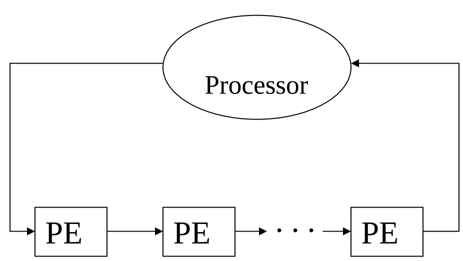
Chapter 7
Systolic Arrays

CSE4210 Winter 2012

Mokhtar Aboelaze

Systolic Architecture

- A number of usually similar processing elements connected together to implement a specific algorithm.
- Data move between PE's in a rhythmic fashion.



Systolic Architecture

- Typically, fully pipelined (all communication between PE's contain delay element (why?). Also communication between neighboring PE's only.
- Some relaxation techniques can get rid of the delay. Also, there may be communication between close bet not neighboring PE's
- Some processors (especially boundary ones may be different than the rest.
- Could be used as a coprocessor

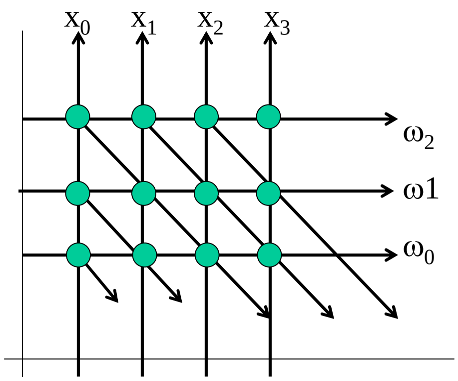
Design Methodology

- Using linear mapping techniques from the dependence space to the space-time
- Usually, algorithm is described by a dependence graph.
- Dependence graph is regular if the presence of any edge connected to a node, means the existence of a similar edge in every node.
- There is no concept of time in the dependence graph.

YORK UNIVERSITY CSE4210

FIR Filter

- $Y(n) = \omega_0\, x(n) + \omega_1\, x(n-1) + \omega_2\, x(n-2)$



- Data is moving in three directions
- X in $[0\ 1]^T$
- ω in $[1\ 0]^T$
- Y in $[1\ -1]^T$

YORK UNIVERSITY CSE4210

Design Methodology

- We map the *N-dimensional* DG to a lower dimension systolic architecture
- Three vectors are introduced
- Projection vector $\mathbf{d} = [d_1\ \ d_2]^T$
- Processor space vector $\mathbf{P}^T = [p_1\ p_2]$
- Scheduling vector $\mathbf{S}^T = [S_1\ S_2]$
- Hardware Utilization Efficiency $= 1/|S^T d|$

Design Methodology

- Projection vector
  - Two nodes are displaced by **d** or multiple of it, are mapped to the same processor
- Processor space vector
  - Any node in the DG ***I*** ***(I^T=(i,j))*** is mapped to processor $P^T I$
- Scheduling vector
  - Any node in the DG ***I*** would be executed at time $S^T I$
- Subject to some constraints

Design Methodology

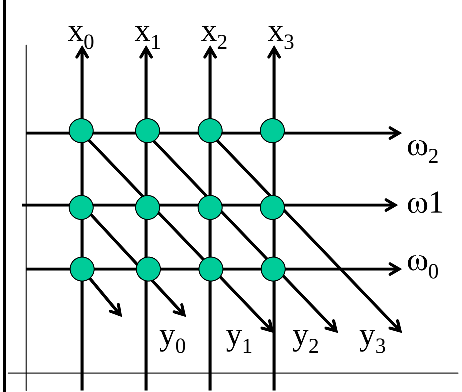
- Steps
  - Represent algorithm as a DG
  - Apply mapping (projection and scheduling)
  - Edge mapping
    - If an edge e exists in the DG, then an edge $P^T e$ is introduced in the systolic array with $S^T e$ delay
  - Construct the systolic array

Design Methodology

- Constraints
  - Processor space vector and the projection vector must be orthogonal $\mathbf{P}^T\mathbf{D}=\mathbf{0}$. if I_A-I_B = multiple of **d**, they are executed by the same processor
  - If A and B are mapped to the same processor, they should not be executed at the same time $\mathbf{S}^T\mathbf{I}_A \neq \mathbf{S}^T\mathbf{I}_B$ i.e. $\mathbf{S}^T\mathbf{d}\neq\mathbf{0}$

