Digital Filter Noise

Why does the textbook tell us not to use direct form 2?

Matt Evans

LIGO-G0900928-v1

Introduction

- Read ref1, in particular chapter 6
- z-transform to go from s-domain to z-domain

$$H(s) = G\frac{s^2 + B_1 s + B_2}{s^2 + A_1 s + A_2} \rightarrow H(z) = g\frac{1 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}}$$

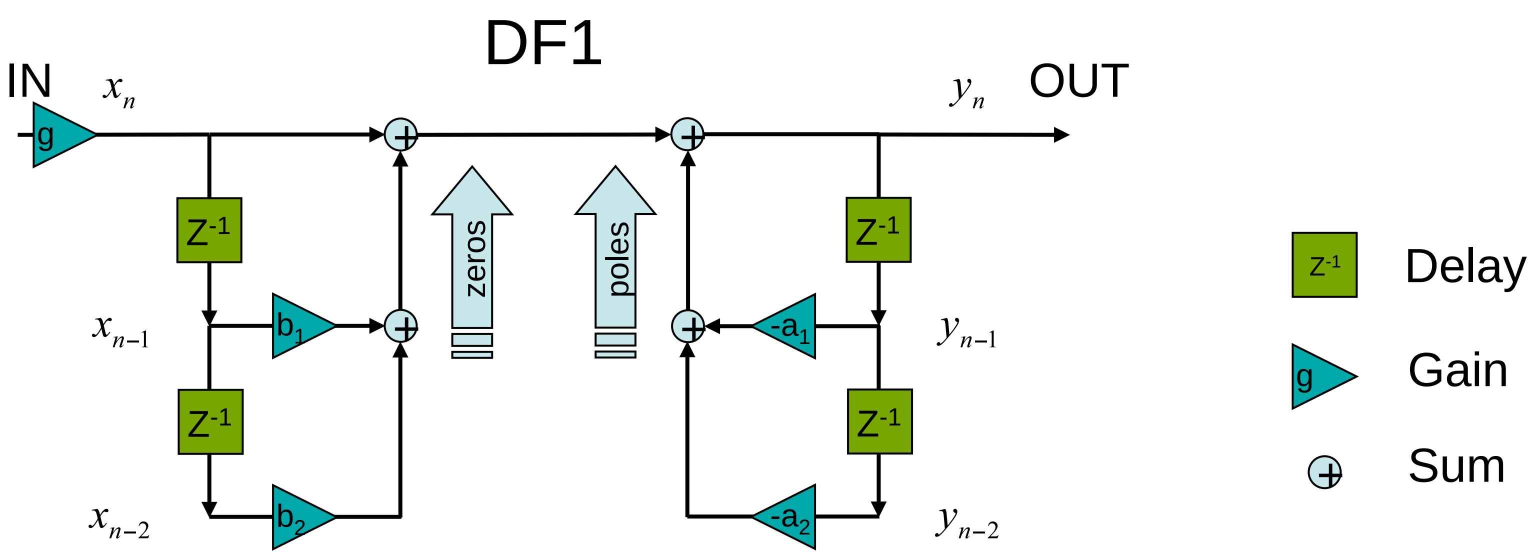
- Once you have the z-domain coefficient, your filter equation is

$$y_n = g\left(x_n + b_1 x_{n-1} + b_2 x_{n-2}\right) - a_1 y_{n-1} - a_2 y_{n-2}$$

"Direct Form 1" Implementation

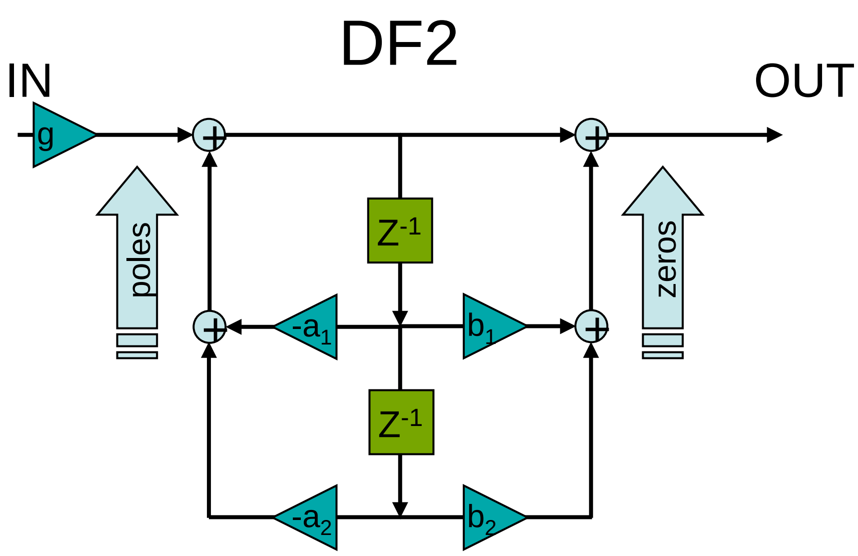
- Direct implementation of equation for y_n
- The TF is performed in two steps

