Finite Wordlength Effects

Binary number representation

•Fixed Point

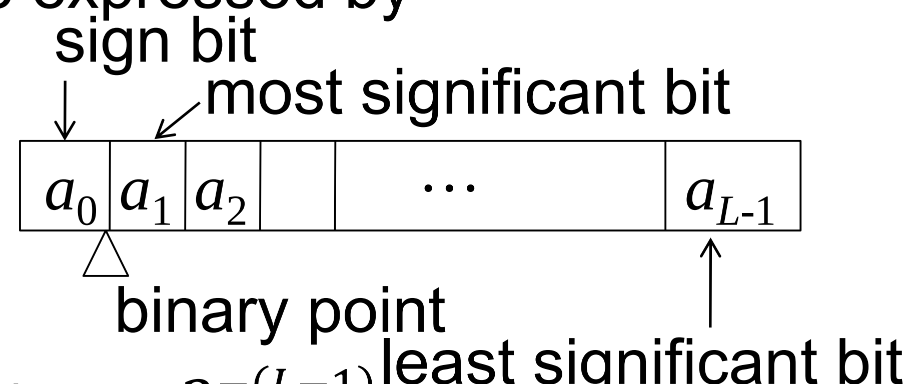
$$f = a_0.a_1a_2\cdots a_{L-1}$$

negative numbers can be expressed by

•two’s complement

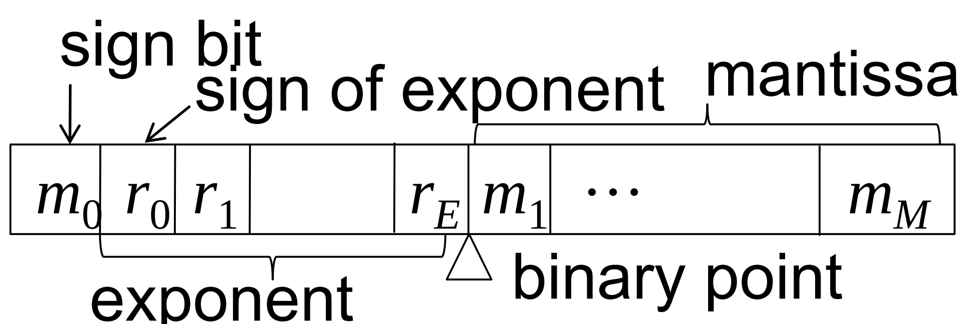
•one’s complement

•sign magnitude



$$f_{(10)} = -a_0 + a_1 2^{-1} + \cdots + a_{L-1} 2^{-(L-1)}$$

•Floating Point

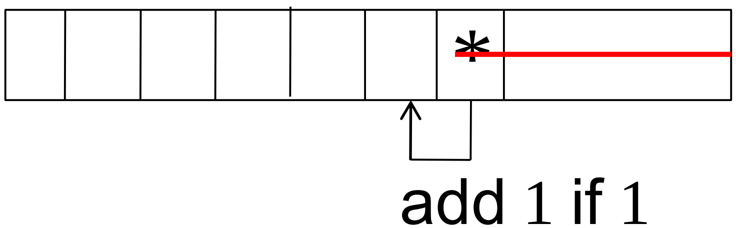


Quantization

- Reduction of Wordlength
- roundoff

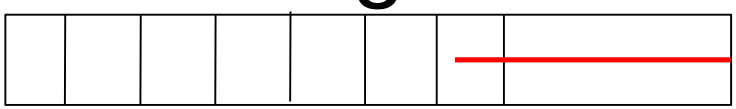
  add 1 to LSB if the next bit is 1

  do nothing if 0



- truncation

  truncate the following bits



Quantization Error

When a number is rounded to *B* bit

- Fixed Point

$$\varepsilon = Y - X$$

$$-\frac{Q}{2} < \varepsilon \le \frac{Q}{2}, \quad Q = 2^{-B}$$

- Floating Point

$$\varepsilon = (Y - X)/X$$

$$-\frac{Q}{2} < \varepsilon \le \frac{Q}{2}$$

Finite Wordlength Effects

- deviation of frequency response
  due to finite-wordlength coefficients

  ⇩

  sensitivity

- round off operation
  - ✧noise ········· uncorrelated error : random
  - ✧limit cycles ········· correlated

- overflow

all are related to **network structure** (direct, cascade ......)

Round off Operation

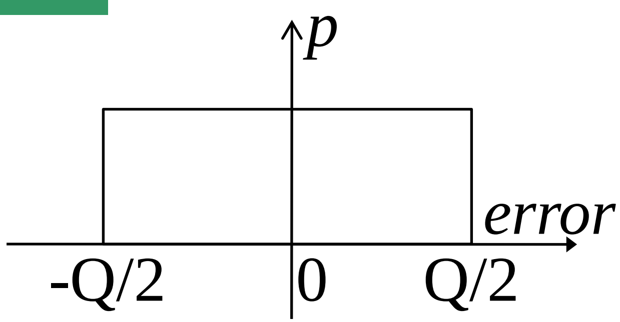
n-bit $\times$ n-bit $=$ $2n$-bit

full precision cannot be kept

$\Downarrow$ n-bit

quantization
- roundoff
- truncation

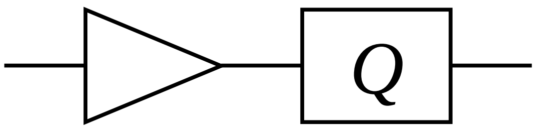
$|\text{error}| \leq \frac{Q}{2}$ $\qquad$ Q: stepsize

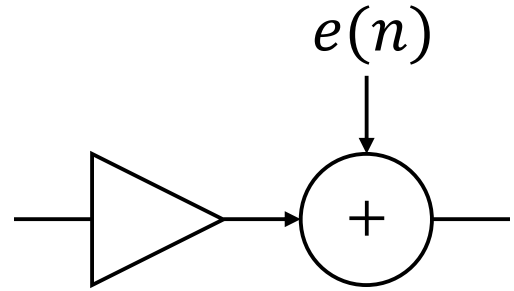


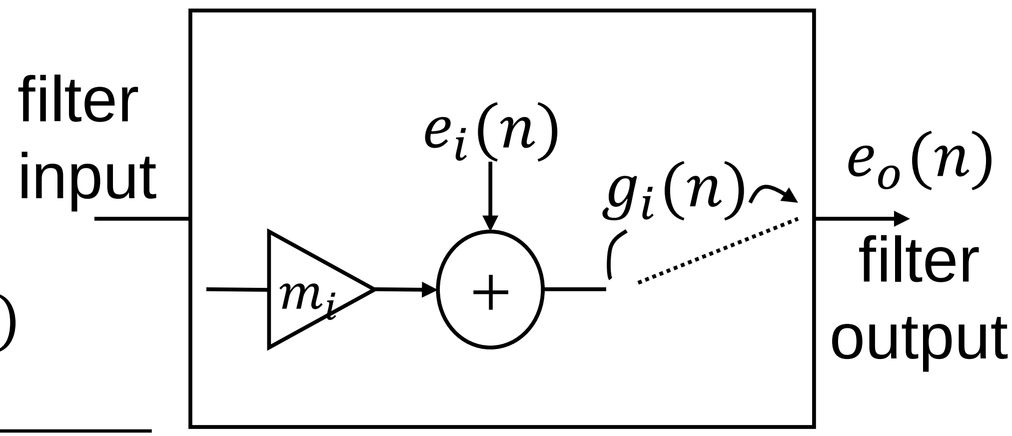
for uniform distribution, zero mean and

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 p(x) dx = \int_{-Q/2}^{Q/2} \frac{1}{Q} x^2 dx = \frac{1}{Q} \left[\frac{x^3}{3} \right]_{-Q/2}^{Q/2} = \frac{Q^2}{12}$$

Linear Model of Quantization







$$e_o(n) = \sum_{k=-\infty}^{\infty} g_i(n-k)e_i(k)$$

$$\overline{e_o^2(n)} = \overline{\left(\sum_{k=-\infty}^{\infty} g_i(n-k)e_i(k)\right)^2}$$

$$= \sum_{n=-\infty}^{\infty} g_i^2(n)\sigma^2$$

Noise Gain

using Parseval's equation

$$\sigma_i^2 = \overline{e_o^2(n)} = \sigma^2 \frac{1}{2\pi j} \oint G_i(z) G_i(z^{-1}) \frac{dz}{z}$$

