AVM867 - VLSI Signal Processing
Assignment - I

SAURABH KUMAR
SC22B146

1. $-2^{N-1} \leq x \leq 2^{+(N-1)} - 1$: N-bit fixed point integer x

(i) $N=12$

$\Rightarrow -2^{11} \leq x \leq 2^{+11} - 1$

$\Rightarrow -2^{11} \leq x \leq 2^{+11} - 2^{-0}$

Q11.0 can be represented in the range $[-2^{11}, 2^{11}-2^{-0}]$

$\therefore$ Most negative no. $= -2^{+11} = -2048$

(ii) $N=12$

Most positive no. $= 2^{+11} - 1$

$= +2047$

2. $N=24$

$-2^{N-1} \leq x \leq 2^{(N-1)} - 1$

$\Rightarrow -2^{23} \leq x \leq 2^{23} - 1$

$$\text{Dynamic range} = \frac{\text{Max value}}{\text{Min value}}$$

$$= \frac{2^{N-1}-1}{1} = \frac{2^{23}-1}{1} \approx 2^{23}.$$

Dynamic range in dB $= 20 \log_{10}(\text{Dynamic Range})$

$= 20 \log (2^{23})$

$= 20 \times 23 \times 0.301$

$= 138.46$ dB

3. N-bit signed fractional $\rightarrow -1 \leq x < 1$

(i) $N=24$

$$DR = \frac{\text{Max value}}{\text{Min value}} = \frac{(1-2^{-(N-1)})}{(2^{-(N-1)})}$$

$(1-2^{-(N-1)})$ → Largest possible positive fraction $= 1 -$ resolution

$(2^{-(N-1)})$ → Smallest possible +ve fraction = resolution

$[2^{-(N-1)}$: Resolution$]$

$$\approx \frac{1}{2^{-(N-1)}} = 2^{N-1}$$

$$= 2^{23}$$

$DR_{dB} = 20 \log_{10}(2^{23}) = 20 \times 23 \times 0.301 = 138.46$ dB

(ii) Precision = Resolution 23 bits

$= 2^{-(N-1)}$

$= \frac{1}{2^{23}} \approx 1.19 \times 10^{-7}$

(iii) $N = 48$,

$DR = 2^{48-1} = 2^{47}$

$\downarrow$

$DR_{dB} = 20 \times 47 \times 0.301$

$= 282.47\ dB \longrightarrow 24 \times 20 \times 0.301 = 6.02N\ dB$ more

Precision $= 2^{-(48-1)} = 2^{-47}$ ~~48~~ 47 bits $\longrightarrow$ 24 bits more

$\approx 7.1 \times 10^{-15} \longrightarrow \frac{1}{2^N}$ times

(iv) N' : no. of bits (new)

$$\frac{2^{-(N'-1)}}{2^{-(N-1)}} = \frac{1}{8}$$

$\Rightarrow 2^{-N'+N} = 2^{-3}$

$\Rightarrow -N' + N = -3$

$\Rightarrow N' = N + 3 \longrightarrow 3$ more bits

④ In $Q_{1.4}$ format,

$(0.4375)_{10} = 0 + 0 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4}$

$\downarrow$

$(0.0111)_2$

ⓐ Signed Representation: (In $Q_{1.4}$ format)

$+0.4375 \longrightarrow 0\ 0\ 0111$

$-0.4375 \longrightarrow 1\ 0\ 0111$

$\downarrow$ signed bit

ⓑ One's complement:

$+0.4375 \xrightarrow{\text{same}} 0\ 0\ 0111$

$-0.4375 \xrightarrow[\text{of } 000111]{\text{flip all bits}} 1\ 1\ 1000$

ⓒ Two's complement:

$+0.4375 \xrightarrow{\text{same}} 0\ 0\ 0111$

$-0.4375 \xrightarrow[\text{1's comp.}]{+1 \text{ to}} 1\ 1\ 1001$

(5) $0.25_{10} \rightarrow 0.0100_2 \longrightarrow 0010_2$

$-0.625_{10} \rightarrow 0.10100_2 \longrightarrow \underbrace{0101_2}_{Q_{1.3}\text{ format}}$

1's complement

1010
+1

1011 → 2's complement $(-0.625)_{10}$

$\therefore$ 0.25 − 0.625

= 0010
+1011 } subtraction

$\underline{1}101$ → Negative no.