$$H(f) = g \times H_{zeros}(f) \times H_{poles}(f)$$



"Direct Form 2" Implementation

- Rearrange DF1 to get DF2
  - Equivalent computation
  - Uses less memory

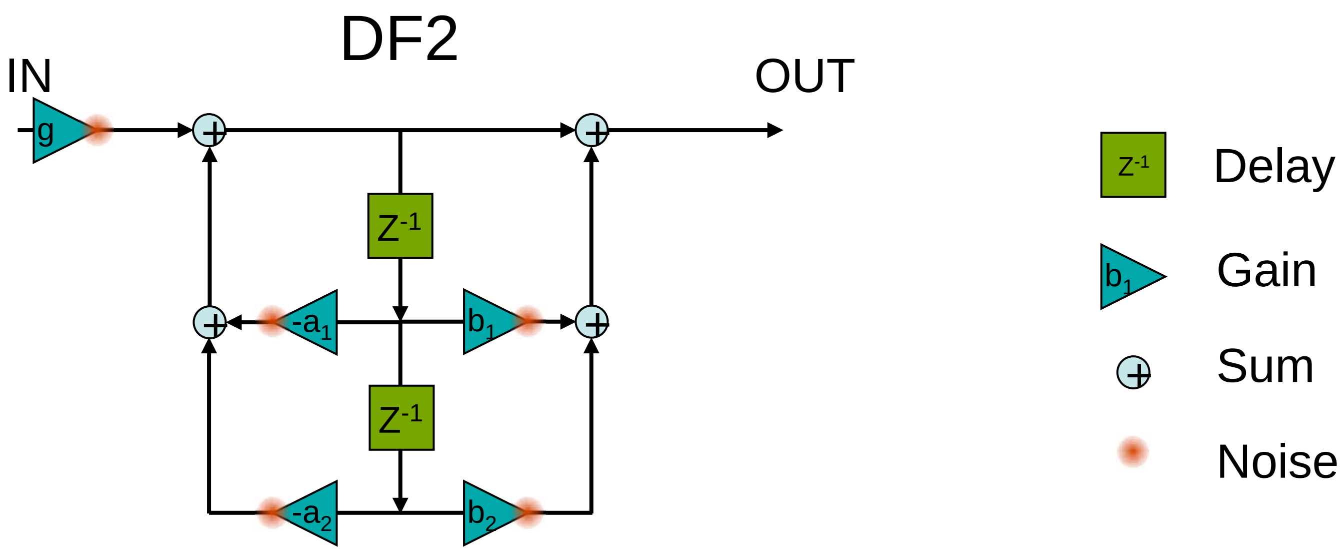


$$H_{poles}(f_s/4) \approx 1$$

$$H_{zeros}(f_s/4) \approx 1$$

Direct Form 2
Fixed Point Noise Analysis

- With fixed point numbers, the quantization noise analysis is relatively easy (ref 1)
- Noise added at each multiplication

