A low-power analog logarithmic map circuit with offset and temperature compensation for use in bionic ears

by

Ji-Jon Sit

B.S. Electrical Engineering, B.A. Computer Science
Yale University, 2000

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Electrical Engineering

at the

Massachusetts Institute of Technology

September 2002

Signature of Author……  ……
Department of Electrical Engineering and Computer Science
August 28, 2002

Certified by……Rahul Sarpeshkar……
Rahul Sarpeshkar
Assistant Professor of Electrical Engineering
Thesis Supervisor

Accepted by……...
Arthur C. Smith
Chairman, Committee on Graduate Students
Department of Electrical Engineering and Computer Science

A low-power analog logarithmic map circuit with offset and temperature compensation for use in bionic ears

by

Ji-Jon Sit

Submitted to the Department of Electrical Engineering and Computer Science on August 28, 2002 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Electrical Engineering

Abstract

Logarithmic map circuits are useful in many applications that require non-linear signal compression, such as in speech recognition and cochlear implants. A logarithmic current-mode A/D converter with temperature compensation and automatic offset calibration is presented in this paper. It employs a dual-slope, auto-zeroing topology with a 60 dB dynamic range and 300 Hz sampling rate, for capturing the envelope of speech signals in a bionic ear. Fabricated in a 1.5 μm process, the circuit consumes only 1 μW of analog power and another 1 μW of digital power, and can therefore run for over 50 years on just a couple of AA batteries. At the current level of power consumption, we have proven that this design is thermal-noise limited to a 6-bit precision, and higher precision is possible only if we expend more power. As such, it is already useful for cochlear implants, as deaf patients can only discriminate 1 dB out of a 30 dB dynamic range in the auditory nerve bundles. For the purpose of using this circuit in other applications, we conclude with several strategies that can increase the precision without hurting the power consumption.

Thesis supervisor: Rahul Sarpeshkar
Title: Assistant Professor of Electrical Engineering

Acknowledgements

The author would like to thank the following (in no order of merit):

1) Micah O'Halloran, for many hours of useful discussion, moral support, and for bringing the WLR's 355 fF well capacitance to my attention,

2) Chris Salthouse, for helping me proofread my first draft, pointing me in the right direction in many aspects of the noise analysis, and for tracking down my "butt" wires mistake in fabrication as soon as he did,

3) Michael Baker, for very helpful instructions on the use of the SRS785 spectrum analyzer, and much late night company,

4) Mary O'Malley, for always being there in the morning (and afternoon, and evening) to help with all the important details of getting a thesis out the door,

5) My advisor, Rahul Sarpeshkar, for truly being our fearless leader, conceiving and spearheading all the exciting projects that are in this lab, but most of all for believing in me and always challenging me to grow,

6) The rest of the AVNSL (now turned AVBS) team, for providing the exciting and challenging camaraderie that I doubt I will find anywhere else,

7) David Diel, for patiently listening to my description of the thermal-noise limitation, and then brightly suggesting offset compensation at the output,

8) My bible study and fellowship, for their faithful prayers throughout my time here at MIT,

9) My parents and sister, for their concern perhaps deeper than my own, their many years of unfailing support and love, and for walking beside me in every step of this thesis, AND

10) My God, without whom I would not even be here.

Table of Contents

1 Introduction and Background

1.1 What is a bionic ear?

People have been known to suffer from deafness for all of recorded history, but only in the last decade has medical engineering been able to restore hearing to the deaf. Thousands of deaf people have now benefited from a prosthetic device called a cochlear implant [1], which uses a surgically implanted array of electrodes to directly stimulate the auditory nerve and recreate the sense of hearing in the brain. The name "bionic ear" also refers to such an implant.

The current generation of cochlear implants performs all its signal processing outside the body, while the implant itself consists only of an electrode array and receiving coil to pick up electrically-encoded impulses.

The bionic ear we are developing, however, is intended to be sufficiently small and power-efficient as to be fully implanted with the electrodes. It exploits a sub-threshold analog CMOS design to achieve the required size and power efficiency, but more than that, also lends itself to the development of a neuromorphic[1] implant that has a higher degree of biological realism than current DSP solutions. Such a neuromorphic cochlea architecture was first presented in [2].

1.2 Why do we need a logarithmic map in a bionic ear?

One of the first things we learn about sound is that loudness is measured in units of "dB." This simply reflects the fact that perceived loudness goes as the log of sound pressure level (spl). Since pressure amplitudes are represented to our brains on a logarithmic scale, it suggests that there must be some mechanism in our ears to perform this compression[2]. Our logarithmic map circuit seeks to replicate this compressive function in the ear.

1.3 Why is compression needed, and how is it done in biology?

A deeper question is why the ear needs to perform logarithmic compression in the first place. The answer lies in observing that our ears are equipped to handle an astonishingly wide dynamic range. When loudness is quoted in "dB," it really refers to dBspl, with 0 dBspl corresponding to a pressure wave of merely 20 μPa in amplitude[3].

---

[1] A term coined by Carver Mead in the 80's, describing how systems can implement the same basic operations, and thereby follow the same organizing principles, as that of the nervous system [3].

[2] The inner hair cells, which perform the last stage of signal transduction from sound pressure waves into electrical impulses in the cochlea, are believed to be (at least partially) responsible for this non-linear compression [4].

[3] A 20 μPa sound is empirically measured to be the threshold of audibility in most people [5].

As a 20 Pa pressure wave is about our threshold of pain in the ear (see Fig. 1), this means that the ear has an input dynamic range of 120 dB! Gain control performed by the outer hair cells is primarily responsible for this astounding spec [6], by actually adding energy to undamp the cochlea mechanics when only quiet sounds are detected, thus performing selective amplification of the lower 50-60 dB.

Type of Sound	dB	Pressure (Pa)
butterfly wings	0	0.00002
rustling leaves	20	0.0002
whispering	25	0.00036
quiet library	30	0.00063
hum of refrigerator	45	0.00356
average home	50	0.00632
normal conversation	60	0.02
dishwasher	65	0.03557
car interior on freeway	70	0.06325
downtown street corner	75	0.11247
outboard motor	80	0.2
electric shaver	85	0.35566
screaming child	90	0.63246
convertible on freeway, top down	95	1.12468
subway train	100	2
jackhammer	105	3.55656
sandblaster	110	6.32456
rock concert	120	20
threshold of pain	120	20
air raid siren at 1m	130	63.24555
jet engine	140	200
instant perforation of eardrum	160	2000
shuttle launch at ground zero	180	20000

Fig. 1. Chart listing the dBspl (shown here as "dB") and equivalent pressure amplitude of various sounds [5].

