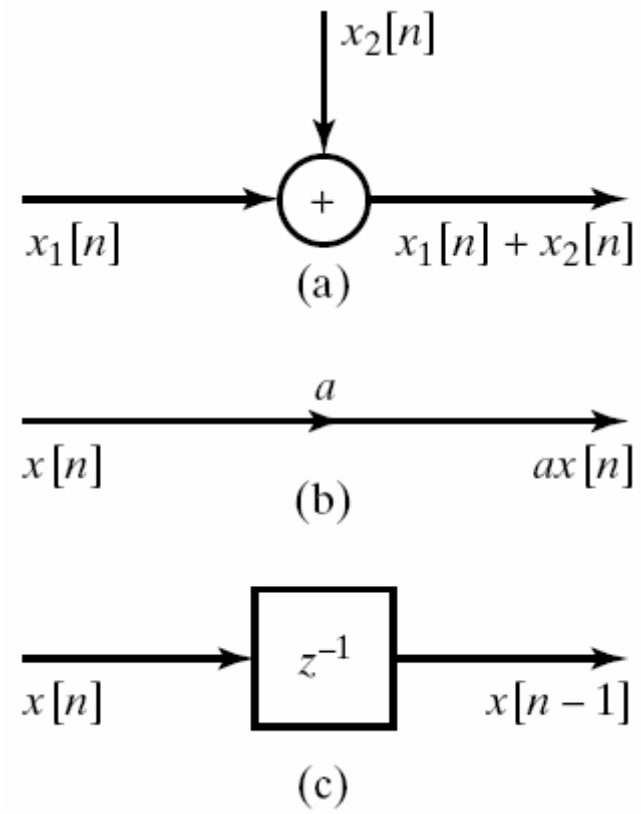
Filter Representation: Block Diagram

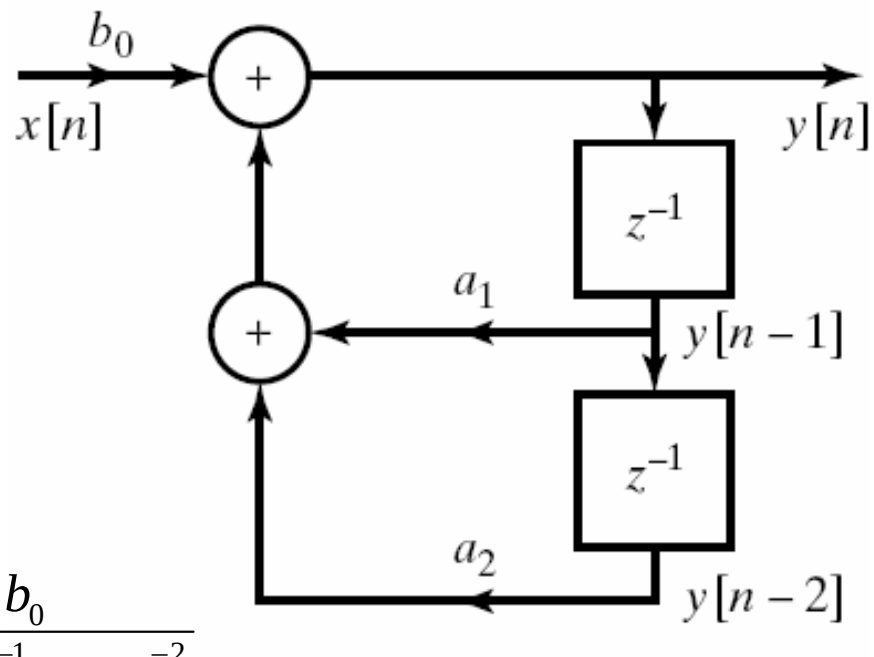
- Three basic block diagrams



(cwliu@twins.ee.nctu.edu.tw)

Example

$$y[n] = a_1 y[n-1] + a_2 y[n-2] + a_3 y[n-3]$$



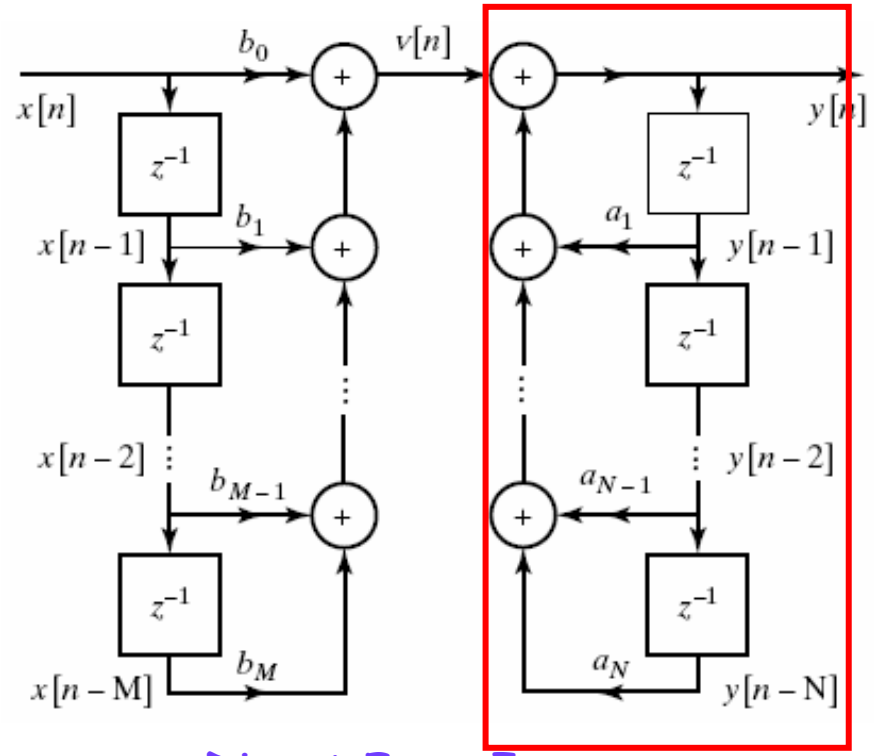
$$H(z) = \frac{Y(z)}{X(z)} = \frac{b_0}{1 - a_1 z^{-1} - a_2 z^{-2}}$$

(cwliu@twins.ee.nctu.edu.tw)

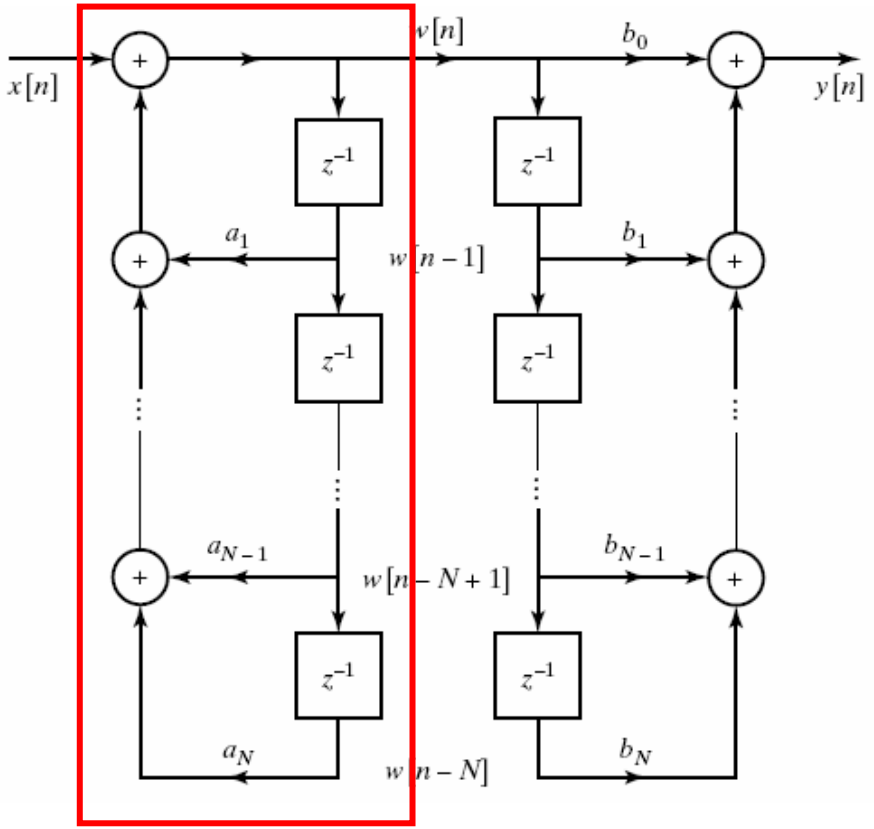
Direct Form: N-th Order Difference Equation