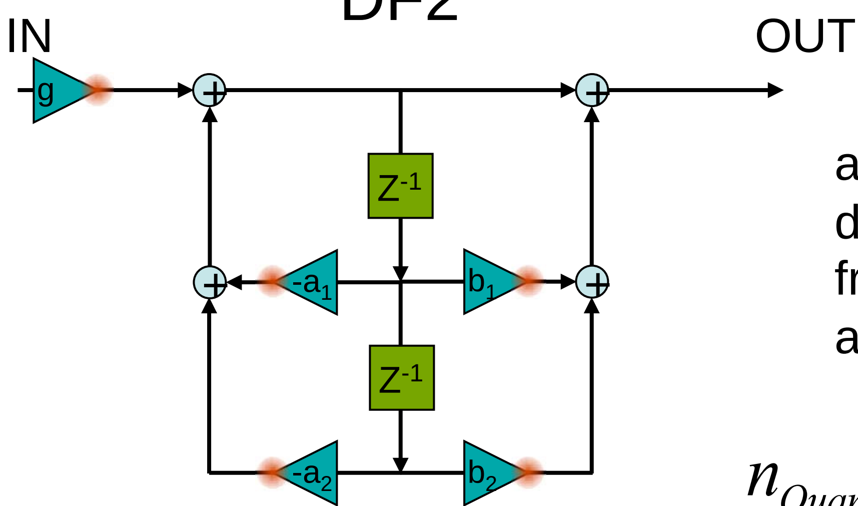


Direct Form 2
Fixed Point Noise Analysis

The output noise power density is

$$N_{OUT}(f) = N_{Quant}\left(3|H(f)|^2 + 2\right)$$

DF2



where

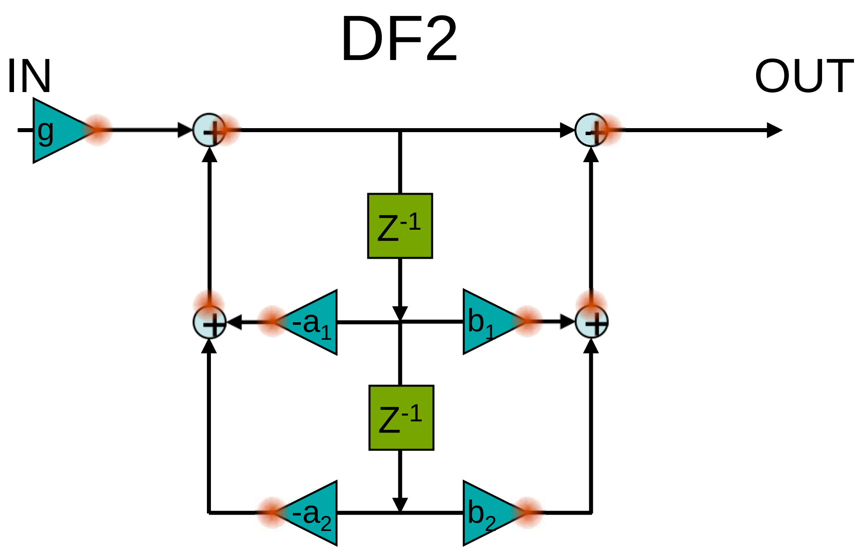
$$N_{Quant} = \frac{2^{-2B}}{12 f_s}$$

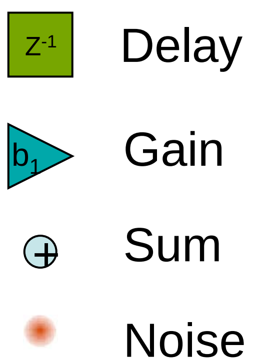
and B is the number of bits after the decimal point and fs the sample frequency. For example, with B=64 and f_s = 16384 Hz,

$$n_{Quant} = \sqrt{N_{Quant}} = \frac{2^{-B}}{\sqrt{12 f_s}} \cong \frac{10^{-22}}{\sqrt{Hz}}$$

Direct Form 2
Floating Point Noise Analysis

- With floating point numbers, things are more complicated
  - Noise added at multiplications and additions
  - *Noise depends on the signal*

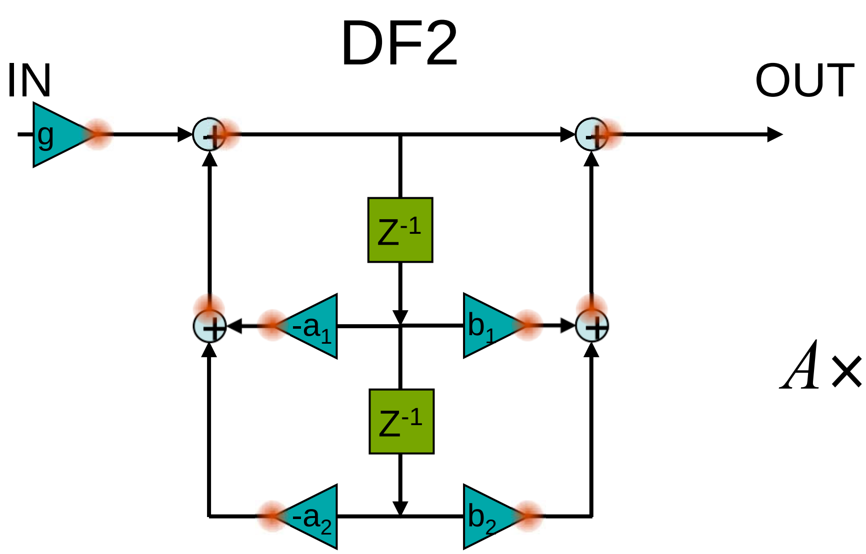




Direct Form 2
Floating Point Noise Analysis

$$N_{OUT}(f) = N_{Quant}\left(\sum_{input} A_i^2 \left|H(f)\right|^2 + \sum_{output} A_j^2\right)$$

Where A_i is the floating point multiplier used in each operation. N_{Quant} is as before, with B the number of bits in the mantissa. For example, a signal between 8 and 16, expressed in double precision, would have