Hence gain control already assists in compressing the 120 dB of input range into a more manageable "internal dynamic range." But however much smaller this internal dynamic range may be, still more compression will be required before stimulation of the auditory nerve fibers, which have at most 30 dB in electrical dynamic range [4]. In fact, the SNR in the Organ of Corti is believed to be at most 25 dB [7], supporting the view that sound must indeed be heavily compressed (and likely distributed) en route to the nerve[4].

[4] In addition to a compressive mapping, there are many theories on how sound is further compressed by neural encodings, but in the absence of a proven model, current implants (including ours) avoid implementing anything more complicated than the basic log compression [8].

Since cochlear implants take on the entire chain of signal transduction to stimulate the auditory nerve directly, it should be clear that they must perform the same signal compression found in the biological cochlea, and fit an output signal onto the limited dynamic range in the nerves.

1.4 Are things any different for deaf patients?

Deaf patients may have much less than 30 dB of electrical dynamic range in their auditory nerve fibers, with a typical level possibly being around 10 dB [9]. One of our design goals is therefore to have a selectable output dynamic range between 3 and 30 dB, so we can custom fit our device to the entire range of patients.

In addition, deaf patients have been found to discriminate a change in intensity of only 1 dB out of 30 dB. Our output, when digitally encoded, should therefore have at least 5 bits of precision.

2 An Overview of the Bionic Ear project

2.1 A brief system-level overview

The bionic ear, as with all cochlear implants, must function as a surrogate middle and inner ear up to the point of nerve stimulation. Hence we must start off with a microphone, which acts in place of the middle ear and tympanic membrane, as a sound sensor.

The microphone provides a current input to a preamplifier, called the analog front end (AFE) [10]. A senseamp in the AFE then drives a bank of band pass filters (BPF), that have their center frequencies log-spaced along the auditory band, just as different regions along the cochlea respond selectively to different frequencies. Thus there will be as many channels of BPFs as there are electrodes, to stimulate different sections of the cochlea.

Each band pass filter then drives an "inner hair cell" (IHC) circuit, that extracts the envelope of the signal by rectification and peak-detection. The envelope detector is designed to have an attack time of 10ms and a release time of 100ms, since the envelope in each band of speech is thought to vary no faster than 100Hz[5].

Finally it is this envelope that arrives at the input of the logarithmic map circuit (logmap). At this point, the signal has 60dB of dynamic range and is bandlimited to 100Hz, thereby setting our input specification.

[5] For this reason, other implants which perform envelope detection by rectification and low-pass filtering have their low-pass cutoff frequency typically between 200 and 400 Hz [8].

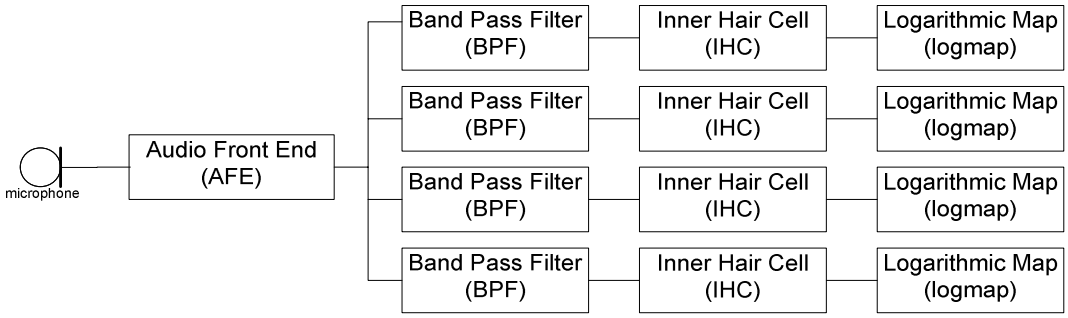


Fig. 2. A basic block diagram of a bionic ear processor as described above, showing just 4 channels for simplicity.

2.2 Specifications on the logmap

The logmap must be a black box which takes in a signal envelope with 60 dB dynamic range and 100 Hz bandwidth, and compresses it logarithmically into the 3-30 dB of nervous dynamic range in a patient.

We decided to log-compress the 60 dB of input range into a fixed bit precision using an analog to digital converter (A/D), and then have those bits shipped out to another chip that is solely responsible for driving the electrodes.

This external "stimulator chip" is therefore responsible for rescaling our bits into electrode currents which match the varying dynamic range in each patient.

The decision to put the stimulation off-chip was made for 3 important reasons:

1) **Scalability Issues**: different patients have vastly differing thresholds of audition, varying in some cases by up to $500\times$[6]. In terms of stimulation currents put out by the electrodes, this means that the minimum output level may vary from 1 μA to as much as 500 μA! If we try to incorporate this degree of scalability into our circuit, it means the circuit must perform the same log-compression function invariant over a 54 dB range in output current, which is much harder problem.
2) **Power consumption**: only the stimulation currents have to be as large as the μA to mA level. By decoupling the output specifications from the log-compression function, we can operate the logmap at a fixed sub-threshold bias (of a few nA), and only scale up to the high powered output when it is actually required. This allows us to save a lot of power in the logmap, and avoid any high power consumption near our sensitive low-power electronics.

[6] The variation in proximity of the electrode to auditory nerve fibers may account for much of this. Other sources of variance are likely pathological (like neuron degeneration in sensorineural deafness) [8].

3) **Compatibility**: Providing output digital bits to an off-chip stimulator allows our system to interface directly with existing DSP implants. Our system can then be easily swapped with a DSP processor, and allows the use of a proven DSP system as a backup to ours.

It was therefore determined that our logmap will be a low-power logarithmic A/D converter. The design goals are listed as follows:

#1. It must have a current mode input that covers 60 dB of dynamic range.
#2. The input bandwidth and sampling frequency must be >200 Hz (the Nyquist rate).
#3. It should have a precision of >5 bits.
#4. It should be temperature and offset insensitive.
#5. It should dissipate a power of 1 μW or less, as the logmap in a 40-channel implant is budgeted to take at most 40 μW.

2.3 Why is temperature and offset compensation important in a bionic device?

I will talk a little more about design goal #4. Firstly, it is true that the implant will be in a very well regulated thermal environment (of 36.9°C). However, heat exchange through the case may be poor, as it is thick and is not necessarily designed to be a good thermal conductor. Thus local thermal gradients within the chip (especially if the high powered electrode drivers are nearby) may wreak havoc with the accuracy of our A/D, if it is not compensated for temperature variation.

In addition, there is an exponential sensitivity to temperature in sub-threshold devices [13], and if we use this device in other applications, e.g. low-power speech recognition, thermal effects become much more important.