$$y[n]-\sum_{k=1}^{N} a_k y[n-k]=\sum_{k=1}^{M} b_k x[n-k]$$

Re-arrange

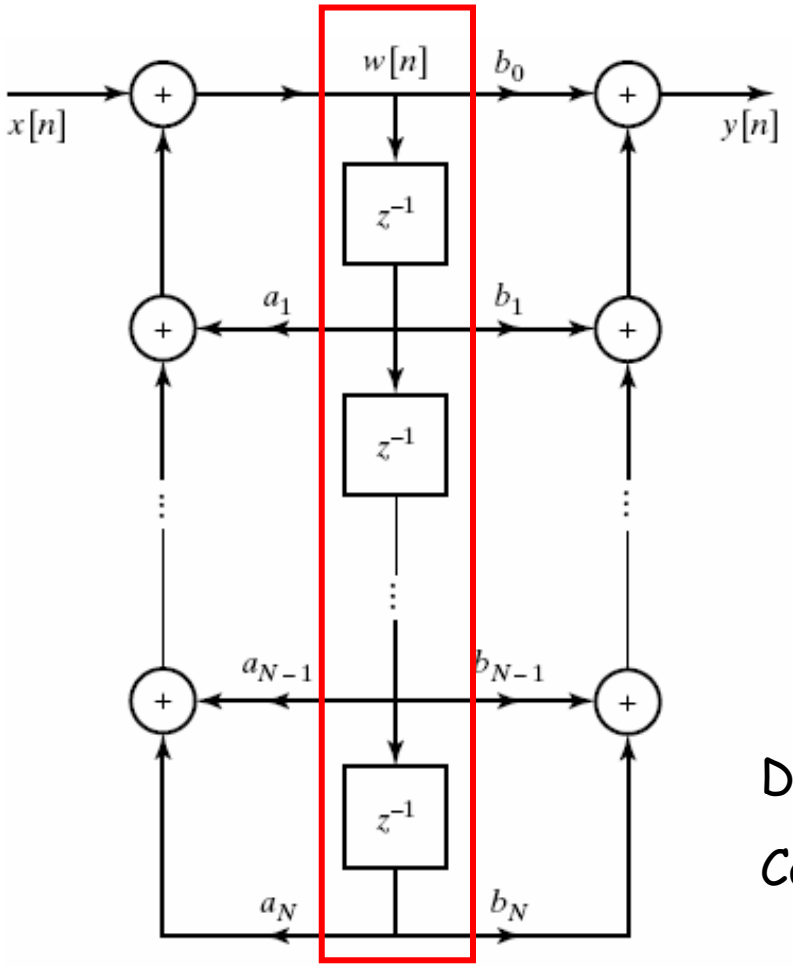


Direct Form I



Canonic Direct Form

$$H(z) = \frac{\sum_{k=1}^{M} b_k z^{-k}}{1 - \sum_{k=1}^{N} a_k z^{-k}}$$



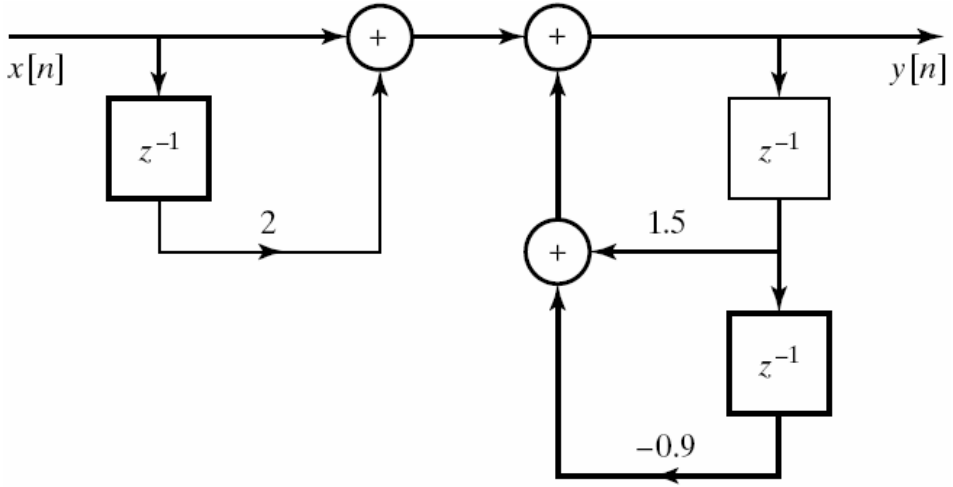
Direct Form II

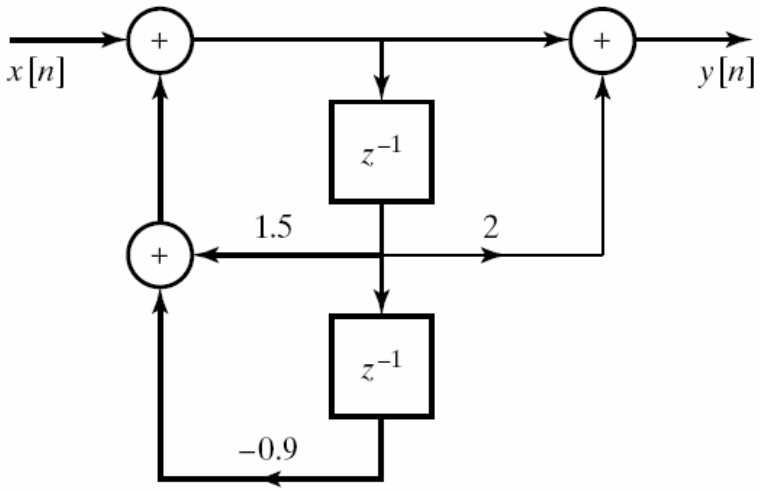
Canonic Direct Form

(cwliu@twins.ee.nctu.edu.tw)

Example

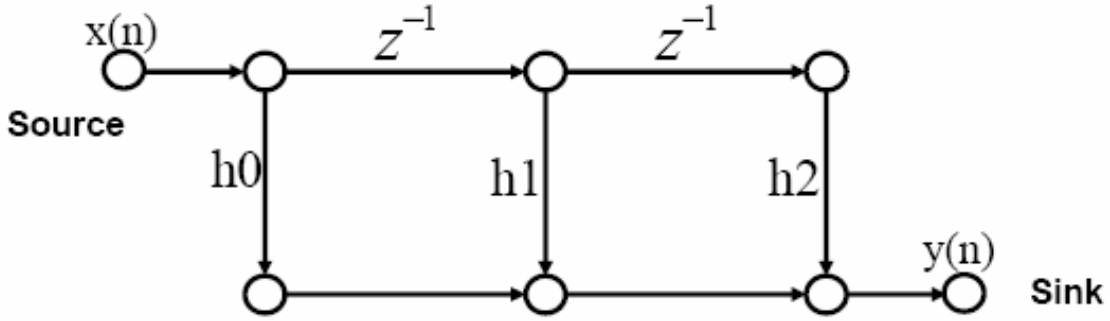
$$H(z)=\frac{1+2z^{-1}}{1-1.5z^{-1}+0.9z^{-2}}$$





Signal-Flow Graph (SFG)

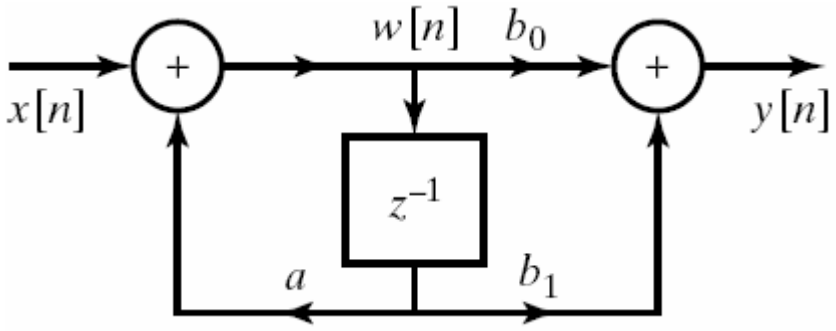
- Nodes: represent computations and/or task
- Directed edge (j,k): denote a linear transformation from the input signal at node j to the output signal at node k
- Source: no entering edge
- Sink: only entering edges

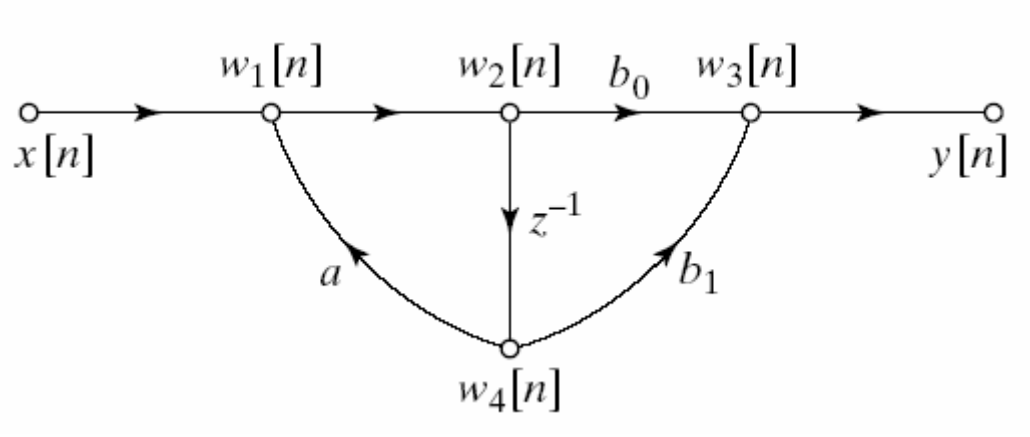


$$y[n] = h_0 x[n] + h_1 x[n-1] + h_2 x[n-2] + h_3 x[n-3]$$

Remarks

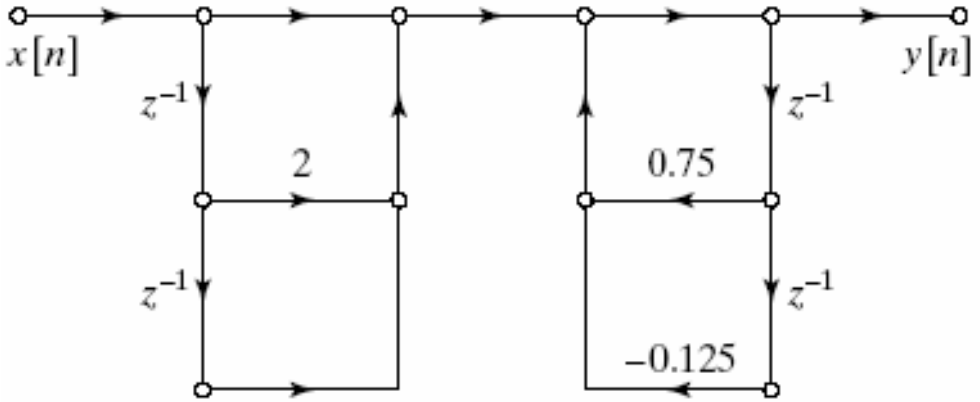
- There is a direct correspondence between branches in the block diagram and branches in the SFG.

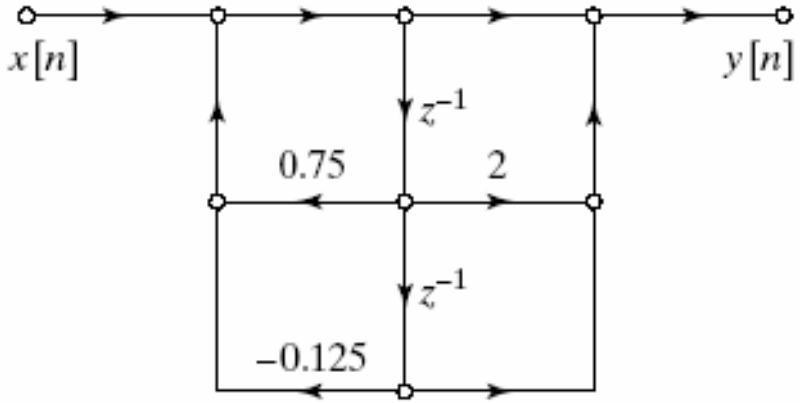




Example: 2-nd Order System

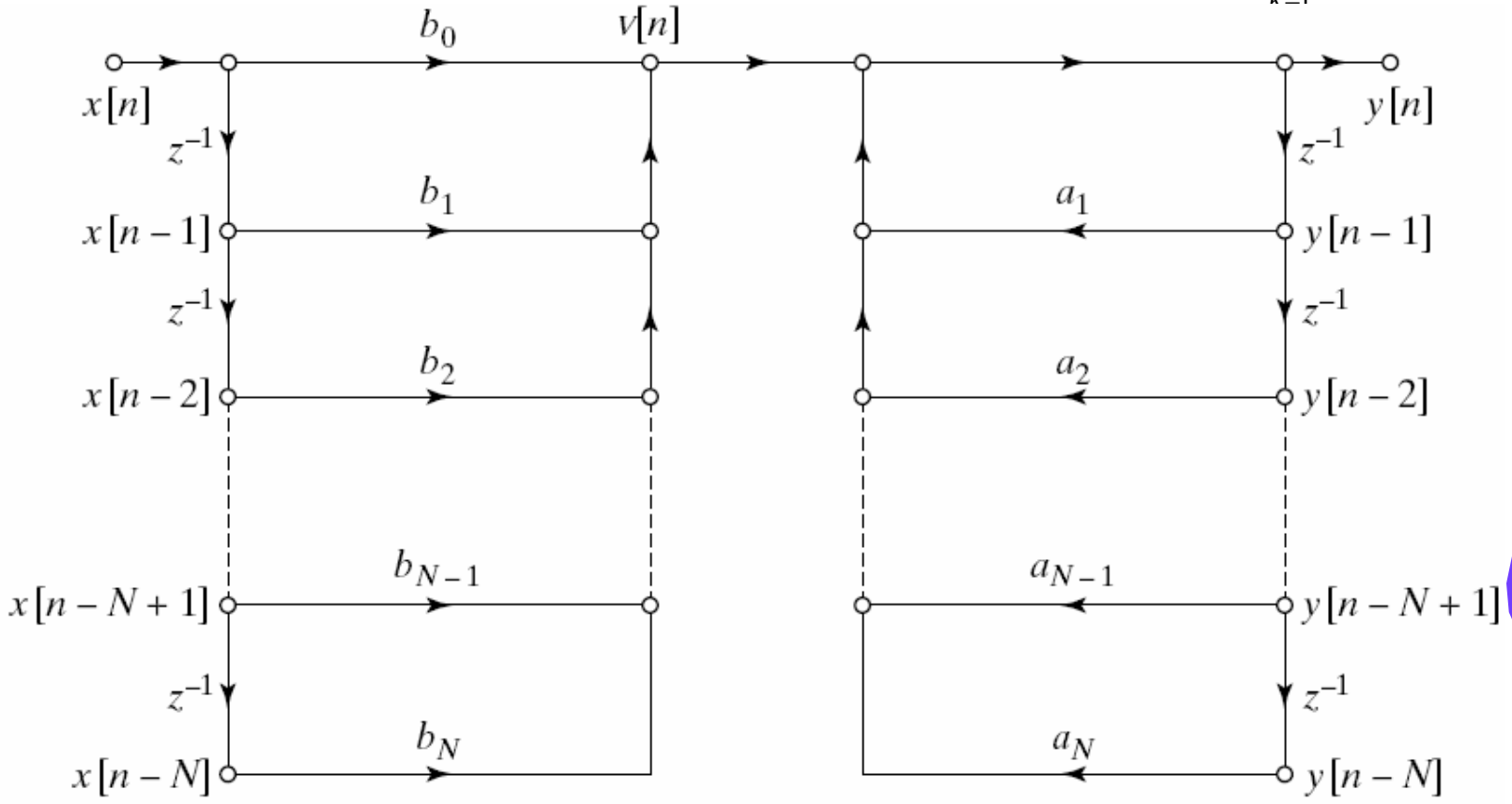
$$H(z)=\frac{1+2z^{-1}+z^{-2}}{1-0.75z^{-1}+0.125z^{-2}}$$