1's comp.

0010
+1

0011 → 2's complement

0.375

$\therefore$ Result = −0.375

(6) −0.72 → $Q_{0.7}$ representation

$0.72_{10} \rightarrow (0.1011100)_2$

In $Q_{0.7}$ representation,

$\underset{\text{signbit}}{\underline{1}}\ \underset{\text{fraction part}}{\underline{1011100}}$

(7) $0xABCD \longrightarrow (10\times16^3) + (11\times16^2) + (12\times16^1) + (13\times16^0)$

$= (43981)_{10}$

$= (1010\ 1011\ 1100\ 1101)_2$

Q.15 format → $\underline{0}.\underline{1010\ 1011\ 1100\ 110}$

Q8.7 format → $\underline{0}\ \underline{1010\ 1011}.\underline{11\ 00\ 110}$

8. $x = -0.5 \xrightarrow{Q_{0.3}} 1100$

$y = 0.875 \xrightarrow{Q_{0.3}} 0111$

$X = -0.5 \times 2^3 = -4$

$Y = 0.875 \times 2^3 = 7$

$Z = XY = -4 \times 7 = -28$

$\Rightarrow z = -\frac{28}{64} = -0.4375$ (for Q.6 format).

and

$xy = -0.5 \times 0.875 = -0.4375$

As $z = xy$, there is no quantization error.

9. Double precision floating pt. : 1 sign bit — 11-exp. bit — 52-frac bit

Min^m: (Subnormal)
exp → 000 - - - (11 zeros)
frac → 000 - - - (52 zeros)

$\hookrightarrow 2^{-1022} \times 2^{-52} = \boxed{2^{-1074}}$

Min^m: (Normal)
exp → 00 - - 1
frac → 000 - - -

$\hookrightarrow 2^{1-1023} \times (1.00)$

$= \boxed{2^{-1022}}$

Max: exp → 1111 - - - -
frac → 000 - - -

$\hookrightarrow 2^{2047-1023} = (2^{1024} \times 1.00\ldots)$

$= \boxed{2^{1024}}$

(10) 32-bit

(a) Fixed-point number:

- Accuracy: Fixed accuracy; loses accuracy for bigger no.s.
- Precision: $2^{(N-1)} = 2^{31}$ bits
- Dynamic range: $20 \log_{10}\left(\frac{\text{Max value}}{\text{Min value}}\right)$

$$= 20 \log_{10}\left(\frac{2^{31}-1}{2^{-31}}\right) \approx 20 \log_{10}(2^{62})$$
$$= 20 \times 62 \times 0.301$$
$$= 373.24 \text{ dB.}$$

(b) Floating point no.:

- Accuracy: Accurate for bigger no.s.
- Precision: 2^{23} bits
- Dynamic Range: $20 \log\left(\frac{2^{128}}{2^{-149}}\right)$

$$= 20 \log(2^{277}) \approx 1667.54 \text{ dB.}$$

(11) $(0.752456)_{10} \rightarrow$ in Q0.4 format.

$\downarrow$

$(0.110000)_2$

In Q0.4 format, $\underline{0}\ \underline{.1100}$

sign bit

$(0.1100)_2 \longrightarrow (0.75)_{10}$

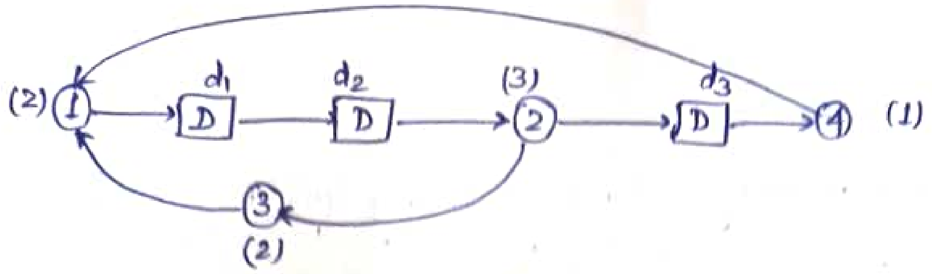
$\therefore$ Error $= 0.752456 - 0.75$
$= 0.002456$

Range:

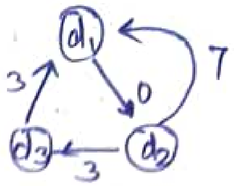
Most positive: $1 - 2^{-4} = 0.9375$

Most negative: $0 - 2^{-4} = -0.0625$

12



$\Rightarrow$



Series of matrices : $L^{(m)}, m = 1, 2, \dots, d$; d = no. of delays
m = delays

$l_{ij}^{(m+1)} = \max\limits_{\substack{k \in K \\ \downarrow \\ [1,d]}} \left(-1, \; l_{i,k}^{(1)} + l_{k,j}^{(m)}\right)$: longest computation time from d_i to d_j

$\underline{L(1)}$:

$l_{11} = l_{22} = l_{33} = -1$ (no path without delay)

$l_{12} = 0$ (no computational delay)

$l_{13} = -1$

$l_{21} = 3 + 2 + 2 = 7$

$l_{23} = 3$

$l_{31} = 1 + 2 = 3$

$l_{32} = -1$

$$\therefore L^{(1)} = \begin{bmatrix} -1 & 0 & -1 \\ 7 & -1 & 3 \\ 3 & -1 & -1 \end{bmatrix}$$

$\underline{L^{(2)}}$:

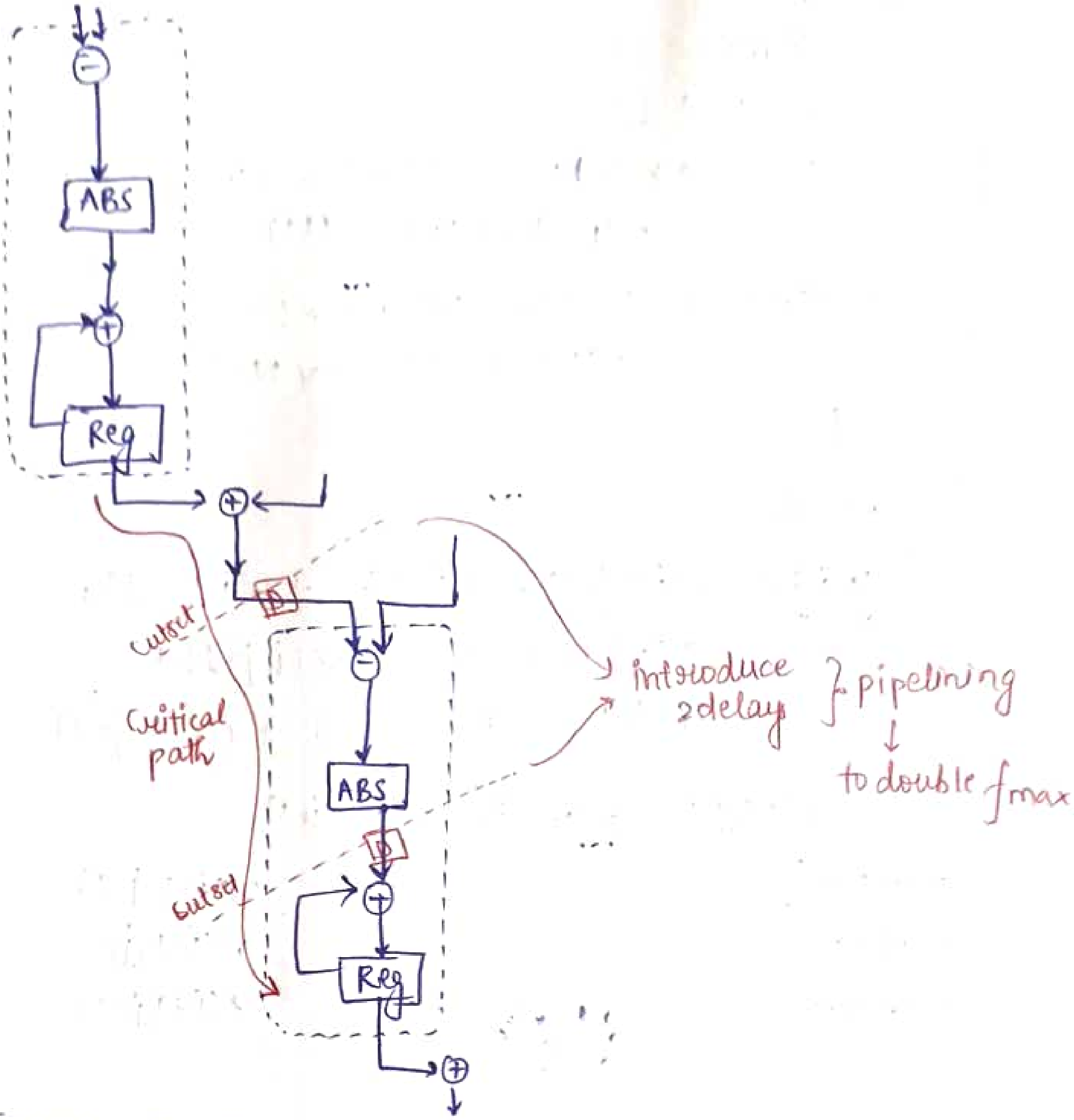
$l_{11} = 3 + 2 + 2 = 7, \; l_{12} = \dots$