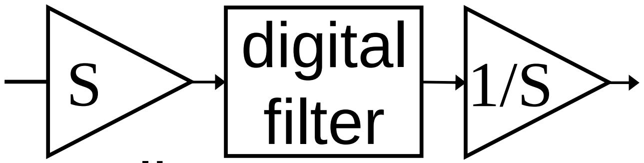
For multiple noise sources

if noises are uncorrelated

$$\sigma_o^2 = \sum_{i=1}^{M} \sigma_i^2 = \sigma^2 \underbrace{\sum_{i=1}^{M} \frac{1}{2\pi j} \oint G_i(z) G_i(z^{-1}) \frac{dz}{z}}_{\text{noise gain}}$$

Scaling

adjust internal signal level so that the output SN ratio is maximized



scaling
multiplier

Scaling strategy

- no overflow for any input sequence

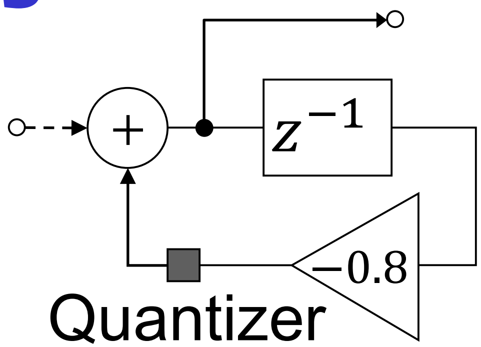
---too pessimistic

- l_p norm scaling $$\|F_i\|_p = \left(\frac{1}{\omega_s}\int_0^{\omega_s}\left|F_i\left(e^{j\omega T}\right)\right|^p d\omega\right)^{1/p}$$

$F_i(z)$: transfer function from the input to the internal node i

Example of Limit Cycles

zero input case



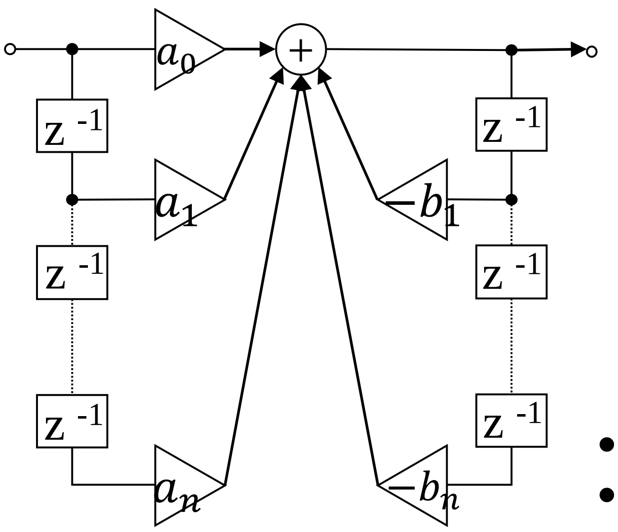
absolute value rounded to integer

$$\boxed{y(n) = [-0.8y(n-1)]_Q, \qquad y(0) = 10}$$

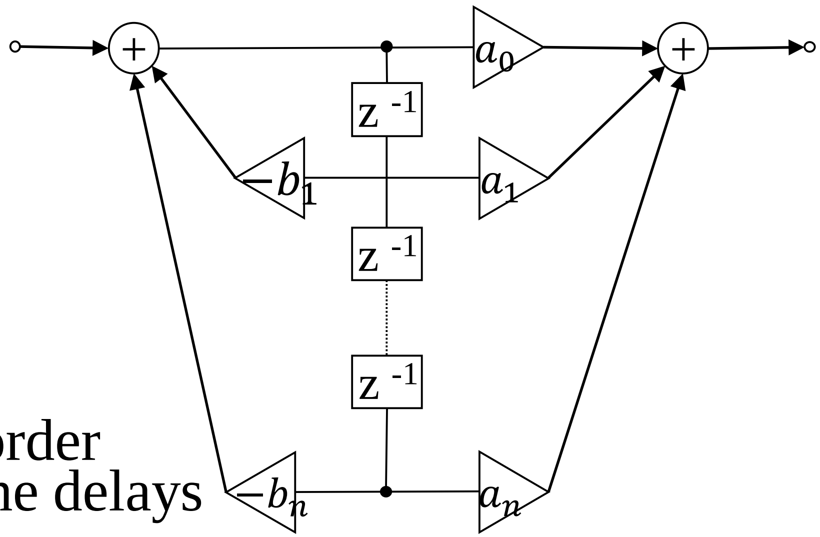
$$y(1) = [-0.8y(0)]_Q = [-0.8 \times 10]_Q = [-8]_Q = -8$$
$$y(2) = [-0.8y(1)]_Q = [-0.8 \times (-8)]_Q = [6.4]_Q = 6$$
$$y(3) = [-0.8y(2)]_Q = [-0.8 \times 6]_Q = [-4.8]_Q = -5$$
$$y(4) = [-0.8y(3)]_Q = [-0.8 \times (-5)]_Q = [4]_Q = 4$$
$$y(5) = [-0.8y(4)]_Q = [-0.8 \times 4]_Q = [-3.2]_Q = -3$$
$$y(6) = [-0.8y(5)]_Q = [-0.8 \times (-3)]_Q = [2.4]_Q = 2$$
$$y(7) = [-0.8y(6)]_Q = [-0.8 \times 2]_Q = [-1.6]_Q = -2$$
$$y(8) = [-0.8y(7)]_Q = [-0.8 \times (-2)]_Q = [1.6]_Q = 2$$