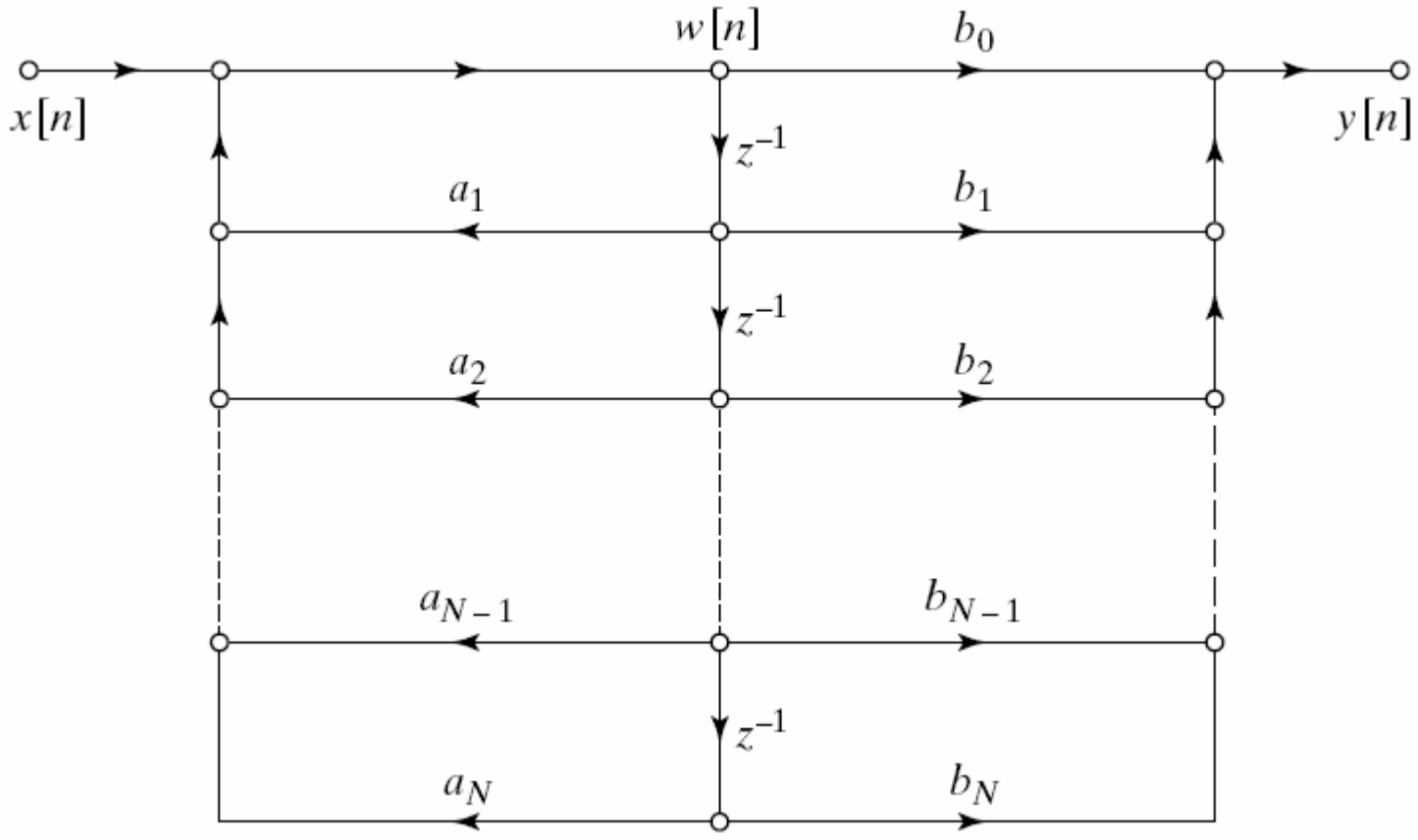


Direct Form: Nth Order System

$$H(z)=\frac{\sum_{k=1}^{M} b_k z^{-k}}{1-\sum_{k=1}^{N} a_k z^{-k}}$$



Direct Form II

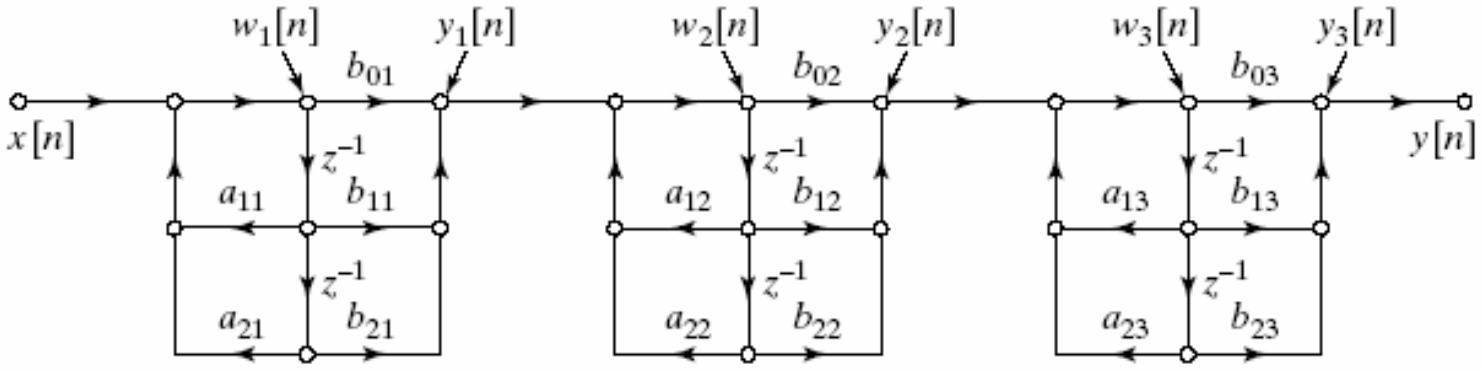


(cwliu@twins.ee.nctu.edu.tw)

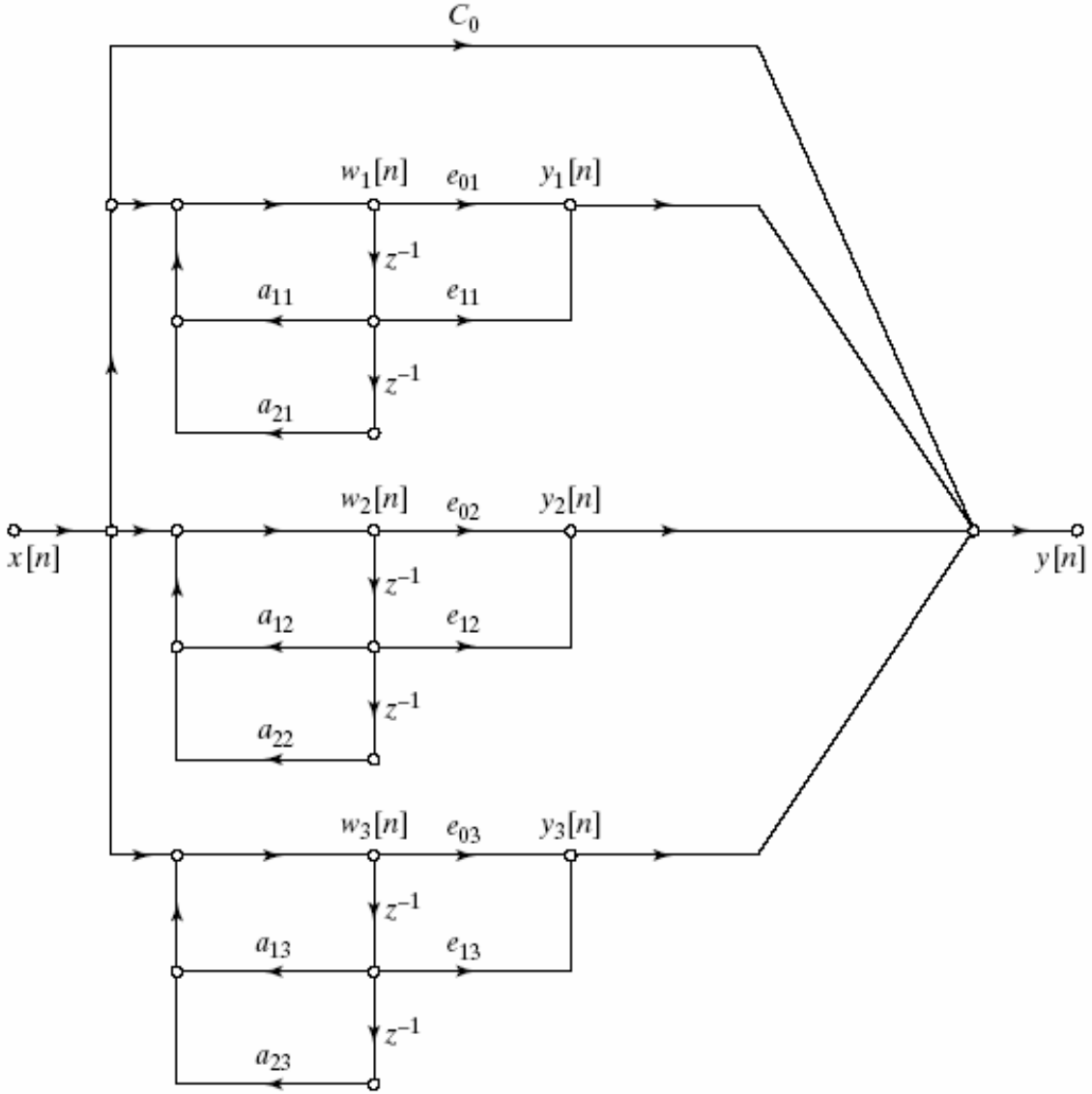
Canonic Cascade Form

- By combining pairs of real factors and complex conjugate pairs into second-order factors
- Example