Example -- IIR

- $Y(n)=\omega_0\ x(n)+\ \omega_1\ x(n-1) + \omega_2\ x(n-2)$



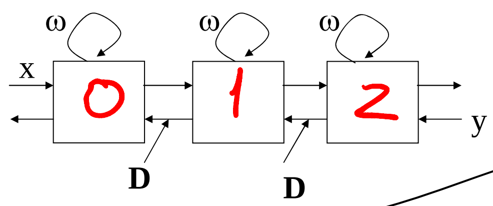
Single assignment format with broadcasting data:

```
Do n=1,2, . . .
  y1(n,-1)=0
  Do k=0,K
   y1(n,k)=y1(n,k-1)
             +w(k)*x(n-k)
  enddo
  y(n)=y1(n,K)
Enddo
```

Design I

$$d = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad P^T = [0 \quad 1] \quad S^T = [1 \quad 0]$$

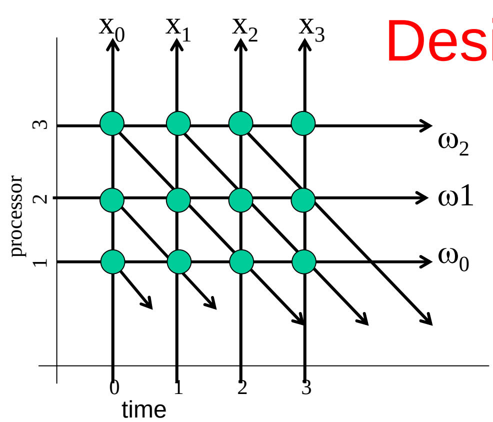
If an edge e, then an edge P^Te is introduced in the array with delay S^Te



Edge e	P^Te	S^Te
ω(1 0)	0	1
X(0 1)	1	0
Y(1 -1)	-1	1

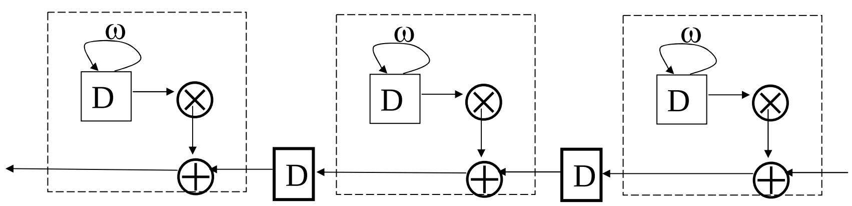
Weights stay, broadcast input, move results

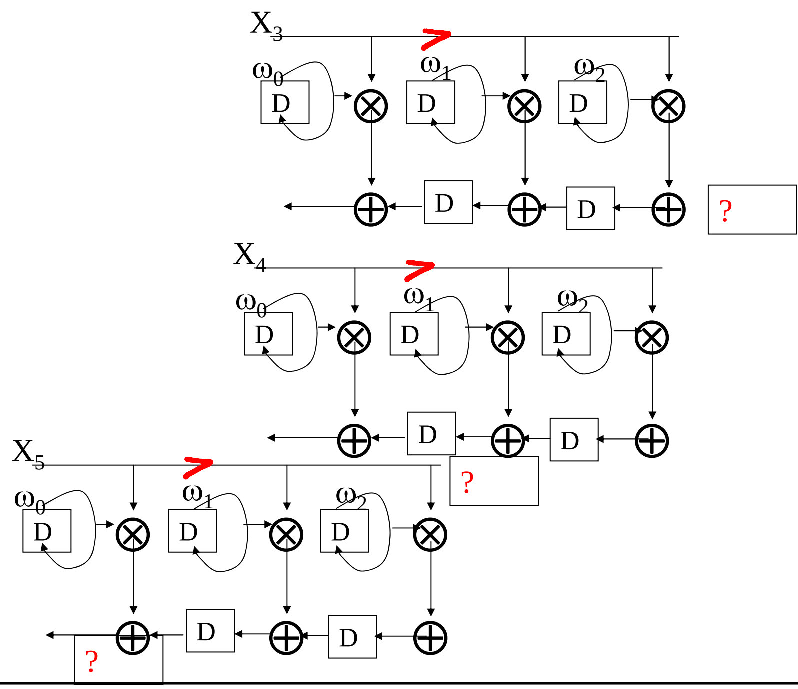
Design I



Point I is executed in PE P^TI at time S^TI

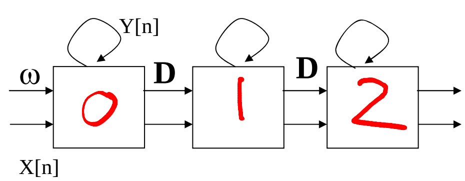
Point (i,j) is executed at PE j at time i