Filter Structures

Direct Form I



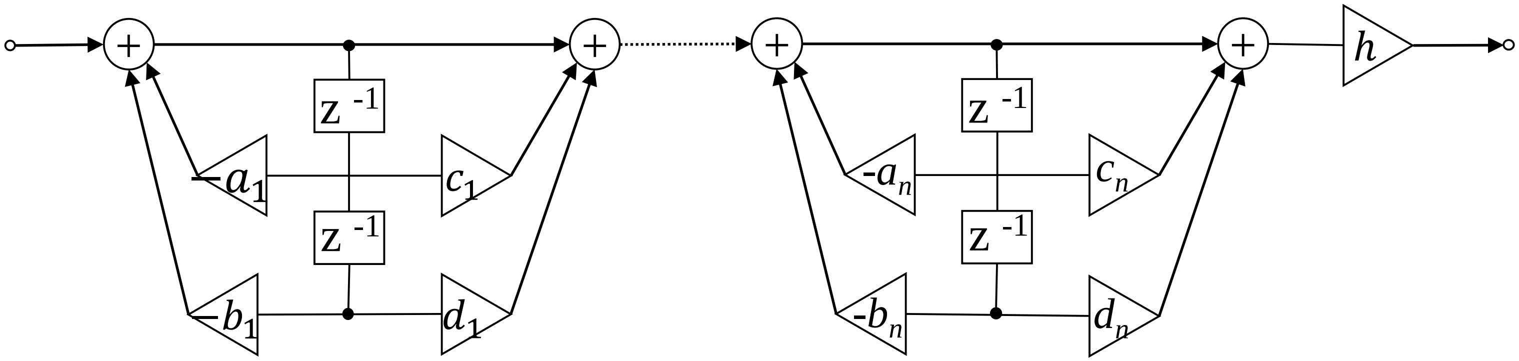
Direct Form II



- exchange of the order
- common use of the delays

Cascade Form

factorization of the transfer function

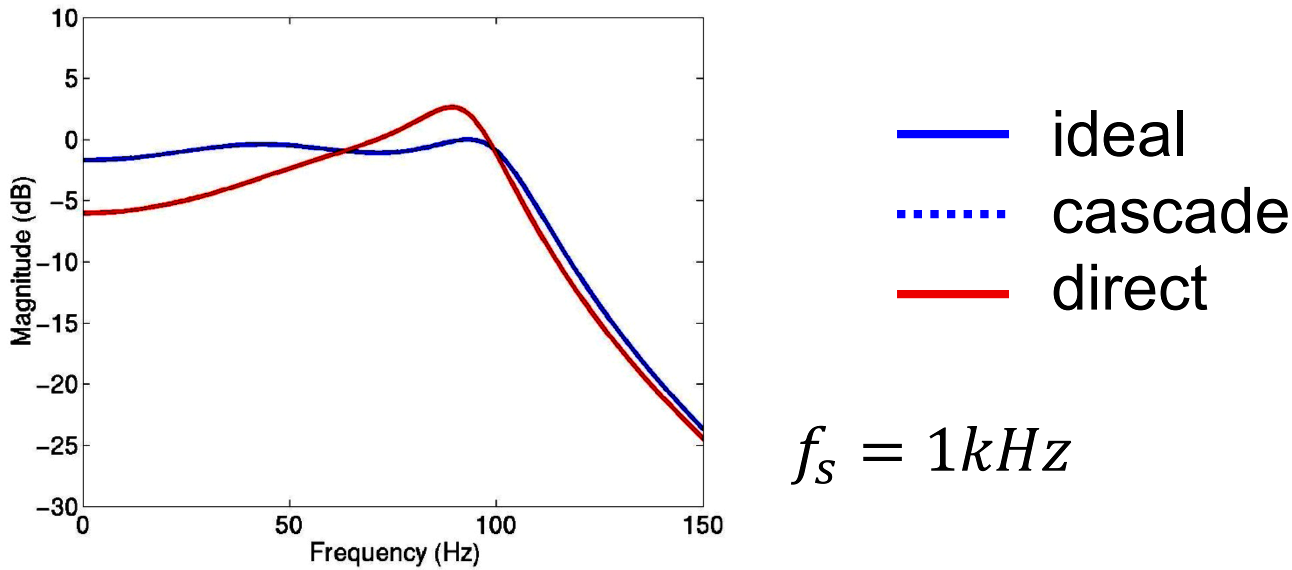


Example of Coefficient Rounding

4th-order LPF

$$H(z) = \frac{0.0018(1+z^{-1})^4}{(1-1.55z^{-1}+0.65z^{-2})(1-1.50z^{-1}+0.85z^{-2})}$$

is realized using 8bit coefficients



$f_s = 1kHz$

ideal response overlaps with the cascade

cf. Numerical Error

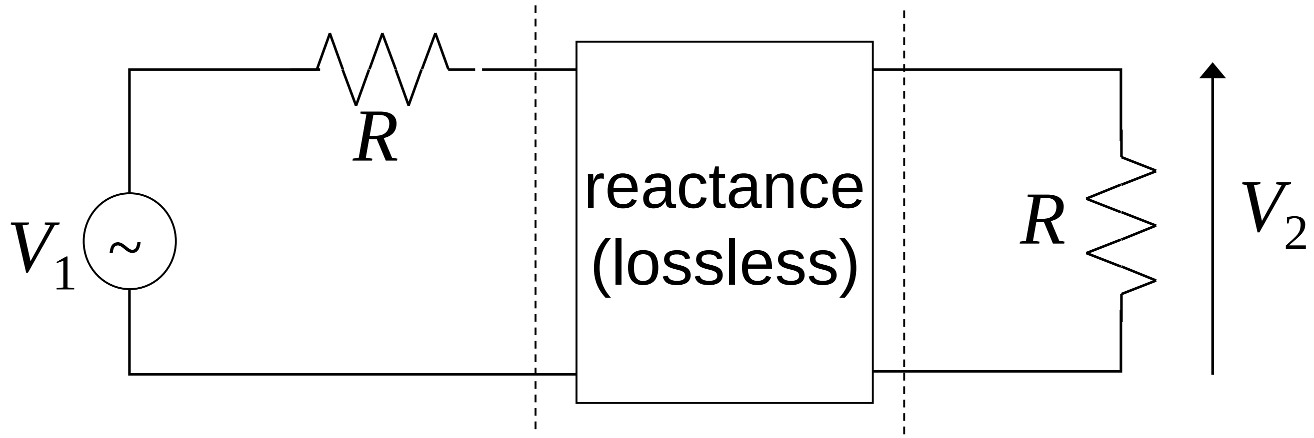
For the solution of linear algebraic equations

- Gauss-Jordan Elimination
- Gaussian Elimination with Back substitution
- LU Decomposition
- etc.

Depending on the algorithm, computational complexity and numerical accuracy are different

