Matrix Multiplication using Systolic Array

Abhipray Singh
Aditya Kumar
Shreya Gulati
Ganesh Bhagwat
Chahak Tyagi
Haoying Tan

Roles and Responsibilities

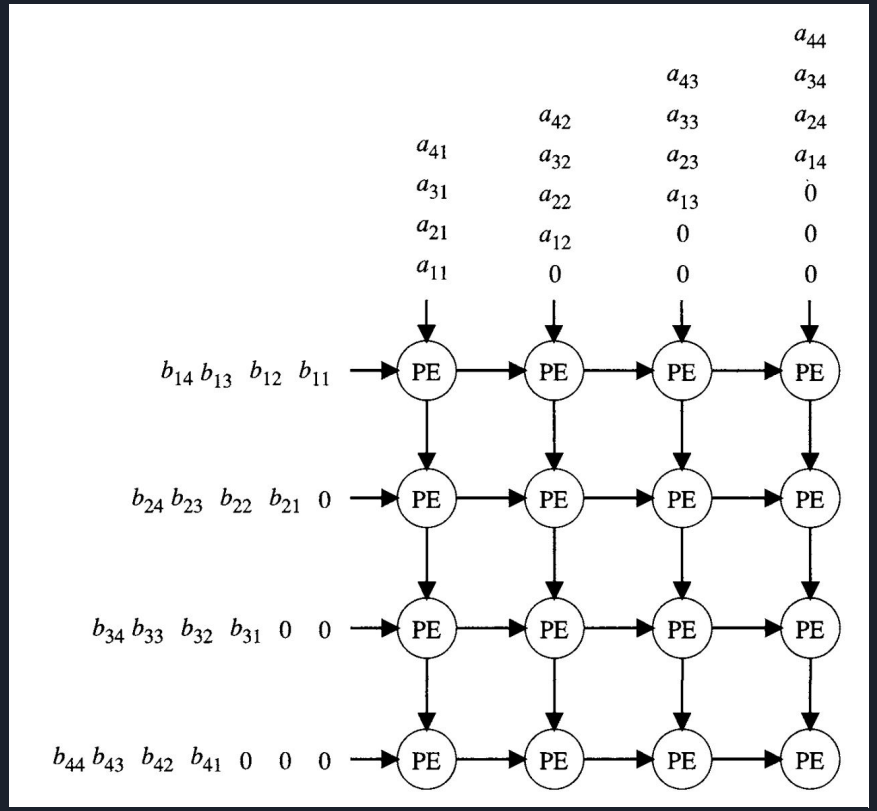
- Team Lead and Design Lead- Abhipray Singh
- Verification Lead - Ganesh Bhagwat
- Physical Design Lead - Aditya Kumar

Introduction

- Matrix multiplication is a compute-intensive operation which is widely used in image processing, data compression, and 3D graphics rendering.
- By having dedicated hardware for this operation, we can speed up all of these applications.
- Systolic arrays are one such approaches to realise this idea.

