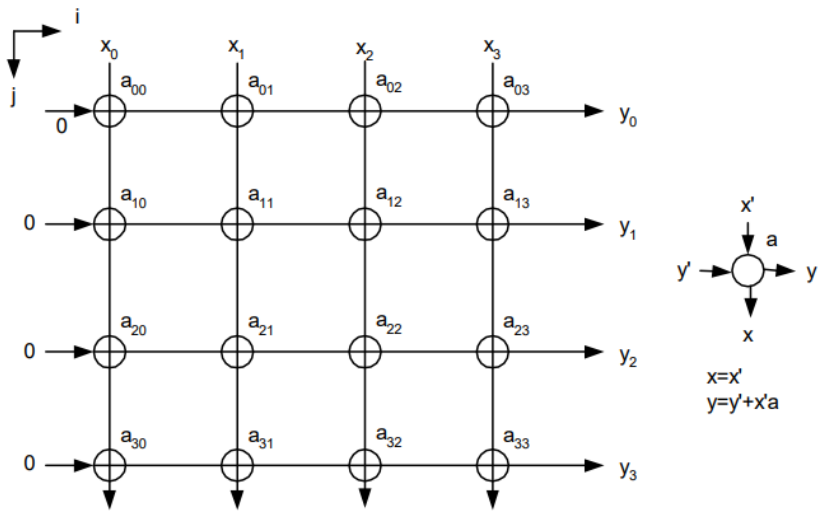
DSP in VLSI Design
Homework (VII)

Systolic Architecture Design

1. This problem addresses systolic architecture design for matrix-vector multiplication. In this problem, we assume that each processing element can only access its local memory and the elements in matrix **A** are constant. Consider the following matrix-vector multiplication:

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ a_{10} & a_{11} & a_{12} & a_{13} \\ a_{20} & a_{21} & a_{22} & a_{23} \\ a_{30} & a_{31} & a_{32} & a_{33} \end{bmatrix} \cdot \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Its data dependency graph is shown in the following figure, where the element a_{ji} of the coefficient matrix **A** is stored at position (i, j).



(a) Draw the space-time representation and the systolic architecture for the matrix-vector multiplication using

$$\mathbf{d}^T=[1\ 0],\ \mathbf{p}^T=[0\ 1],\ \mathbf{s}^T=[1\ 1].$$

Assume that storage is localized to each processing element. Explicitly show the contents of local memory.

(b) Map the same DG with different mapping vectors:

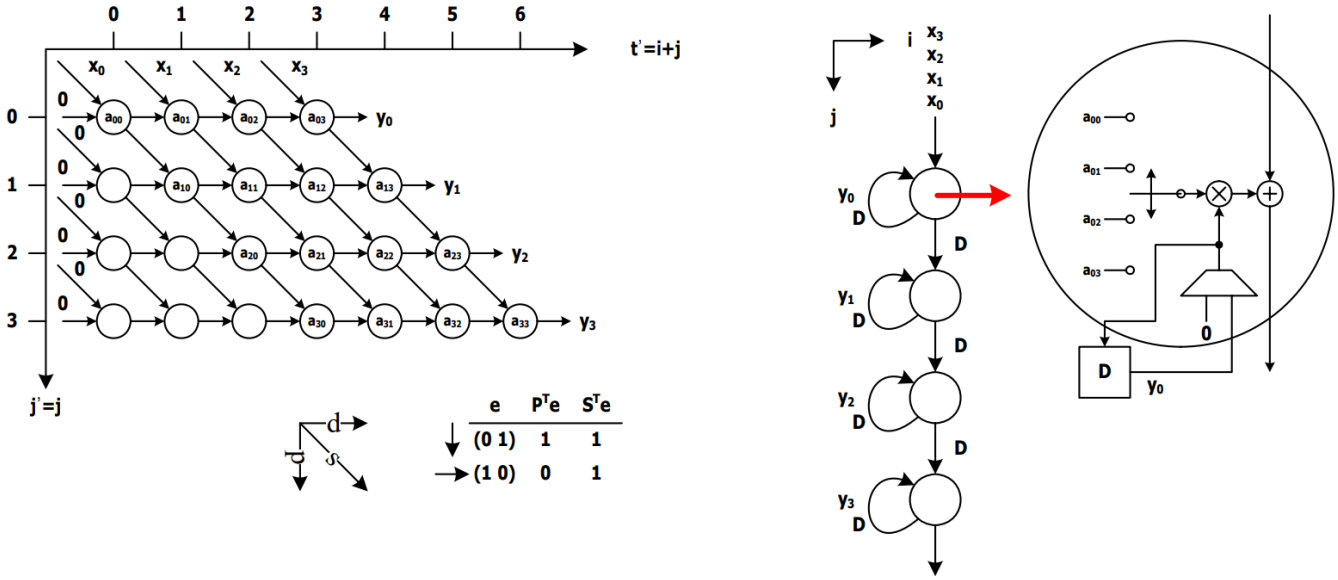
i. $\mathbf{d}^T$=[1 1], $\mathbf{p}^T$=[1 -1], $\mathbf{s}^T$=[1 0].