Choice of filter structures ~ choice of computational algorithms

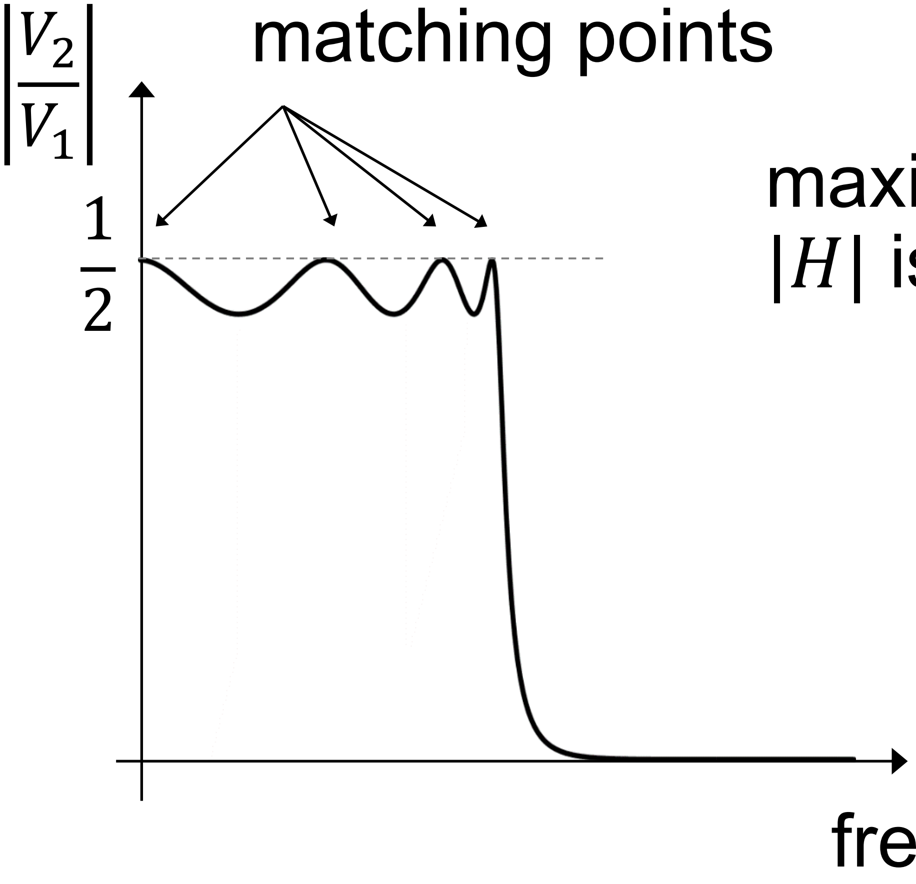
Simulation of reactance circuit



available power $\qquad\qquad$ consumed power

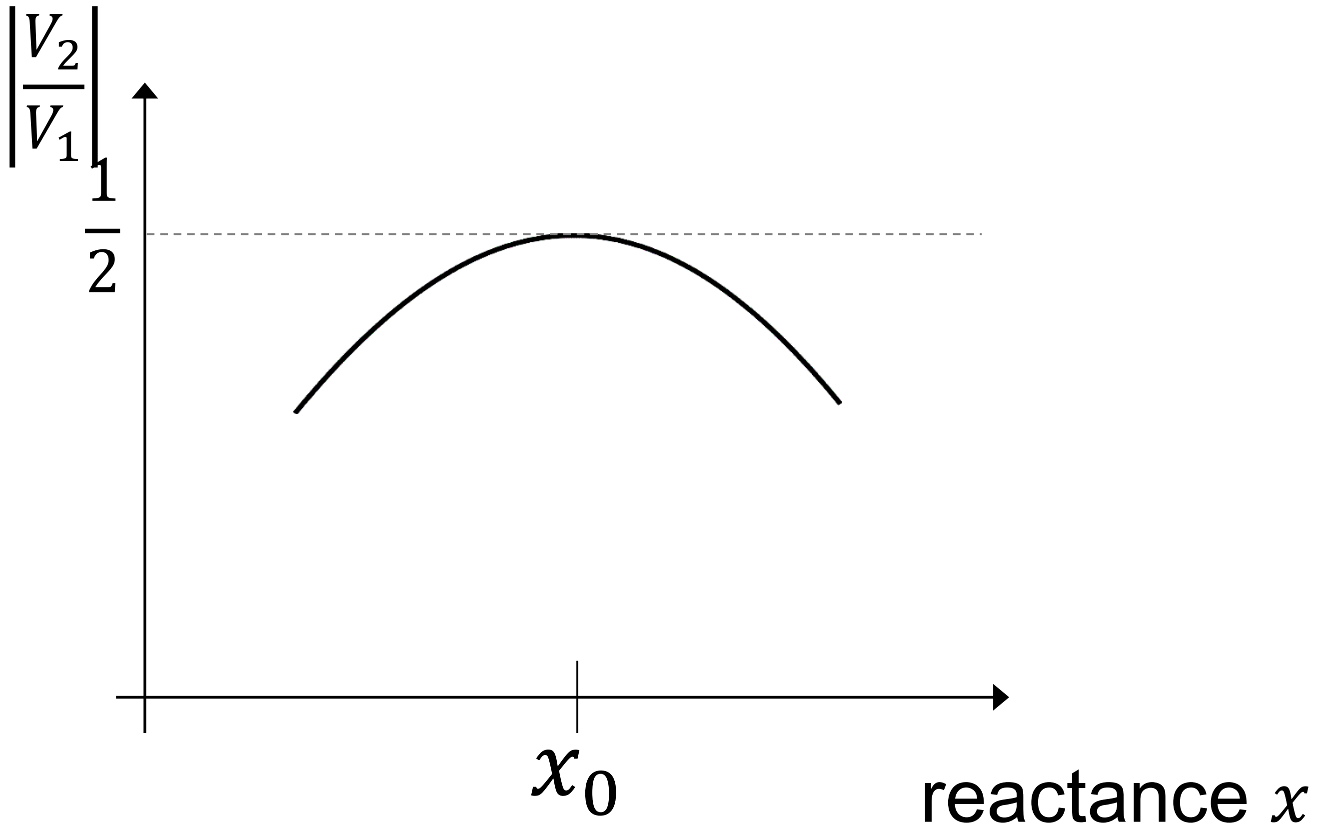
$$\frac{|V_1|^2}{4R} \geq \frac{|V_2|^2}{R}$$

Matching



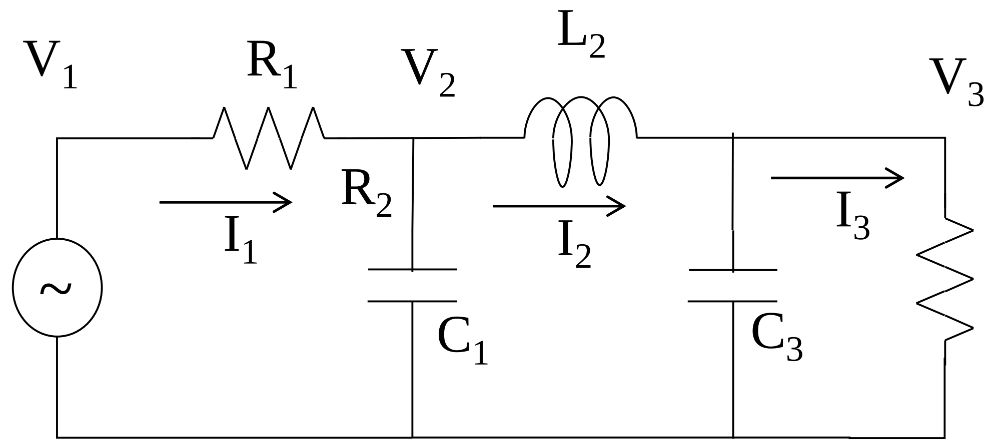
maximum power transfer
$|H|$ is bounded

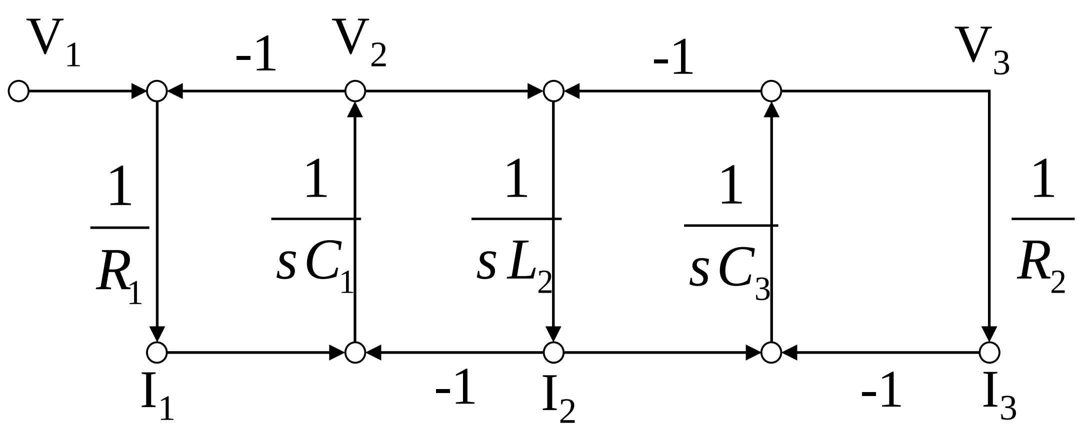
at a matching point



sensitivity $\frac{\partial |H|}{\partial x}$ is zero at matching frequencies
and low in the passband

voltage-current simulation





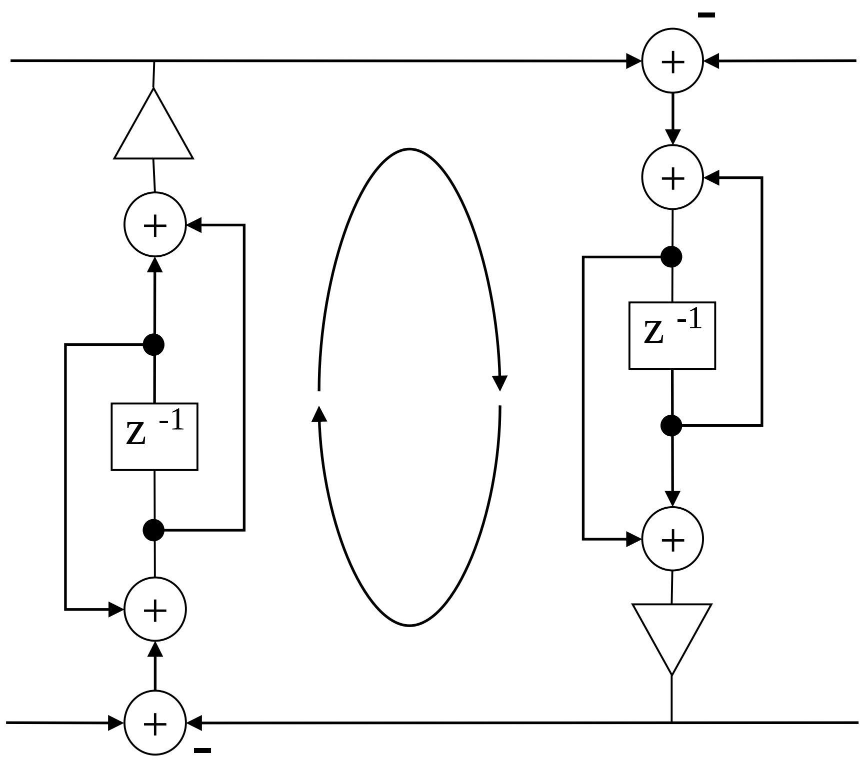
<u>leap frog</u>

widely used in RC active and switched capacitor filters

Delay-free Loops

integrator $\qquad$ bilinear transformation

$$\frac{1}{s} \Rightarrow \frac{1+z^{-1}}{1-z^{-1}}$$



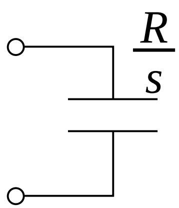
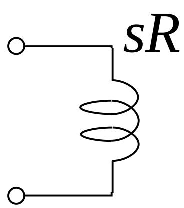
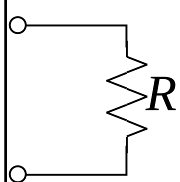
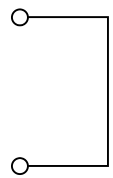
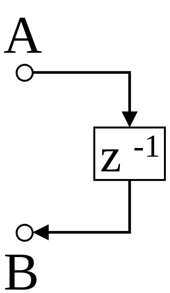
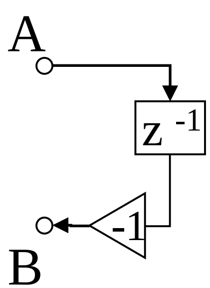
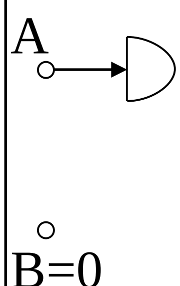
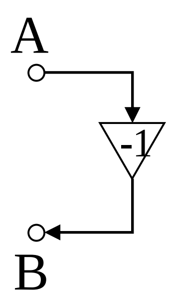
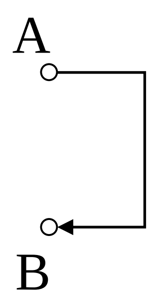
Simulation in terms of Wave Quantities