In addition to temperature effects, a related concern is how these implants perform over long periods of time. It is inevitable that a patient's thresholds to electrode stimulation will drift slightly, but if we can eliminate as much component and electronic drift as possible, their visits to the audiologist for tuning of the implant may not be as frequent. Hence it would be desirable for us to perform continual offset compensation to enforce a zero input, zero output condition at all times.

3 A/D design choices, leading up to a dual slope architecture

3.1 Why choose a time-based topology?

The ultra-low power budget is our tightest constraint. Flash and oversampling A/Ds are immediately thrown out because they are inherently high-speed designs, and do not have topologies that lend themselves to being exploited at low power levels. For example, there is no

running away from static power dissipation in a flash, nor is there a way to avoid a fast clock in a Σ-Δ converter.

However, a time-based (integrating) A/D seemed possible because it could take advantage of the slow sampling requirement to churn out a large number of bits when given a large amount of time, and thus win back a high precision even when faced with stringent power constraints.

This insight also arises from the general principle that the product of an A/D's speed and precision roughly works out to be a constant limited by power dissipation. This was implicitly observed by Bob Walden in his survey of modern A/Ds, when he described the A/D figure of merit, F, as $2^{SNRbits} \times$ sampling frequency / power dissipation [11].

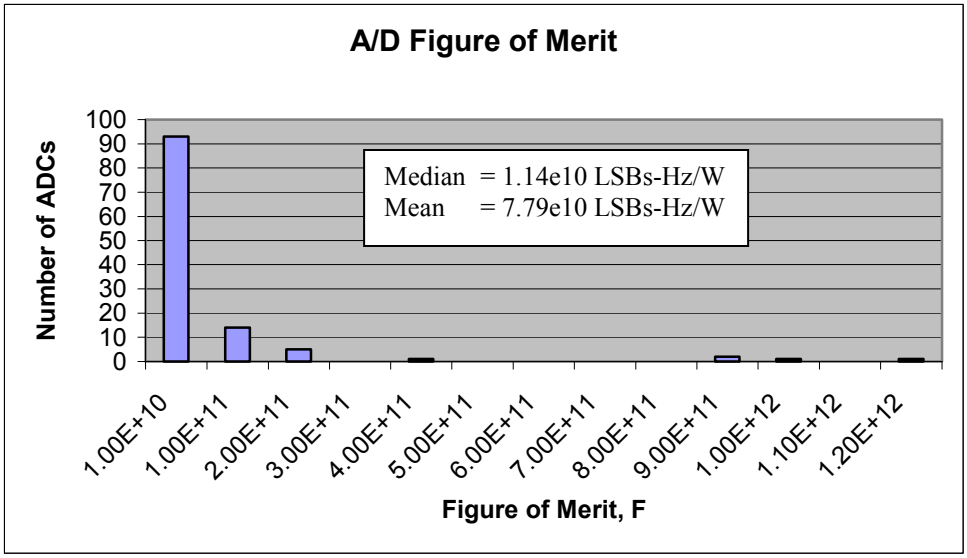


Fig. 3. Walden compiled the figure of merit for 117 modern A/Ds [11], which I have plot here as a histogram. Note that almost all (94 out of 117) the A/Ds have a value of F between 1e10 and 1e11. All but 5 (96%) fall between 1e10 and 3e11. This shows that the speed×precision product *per unit power* is practically bounded by a window only 30dB wide.

Consequently, it makes sense to exploit the requirements of low power and slow sampling frequency with a topology that is inherently low speed, in order to gain higher precision.

Now between topologies that "unroll" a bit precision out in time, we have the choice between the family of successive approximation converters and integrating converters. I chose to go with integrating converters.

3.2 Why choose an integrating topology?

If I chose a successive approximation topology, I would basically have to precede it with a sample and hold (S&H), since it is crucial that the input is held steady while the A/D settles and converges on an output. If the input is a moving target, the successive approximator can be completely thrown off[7].

Now we will compare this with an integrating A/D. In signal processing terms, the choice between a S&H and an integrator boils down to choosing between a delta function or a sinc to do our sampling.

A S&H basically allows us to take digital samples by multiplying the input signal with a train of delta functions, spaced by the sampling period Ts.

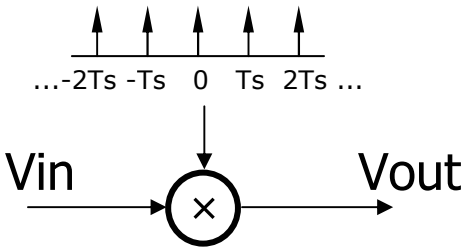


Fig. 4. Signal representation of a sample and hold.

An integrating A/D however will generate digital samples by convolving with a box (the width of the box being the duration of the integration), and then sampling with a delta function train running at the sampling frequency (f_s).

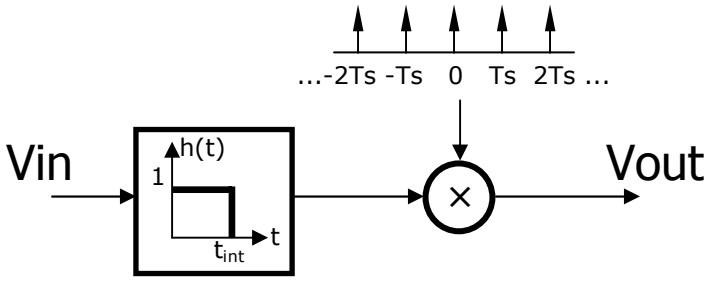


Fig. 5. Signal representation of an integrating sampler, that integrates for time t_{int} and then samples that integral immediately thereafter; repeating every Ts seconds.

[7] Take for example a 5 bit output of 01111. A successive approximator would determine that the MSB is 0 on its first step, but if the input then moves up by just one LSB (to 10000), the successive approximator may conclude the final output to be 00000, which is a whole MSB/2 off the mark.

The only difference is that the integrating A/D first convolves the input signal with a box, which is multiplication by a sinc in the frequency domain. This has the benefit of acting as a noise attenuation filter as well as an anti-aliasing filter!

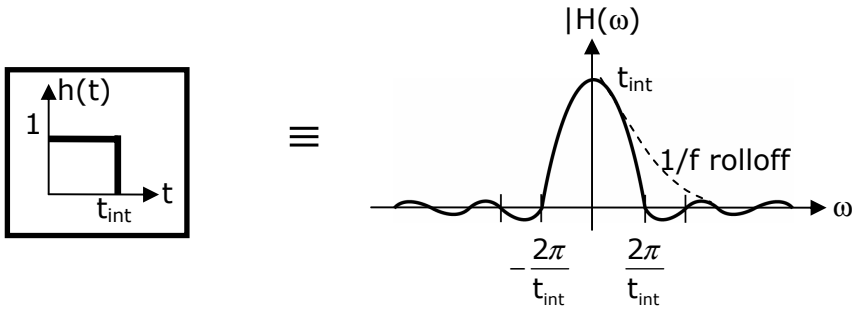


Fig. 6. Equivalence between an integration for time t_{int} in the time domain and multiplication by a sinc in the frequency domain.