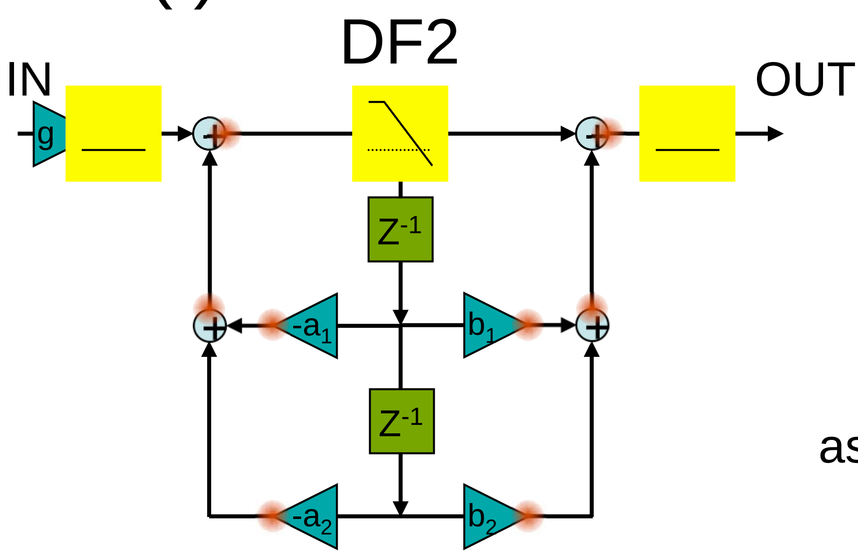


$$A \times n_{Quant} = 8\frac{2^{-53}}{\sqrt{12 f_s}} \cong \frac{2\times10^{-18}}{\sqrt{Hz}}$$

for f_s = 16384 Hz as before.

Direct Form 2
Noise Analysis: Example 1

- Input signal white (BW = 8kHz)
- 2 poles and 2 zeros 1Hz
  - Roughly: a_1=-2+ε, a_2=1-ε, b_0=1, b_1=a_1, b_2=a_2
  - H(f) = 1



All of the operations involve the signal filtered by the poles. The amplitude spectral density is about 10^7 in DC. Taking A_{MID} to be the RMS of this signal,

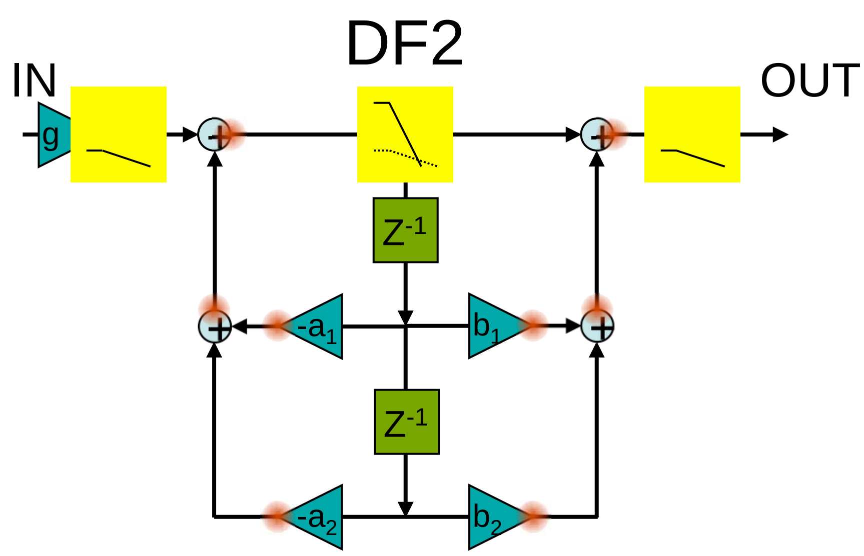
$$n_{OUT}(f) = \sqrt{8 A_{MID}^2 N_{Quant}} \cong \frac{6\times 10^{-12}}{\sqrt{Hz}}$$

as compared to

$$n_{IN}(f) = \sqrt{A_{IN}^2 N_{Quant}} \cong \frac{2\times 10^{-15}}{\sqrt{Hz}}$$

Direct Form 2
Noise Analysis: Example 2

- Input signal pink
  - RMS dominated by low-frequency signal
  - typical of LIGO signals
  - Assume RMS of input A_{IN}=1



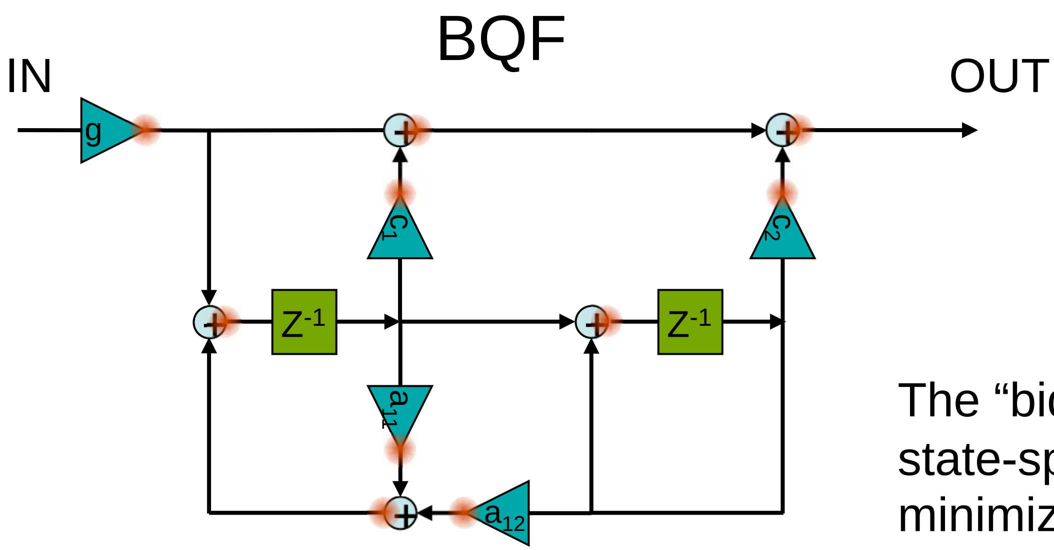
$$n_{IN}(f) = \sqrt{A_{IN}^2 N_{Quant}} \cong \frac{2\times 10^{-19}}{\sqrt{Hz}}$$

