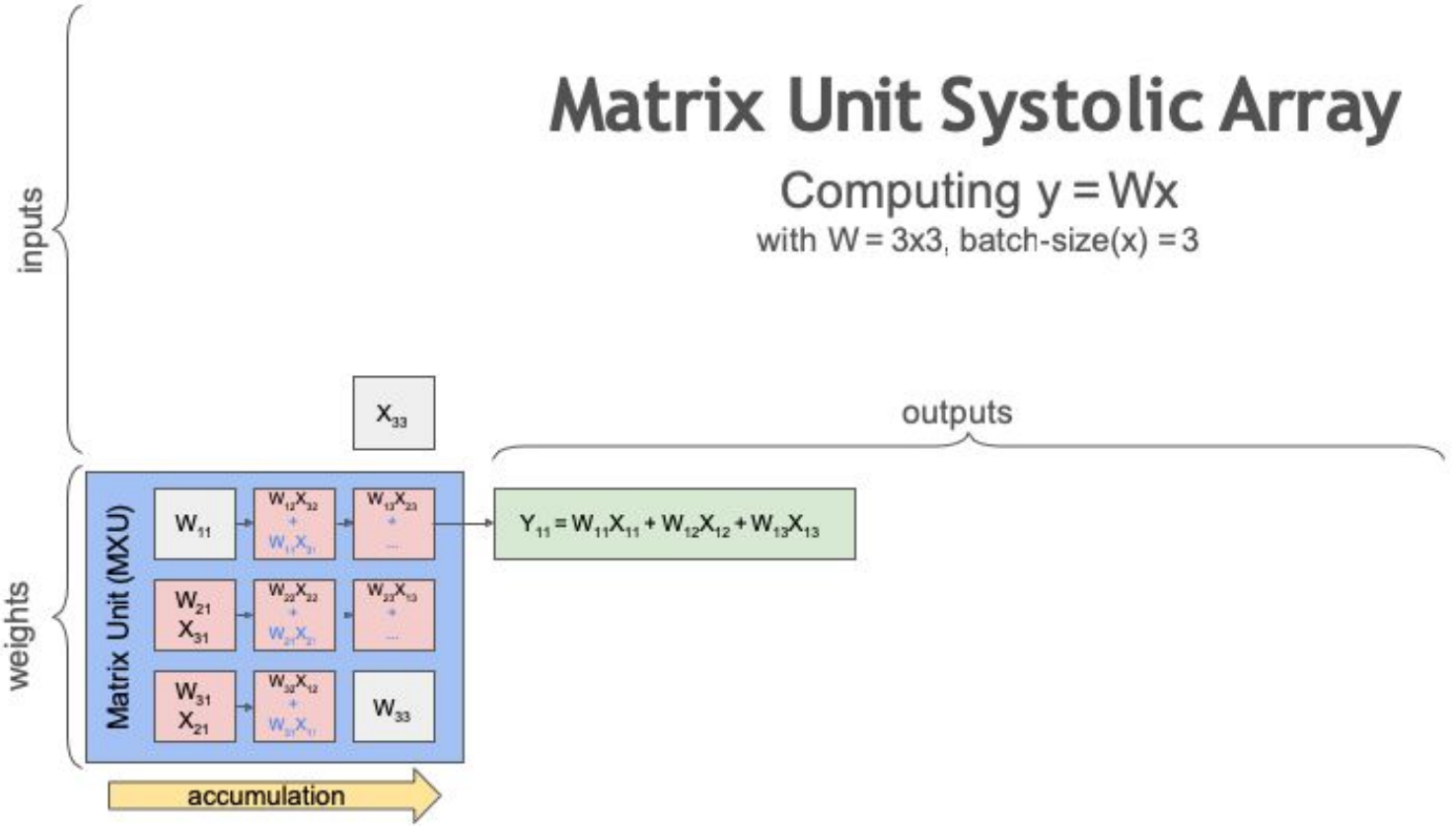


Matrix Unit Systolic Array

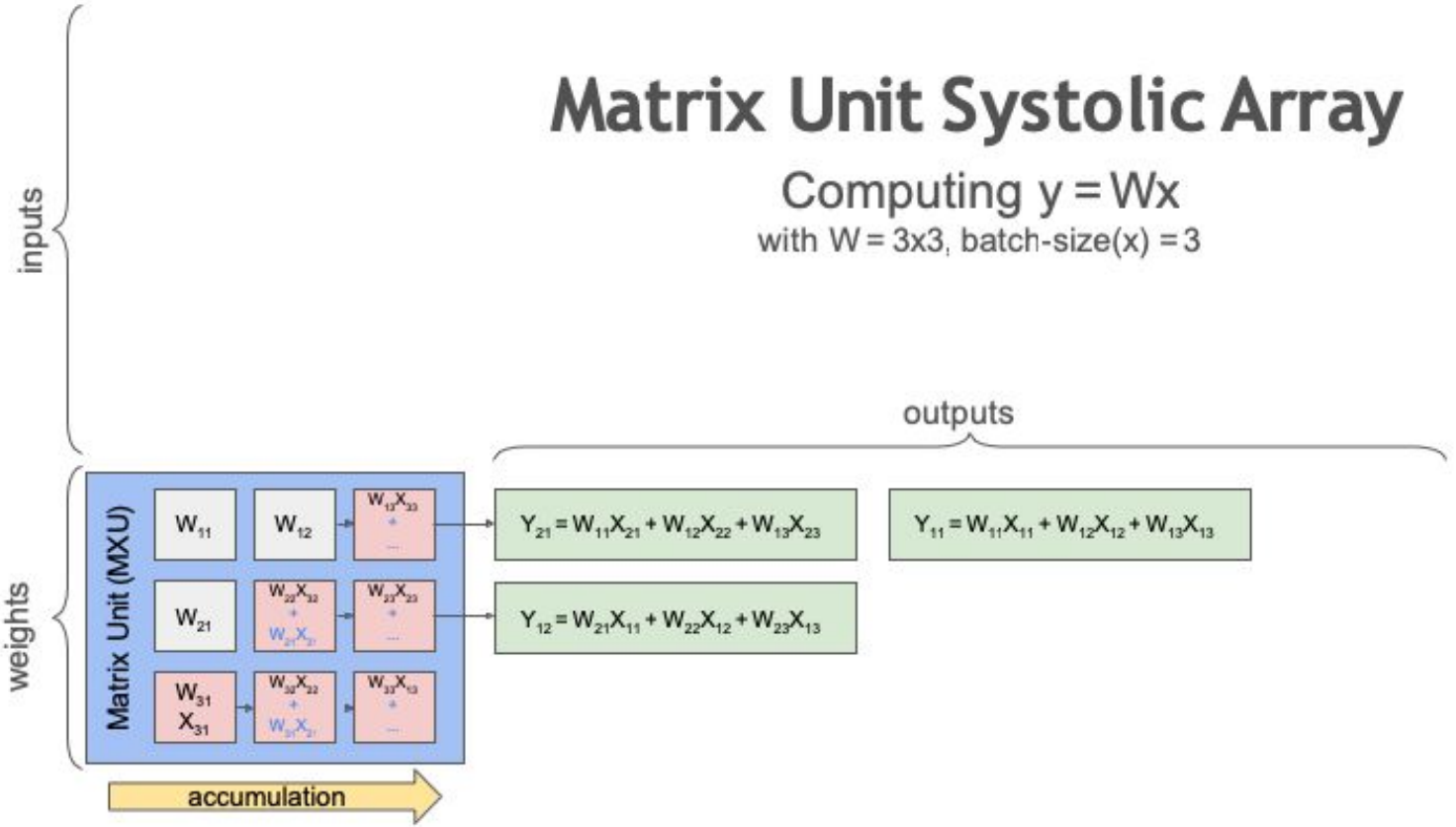
Computing $y = Wx$