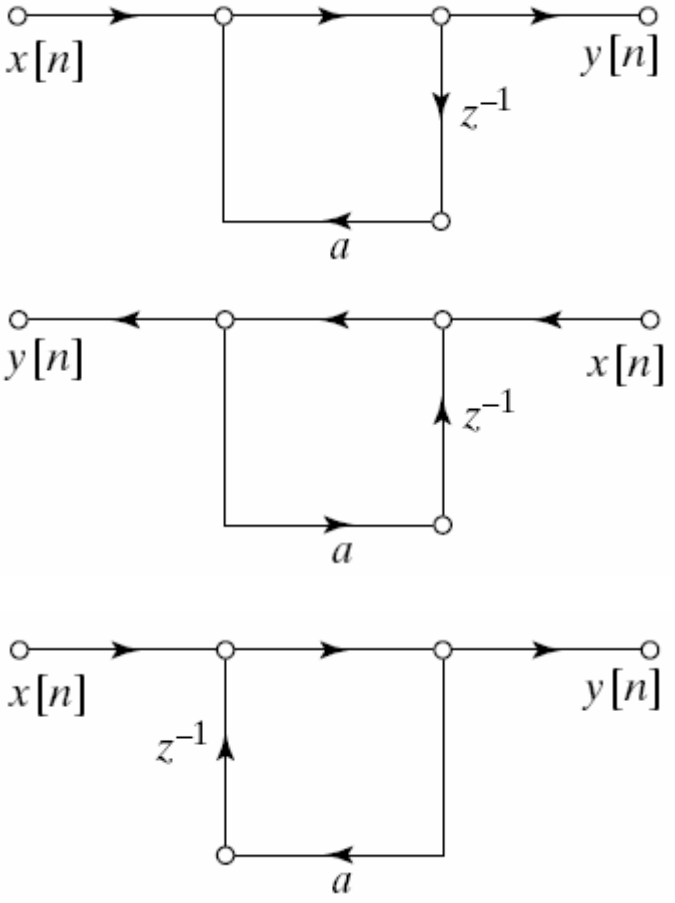
$$H(z)=\prod_{k=1}^{N}\frac{b_{0k}+b_{1k}z^{-1}+b_{2k}z^{-2}}{1-a_{1k}z^{-1}-a_{2k}z^{-2}}$$



Canonic Parallel Form



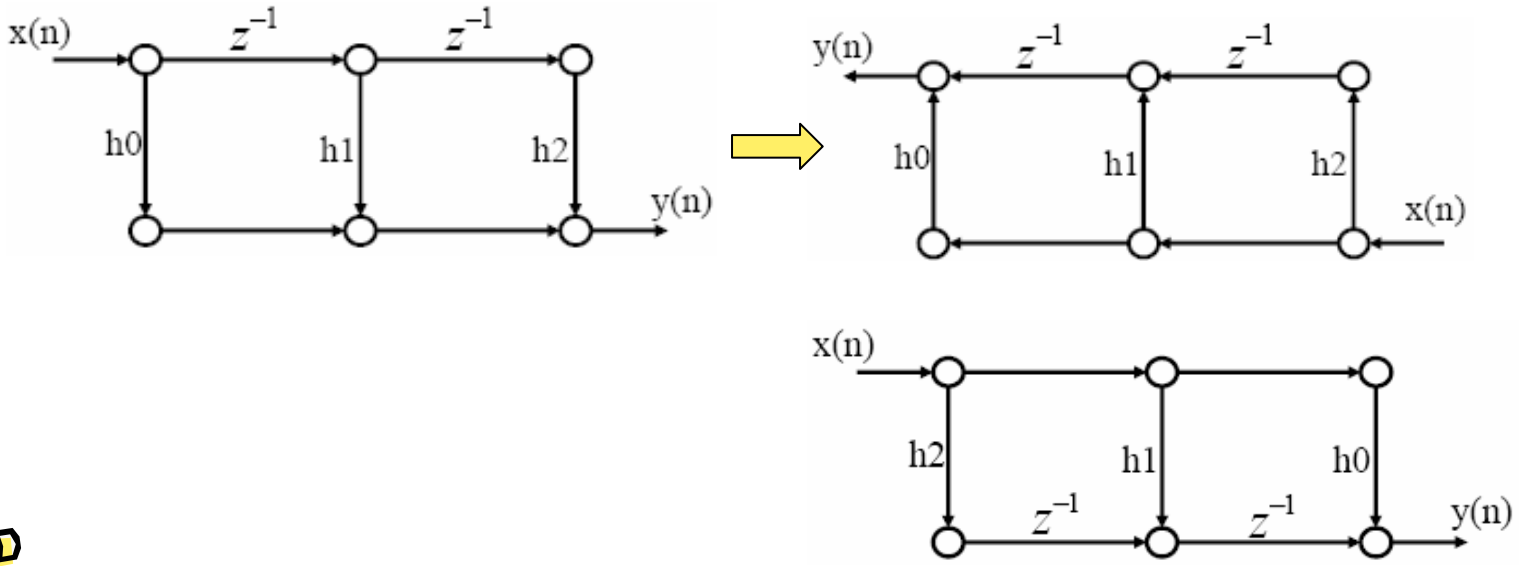
Flow Graph Reversal



(cwliu@twins.ee.nctu.edu.tw)

Transposition of an SFG

- Applicable to linear SISO systems
- Flow graph reversal
  - Reserve the direction of all edges
  - Exchange input and output



Review of Quantization

- (B+1)-bit 2's complement number representation

$$\hat{x}[n] = X_m \hat{x}_B[n]$$

Note: A real number can be represented with infinite precision with $B \to \infty$

$$\hat{x}_B = a_0 \circ a_1 a_2 \cdots a_B$$

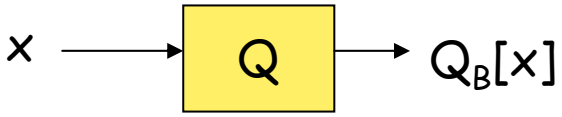
$$\hat{x}_B[n] = -a_0 2^0 + a_1 2^{-1} + \cdots + a_B 2^{-B} \qquad -1 \le \hat{x}_B[n] \le 1$$

$\Delta = \dfrac{2X_m}{2^{B+1}}$ denotes the resolution

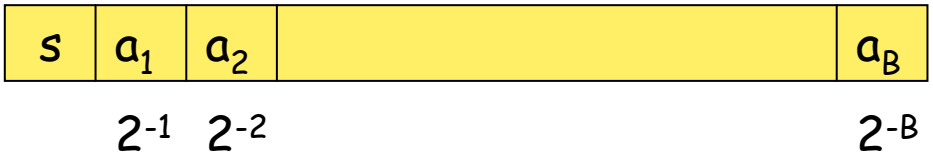
$e = \hat{x} - x = Q_B[x] - x$ denotes the quantization error

X_m denotes the scale factor

Quantization of Fixed-Point Number

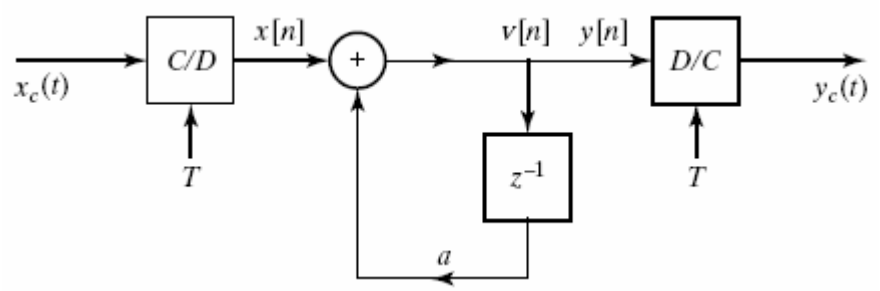


$$e = \hat{x} - x = Q_B[x] - x$$



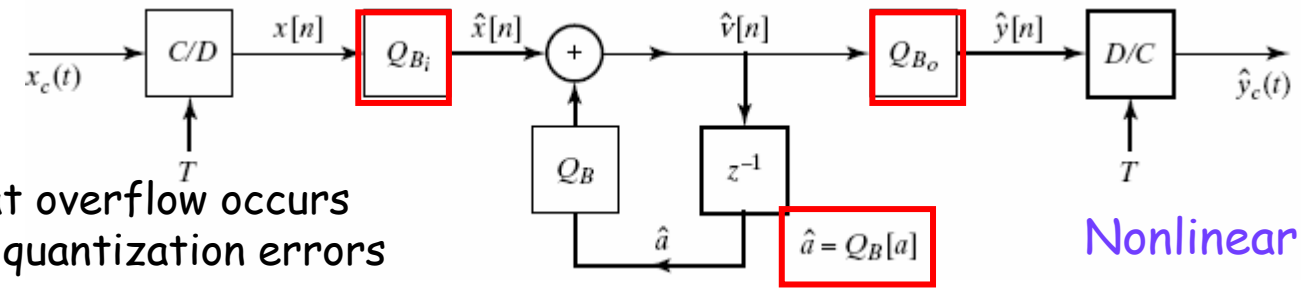
$$0 \le \varepsilon_t \le 2^{-B} - 2^{-\beta}$$

Quantization in Real Example



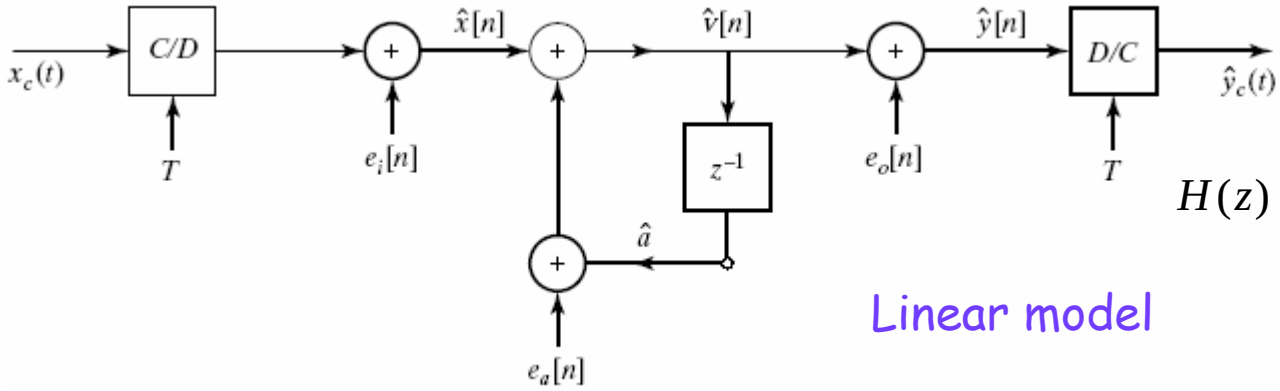
$$H(z) = \frac{1}{1 - az^{-1}}$$

$$\hat{H}(z) = \frac{1}{1 - \hat{a}z^{-1}}$$



Ensure that overflow occurs rarely and quantization errors are small

Nonlinear model



$$H(z) = \frac{1}{1 - \hat{a}z^{-1}}$$

Linear model

Filter Implementation

- Determine the coefficients of the FIR/IIR filter with high accuracy (32-bit floating point).
- Choose an implementation structure
- Coefficient quantization
  - System implementation structure may be highly *sensitive* to perturbations of the coefficients, the resulting system may no longer meet the original design specifications, or IIR system might even become *unstable*.