which makes the unchanged output noise seem very large

$$n_{OUT}(f) = \sqrt{8 A_{MID}^2 N_{Quant}} \cong \frac{6\times 10^{-12}}{\sqrt{Hz}}$$

Biquad Form

- Biquad form avoids large internal values
  - A null filter (as in previous example) leads to no added noise
  - Requires one additional summation



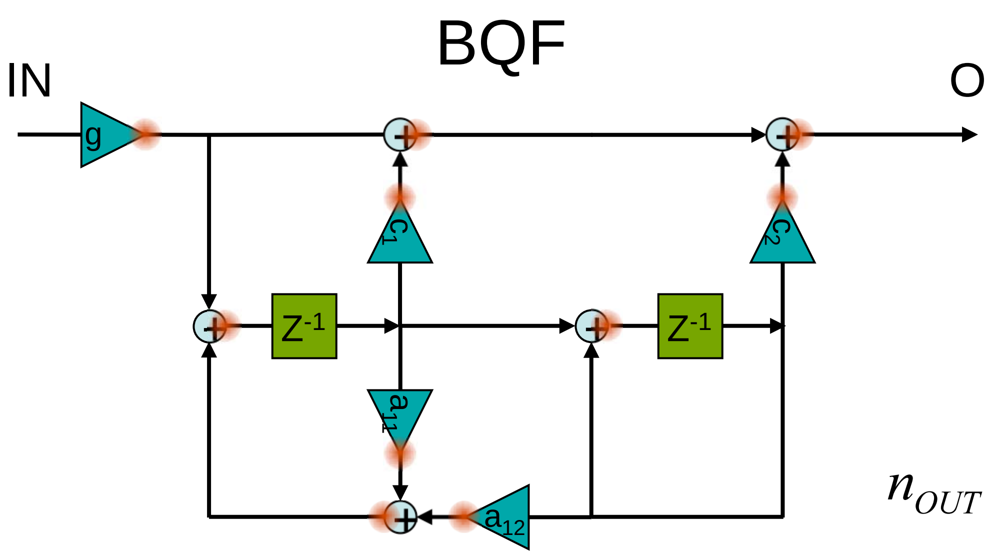
$$a_{11} = -a_1 - 1$$
$$a_{12} = -a_2 - a_1 - 1$$
$$c_1 = b_1 - a_1$$
$$c_2 = b_2 - a_2 + b_1 - a_1$$

The “biquad” form is a particular state-space arrangement which minimizes flops. Its layout is very similar to the analog biquad filter.

Biquad Form
Noise Analysis

- For non-null filters, the internal signal RMS is similar to the output RMS, so

$$N_{OUT}(f) \approx 8 N_{Quant} \max\left(A_{input}^2, A_{output}^2\right)$$



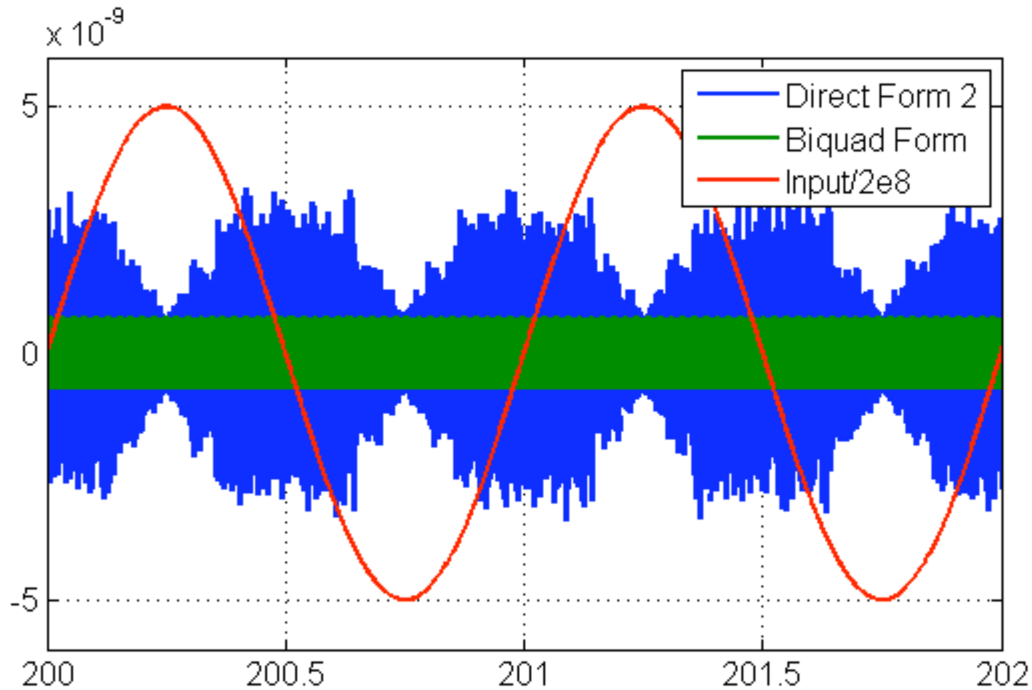
a notch filter applied to a signal with an RMS of 1, for example, would give