with $W = 3x3$, batch-size(x) = 3

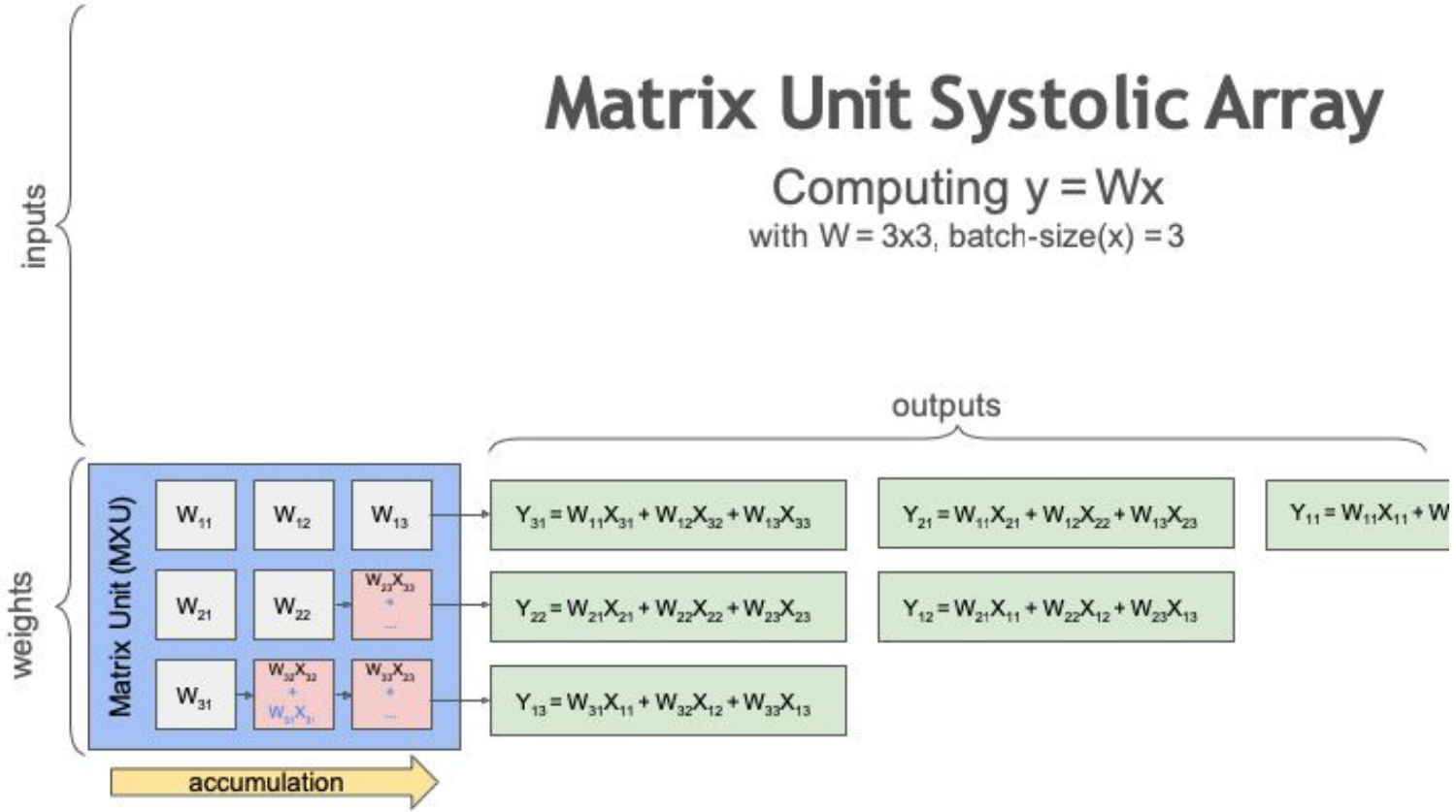
UCLA Neural Networks: Computation Example