Effect of coefficient quantization in FIR system

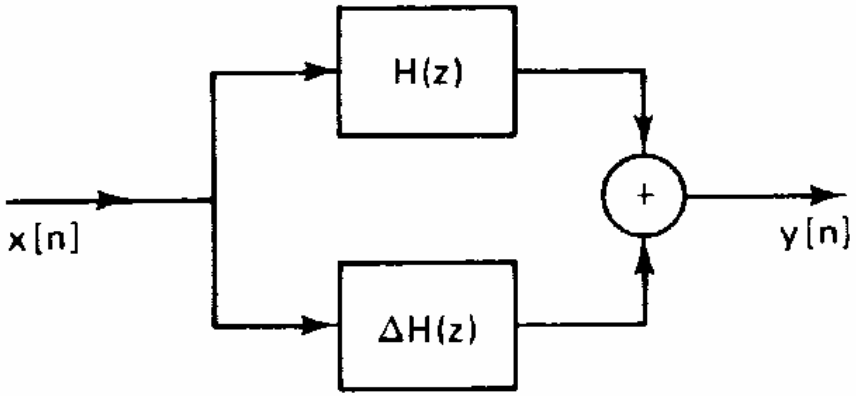
$$H(z) = \sum_{n=0}^{M} h[n] z^{-n}$$

$$\hat{H}(z) = \sum_{n=0}^{M} \hat{h}[n] z^{-n}$$

$$= \sum_{n=0}^{M} (h[n] + \Delta h[n]) z^{-n}$$

$$= \sum_{n=0}^{M} h[n] z^{-n} + \sum_{n=0}^{M} \Delta h[n] z^{-n}$$

$$= H(z) + \Delta H(z)$$

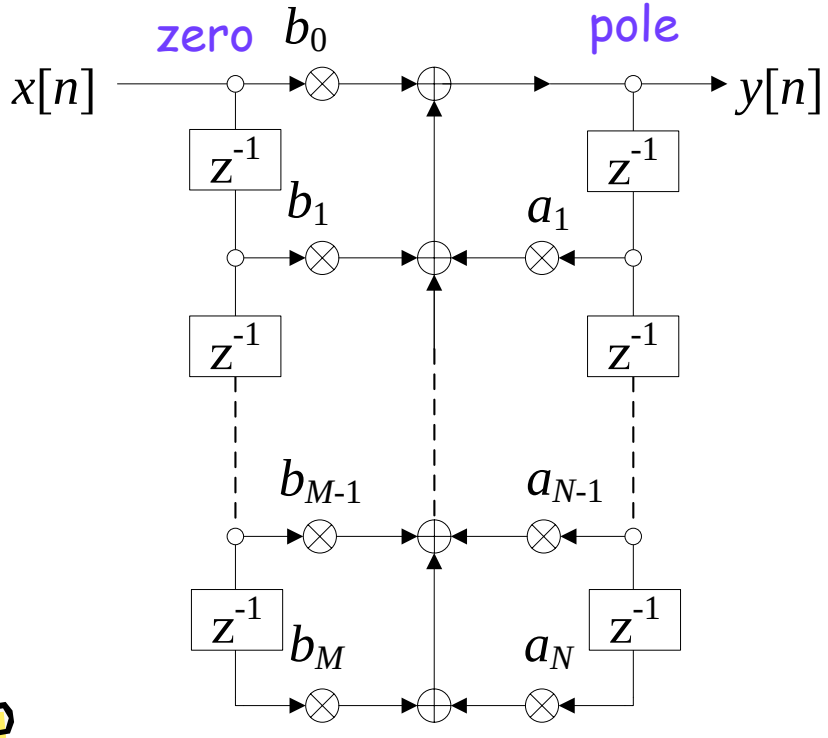


consider as a parallel FIR structure

Effect of Coefficient Quantization in IIR system

$$H(z)=\frac{\sum_{k=0}^{M} b_k z^{-k}}{1-\sum_{k=1}^{N} a_k z^{-k}}$$

$$\hat{H}(z)=\frac{\sum_{k=0}^{M} \hat{b}_k z^{-k}}{1-\sum_{k=1}^{N} \hat{a}_k z^{-k}}$$



$$\hat{a}_k = a_k + \Delta a_k$$

$$\hat{b}_k = b_k + \Delta b_k$$

(cwliu@twins.ee.nctu.edu.tw)

Estimation of Simple Root Displacements

- Consider the Nth-degree polynomial B(z) with simple roots (b_N=1)

$$B(z) = \sum_{i=0}^{N} b_i z^i = \prod_{k=1}^{N} (z - z_k)$$

- Consider the quantized polynomial

$$\hat{B}(z) = \sum_{i=0}^{N} (b_i + \Delta b_i) z^i = \prod_{k=1}^{N} (z - \hat{z}_k)$$

- Now Consider the partial-fraction expansion

$$\frac{1}{B(z)} = \sum_{k=1}^{N} \frac{\rho_k}{z - z_k}, \text{ where } \rho_k = \left. \frac{(z - z_k)}{B(z)} \right|_{z=z_k} = R_k + jX_k$$

Conti'

If we assume that $\hat{z}_k$ is very close to z_k, then $\frac{1}{B(\hat{z}_k)} \cong \frac{\rho_k}{\hat{z}_k - z_k}$ or $\Delta z_k = \rho_k B(\hat{z}_k)$

But, by definition $\hat{B}(z) = \sum_{i=0}^{N} (b_i + \Delta b_i) z^i = B(z) + \sum_{i=0}^{N} (\Delta b_i) z^i$. Then

$0 \approx \hat{B}(\hat{z}_k) = B(\hat{z}_k) + \sum_{i=0}^{N} (\Delta b_i) \hat{z}_k^{\,i}$. Consequently,

$$\Delta z_k = -\rho_k \left(\sum_{i=0}^{N} (\Delta b_i) \hat{z}_k^{\,i} \right) \approx -\rho_k \left(\sum_{i=0}^{N} (\Delta b_i) z_k^{\,i} \right)$$

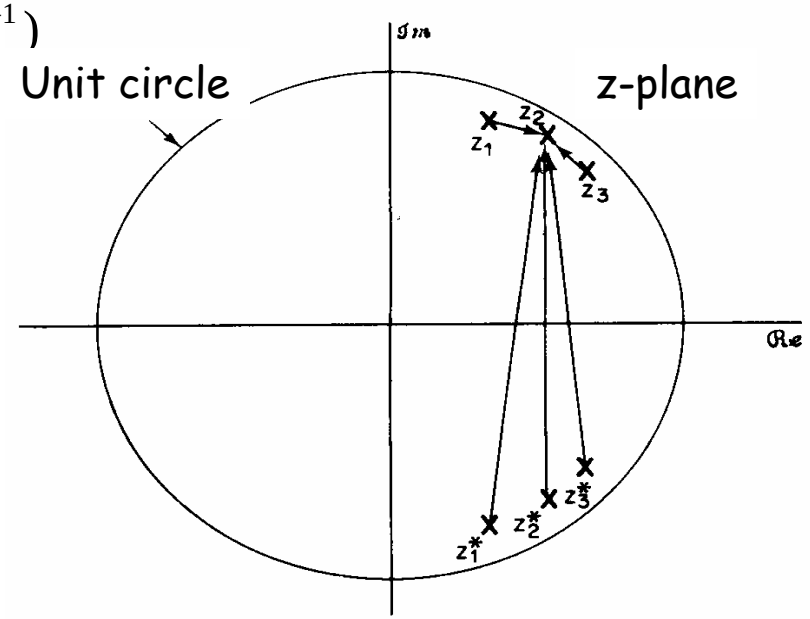
Sensitivity

$$A(z) = 1 - \sum_{k=1}^{N} a_k z^{-k} = \prod_{j=1}^{N} (1 - z_j z^{-1})$$

New pole of $\hat{H}(z)$ are

$$\hat{z}_i = z_i + \Delta z_i \,,\ i = 1,2\ldots N$$

$$\Delta z_i = \sum_{k=1}^{N} \frac{\partial z_i}{\partial a_k} \Delta a_k$$



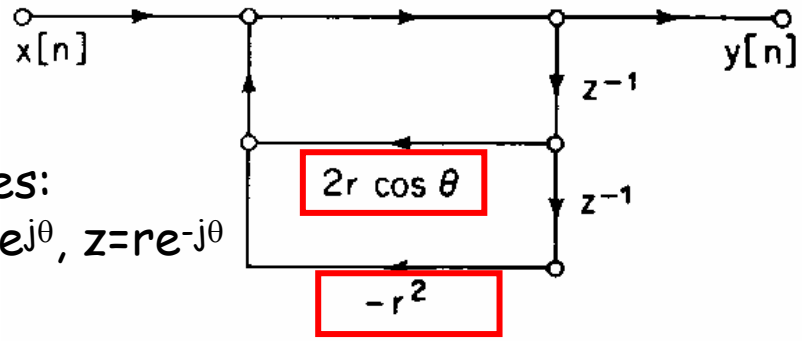
$$\left(\frac{\partial A(z)}{\partial z_i} \right)_{z=z_i} \frac{\partial z_i}{\partial a_k} = \left(\frac{\partial A(z)}{\partial a_k} \right)_{z=z_i}$$

$$\frac{\partial z_i}{\partial a_k} = \frac{z_i^{N-k}}{\prod_{j=1,\, j\neq i}^{N} (z_j - z_i)}$$

Remarks

- The factor z_i-z_j is just a vector in the z-plane
- The magnitude of the denominator is equal to the product of the lengths of the vectors from all the remaining poles to the pole location z_i
- If the poles (or zeros) are tightly clustered, it is possible that small errors in the denominator (or numerator) coefficients may cause large shifts of the poles (or zeros) for the direct form structure.
- Direct form is quite sensitive to quantization errors in the coefficients
- The cascade and parallel 2nd order system forms, however, are less sensitive to quantization errors in the coefficients
- For most linear phase FIR filters, the zeros are more or less uniformly spread in the z-plane

Example: 2nd-order IIR Filter (I)

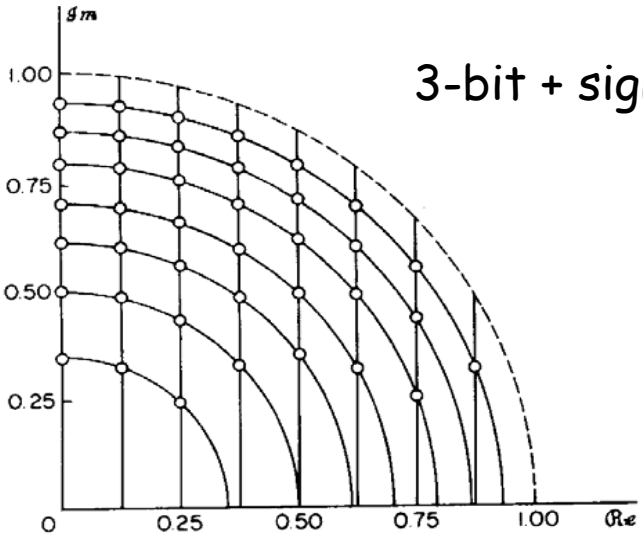


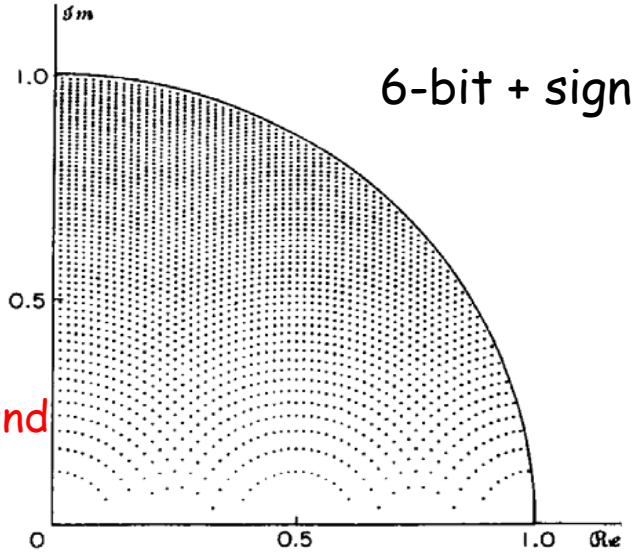
Poles:
$z=re^{j\theta}$, $z=re^{-j\theta}$