$$n_{OUT}(f) = \sqrt{8 A_{IN}^2 N_{Quant}} \cong \frac{6 \times 10^{-19}}{\sqrt{Hz}}$$

DF2 vs. BQF Empirical Results

- High and low frequency input
- 4^{th} order notch
  - fp=fz=1Hz
  - Qp=1
  - Qz=1e6

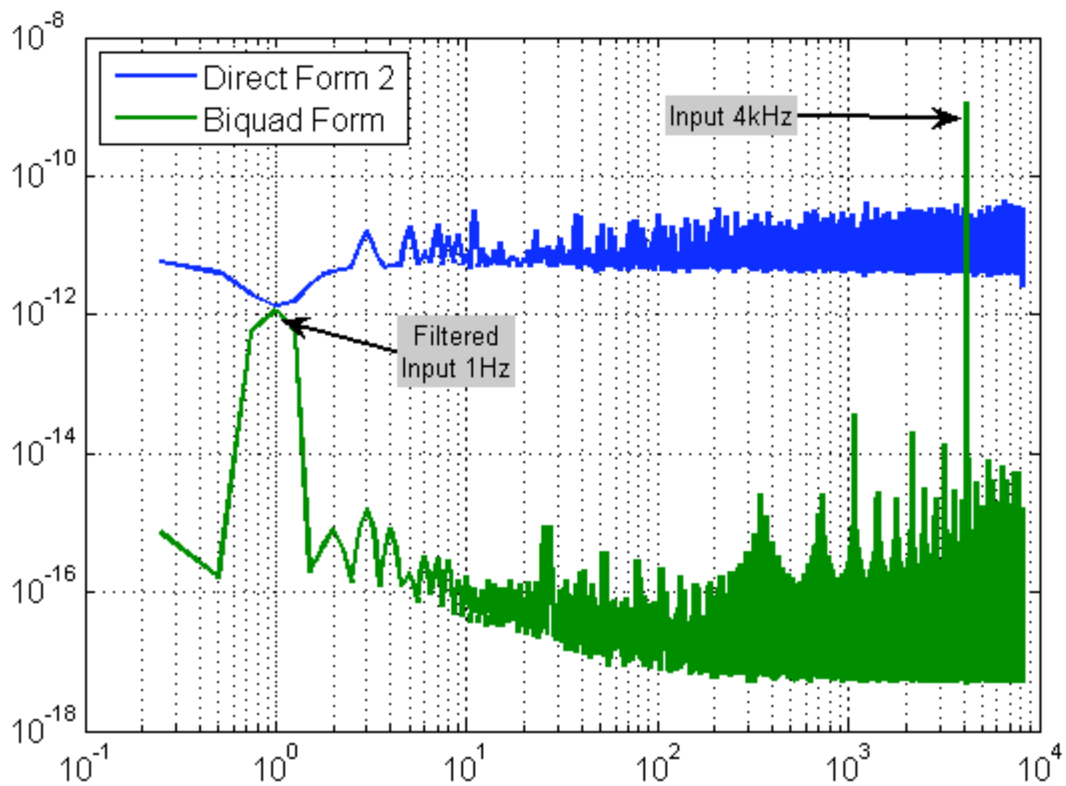
$$x_{input} = \sin(2\pi \times t) + 10^{-9} \sin(2\pi \times t \times f_s / 4)$$



DF2 vs. BQF
Empirical Results

- Output noise roughly as expected in both cases
- Biquad reveals quant noise not well modeled by white noise