Design II

$$d = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad P^T = [1 \quad 1] \quad S^T = [1 \quad 0]$$



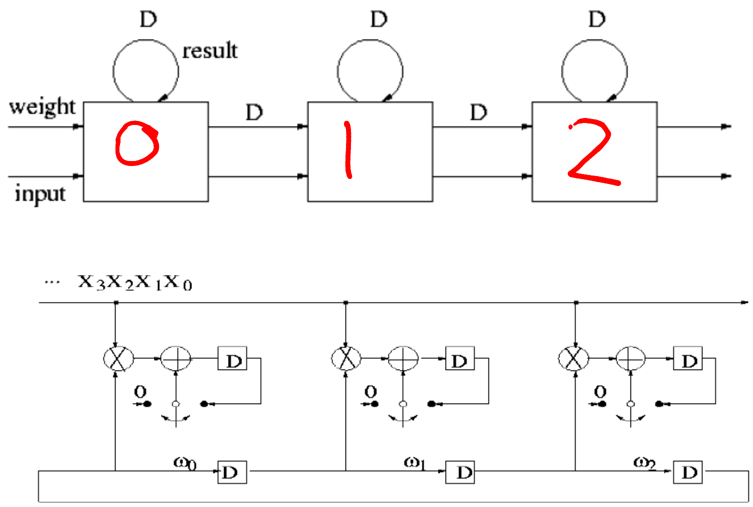
Edge e	P^Te	S^Te
W(1 0)	1	1
X(0 1)	1	0
Y(1 -1)	0	1

Point I is executed in PE $P^T I$ at time $S^T I$

Point (i,j) is executed at PE i+j at time i

Broadcast input, move weights, result stay

Design II

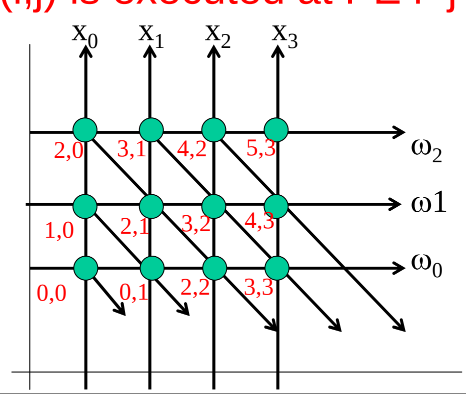


YORK UNIVERSITY CSE4210

Design II

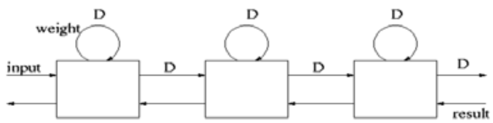
- How can you square the previous design with.
- Point (i,j) is executed at PE i+j at time i

x,y means at time x in processor y



Design III

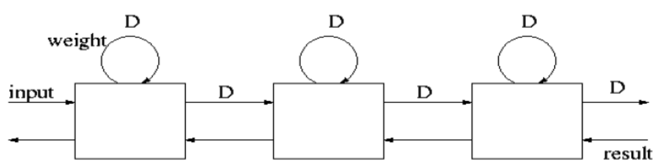
$$d = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad P^T = [0 \quad 1] \quad S^T = [1 \quad 1]$$