more precisely : voltage waves

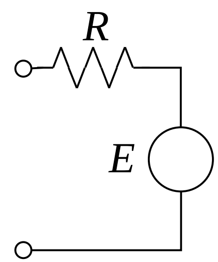
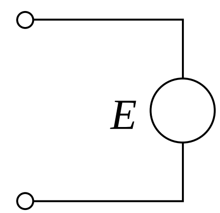
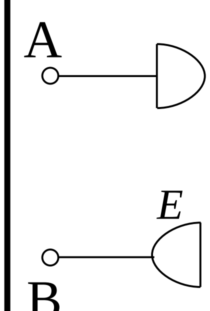
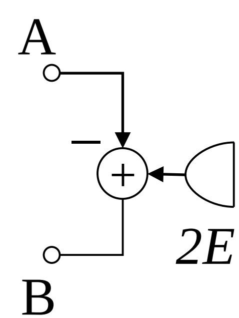
incident wave A=V+RI
reflected wave B=V-RI

R : port resistance

one-part elements

				
$B = z^{-1}A$	$B = -z^{-1}A$	$B = 0$	$B = -A$	$B = A$
				

sources

	
$B = E$	$B = 2E - A$
	

Interconnections and Adaptors

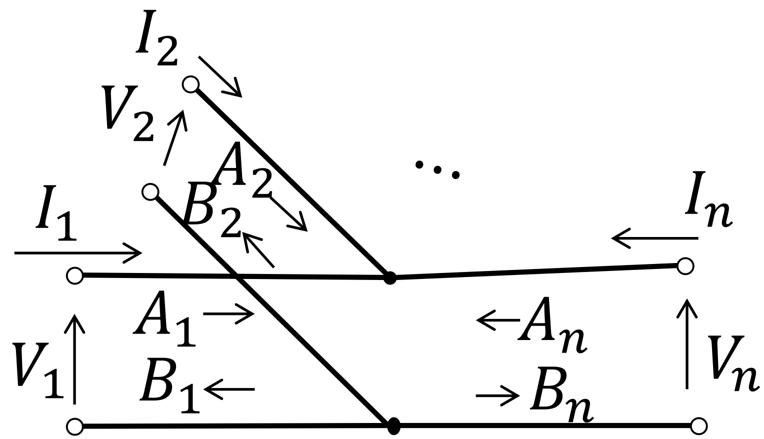
$$\begin{cases} \text{Parallel connection of } n \text{ ports} \\ \text{Series connection of } n \text{ ports} \end{cases}$$

- Kirchhoff 's voltage law
- Kirchhoff 's current law

are interpreted in terms of waves using

$$A_k = V_k + R_k I_k$$
$$B_k = V_k - R_k I_k \qquad k = 1, 2, \cdots, n$$

Parallel Connection



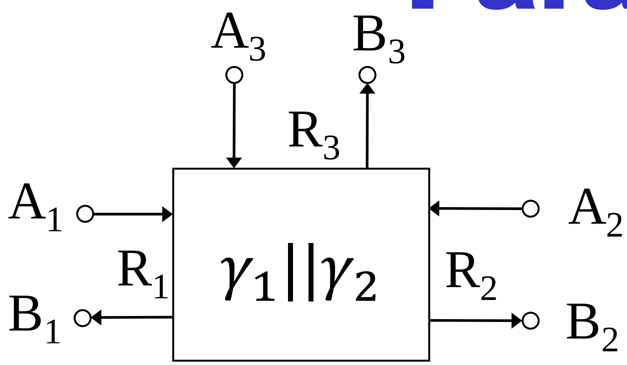
$$I_1 + I_2 + \cdots + I_n = 0, \qquad V_1 = V_2 = \cdots = V_n$$

$$B_k = (\gamma_1 A_1 + \gamma_2 A_2 + \cdots + \gamma_n A_n) - A_k, \qquad k = 1, 2, \cdots, n$$

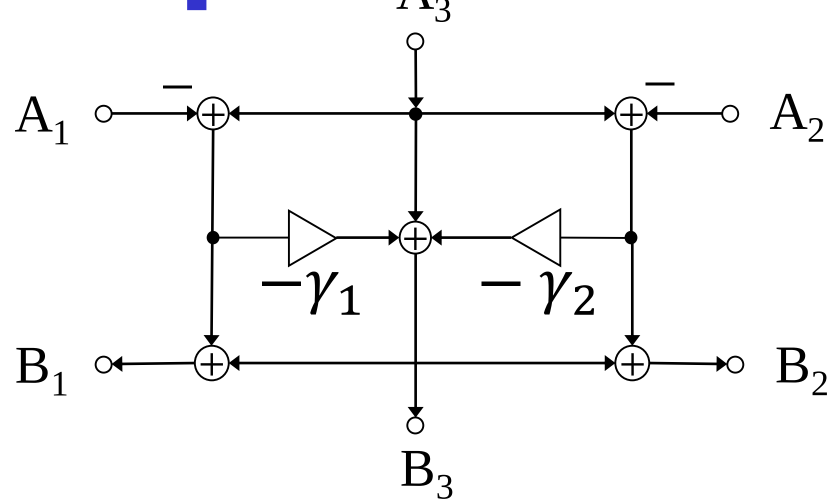
where $\gamma_k = \frac{2G_k}{G_1 + G_2 + \cdots + G_n}, \quad G_k = \frac{1}{R_k}$