This can be seen because the tail of a sinc rolls off as 1/f, hence it has a low-pass filter effect that gets rid of high frequency noise past the $2\pi/t_{int}$ cutoff frequency. In addition, we know that any frequency content in the input $>f_s/2$ will cause aliasing in our digital representation. The sinc low-pass thus acts as an anti-aliasing filter as well, while preserving the low frequency content of our signal.

In addition, if there is a stable noise frequency in our circuit, we can position the nulls of the sinc to coincide with those frequencies, to kill the noise.

We can also calculate the distortion (frequency-specific attenuation) that the integration (i.e., the sinc) will induce. In this converter I chose t_{int} = ¼ Ts, so the maximum attenuation at $f_s/2$ is only 2.6%. If this is too high, t_{int} can be decreased as a fraction of Ts, which broadens the flat portion of the sinc, but at the cost of increasing the high frequency noise. This relationship is quantitatively derived in the following equations:

$$\begin{aligned}
\text{Sinc}(\omega) &= \frac{2\cdot\text{Sin}(\omega\cdot\frac{t_{int}}{2})}{\omega} \\
\text{Attenuation}(\omega) &= \frac{\text{Sinc}(\omega)}{t_{int}} \\
\text{Attenuation}(\frac{\omega_s}{2}) &= \frac{2\cdot\text{Sin}(2\pi\cdot\frac{t_{int}}{4Ts})}{\pi\cdot\frac{t_{int}}{Ts}} \qquad (1) \\
\text{when } t_{int} = \frac{1}{4}Ts, &= \frac{8\cdot\text{Sin}(\frac{2\pi}{16})}{\pi} = 0.974
\end{aligned}$$

3.3 Why choose a dual-slope?

It turns out that the twin goals of offset and temperature stability can be most easily accomplished with a dual-slope converter:

1. There are standard techniques in the literature on auto-zeroing a dual-slope converter as part of the A/D conversion cycle [12].
2. The ratiometric nature of the dual-slope conversion allows us to divide out all the temperature dependent terms in the transfer function.
3. The dual-slope architecture is independent of component drift (such as in the size of the integrating capacitor), in contrast to successive approximation converters that often rely on matching a string of capacitors (in charge redistribution D/As) or resistors (in resistor divider D/As).

To explain how offset and temperature stability are achieved, I will first explain the basic principles of operation in a dual-slope converter.

The dual-slope topology that we will use is shown below in Fig. 7. It consists of an integrator, a comparator and a storage capacitor for offset compensation.

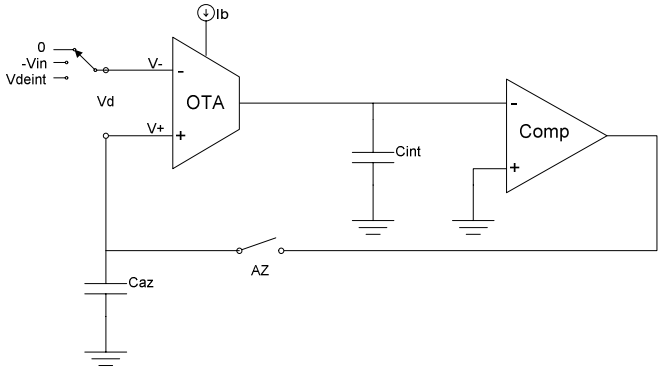


Fig. 7. A simplified representation of our dual-slope converter. The OTA together with Cint function as the integrator, and Caz is an offset storage capacitor for use in an auto-zeroing phase.

This dual-slope converter operates in three phases, as follows:

Phase I is an auto-zeroing phase, which happens when switch AZ is closed and V- is tied to some reference (shown as 0V in Fig. 7). Negative feedback then forces the offset of the OTA and comparator to be stored on Caz and Cint, respectively.

Phase II is the integration phase, which starts when switch AZ opens, and V- is switched to the input (shown as -Vin in Fig. 7). The OTA then charges Cint with a current I_{int} proportional to Vd, for a fixed amount of time.

Phase III is the deintegration phase. The polarity of Vd is flipped by switching V- to a reference input (shown as Vdeint in Fig. 7), causing the OTA to discharge Cint with a current I_{deint}, back to its original state. When this happens, the comparator trips and signals the end of deintegration.

The three stages are diagrammed below in Fig. 8.

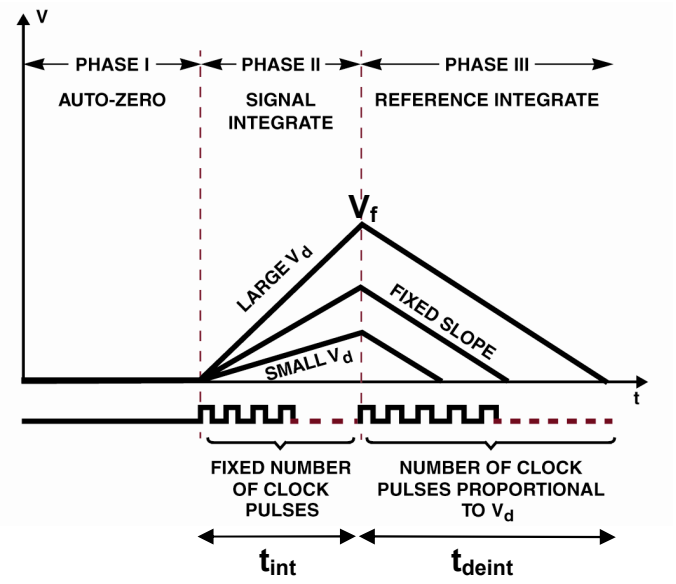


Fig. 8. A single dual-slope conversion cycle. The y-axis shows how the voltage on Cint varies over one cycle. Cint stores the offset voltage of the comparator during phase I. During phase II, Cint charges up for a fixed duration t_{int}, and reaches a final voltage V_f based on the size of Vd. During phase III, Cint is discharged with a fixed current, back down to the trip point of the comparator. The duration of this phase, t_{deint} is then proportional to Vd.