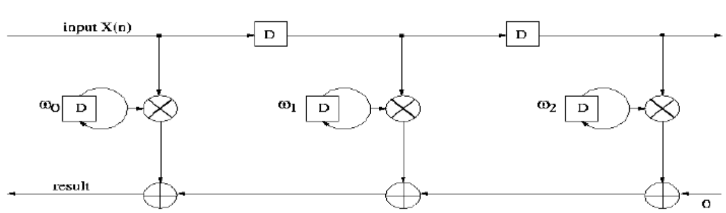


Edge e	$P^T e$	$S^T e$
W(1 0)	0	1
X(0 1)	1	1
Y(1 -1)	-1	0

Weights stay, move input,
fan in output

Design III





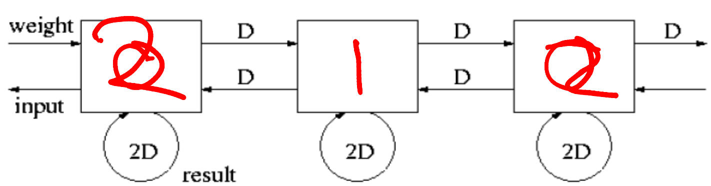
Design IV

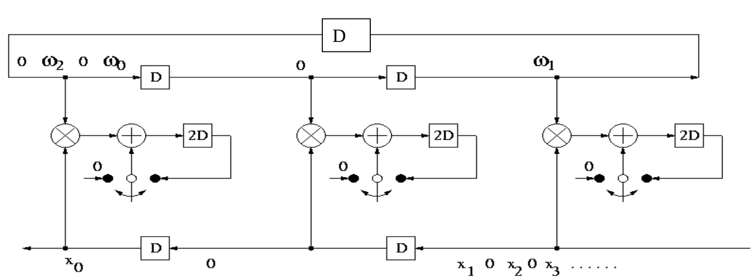
$d = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ $P^T = [1 \quad 1]$ $S^T = [1 \quad -1]$

Edge e	$P^T e$	$S^T e$
W(1 0)	1	1
X(0 1)	-1	1
Y(1 -1)	0	2

$s^T d = 2$, HUD = $\frac{1}{2}$

Design IV





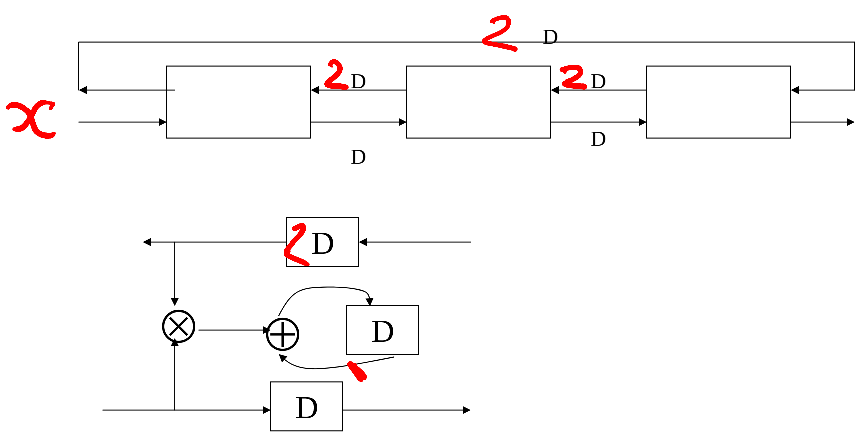
Design V

$$d=\begin{bmatrix}1\\-1\end{bmatrix}\quad P^T=[1\ \ 1]\quad S^T=[2\ \ 1]$$

Edge e	P^Te	S^Te
W(1 0)	1	2
X(0 1)	1	1
Y(1 -1)	0	1

YORK UNIVERSITY CSE4210

Design V



Dual

$$d = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad P^T = [1 \quad 1] \quad S^T = [1 \quad 2]$$

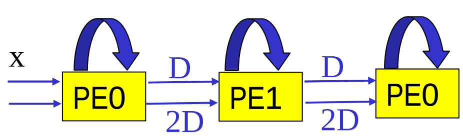
Dual of the previous design.

X and w are exchanged

Edge e	P^Te	S^Te
W(1 0)	1	1
X(0 1)	1	2
Y(1 -1)	0	1

Design VI

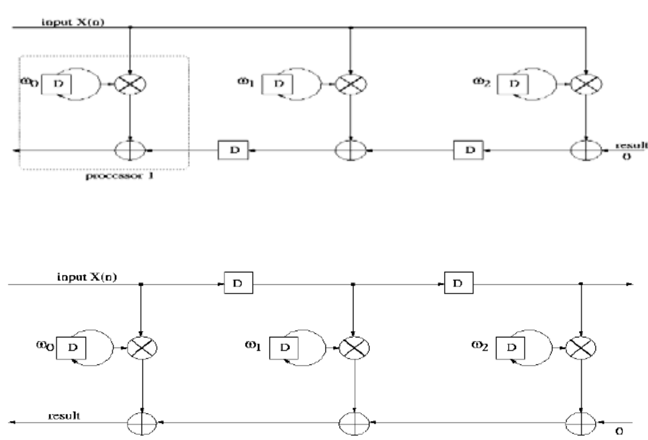
$$d = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad P^T = [0 \quad 1] \quad S^T = [1 \quad -1]$$