$$(\gamma_1 + \gamma_2 + \cdots + \gamma_n = 2)$$

Parallel Adapters

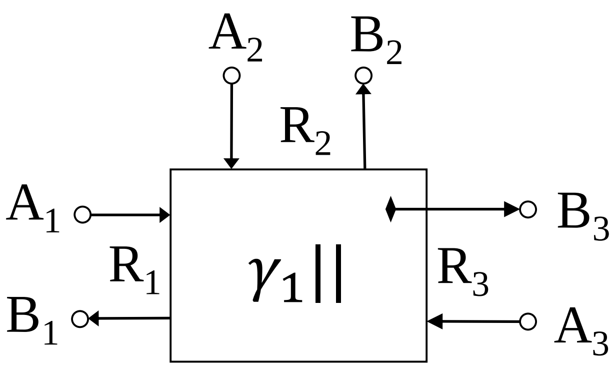


$$\gamma_k = \frac{2G_k}{G_1 + G_2 + G_3}, k = 1,2$$

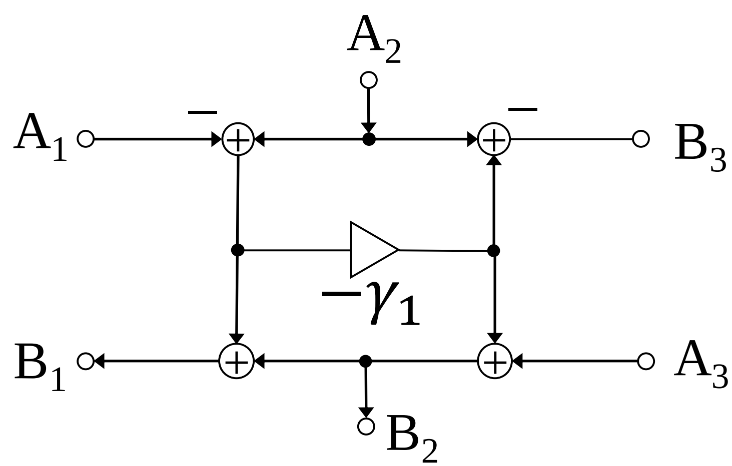


port 3 : dependent port

- if $\gamma_3 = 1$, i.e. $G_3 = G_1 + G_2$



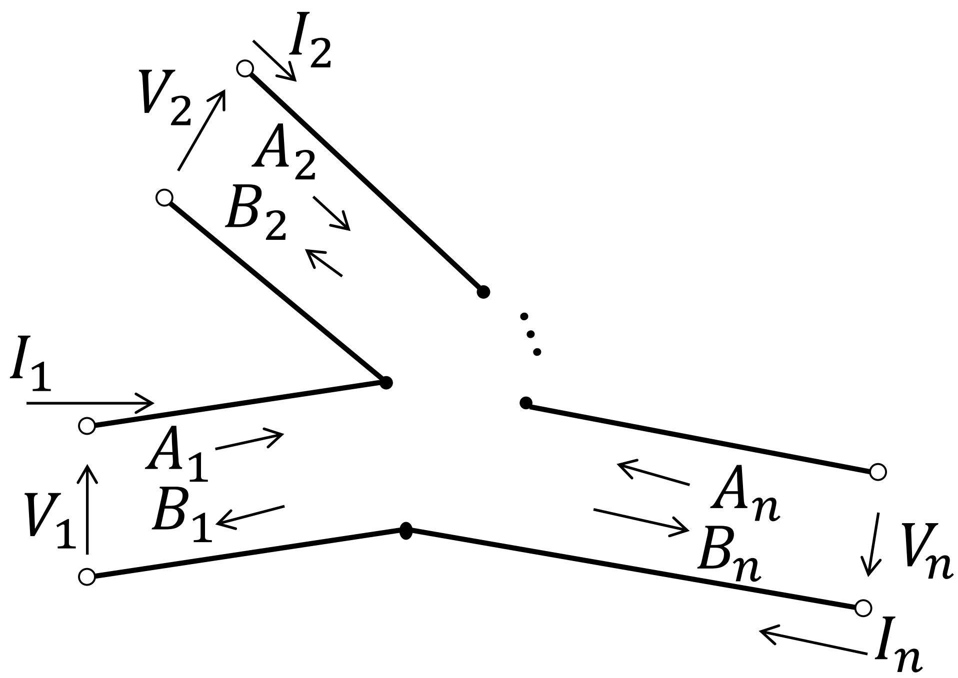
$$\gamma_1 = \frac{G_1}{G_1 + G_2} = \frac{G_1}{G_3}$$



constrained three-port parallel adaptor

- with port3 reflection-free
- and port2 dependent

Series Connection



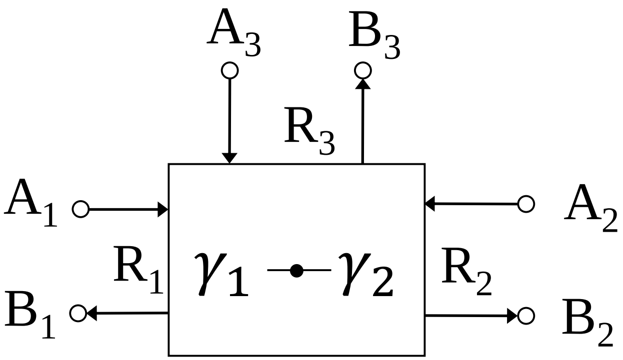
$$V_1 + V_2 + \cdots + V_n = 0, \qquad I_1 = I_2 = \cdots = I_n$$

$$B_k = A_k - \gamma_k (A_1 + A_2 + \cdots + A_n), k = 1, 2, \cdots, n$$

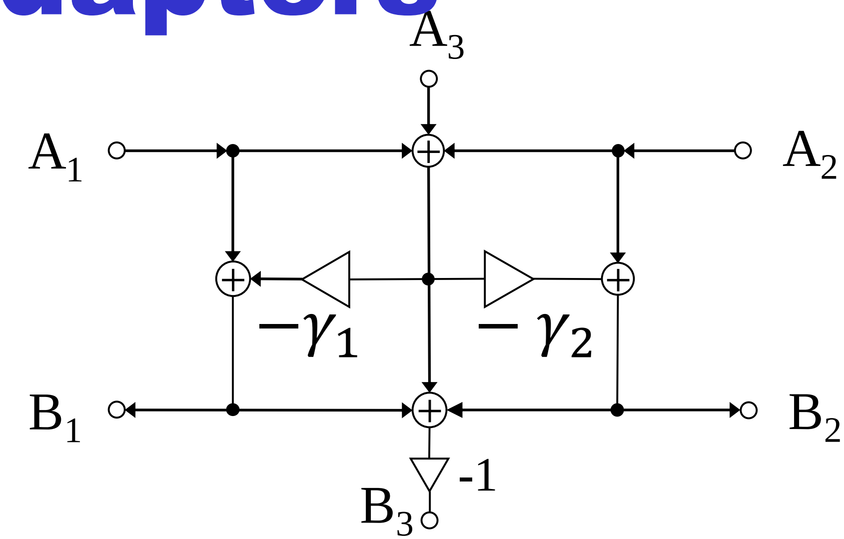
where $\gamma_k = \frac{2R_k}{R_1 + R_2 + \cdots + R_n}$