$$H(z)=\frac{1}{1+k_1z^{-1}+k_2z^{-2}}$$

$$=\frac{1}{1-2r\cos\theta z^{-1}+r^2z^{-2}}$$

$$\rightarrow\begin{cases}k_1=2\cos\theta\cdot R\\ k_2=R^2\end{cases}$$

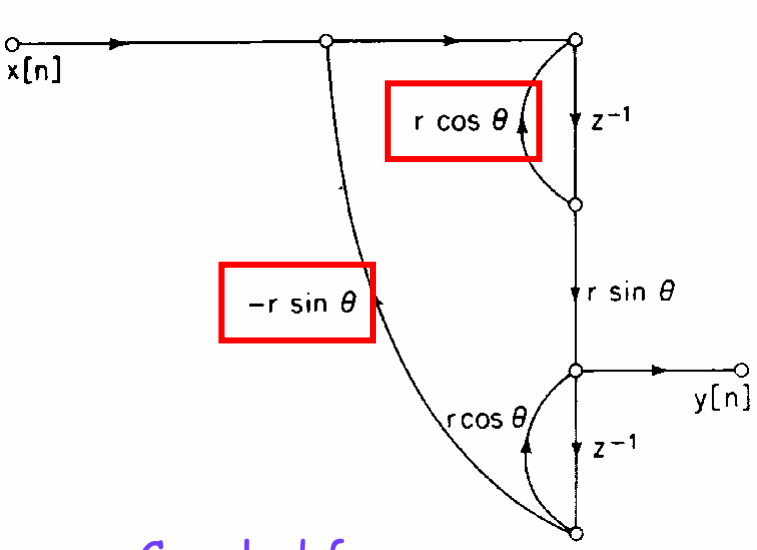




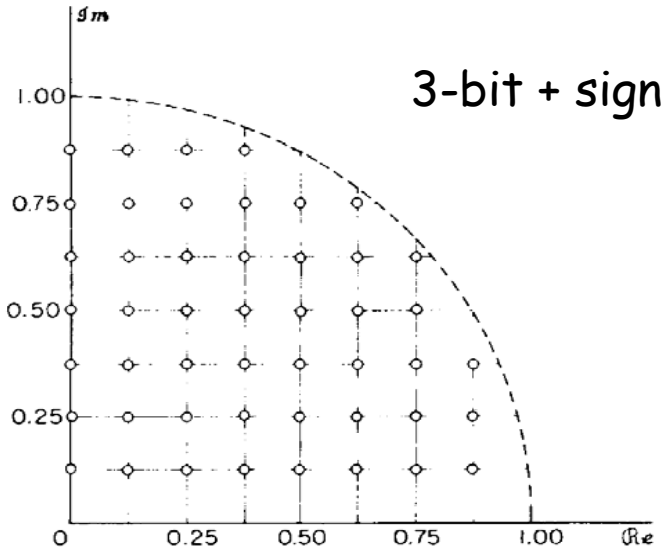
Be spare around the real axis

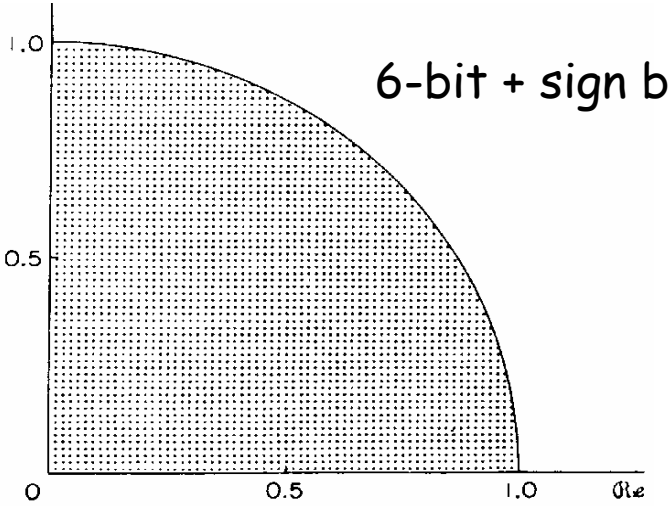
Example:

2nd-order IIR Filter (II)



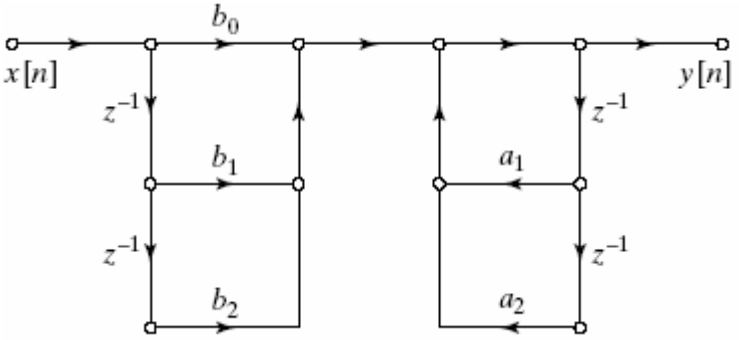
Coupled form

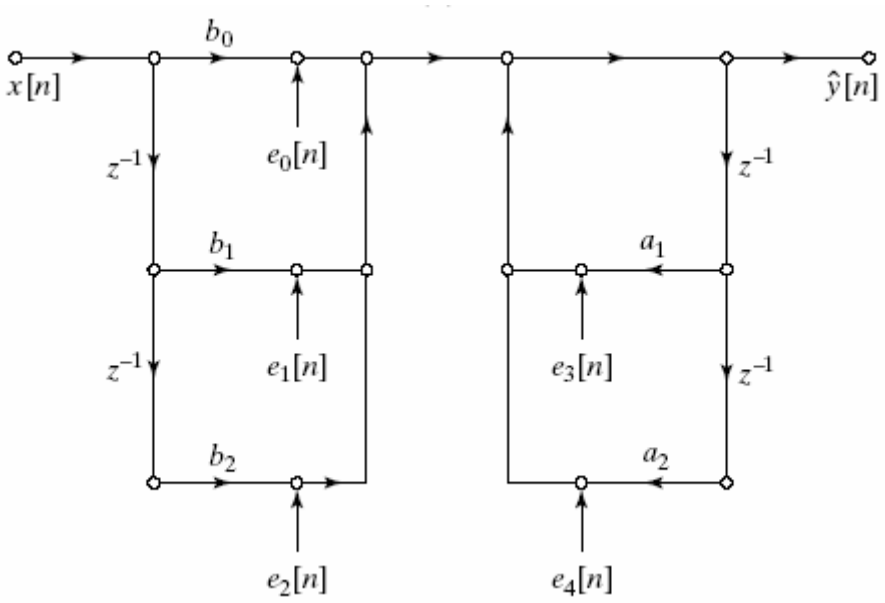




Quantization Noise Sources

- Each quantization noise source is a wide-sense stationary white-noise process
- Each noise has a uniform distribution of amplitudes over one quantization interval
- Each quantization noise source is uncorrelated with the input to the corresponding quantizer, all other quantization noise sources, and the input to the system

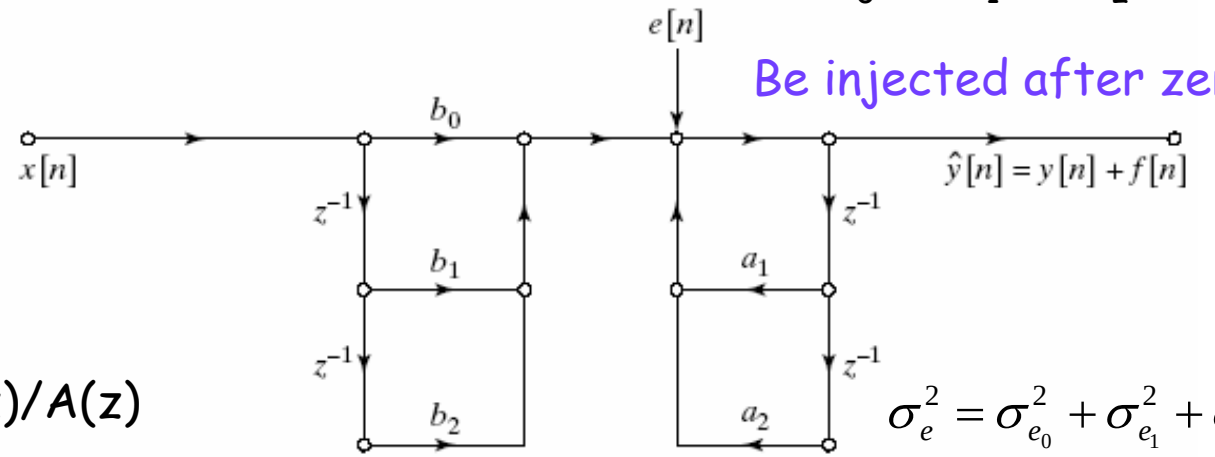




(cwliu@twins.

Linear Noise Model: Direct form I

$e[n]=e_0[n]+e_1[n]+e_2[n]+e_3[n]+e_4[n]$



Be injected after zeros before poles

H(z)=B(z)/A(z)

$$\sigma_e^2=\sigma_{e_0}^2+\sigma_{e_1}^2+\sigma_{e_2}^2+\sigma_{e_3}^2+\sigma_{e_4}^2$$

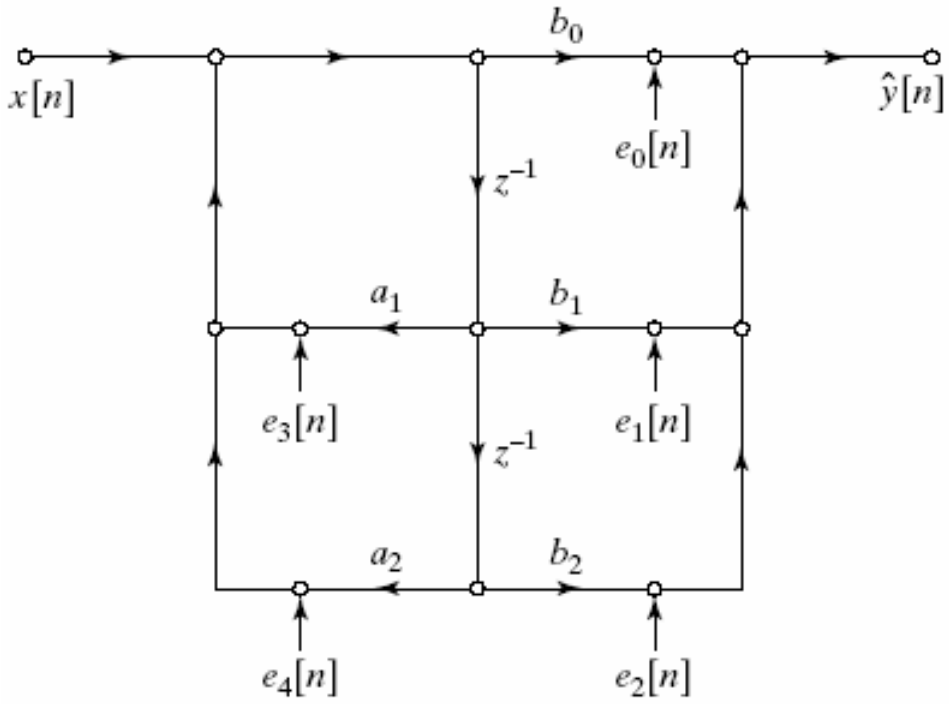
General Case : if $y[n]=\sum_{k=1}^{N}a_k y[n-k]+\sum_{k=0}^{M}b_k x[n-k]$, then