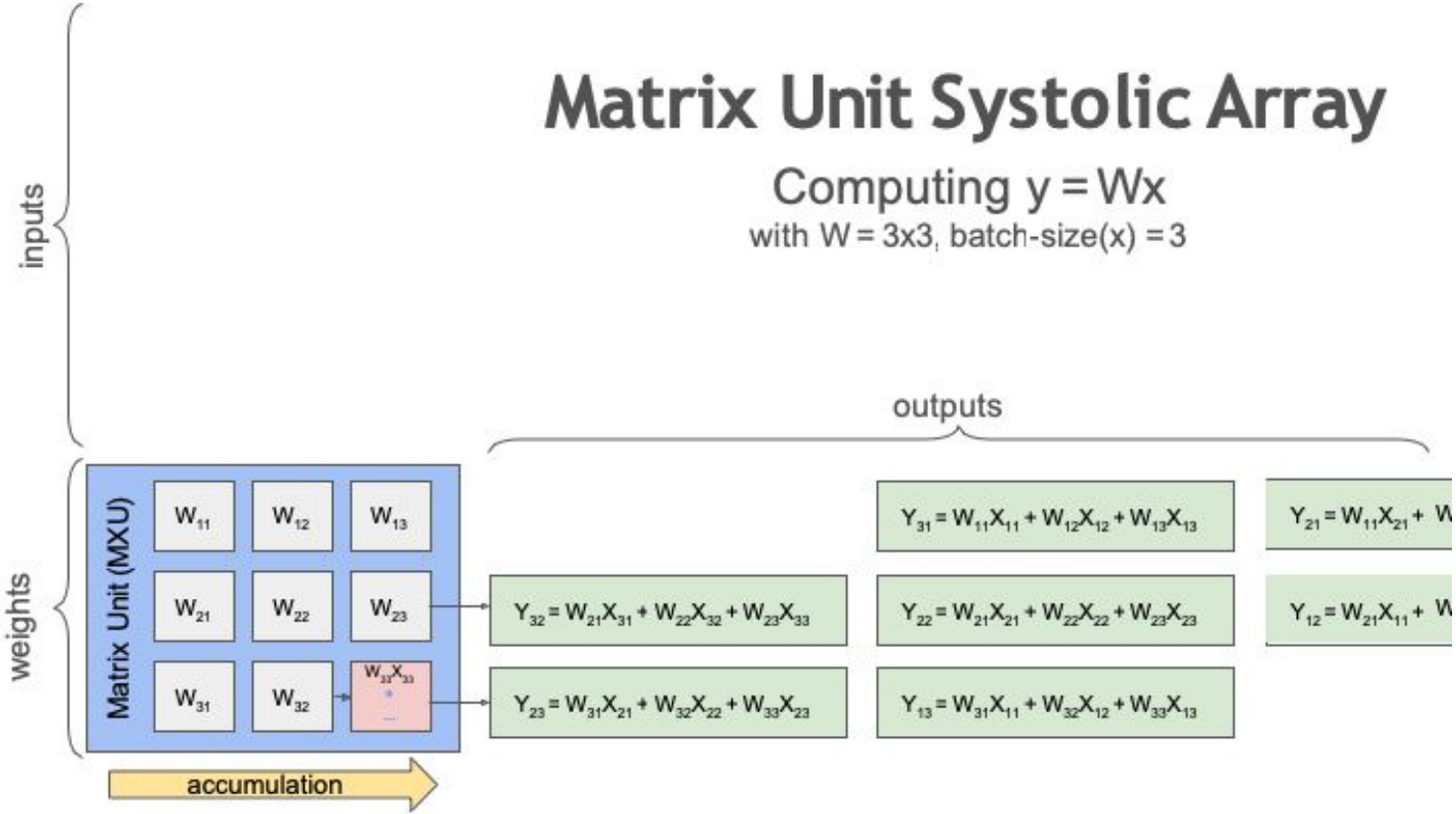
UCLA Neural Networks: Computation Example



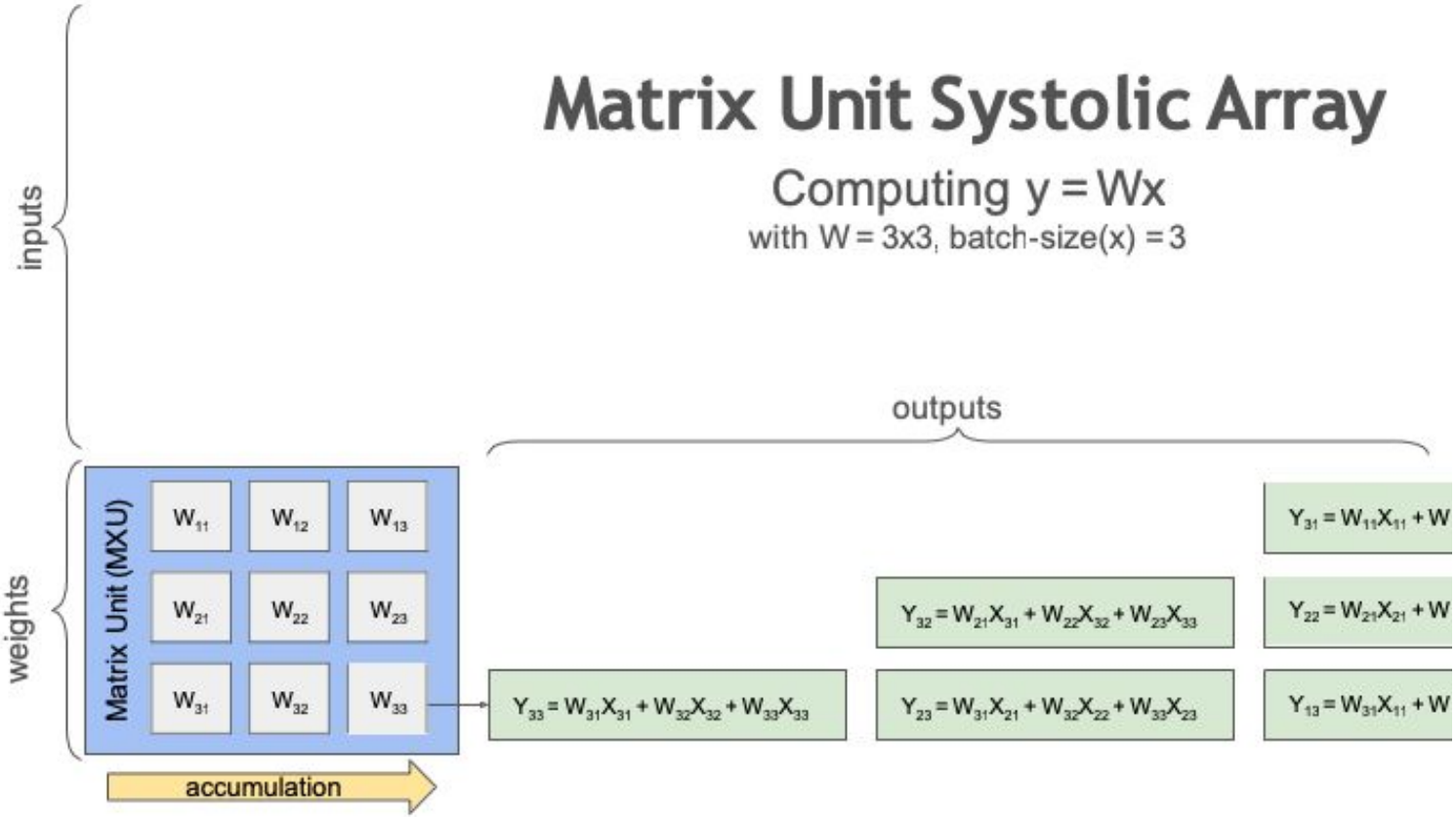