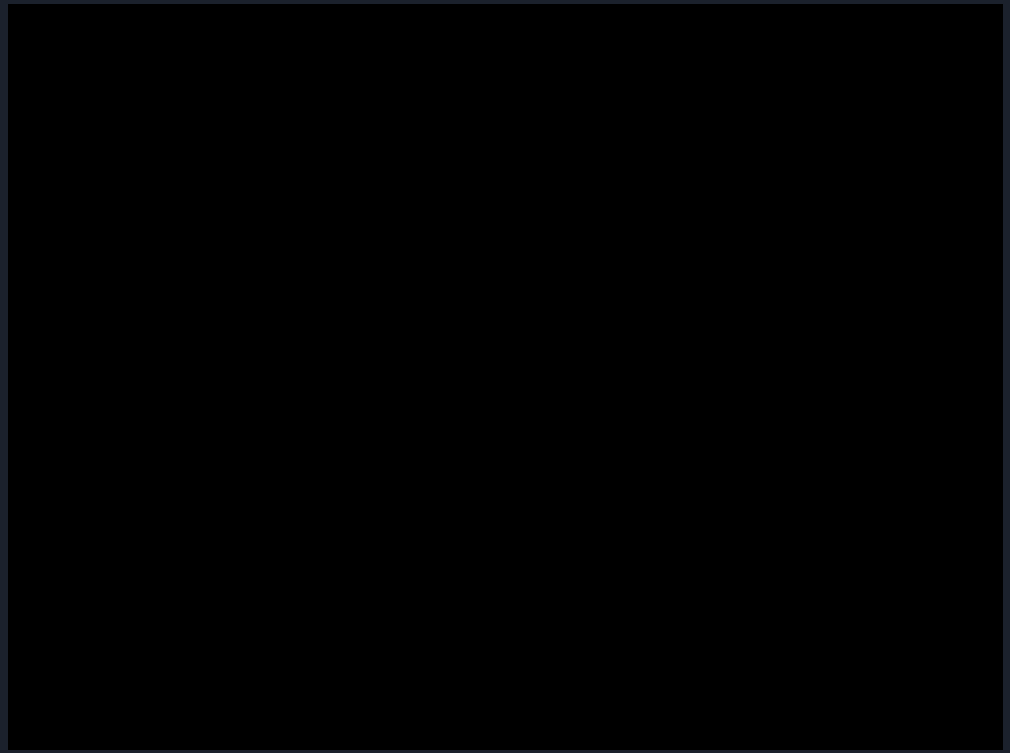
Systolic Array

- A grid-like arrangement of simple processing elements that work in a synchronized manner.
- Particularly effective in accelerating matrix multiplication operations.
- We will initially focus on the implementation of systolic arrays for 4x4 matrix multiplication, and then extend to matrices of other sizes.