Edge e	P^Te	S^Te
W(1 0)	0	1
X(0 1)	-1	1
Y(1 -1)	-1	2

Dual

Transformation



Scheduling Vector

- Consider the dependence X → Y
- Y can start after X has started and completed.
- We also have to take into consideration the time it will take the data to travel from X to Y
- Constraints on the scheduling vector.

YORK UNIVERSITY CSE4210

$$X : I_x = \begin{pmatrix} i_x \\ J_x \end{pmatrix} \rightarrow Y : I_Y = \begin{pmatrix} i_y \\ J_y \end{pmatrix}$$

X → Y

$$S_y \geq S_x + T_x$$

$$S_x = S^T I_x = (s_1 \quad s_2)\begin{pmatrix} i_x \\ J_x \end{pmatrix}$$

$$S_y = S^T I_y = (s_1 \quad s_2)\begin{pmatrix} i_y \\ J_y \end{pmatrix}$$

Linear scheduling

$$S_x = S^T I_x = (s_1 \quad s_2)\begin{pmatrix} i_x \\ J_x \end{pmatrix} + \gamma_x$$

$$S_y = S^T I_y = (s_1 \quad s_2)\begin{pmatrix} i_y \\ J_y \end{pmatrix} + \gamma_y$$

Affine scheduling

Assume that $e_{x \rightarrow y} = I_y - I_x$

X → Y

Using affine scheduling,

$$S^T I_y + \gamma_y \geq S^T I_x + \gamma_x + T_x + T_{com}$$

The scheduling inequality for an edge

$$S^T e_{x \rightarrow y} + \gamma_y - \gamma_x \geq T_x$$

YORK UNIVERSITY CSE4210

Scheduling Vector

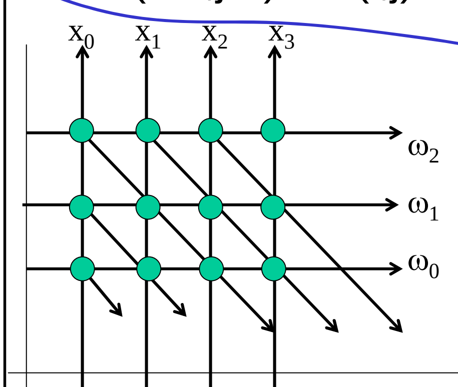
- Capture all the fundamental edges (Reduced Dependence Graph RDG).
- Use the Regular Iterative Algorithm (RIA) to describe the problem.
- Construct the scheduling inequalities and solve them for a possible S^T

RIA Description

- The regular iterative algorithm has two standard forms
- *Standard Input* if the index of the inputs are all the same for all equations
- *Standard Output* if the index of the output are all the same for all equations

RIA Description

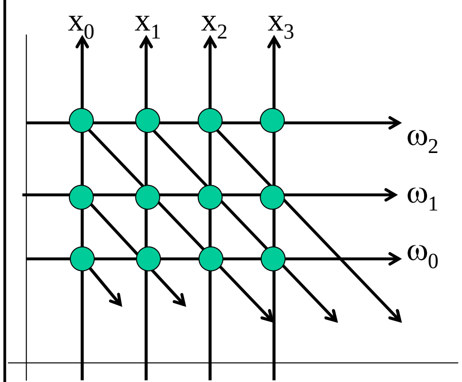
- W(i+1,j) = W(i,j)
- X(i,j+1) = X(i,j)
- Y(i+1,j-1) = Y(i,j)+W(i+1,j-1)X(i+1,j-1)



Not RIA Indices are not the same

RIA Description

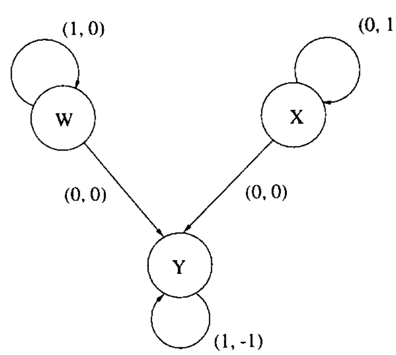
- W(i,j) = W(i-1,j)
- X(i,j) = X(i,j-1)
- Y(i,j) = Y(i-1,j+1)+W(i,j)X(i,j)