$$x_{input} = \sin(2\pi \times t) + 10^{-9} \sin(2\pi \times t \times f_s / 4)$$



Conclusion

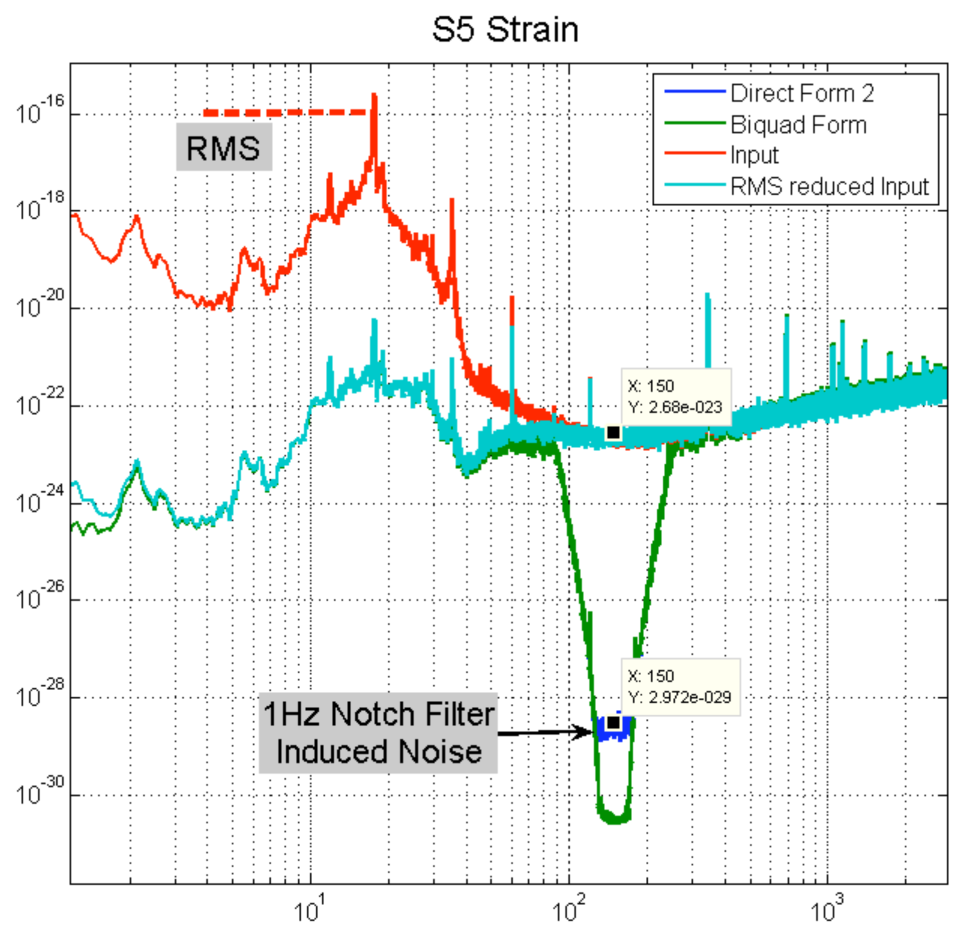
- Direct Form 2, used by LIGO, is not a good choice for low-noise filtering
- Noise in floating point DSP has been studied extensively for high-quality audio applications
- Many low-noise implementations are available
  - State-space second-order sections are general
  - Noise optimized forms usually involve more flops
- For a very modest increase in computational time, we can improved noise performance by many orders of magnitude

References

1. Discrete-Time Signal Processing Oppenheim and Schafer, 2^{nd} Ed 1999
2. "Floating-point roundoff noise analysis of second-order state-spacedigital filter structures" Smith, L.M.; Bormar, B.W.; Joseph, R.D.; Yang, G.C.-J.
   Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on Volume 39, Issue 2, Feb 1992 Page(s):90 – 98

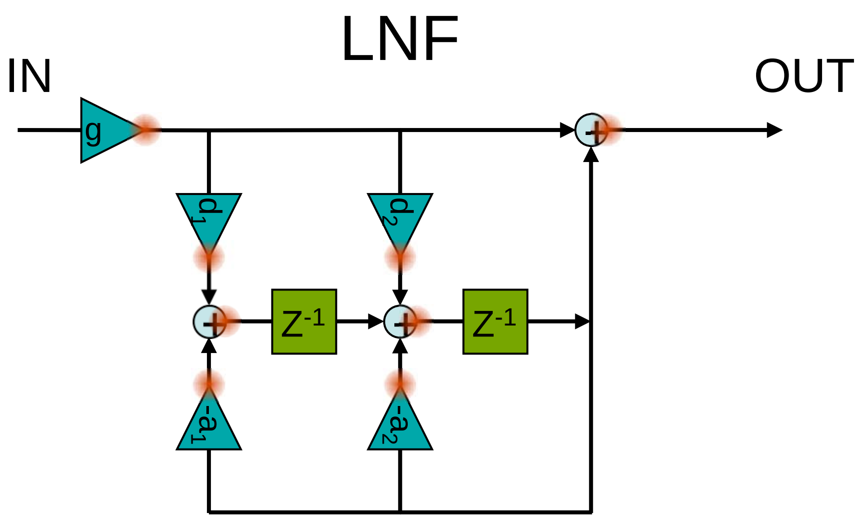
A LIGO Signal

- Strain
  - RMS dominated by 18Hz peak
  - Added 160dB band-stop around 150Hz
- A notch at 1Hz induces noise in the stop-band