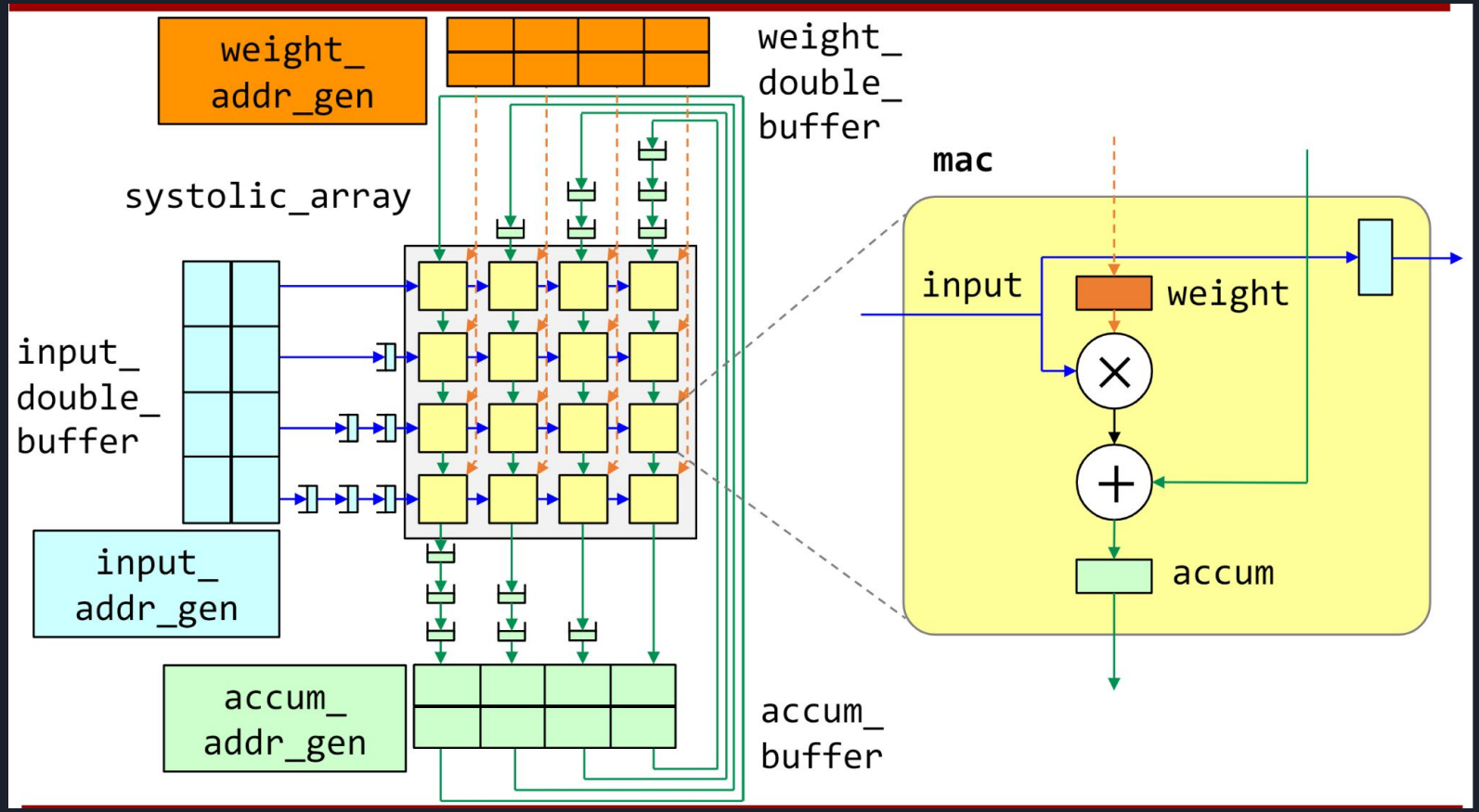


Lars Wanhammar, 8 - DSP Architectures, Editor(s): Lars Wanhammar, In Academic Press Series in Engineering, DSP Integrated Circuits, Academic Press, 1999, Pages 357-385, ISBN 9780127345307,https://doi.org/10.1016/B978-012734530-7/50008-8.

Systolic Array



https://www.youtube.com/watch?v=cmy7LBaWuZ8


weight_
addr_gen
weight_
double_
buffer
systolic_array
mac
input
weight
input_
double_
buffer
input_
addr_gen
accum
accum_
addr_gen
accum_
buffer

Our approach

- Target is to implement a systolic array.
- It is being exercised by implementing it with five stage in order pipeline.
- Two approaches of data loading:

>> approach 1: load all matrices data in buffer before calculation commences

>> approach 2 : parallely load data and perform calculation to hide load latency

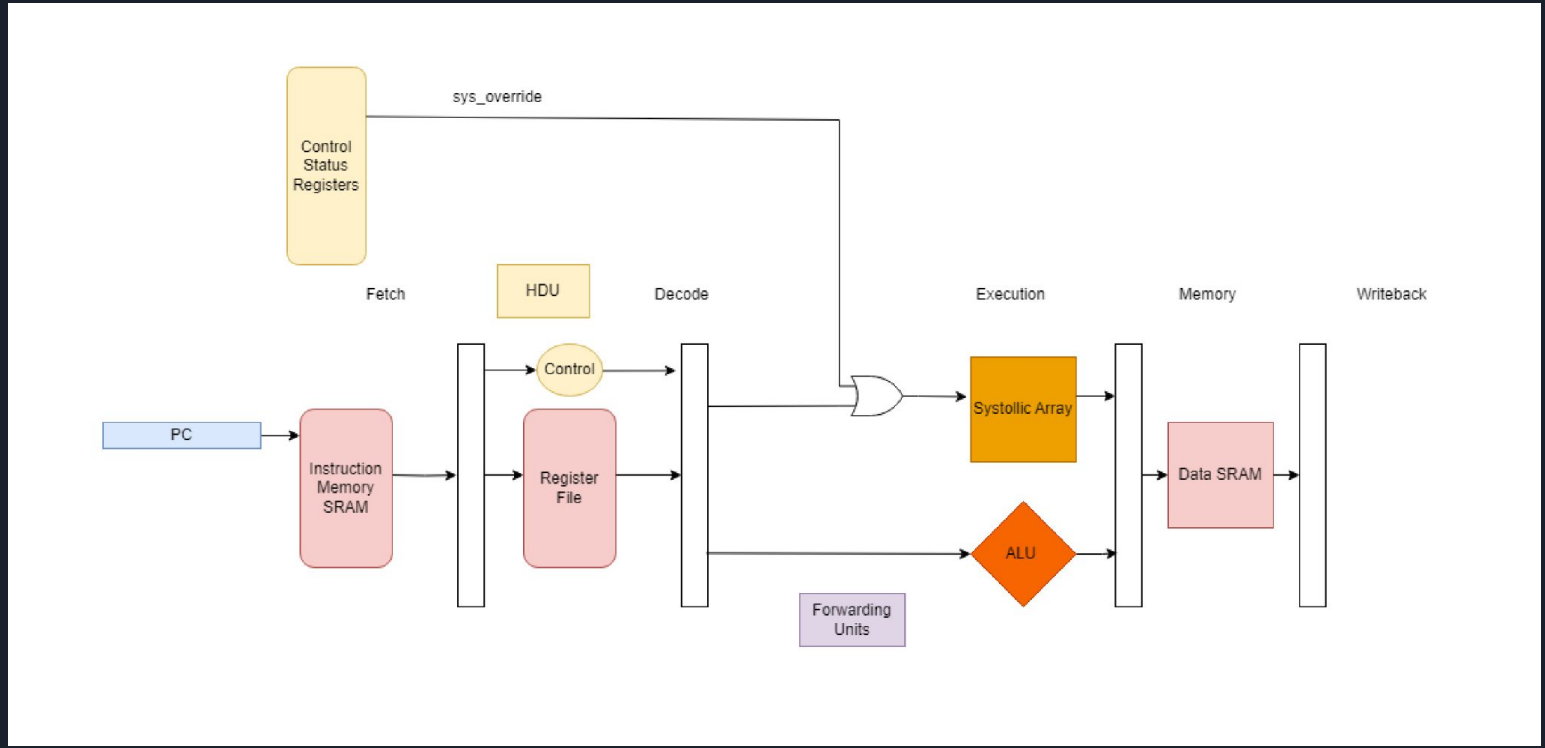
- Plan is to arrange data memory in a 8 banked fashion.
- This enables us to fetch all inputs from a single cache line , which enables us aligned fetch