RIA

RIA Description

$$W(i,j) = W(i-1,j)$$
$$X(i,j) = X(i,j-1)$$
$$Y(i,j) = Y(i-1,j+1) + W(i,j)X(i,j)$$



Reduced RIA graph for the FIR filter

$$s^T e_{x \to y} + \gamma_y - \gamma_x \geq T_x$$

$T_{mul}=5, T_{ad}=2$

$T_{comm}=1$

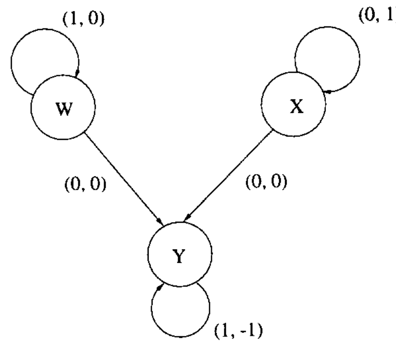
$$W \to Y : e = [s_1 \quad s_2]\begin{pmatrix}0\\0\end{pmatrix}, \gamma_y - \gamma_w \geq 0$$

$$X \to X : e = [s_1 \quad s_2]\begin{pmatrix}0\\1\end{pmatrix}, s_2 + \gamma_x - \gamma_x \geq 1$$

$$W \to W : e = [s_1 \quad s_2]\begin{pmatrix}1\\0\end{pmatrix}, s_1 + \gamma_w - \gamma_w \geq 1$$

$$X \to Y : e = [s_1 \quad s_2]\begin{pmatrix}0\\0\end{pmatrix}, \gamma_y - \gamma_x \geq 0$$

$$Y \to Y : e = [s_1 \quad s_2]\begin{pmatrix}1\\-1\end{pmatrix}, s_1 - s_2 + \gamma_y - \gamma_y \geq 5+2+1$$

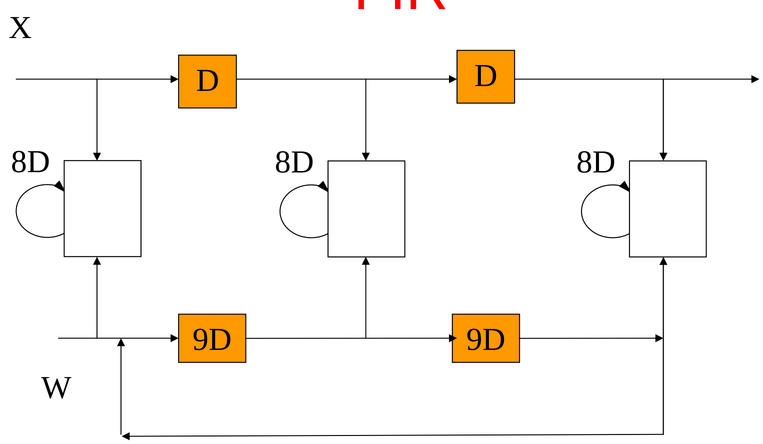


FIR

- Solving the set of equation assuming all γ's to be zero.
- A possible solution is s=[9 1]
- A possible selection for d=[1,-1] and p = [1 1]
- $s^T d$=8,HUE =1/ 8

e^T	$P^T e$	$S^T e$
W(1,0)	1	9
X(0,1)	1	1
Y(1,-1)	0	8

FIR



Matrix Multiplication

$$\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

$$c_{11} = a_{11}b_{11} + a_{12}b_{21}$$

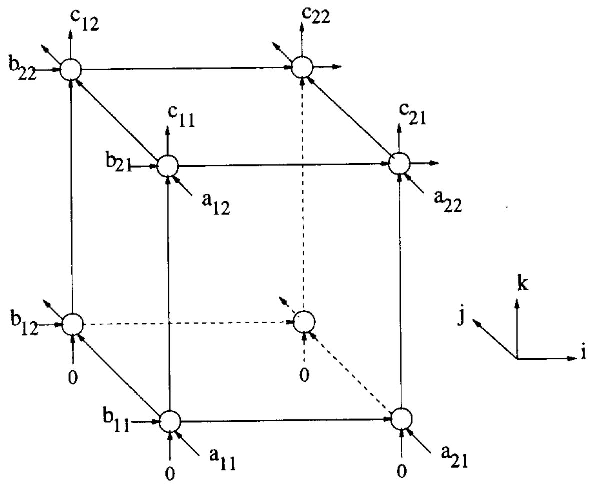
$$c_{12} = a_{11}b_{12} + a_{12}b_{22}$$