$$(\gamma_1 + \gamma_2 + \cdots + \gamma_n = 2)$$

Series Adapters

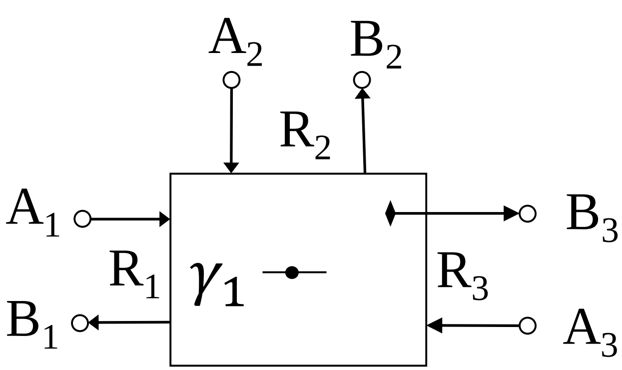


$$\gamma_k = \frac{2R_k}{R_1 + R_2 + R_3} \qquad k = 1,2$$

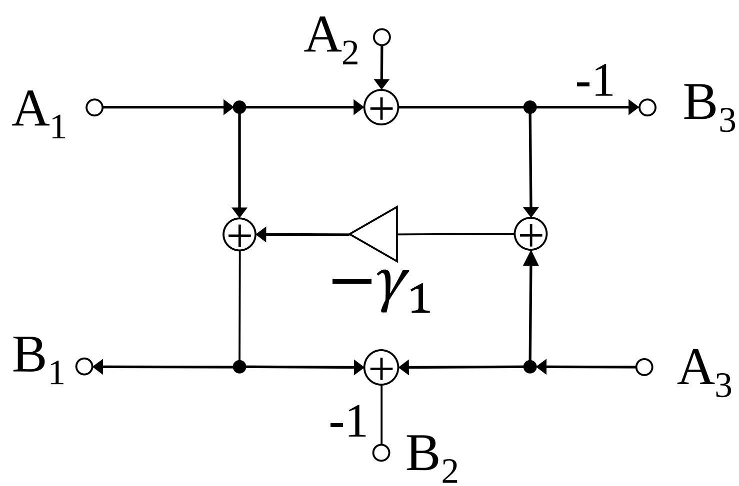


port 3 : dependent port

- if $\gamma_3 = 1$, i.e. $R_3 = R_1 + R_2$



$$\gamma_1 = \frac{R_1}{R_1 + R_2} = \frac{R_1}{R_3}$$



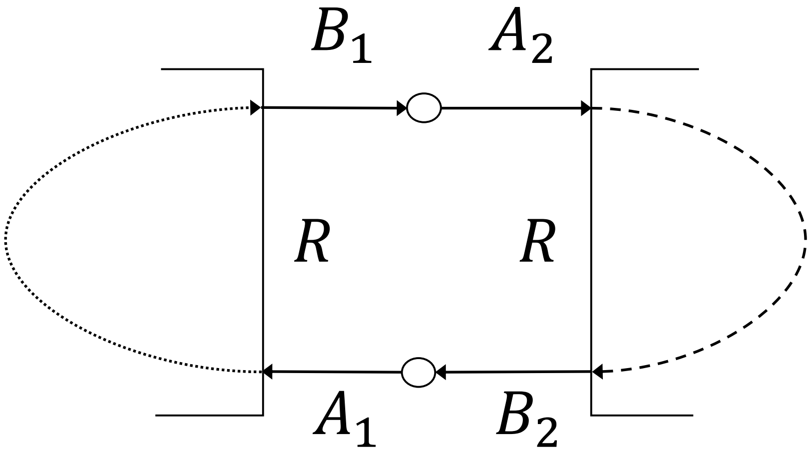
constrained three-port parallel adaptor

•with port3 reflection-free

•and port2 dependent

Interconnection Rules

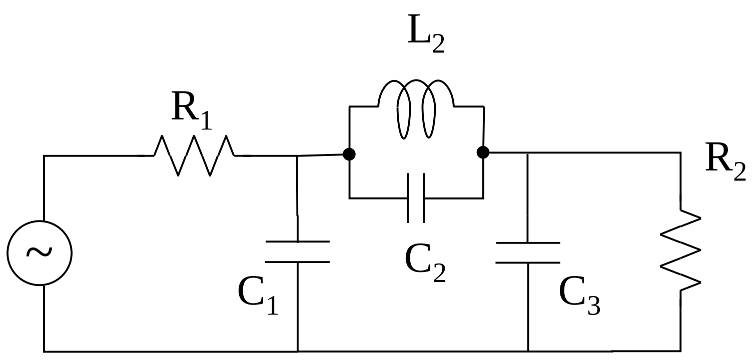
1. grouping of terminals into ports must remain respected
2. waves must flow in the same direction
3. two port resistances must be the same
4. connection must not make delay-free loops

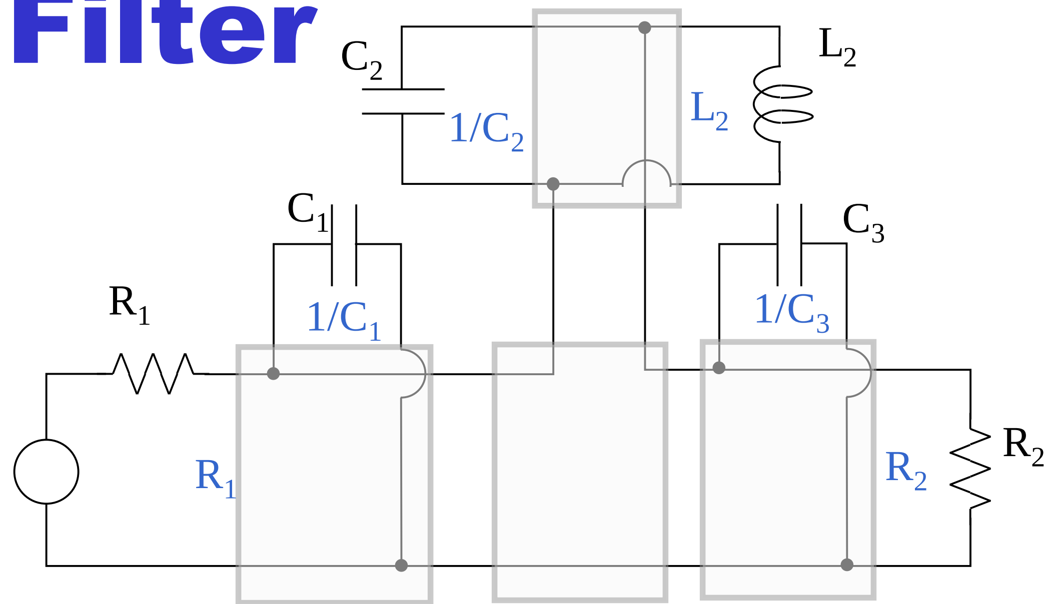


Wave Digital Filters

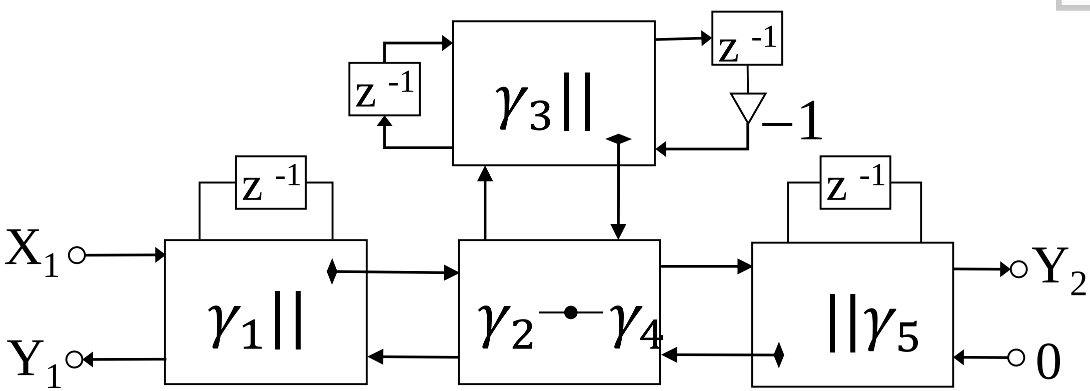
- low sensitivity
- limit cycles can easily be suppressed
- forced response stability
- stability under looped condition

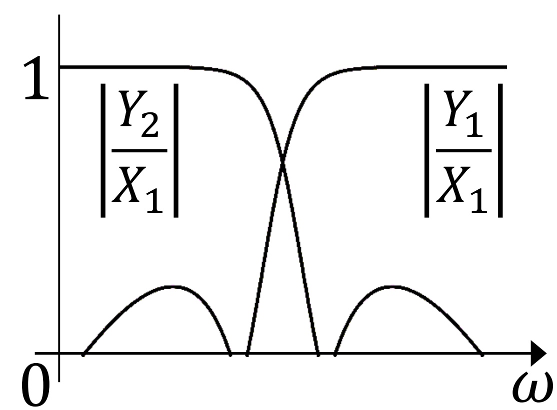
Third-order Filter



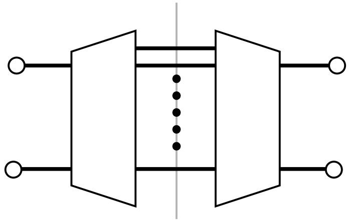


$$\left|\frac{Y_1}{X_1}\right|^2 + \left|\frac{Y_2}{X_1}\right|^2 = 1$$





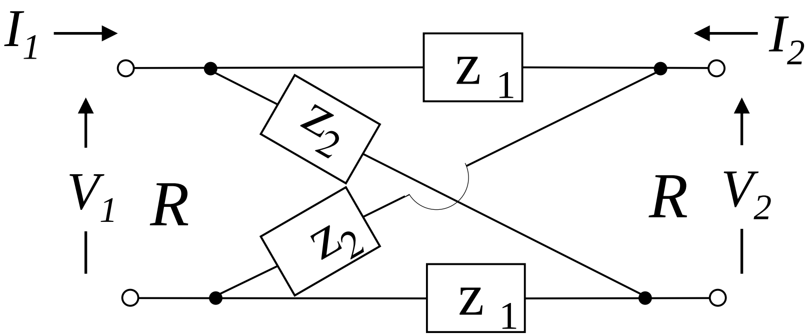
five multipliers in general cases
four multipliers if symmetric

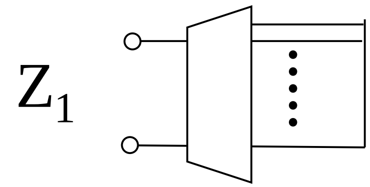


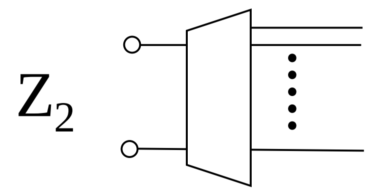
Symmetric Case

Bartlett's bisection theorem

||







$$\begin{pmatrix} V_1 \\ V_2 \end{pmatrix} = \begin{pmatrix} \frac{Z_1 + Z_2}{2} & \frac{Z_2 - Z_1}{2} \\ \frac{Z_2 - Z_1}{2} & \frac{Z_1 + Z_2}{2} \end{pmatrix} \begin{pmatrix} I_1 \\ I_2 \end{pmatrix} \quad \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \begin{pmatrix} \frac{S_1 + S_2}{2} & \frac{S_1 - S_2}{2} \\ \frac{S_2 - S_1}{2} & \frac{S_1 + S_2}{2} \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$$