Project Progression

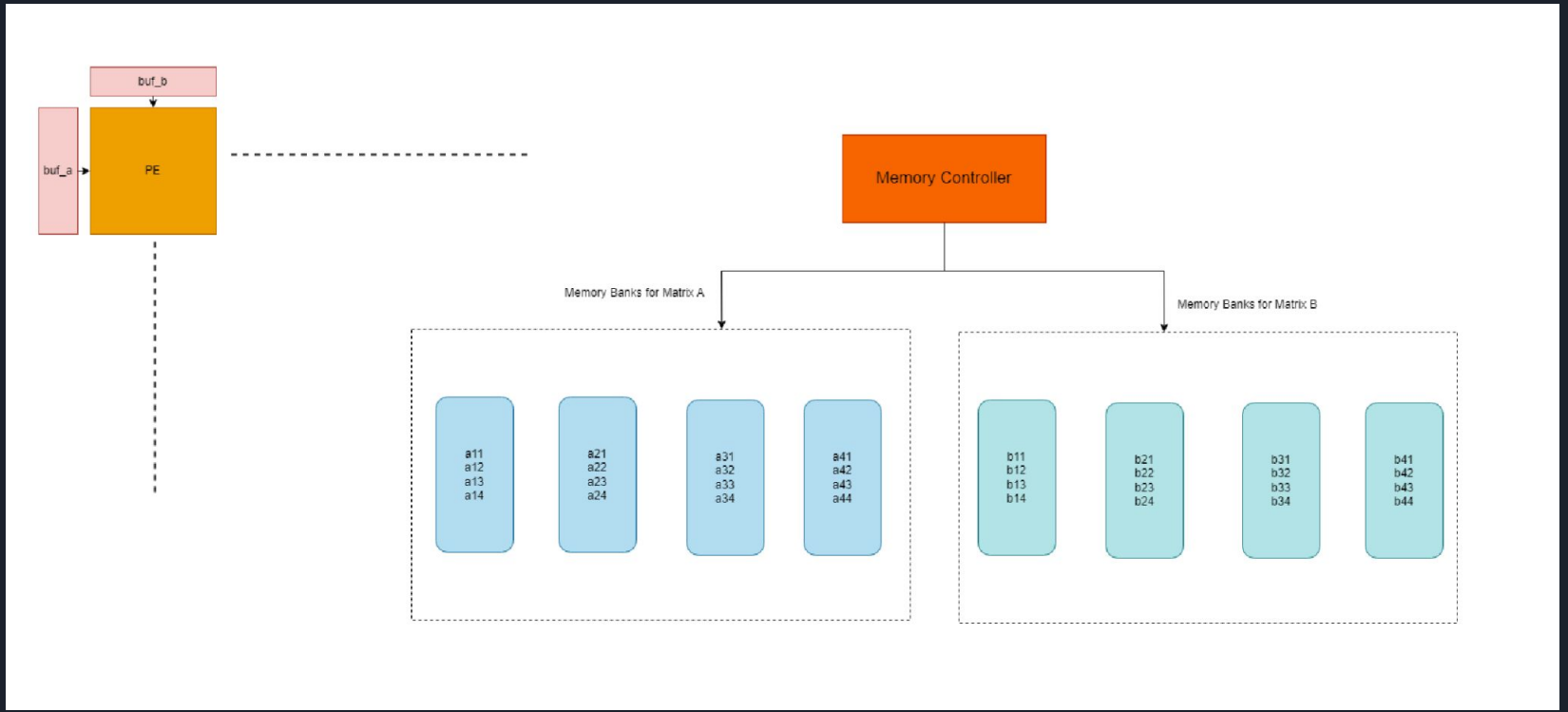
- Our first milestone is to have a working 4x4 systolic array that can do basic 4x4 matrix multiplication. This also serves as our fallback design in case future iteration fail.
- Next , we develop the 5-stage pipeline and integrate with the systolic array and so we can use the instructions to control the array.
- Finally, we plan to have a 5-stage pipeline systolic array that can perform any dimension of matrix multiplication like 4x1 * 1x4 , 2x4 * 4x2 etc.

Project Plan

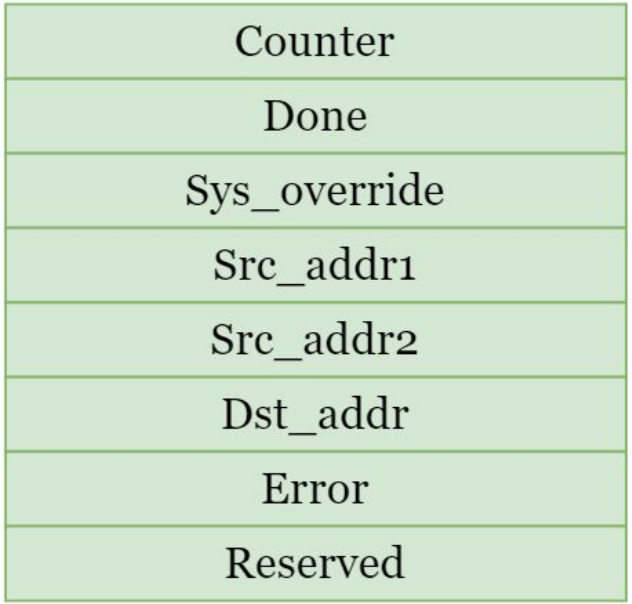
Proposed high-level block diagram



Proposed high-level block diagram



Control Status Register



Proposed Instructions

- **Add**
- **Lw**
- **Sw**
- **Beq**
- **jump**
- **Sys dest_matrix_addr , source_1_addr , source_2_addr , size**
- **store_counter - Write into memory number of cycles it took for computation**
- **TBD**

Future considerations

- We start by supporting only the 4x4 multiplications, but we plan to support customizing the matrix dimensions.
- Make the hardware do other operations, such as convolutions, etc., along with the matrix multiplications.

Thank you!