Low-Noise Form

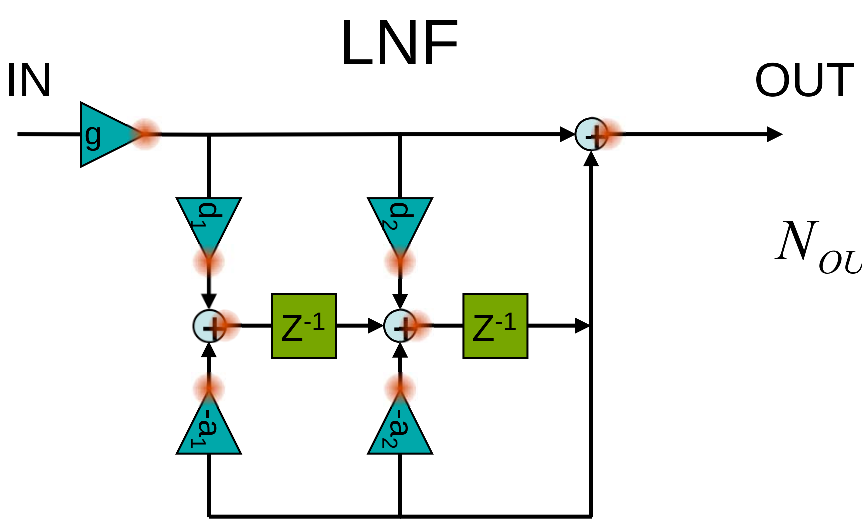
- Proposed form avoids large internal values
  - A null filter (as in previous example) leads to no added noise
  - Requires no additional flops



$$\boxed{d_n = b_n - a_n}$$

Low-Noise Form
Noise Analysis

$$N_{OUT}(f) = N_{Quant}\left(A_{input}^2 \left|H(f)\right|^2 + \sum_{loop} A_i^2 \left|H_{poles}(f)\right|^2 + A_{output}^2 \right)$$



For non-null filters, the internal signal RMS is similar to the output RMS, so

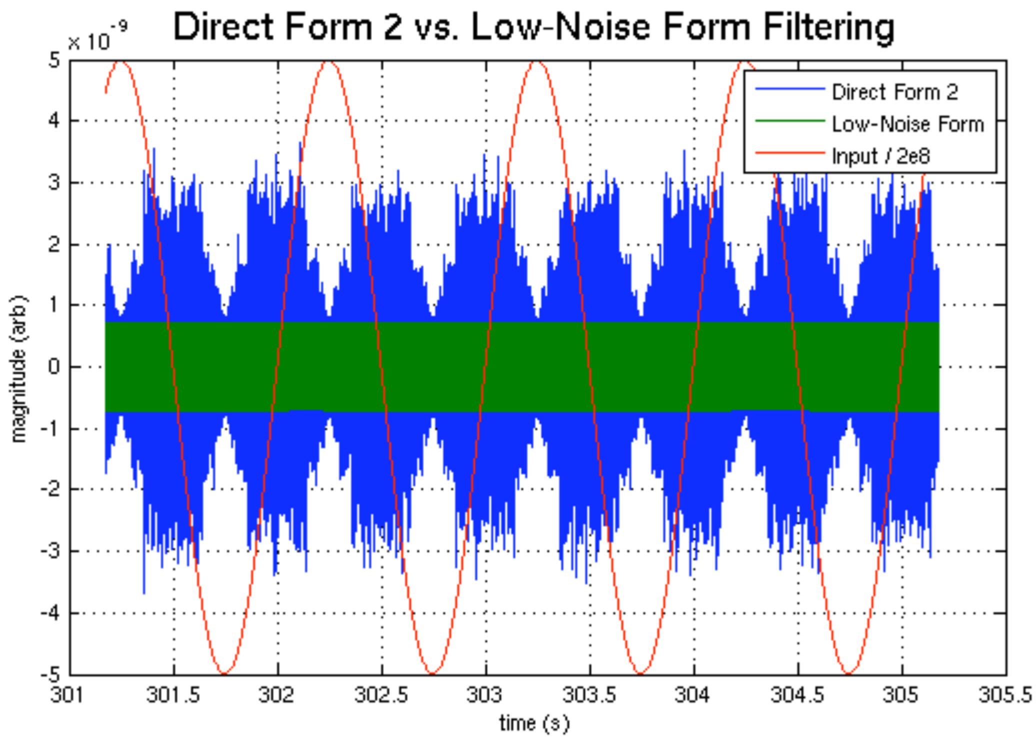
$$N_{OUT}(f) \approx N_{Quant} A_{output}^2 \left(2 + 6\left|H_{poles}(f)\right|^2 \right)$$

which is similar to DF2 below the pole frequency, but lower above that frequency.

DF2 vs. BQF
Empirical Results

- High and low frequency input
- 4th order notch
  - fp=fz=1Hz
  - Qp=1
  - Qz=1e6

$$x_{input} = \sin(2\pi \times t) + 10^{-9} \sin(2\pi \times t \times f_s / 4)$$



DF2 vs. LNF
Empirical Results

- Output noise close to expected in both cases