$$A_k = V_k + R_k I_k$$
$$B_k = V_k - R_k I_k, \quad k = 1,2$$

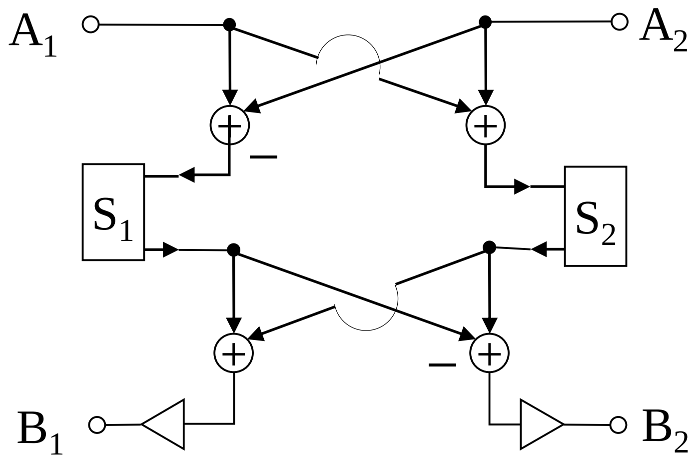
where $S_k = \frac{Z_k - R}{Z_k + R}$, $k = 1,2$

reflection coefficient

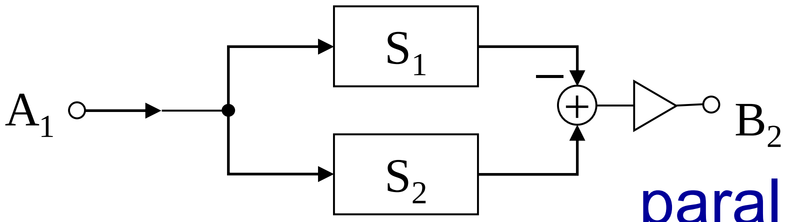
if Z_1 and Z_2 are reactances

S_1 and S_2 are all-paass fucntions.

Lattice Wave Digital Filters



for A_2=0

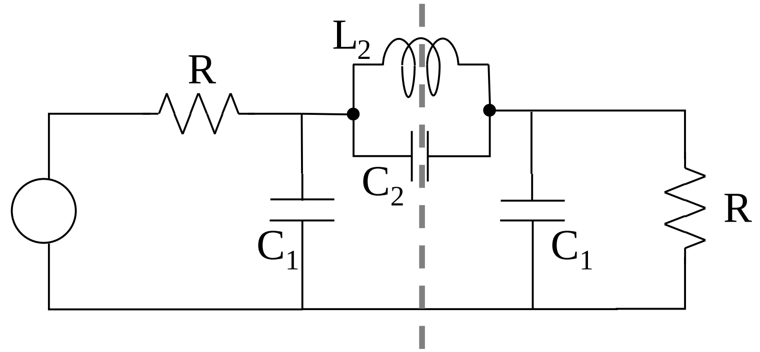


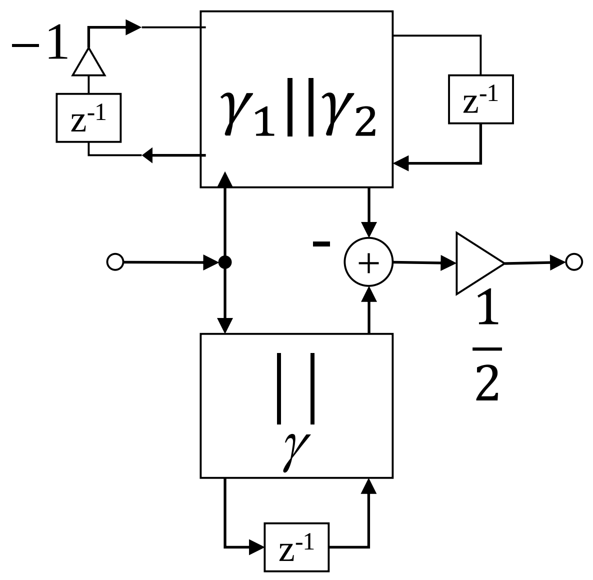
parallel of two all-pass filters

for the realization of S_1 and S_2 themselves,
any methods can be used.

for example, Cauer
cascade of all-pass sections

example





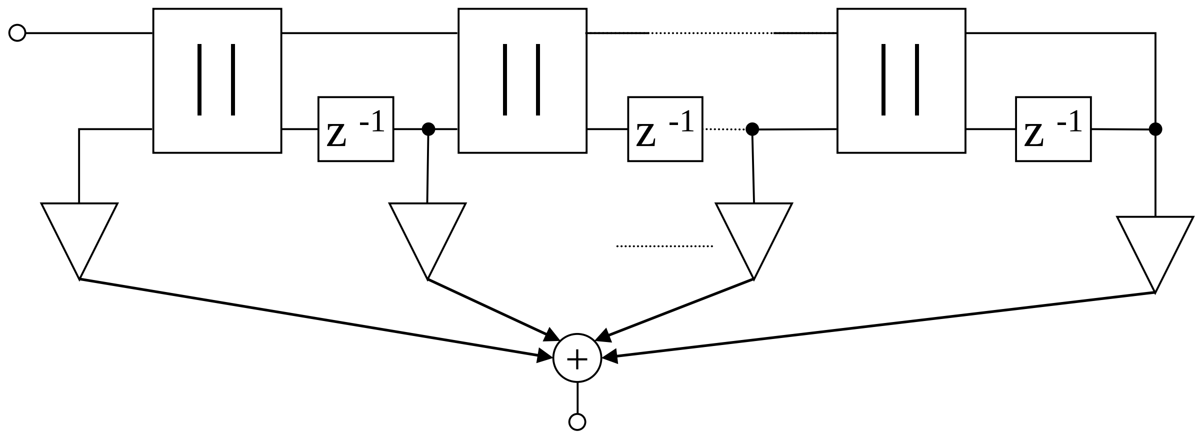
three multipliers

symmetry

odd-order Butterworth, Chebyshev and elliptic

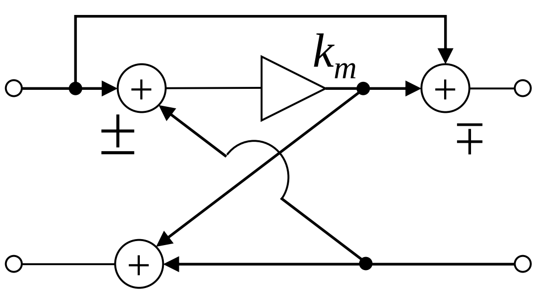
WDF structures may become very complicated when general transfer functions are realized.

Simple recursive structures + weighted taps cascade unit elements (vocal tract)

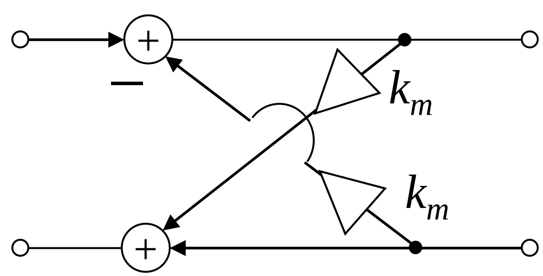


Digital Lattice Filters

internal circuit of two-port adaptor



one-multiplier lattice



two-multiplier lattice

Simulation of Physical Systems

- Wave digital filters physically model passive RLC circuits
- Can model any systems described by (ordinary or partial) differential equations (both linear and non-linear) with high degrees of parallelism and locality
- A “traveling wave” view (at the speed of light) is much closer to underlying physical reality than any instantaneous models
- The bilinear transform is equivalent in the time domain to the trapezoidal rule for numerical integration
- In discretizing space-time continuum, we can get stability and robustness, unlike Courant-Friedrichs-Lewy Condition which is necessary but not strictly sufficient

Courant, R., Friedrichs, K., and Lewy, H., "Über die partiellen Differenzengleichungen der mathematischen Physik", *Mathematische Annalen* **100** (1): 32–74 1928 (English versions in 1956 and 1967)

Exercise 3

1. For the zero-input first-order digital filter in Slide 9, what is a multiplier coefficient range that does not cause any limit cycle oscillation?
2. Show that the wave digital filter for a capacitor with its impedance $1/sC$ is a simple delay, where the port resistance is $R = 1/C$.
3. Derive a wave digital filter structure for a voltage source E with a series inductive impedance sL, where the port resistance is $R = L$.
4. Read the following paper;

   A.Nishihara & M.Murakami, Signal-Processor-Based Digital Filters Having Low Sensitivity, Electronics Letters, 20, 8, pp.325-326, Apr. 1984