ii. $\mathbf{d}^T$=[1 1], $\mathbf{p}^T$=[1 -1], $\mathbf{s}^T$=[0 1].

We have three architectures now, including these two architectures and the one in (a). Do you think which one is better? Why?

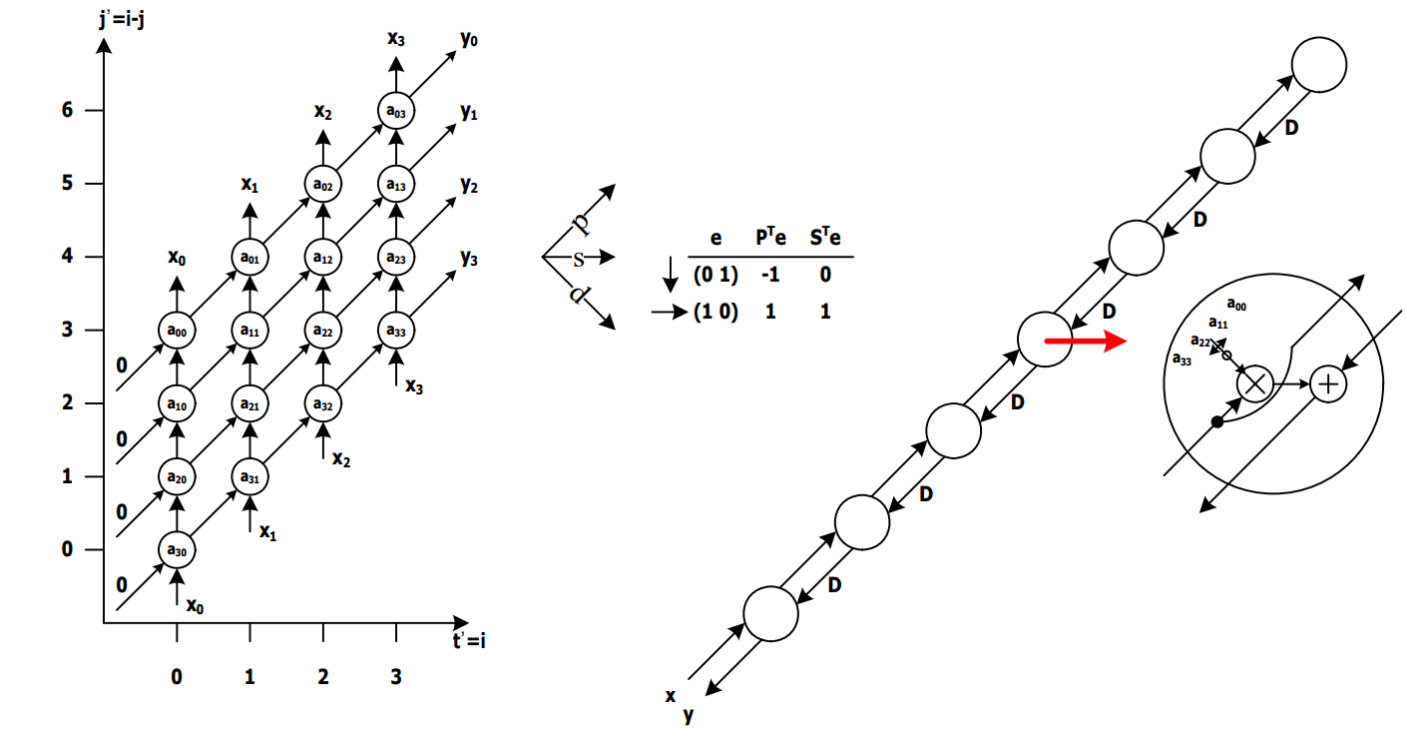
(c) Use the mapping vector in (a). If the elements in matrix **A** are not constant. Derive the systolic architecture.