Thus the conversion can be written out as follows:

$$\begin{aligned} V_f &= \frac{I_{int}}{Cint} \times t_{int} \quad (I \rightarrow V \text{ transformation}) \\ t_{deint} &= \frac{V_f}{I_{deint}/Cint} \quad (V \rightarrow t \text{ transformation}) \\ &= \frac{I_{int}}{I_{deint}} \cdot t_{int} \end{aligned} \tag{2}$$

t_{deint} is thus a proportional measure of I_{int}, computed through the intermediate variable V_f.

This relation can be derived more simply if we note that no net charge was added to the capacitor over the conversion cycle, since Cint starts and ends in the same state. Then by conservation of charge:

$$\begin{aligned} Q_{added} &= Q_{removed} \\ \mathrm{I}_{\mathrm{int}} \cdot t_{\mathrm{int}} &= \mathrm{I}_{\mathrm{deint}} \cdot t_{\mathrm{deint}} \\ t_{\mathrm{deint}} &= \frac{\mathrm{I}_{\mathrm{int}}}{\mathrm{I}_{\mathrm{deint}}} \cdot t_{\mathrm{int}} \end{aligned} \qquad (3)$$

Dual-slope conversion therefore has the property of path-independence (in voltage), which supports the fact that both capacitance and voltage must be intermediate variables which cancel out in Eq. (2). Any non-linearity or hysteresis in the capacitance thus turns out not to matter, which is a nice side benefit.

The main benefit, however, is that t_{deint} is proportional to a ratio of currents, rather than the input current alone. Since I_{int} and I_{deint} are both generated by the same OTA with bias current Ib (see Fig. 7), Ib (and any temperature coefficient that it may have) is basically divided out when we take the ratio of the slopes.

4 Making the converter logarithmic

One detail that I left out till now is how the input voltage Vin is generated. So far, this dual-slope can very well function as a linear A/D (for as long as the transconductor is operating linearly). To be a logmap, Vin must therefore take the log of the input current.

4.1 Logarithmic current-to-voltage conversion

If we send a current through a diode, the voltage developed will be the log of that current. Equivalently, we can send a sub-threshold current into the source of a source-follower, which also has an exponential V-to-I relation [13]:

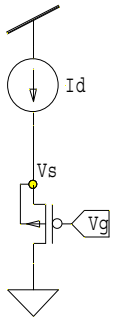


Fig. 9. A PMOS in a source-follower configuration, taking as its input a sub-threshold input current Id.

And the voltage developed on Vs is given by:

$$V_{sg} = \frac{\phi_T}{\kappa} \ln \frac{I_d}{I_o} \tag{4}$$

In addition, the value of V_s can be level-shifted by applying any voltage we want on V_g. This will be useful in performing offset compensation, as the gate can be connected to an intentional capacitor and charged with offset information as shown in Fig. 10.

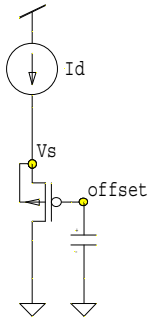


Fig. 10. The source voltage Vs, level shifted by an offset stored on the gate.

4.2 Division of currents by taking a differential voltage

In order to define the logarithmic baseline of 0dB for our input current, the logarithmic voltage dropped across the PMOS diodes will be taken differentially. Therefore a second PMOS diode is designated to carry a reference current, I_{ref}, equal in value to the minimum current level seen at the input:

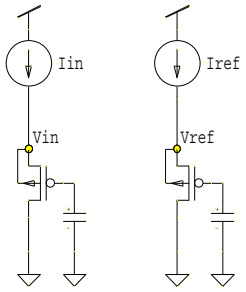


Fig. 11. Differential voltage inputs, Vin and Vref.

Now by taking the difference between these two logarithmic voltages,

$$
\begin{aligned}
V_d &= V_{in} - V_{ref} \\
&= \frac{\phi_T}{\kappa}\left(\ln\frac{I_{in}}{I_o} - \ln\frac{I_{ref}}{I_o} \right) \\
&= \frac{\phi_T}{\kappa} \ln\frac{I_{in}}{I_{ref}}
\end{aligned}
\tag{5}
$$

Note that I_o, which is a highly temperature-dependent term, is thereby divided out. Vd is thus a measure of I_{in} in dB, with units of ϕ_T / κ volts per e-fold.

4.3 Linear voltage-to-current conversion, canceling ϕ_T in the process!

Vd is then presented as the input to our OTA, as shown in Fig. 12.

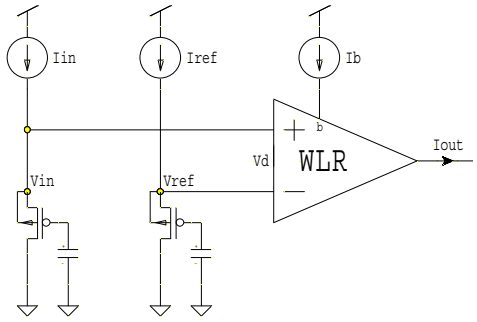


Fig. 12. Production of Iout proportional to Vd using a wide-linear-range OTA (WLR).

Now as long as Vd lies well within the linear range of the OTA (V_L), the following linear relationship will hold:

$$
\begin{aligned}
I_{out} &= G_m \cdot V_d \\
&= \frac{I_b}{V_L} \cdot \frac{\phi_T}{\kappa} \ln\frac{I_{in}}{I_{ref}}
\end{aligned}
\tag{6}
$$

However, a vanilla 5-transistor OTA only has a V_L of about 80mV. In order to achieve 60dB of dynamic range, our OTA must be linear over an input range of 40mV · 3 · 2.3 ≈ 280mV. Known degeneration techniques were therefore applied to the design of a wide-linear-range OTA (WLR) to increase its linear range [14].

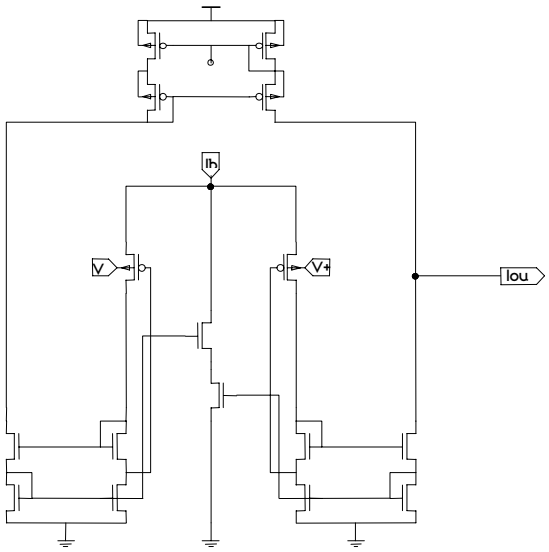


Fig. 13. The WLR used in our design, with well inputs, gate-degeneration and bump linearization to achieve a linear range of >1V. The Wilson mirrors help us achieve higher output impedance for more ideal integration, and lowers the number of thermal noise sources per mirror from 2 to ~2/3 [derived in appendix].