$$c_{21} = a_{21}b_{11} + a_{22}b_{21}$$

$$c_{22} = a_{21}b_{12} + a_{22}b_{22}$$

YORK UNIVERSITY CSE4210

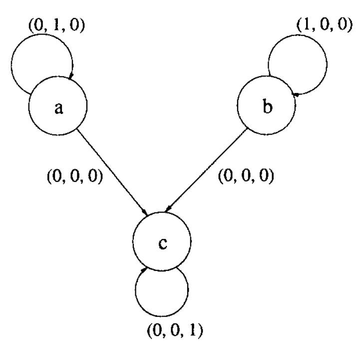
Matrix Multiplication



YORK UNIVERSITY CSE4210

Matrix Multiplication

$$\begin{aligned} a(i,j,k) &= a(i,j-1,k) \\ b(i,j,k) &= b(i-1,j,k) \\ c(i,j,k) &= c(i,j,k-1) + a(i,j,k)b(i,j,k). \end{aligned}$$



Matrix Multiplication

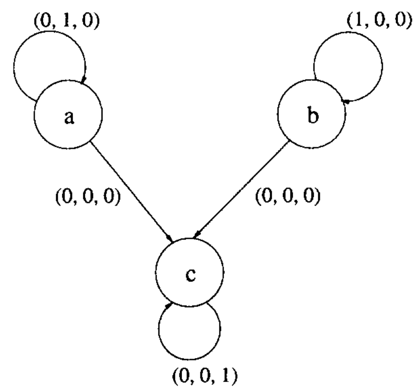
$a \to a: \quad \mathbf{e} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad s_2 \geq 0$

$b \to b: \quad \mathbf{e} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad s_1 \geq 0$

$c \to c: \quad \mathbf{e} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad s_3 \geq 1$

$a \to c: \quad \mathbf{e} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \gamma_c - \gamma_a \geq 0$

$b \to c: \quad \mathbf{e} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \gamma_c - \gamma_b \geq 0.$



$T_{accu} = 1$ T com = 0

Matrix Multiplication

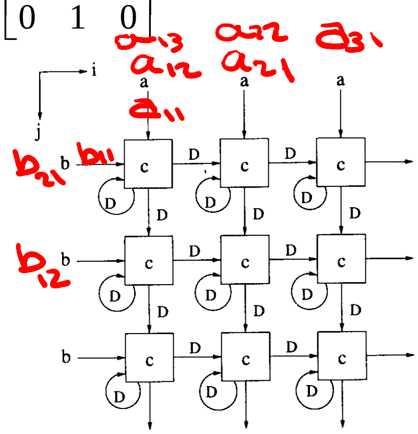
$$\mathbf{P}^T\mathbf{d} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\mathbf{s}^T\mathbf{d} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = 1.$$