(a)

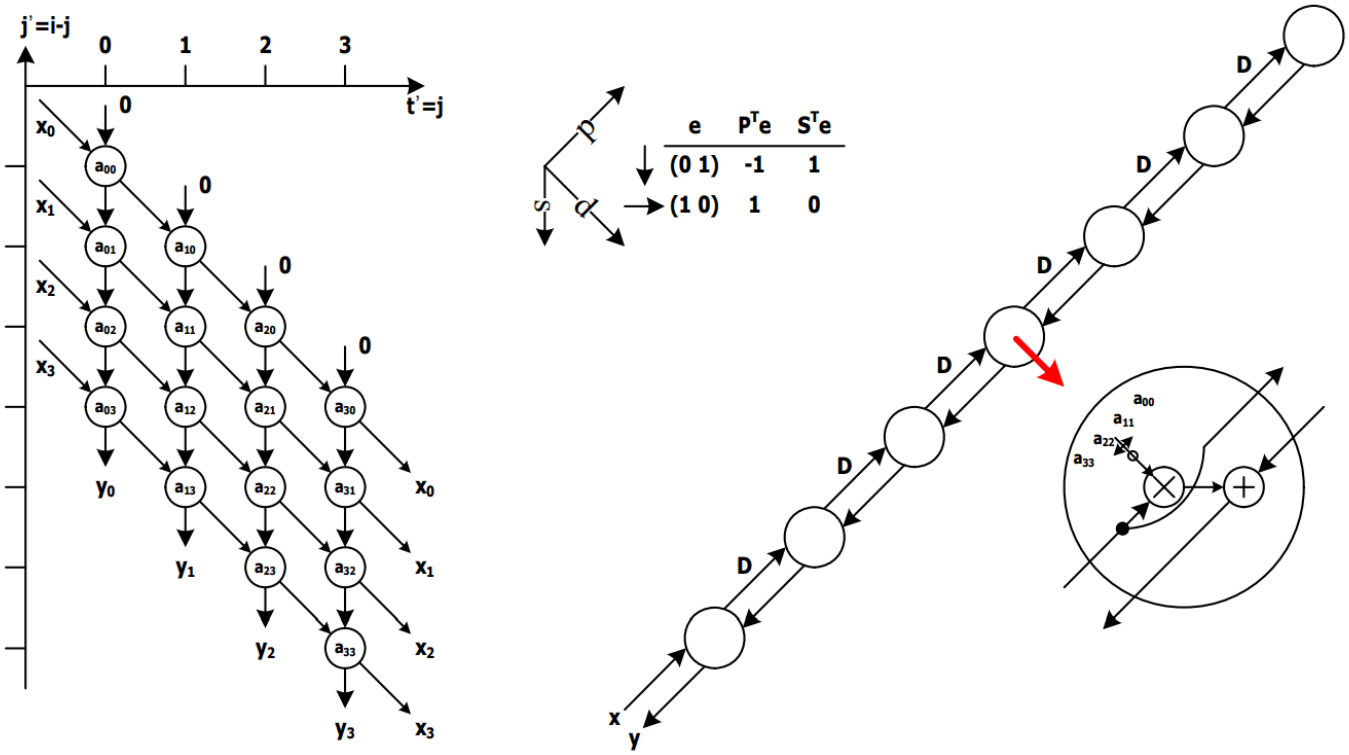


(b)

(i)



(ii)



After folding (b-i) and (b-ii), the resources used by them are less than that used by (a). Besides, critical paths of (a) and (b-i) are shorter than (b-ii). From the two points of view, (b-i) is better. But we should choose one of them not only according to the performance but also to our design specifications like I/O constraints.

(c) The only difference between this one and architecture in (a) is the registers put on the path of a_{xx} to store arbitrary data.