$$L^{(2)} = \begin{bmatrix} 7 & -1 & 3 \\ 6 & 7 & -1 \\ -1 & 3 & -1 \end{bmatrix}, \quad L^{(3)} = \begin{bmatrix} 6 & 7 & -1 \\ 14 & 6 & 10 \\ 10 & -1 & 6 \end{bmatrix}$$

Iteration bound, $T_\infty = \max\left\{\frac{7}{2}, \frac{7}{2}, \frac{6}{3}, \frac{6}{3}, \frac{6}{3}\right\}$

$= 3.5$ ut.

(13) (a) $T_S = 5ns$, $T_{AB} = 7ns$, $T_A = 6ns$.



$$T_{critical} = T_S + T_B + 2T_A$$
$$= 24\ ns = T_{clk}.$$

Max. frequency : $f_{max} = \frac{1}{T_{critical}} = \frac{1}{24ns} = 41.67\ MHz$

(b) To double the working frequency,

$$T_{crit.}' = \frac{T_{crit.}}{2} = 12\ ns.$$

↓

Introduce 2 pipelining delays in every critical path.

$\therefore T_{critical}' = T_S + T_{ABS} = 12\ ns.$

$\hookrightarrow f_{max}' = \frac{1}{T_{crit.}'} = \frac{1}{12ns} = 83.33\ ns$

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

$$i, i+m \in [0, W]$$
$$j, j+n \in [0, H]$$

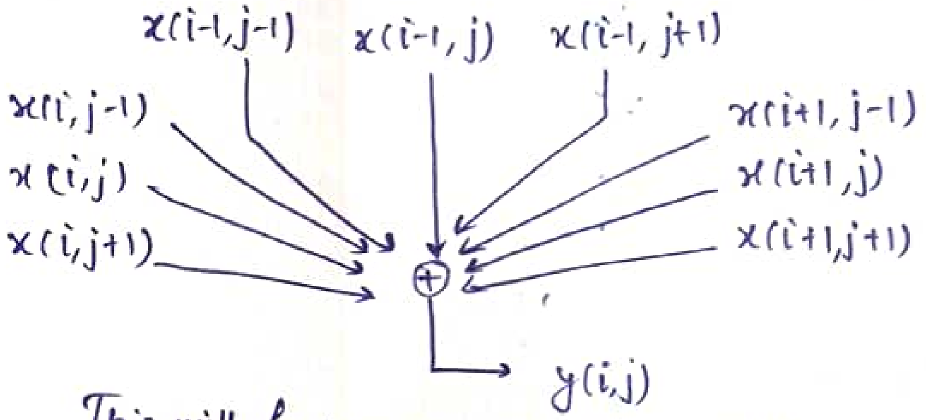
$$y(0,0) = x(-1,-1) + x(-1,0) + x(-1,+1) + x(0,-1) + x(0,0) + x(0,1) + x(1,-1) + x(1,0) + x(1,1)$$

$$y(0,1) = x(-1,0) + x(-1,1) + x(-1,2) + x(0,0) + x(0,1) + x(0,2) + x(1,0) + x(1,1) + x(1,2)$$

$\vdots$

In general,

$$y(i,j) = x(i-1,j-1) + x(i-1,j) + x(i-1,j+1) + x(i,j-1) + x(i,j) + x(i,j+1) + x(i+1,j-1) + x(i+1,j) + x(i+1,j+1).$$



This will happen for all i, j, and the actual dependence graph will be 3D.