$S^T = [1 \quad 1 \quad 1] \quad d^T = [0 \quad 0 \quad 1]$

$$p = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

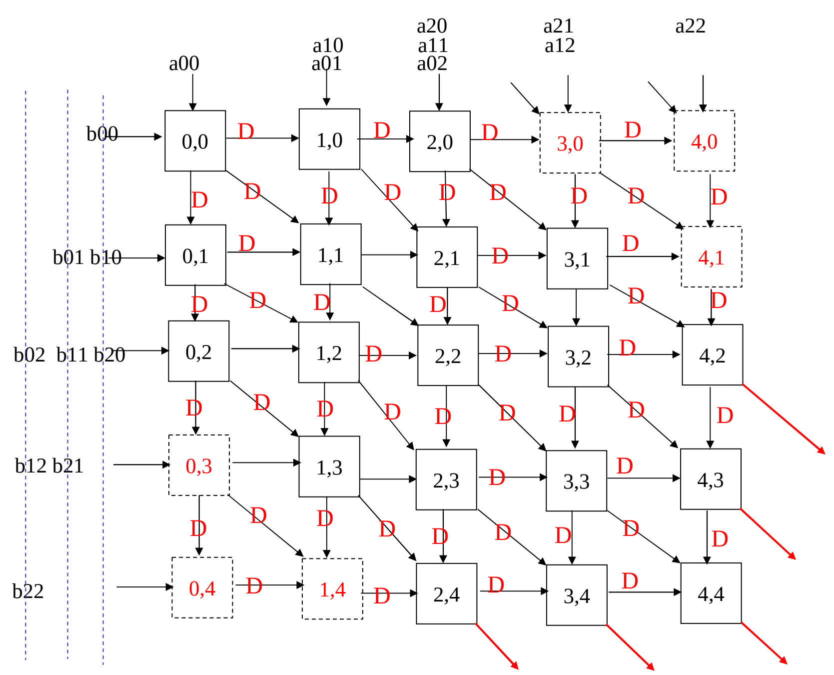
	Sol. 1		Sol. 2	
e	$\mathbf{P}^T\mathbf{e}$	$\mathbf{s}^T\mathbf{e}$	$\mathbf{P}^T\mathbf{e}$	$\mathbf{s}^T\mathbf{e}$
a(0, 1, 0)	(0, 1)	1	(0, 1)	1
b(1, 0, 0)	(1, 0)	1	(1, 0)	1
C(0, 0, 1)	(0, 0)	1	(1, 1)	1



Solution 2

HUE = 1

$$s^T = [1 \quad 1 \quad 1] \quad d^T = [1 \quad 1 \quad -1],\ p = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$



Solution 3

$$S^T = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}, \quad d = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad P^T = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$$

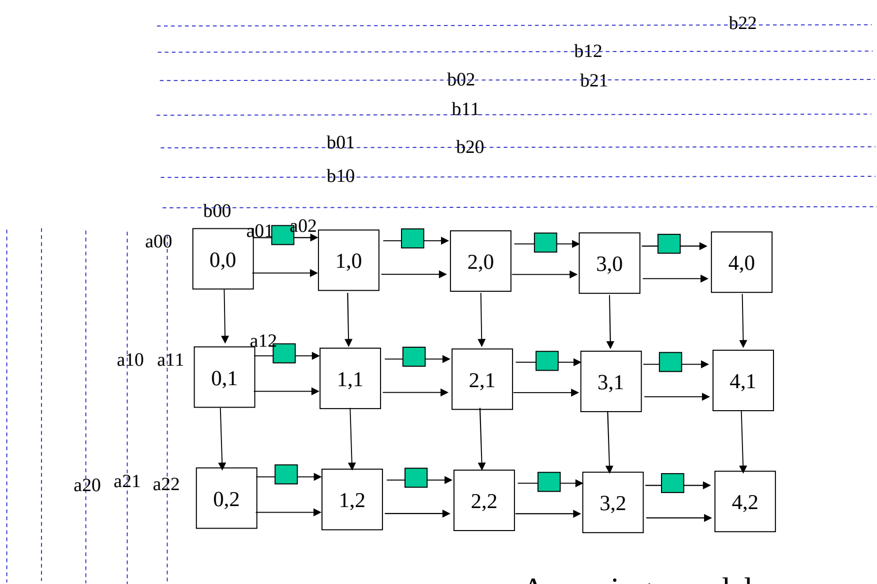
Solution 4

$$S^T = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}, \quad d = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad P^T = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}$$

Solution 5

$$s^T = (1 \quad 2 \quad 1), \quad d = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}, \quad P^T = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Vector	$P^T e$	$s^T e$
a(0,1,0)	1,0	2
b(1,0,0)	0,1	1
c(0,0,1)	1,0	1



Assuming one delay element in every PE

Solution 6

$$S^T = (1 \quad 1 \quad 1), \quad d = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}, \quad P^T = \begin{pmatrix} 1 & -1 & -1 \\ 0 & 1 & -1 \end{pmatrix}$$

Solution 7

$$S^T = (1 \quad 2 \quad 1), \quad d = \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}, \quad P^T = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \end{pmatrix}$$

Solution 4

- Solution 3:

$$\mathbf{s}^T = (1,1,1), \quad \mathbf{d} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{P}^T = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$$

This solution leads to the *Schreiher-Rao* 2D systolic array.

- Solution 4:

$$\mathbf{s}^T = (1,1,1), \quad \mathbf{d} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{P}^T = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}$$

This solution leads to the *Kung-Leiserson* systolic array.