2. Remember in homework 1. You have drawn the DG of a 2D moving average filter for an image, which can be shown as the following equation:

$$y(i,j) = \sum_{m=-1}^{1}\sum_{n=-1}^{1} x(i+m, j+n),$$

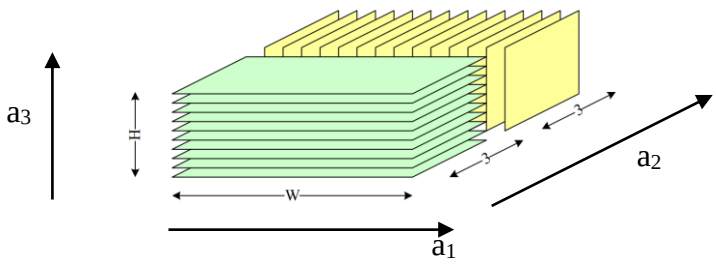
$$i, i+m \in [0, W]$$

$$j, j+n \in [0, H]$$

where x(i,j) is the input image, y(i,j) is the filtered image, and W and H are the width and height of the image, respectively.

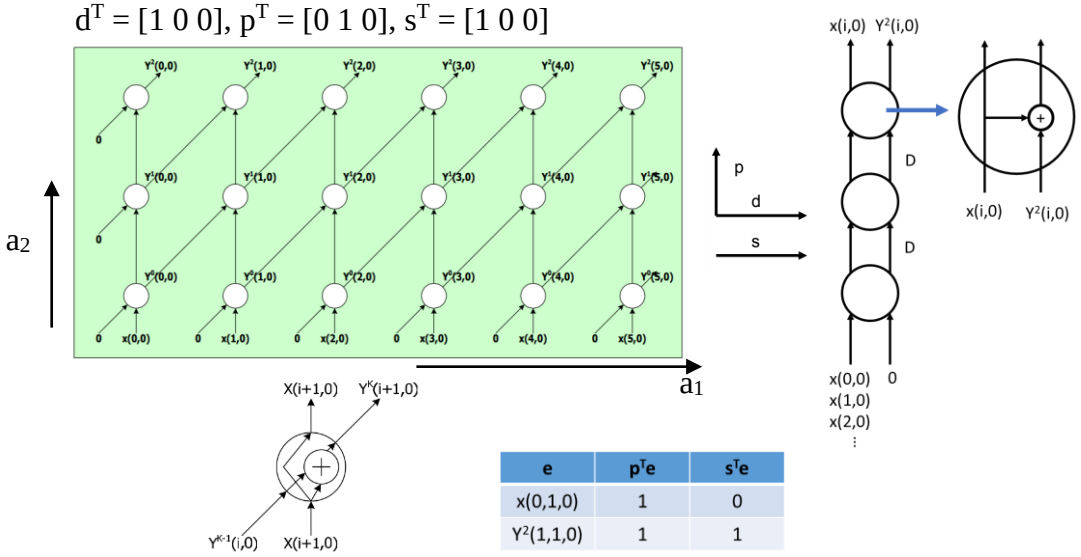
(a) Now, please choose two set of mapping vectors (projection vector and scheduling vector) and draw the associated hardware architecture.

In homework1, we separate the DG to 2 directional DG.