UCLA Neural Networks: Computation Example



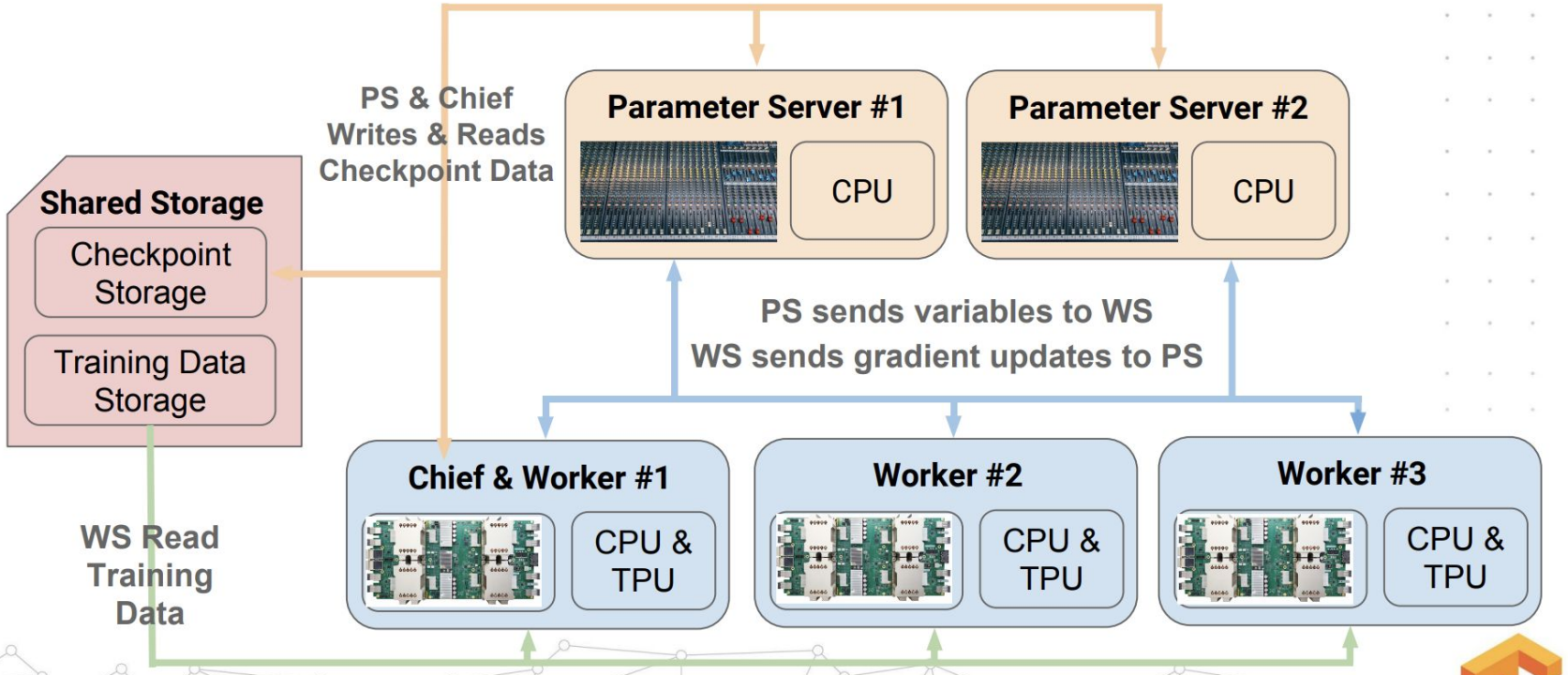
UCLA Neural Networks: Computation Example



UCLA Neural Networks: Computation Example



Neural Networks: Distributed Training



https://qcon.ai/system/files/presentation-slides/180411_qcon_presentation_-_2.pdf?fbclid=IwAR38kESwm8e2NAhkj5JFqgz0F0VtnCpFyBp1HH5itsoSQllYvkyYEwsc9uY