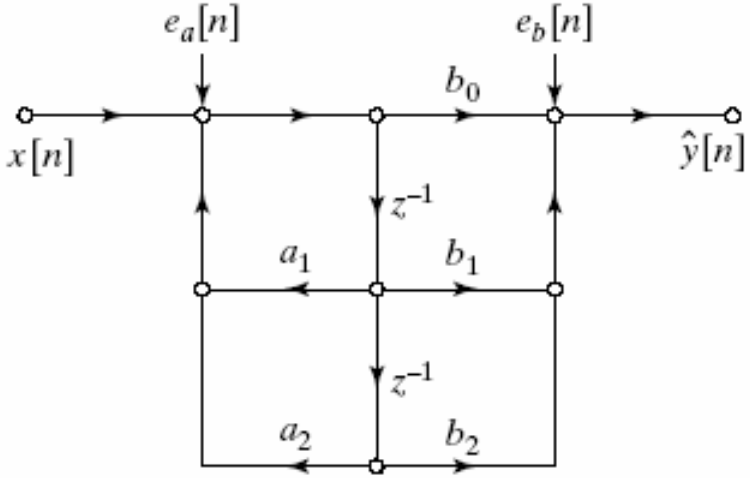
(1) $\sigma_e^2=(M+1+N)\frac{\Delta^2}{12}$

$H_{ef}(z)=1/A(z)$

(2) $f[n]=\sum_{k=1}^{N}a_k f[n-k]+e[n],\quad \sigma_f^2=(M+1+N)\frac{\Delta^2}{12}\sum_{n=-\infty}^{\infty}\left|h_{ef}[n]\right|^2$

Linear Noise Model: Direct Form II





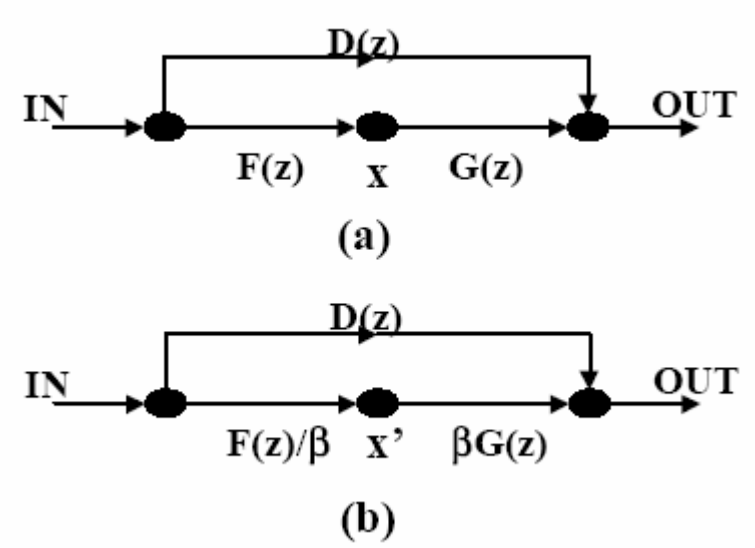
$$\sigma_f^2 = N\frac{\Delta^2}{12}\sum_{n=-\infty}^{\infty}|h[n]|^2 + (M+1)\frac{\Delta^2}{12}$$

Quantization Effects

- Quantization error in a given coefficient affects all the poles/zeros of the system function.
- Signal rounding
  - Limit-cycle oscillation: undesirable periodic components at the filter output (e.g. oscillations when there exist nonlinear operations in feedback paths)
  - Roundoff noise: random disturbance
- Signal scaling
  - Internal overflows
  - Scaling constraints the numerical values of the internal filter variables to remain in a range appropriate to the hardware

Scaling Operation

- To readjust certain internal gain parameters



$H(z)=D(z)+F(z)G(z)$

x': scaled node

- Example: to prevent overflow at x, one may choose (if the input is bounded by $|u(n)|\le 1$)

$$\beta=\sum_{i=0}^{\infty}|f(i)| \qquad |x(n)|=\left|\sum_{i=0}^{\infty}f(i)u(n-i)\right|\le\sum_{i=0}^{\infty}|f(i)|$$

Overflow and Scaling Factor

Let $w_k[n]$ denote the value of the kth node variable and $h_k[n]$ denote the impulse response from the input x to the node w, then

$$|w_k[n]| = \left| \sum_{m=-\infty}^{\infty} x[n-m]h_k[m] \right| \le x_{\max} \sum_{m=-\infty}^{\infty} |h_k[m]| < 1$$

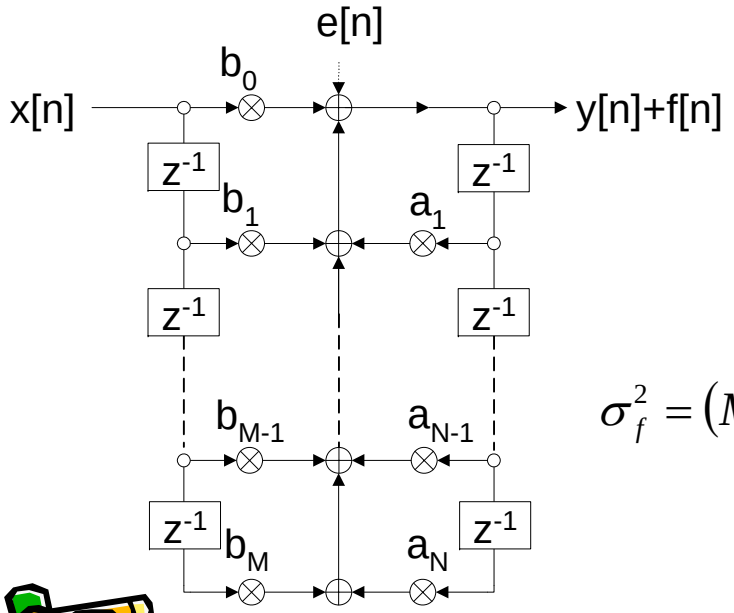
to prevent overflow

Choose scaling factor s, such that

$$sx_{\max} < \frac{1}{\max_k \left[\sum_{m=-\infty}^{\infty} |h_k[m]| \right]} \qquad sx_{\max} < \frac{1}{\max_{k,|\omega| \le \pi} |H_k(e^{jw})|}$$

Improve the Noise Performance

- The use of a double-length accumulator
- For direct form I architecture
  - The sums of produces are accumulated with 2B+1 or 2B+2-bit accuracy and the result is quantized to B+1 bits



$$\hat{y}[n] = Q\left[\sum_{k=1}^{N} a_k \hat{y}[n-k] + \sum_{k=0}^{M} b_k x[n-k]\right]$$

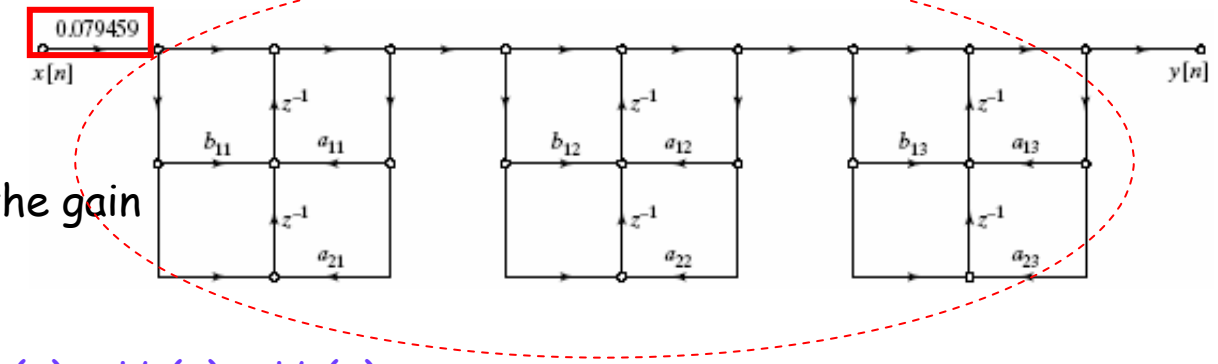
$$\sigma_f^2 = (M+1+N)\frac{\Delta^2}{12}\sum_{n=-\infty}^{\infty}\left|h_{ef}[n]\right|^2 \qquad \Rightarrow \sigma_f^2 = \frac{\Delta^2}{12}\sum_{n=-\infty}^{\infty}\left|h_{ef}[n]\right|^2$$

Example: 6th-Order Cascade System

prevent overflow

$$s \max_{|\omega|\le\pi} |H(e^{j\omega})| < 1$$

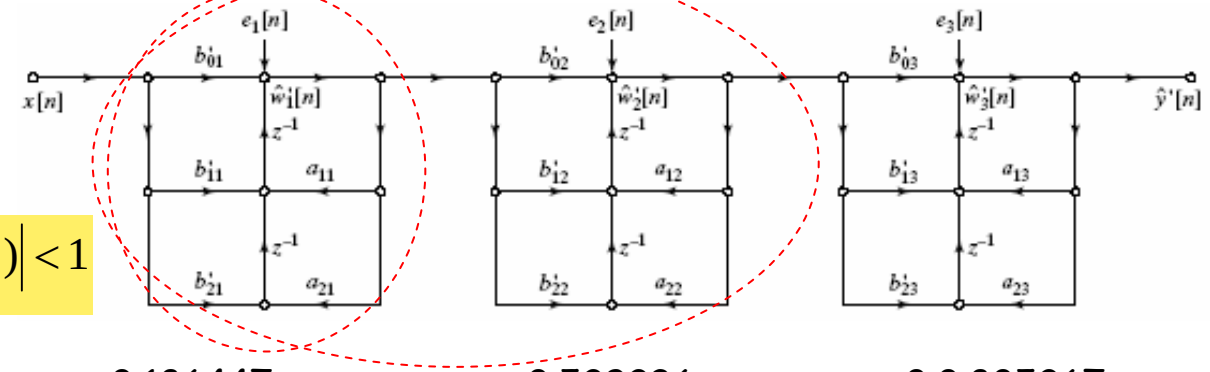
Distribute the gain



$H(z)=s_1H_1(z)\ s_2H_2(z)\ s_3H_3(z)$

$s_1s_2s_3=0.079459$

$$s_1 s_2 \max_{|\omega|\le\pi} |H_1(e^{j\omega})H_2(e^{j\omega})| < 1$$



$$s_1 \max_{|\omega|\le\pi} |H_1(e^{j\omega})| < 1$$

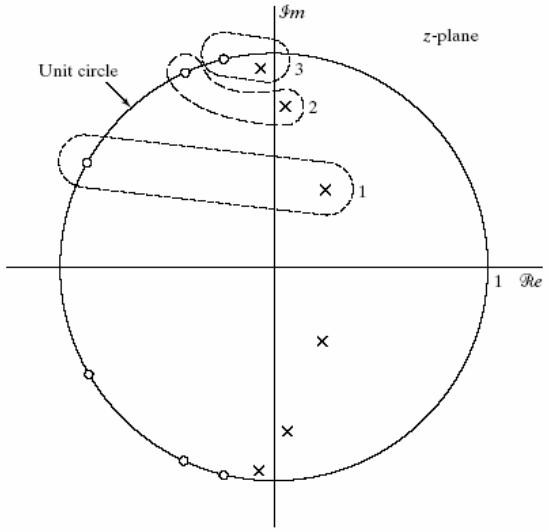
s_1=0.186447 s_2=0.529236 s_3=0.0.805267