Step 1

$d^T = [1\ 0\ 0]$, $p^T = [0\ 1\ 0]$, $s^T = [1\ 0\ 0]$

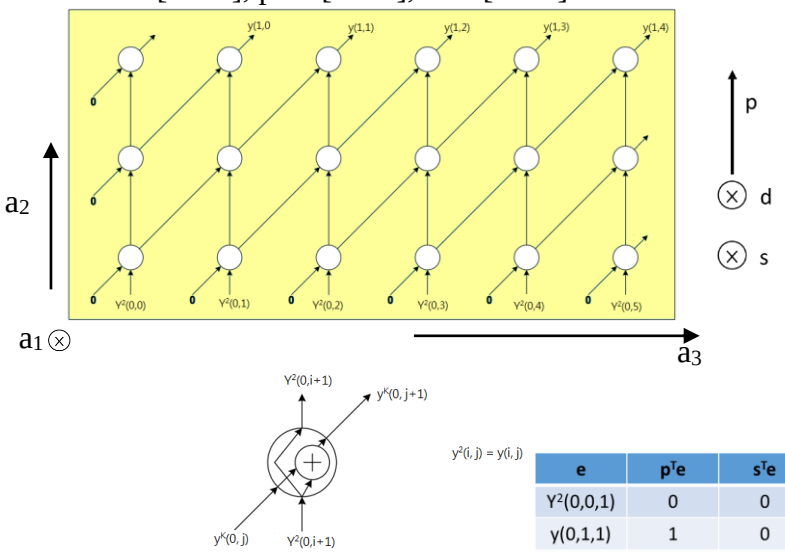


e	p^Te	s^Te
x(0,1,0)	1	0
$Y^2(1,1,0)$	1	1

Step 2

Similar to step 1

$d^T = [1\ 0\ 0]$, $p^T = [0\ 1\ 0]$, $s^T = [1\ 0\ 0]$

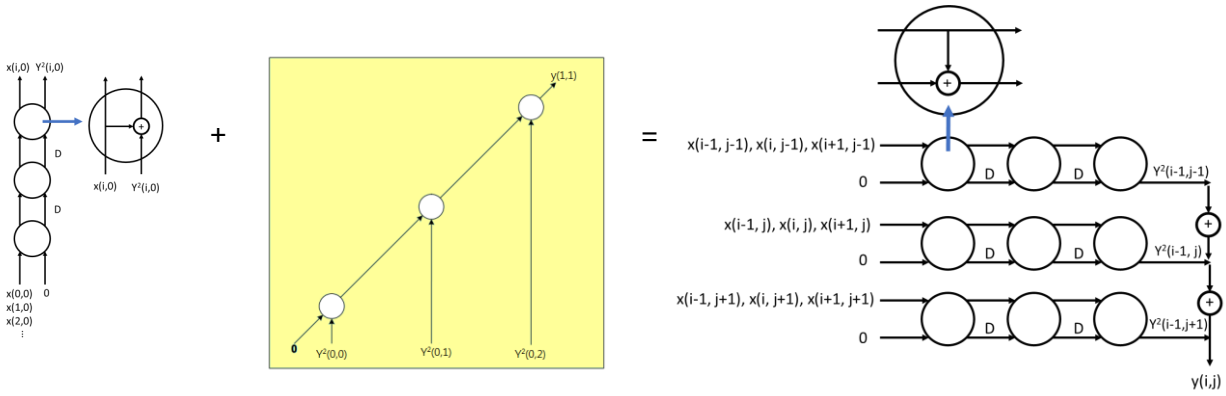


e	p^Te	s^Te
$Y^2(0,0,1)$	0	0
y(0,1,1)	1	0

The systolic architecture is the same.

Step 3

If we compute $Y^2(i, j)$, $Y^2(i, j+1)$, $Y^2(i, j+2)$ at the same time, we can merge 2 architecture.



(b) Show the scheduling of our hardware to prove its function is correct.

Cycle 1

x(0, 0)	x(0, 0)	x(0, 0)
x(0, 1)	x(0, 1)	x(0, 1)
x(0, 2)	x(0, 2)	x(0, 2)

Cycle 2

x(1, 0)	x(0, 0)+ x(1, 0)	x(0, 0)+ x(1, 0)
x(1, 1)	x(0, 1)+ x(1, 1)	x(0, 1)+ x(1, 1)
x(1, 2)	x(0, 2)+ x(1, 2)	x(0, 2)+ x(1, 2)

Cycle 3

x(2, 0)	x(1, 0)+ x(2, 0)	x(0, 0)+ x(1, 0)+ x(2, 0)
x(2, 1)	x(1, 1)+ x(2, 0)	x(0, 1)+ x(1, 1)+ x(2, 0)
x(2, 2)	x(1, 2)+ x(2, 0)	x(0, 2)+ x(1, 2)+ x(2, 0)
		Sum: y(1,1)

Cycle 4

x(3, 0)	x(2, 0)+ x(3, 0)	x(1, 0)+ x(2, 0)+ x(3, 0)
x(3, 1)	x(2, 1)+ x(3, 1)	x(1, 1)+ x(2, 0)+ x(3, 1)
x(3, 2)	x(2, 2)+ x(3, 2)	x(1, 2)+ x(2, 0)+ x(3, 2)
		Sum: y(2,1)