The linear range of this WLR can be derived to be:

$$V_L = \frac{2\phi_T}{1-\kappa_p} \cdot \left(1+\frac{\kappa_p}{\kappa_n}\right) \cdot \frac{3}{2}\Bigg|_{\substack{\kappa_p=0.85 \\ \kappa_n=0.6}} \tag{7}$$

$$= 1.2\text{V}$$

The value of κ_p that we have used is slightly higher than usual (around 0.75), because κ_p increases with well-to-gate potential [14], and Vwg happens to be large in our case since we operate the WLR at a high common-mode voltage (of around 2.2V).

It can now be noted that Iout will have the original PTAT dependence in Vd cancelled out, since V_L is also proportional to ϕ_T.

4.4 Experimental DC transfer curves for our WLR

V-to-I transfer curves with 2 different bias currents were taken for this WLR, and are shown below in Fig. 14. One input is held at 2.2V (our ~common mode during normal operation) while the other was swept over ±300mV. The output current measured with a Keithley 6514 electrometer is then plot against the input.

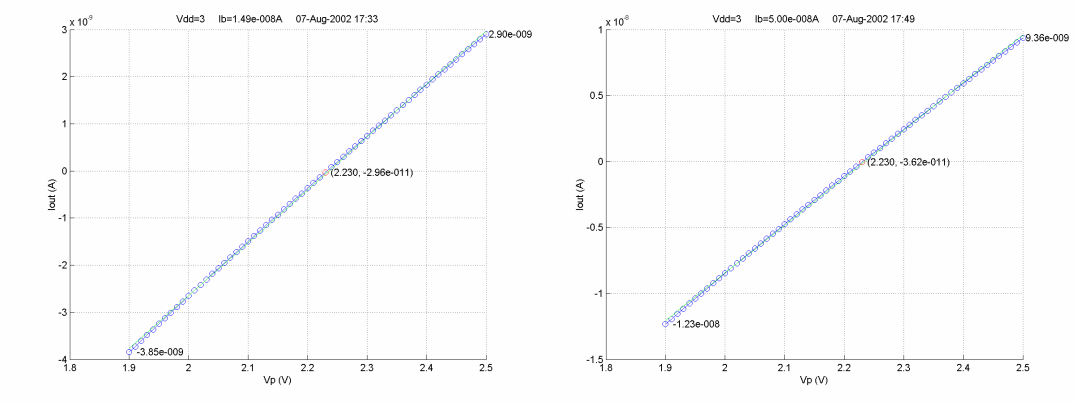


Fig. 14. V-to-I output curve for our WLR over ±300mV when biased at 15nA and 50nA. We can observe that the output is sufficiently linear for the dynamic range we need.

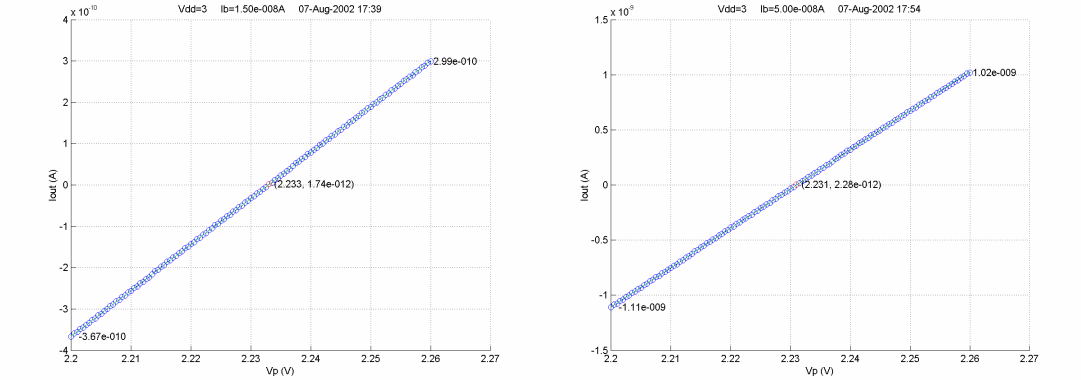


Fig. 15. Zooming in to the origin to get the value of Gm. By definition, Gm is the slope at the origin, and equal to Ib / V_L. For 15nA and 50nA biases levels, Gm is measured to be 11.1nA/V and 35.6nA/V. Hence V_L = 1.35V and 1.40V respectively. The increase in V_L with bias current is to be expected as we approach moderate inversion, since the Gm per unit Ib starts to fall [14].

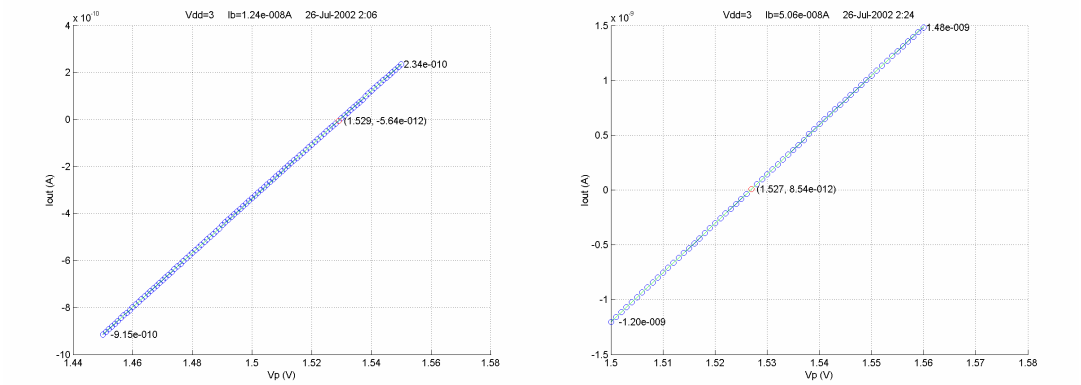


Fig. 16. Performing the same sweeps at a lower common mode voltage of 1.5V show a noticeable increase in Gm. For 12.4nA and 50.6nA biases, the Gm is measured to be 11.5nA/V and 44.9nA/V. Thus V_L = 1.08V and 1.13V respectively. This is also to be anticipated due to the decrease in Vwg and hence κ_p [14].

We can thus predict that our WLR should have a V_L of at least 1.1V during normal operation, which should be sufficiently linear for our input range of 280mV.