Shape the Output Noise Power

- Depend on the way to separate the pairs of zeros and poles to form the 2nd-order sections
- Depend on the order of the 2nd-order sections in the cascade form
- N_s sections: a total of $(N_s!)^2$ different systems
- Simple rule:
  - The pole that is closet to the unit circle should be paired with the zero that is closet to it in the z-plane
  - Rule 1 should be repeatedly applied until all the poles and zeros have been paired
  - The resulting 2^{nd}-order sections should be ordered according to the closeness of the poles to the unit circle, either in order of increasing closeness to the unit circle or in order of decreasing closeness to the unit circle

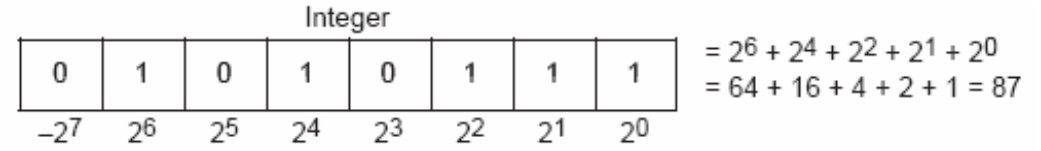
Cascade IIR Structure

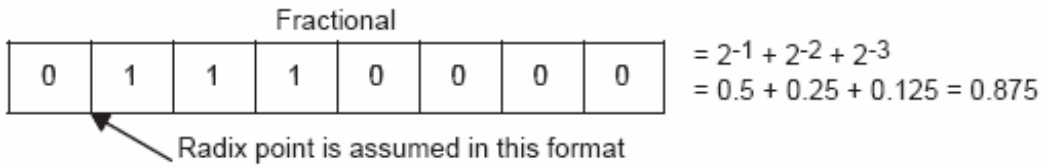
- The subsystem with high peak gain are undesirable because they can cause overflow (also amplify quantization noise)
- Moving high Q poles to the beginning of the cascades
- To avoid excessive reduction of the signal level in the early stages of cascade, we should place the poles that are close to the unit circle last in order



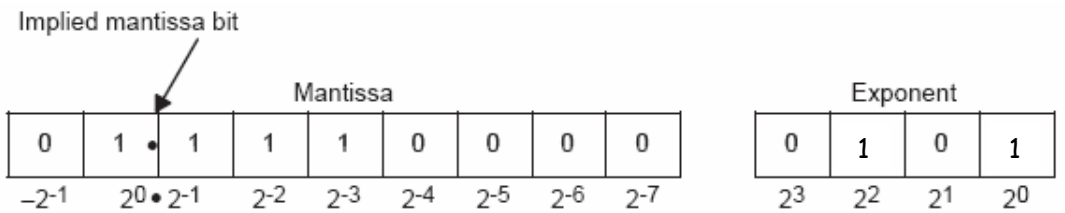
Floating-Point Realization

- Fixed-point representation
  - Integer
  - Fractional





- Floating-point representation
  - Mantissa
  - Exponent



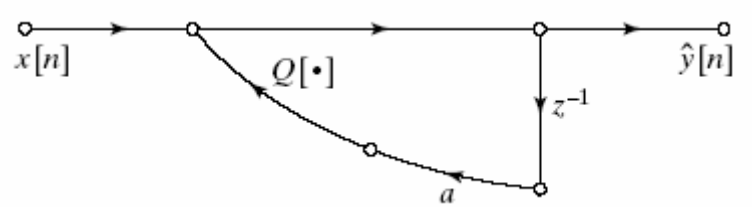
Mantissa = $2^0 + 2^{-1} + 2^{-2} + 2^{-3} = 1 + 0.5 + 0.25 + 0.125 = 1.875$

Exponent = $2^2 + 2^0 = 4 + 1 = 5$

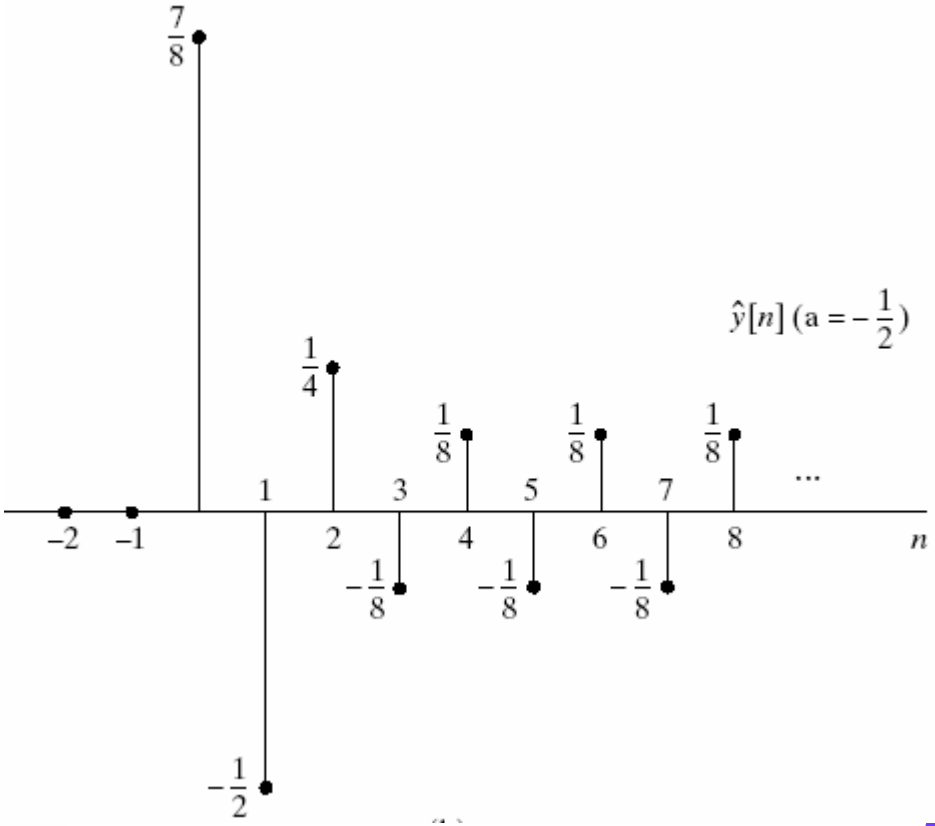
Decimal value = $1.875 \times 2^5 = 60$

a wide dynamic range

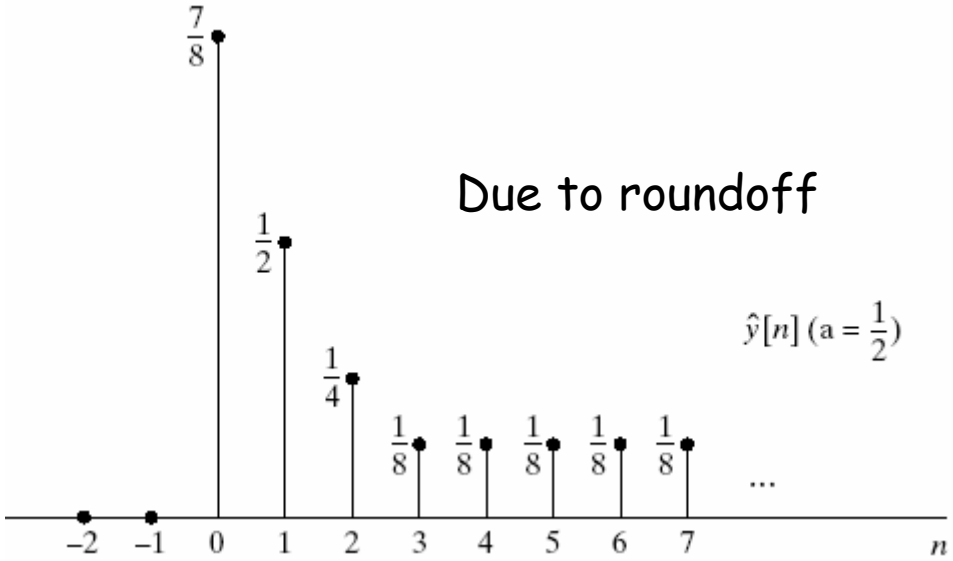
Limit Cycles Oscillation



Rounded to 4-bit



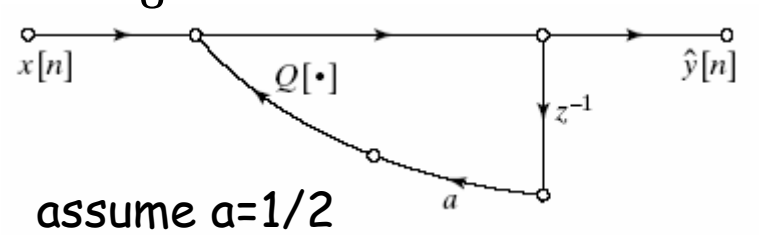
Due to roundoff



steady state periodic outputs

Limit Cycles due to Roundoff

$$x[n]=\frac{7}{8}\delta[n]$$



assume a=1/2

$$\hat{y}[n]=Q\left[a\hat{y}[n-1]\right]+x[n]$$

$$a=\frac{1}{2}=0\circ 100$$

$$x[n]=(0\circ 111)\delta[n]$$

3/8

7/16 → 4/8 to round upward

$$\hat{y}[0]=0\circ 111$$

$$\hat{y}[1]=Q\left[a\hat{y}[0]\right]=Q[0\circ 011100]=0\circ 100$$

$$\hat{y}[2]=Q\left[a\hat{y}[1]\right]=Q[0\circ 010000]=0\circ 010$$

$$\hat{y}[3]=Q\left[a\hat{y}[2]\right]=Q[0\circ 001000]=0\circ 001$$

$$\hat{y}[4]=Q\left[a\hat{y}[3]\right]=Q[0\circ 000100]=0\circ 001$$

to round upward

This is a first order pole at z=1 instead of z=1/2

Limit Cycles due to Overflow

2nd order IIR system

$$\hat{y}[n] = x[n] + Q[a_1 \hat{y}[n-1]] + Q[a_2 \hat{y}[n-2]]$$

Assumptions

$$a_1 = 3/4 = 0 \circ 110$$
$$a_2 = -3/4 = 1 \circ 010$$
$$\hat{y}[-1] = -3/4 = 1 \circ 010$$
$$\hat{y}[-2] = -3/4 = 1 \circ 010$$
$$x[n] = 0, \ \forall n \geq 0$$

$$\hat{y}[0] = 0 \circ 111 \times 0 \circ 111 + 1 \circ 010 \times 1 \circ 010$$
$$= 0 \circ 100100 + 0 \circ 100100$$
$$= 0 \circ 101 + 0 \circ 101 \quad \text{round upward}$$
$$= 1 \circ 010 \quad \text{2's complement addition}$$

$$\hat{y}[1] = 1 \circ 011 + 1 \circ 011 = 0 \circ 110$$

$\vdots$

Conclusions

- Fixed-point analysis is required in converting floating-point based DSP algorithm into fixed-point based implementation (e.g. VLSI circuits & fixed-point Programmable DSP processors)
- Usually it is done by doing extensive simulations
- Closed-form analytical results help to see the effectiveness of each design parameters (W, S, etc.)
- Each algorithm has its own numerical property in fixed-point implementation