5 Details behind autozeroing the OTA and comparator

5.1 Stability concerns

The observant reader may have noted that the circuit shown in Fig. 7 is probably unstable, as it contains two high gain nodes loaded with large capacitances! Indeed when we hook it up in simulation, it oscillates wildly. The loop was stabilized by using a standard miller compensation technique:

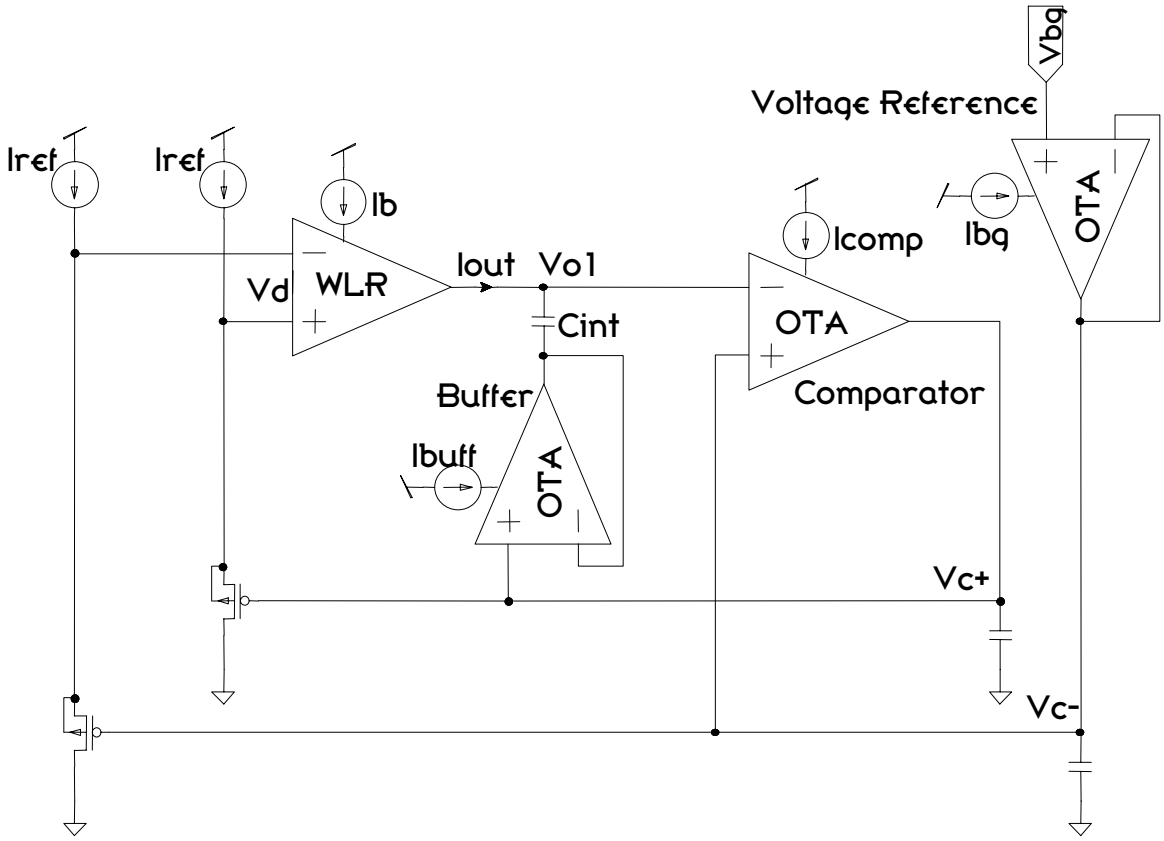


Fig. 17. The miller compensated autozero loop. The key step in stabilizing the loop was to transform the integrating capacitor Cint into a miller capacitor during the autozero phase. The 3 blocks labeled as 'OTA' are just high-gain amplifiers serving as a buffer, comparator and voltage reference. The voltage reference Vbg sets up some midrail DC voltage for the loop to servo to, and is biased identical to the comparator for matching charge injection to Vc+ and Vc-, which will be covered further later.

The buffer in Fig. 17 also does double duty in both the autozeroing and integration stages to prevent feedforward through Cint. During integration, it acts as a level shifter and virtual ground, preventing corruption of the voltage on Vc+. During autozeroing, it acts as a series resistor that inserts a LHP zero via the feedback path by forcing the series conductance to roll off from being

capacitive (sCint) to resistive (1/Gm). This further stabilizes the loop by adding positive phase near crossover.

If we perform a small-signal and block diagram analysis of this loop, we discover that it is almost identical to the compensation of a 2-stage opamp:

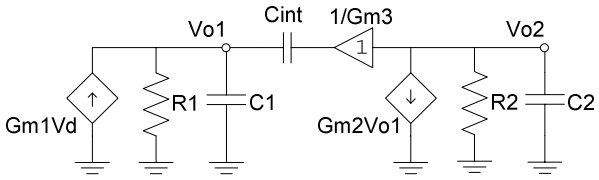
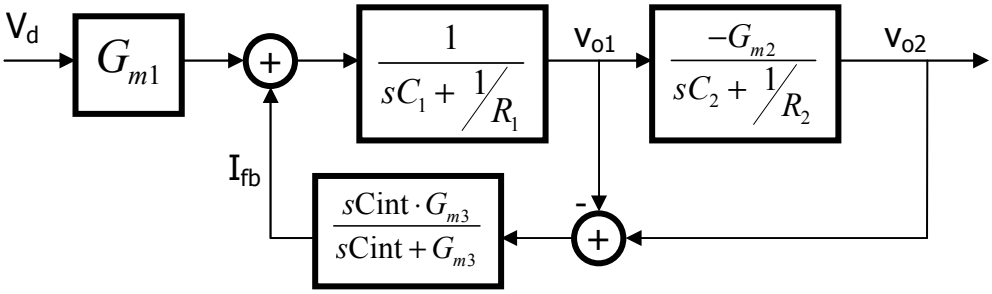


Fig. 18. Small-signal model of the minor loop.



$$\textit{Defining}: \quad A_1 = G_{m1}R_1 \qquad A_2 = G_{m2}R_2 \qquad \tau_c = G_{m1}/\text{Cint}$$
$$\tau_1 = R_1C_1 \qquad \tau_2 = R_2C_2 \qquad \tau_3 = G_{m3}/\text{Cint} \tag{8}$$

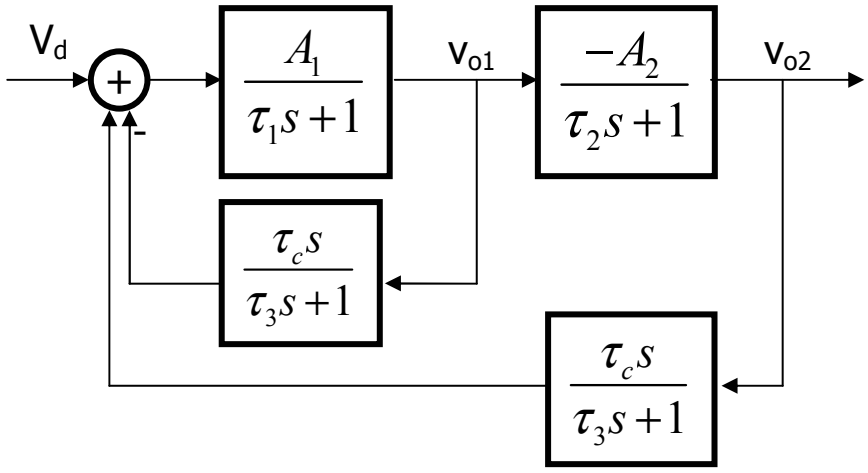


Fig. 19. Block diagrams showing the miller compensation and zero at τ_3 created by the buffer.

The transfer function then falls straight out of the block diagram as:

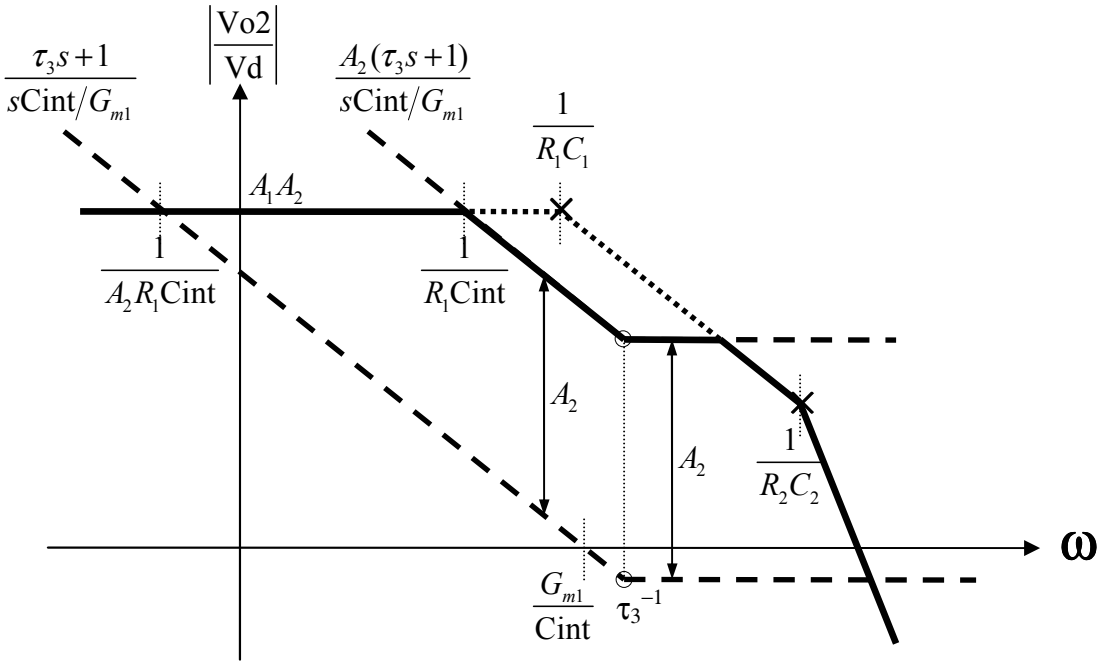


Fig. 20. The input-output transfer function |Vo2/Vd| for the autozero loop. The forward transmission is shown with a solid line, while the effect of the feedback blocks $\tau_c s / (\tau_3 s+1)$ are shown with dashed lines.

The bandwidth of the loop is thus set by G_{m1}/Cint with τ_3 providing extra positive phase near crossover to improve stability.

We observe the effect of the differenced feedback current (Ifb, in Fig. 19) on the forward transmission in Fig. 20. The poles $1/\tau_1$ and $1/\tau_3$ get split, creating a slower pole at $1/(R_1\text{Cint})$ and a faster pole beyond $1/\tau_3$, together with the appearance of a zero at $1/\tau_3$. By effectively moving the first order roll-off to a higher frequency beyond $1/\tau_3$, this has the advantage of decreasing the negative phase contribution from the forward transmission. Since the forward transmission now breaks at a lower frequency but recovers with a zero at a higher frequency, this effect can also be regarded as classic lag compensation.

An interesting observation we can make is that the minor loop stability is not adversely affected by the appearance of τ_3, although it would appear to be adding a pole in the loop transmission. This is because the impedance seen at Vo1 goes up at precisely the same frequency when the capacitive feedback from Vo2 rolls off to become resistive. Hence the pole in the capacitive feedback term cancels with the zero in the forward transmission, and contrary to our first guess, the minor loop is not destabilized.

5.2 Settling concerns

It is important to determine whether the autozero loop has enough time to settle before the start of conversion. This calculation goes as follows:

1) We know that the speed of settling is given by the bandwidth of the loop, and therefore the settling time constant is τ_c = Cint / Gm1.

2) We shall now calculate the minimum amount of time available for settling. As described previously, I had chosen t_{int} to be ¼ Ts, and at full scale, t_{deint} was set to take ½ Ts. Thus this leaves a minimum of ¼ Ts for the AZ phase.

3) Thirdly, we know that we must set the maximum slope of integration such that the final voltage, V_f, does not exceed the maximum swing at the output of the WLR, which we shall call Vfs. This maximum slope will be generated when the WLR is presented with its full scale input, Vd-max. The maximum slope is therefore the maximum output current from the WLR, Iout-max, divided by Cint, where Iout-max = Gm1 · Vd-max.

4) We can then express the minimum time given to autozeroing = ¼ Ts = t_{int} as the ratio Vfs / (Iout-max / Cint).

We can therefore conclude the following:

$$\begin{aligned} \min \#\tau_c \text{ for settling} &= \frac{\text{Vfs}}{\text{Iout-max}/\text{Cint}} \div \frac{\text{Cint}}{G_{m1}} \\ &= \frac{\text{Cint} \cdot \text{Vfs}}{G_{m1}(\text{Vd-max})} \times \frac{G_{m1}}{\text{Cint}} \\ &= \frac{\text{Vfs}}{\text{Vd-max}} \simeq \frac{1.2\text{V}}{280\text{mV}} = 4.3 \end{aligned} \tag{9}$$

This is an interesting result because the number of time constants available for settling is independent of the sampling period Ts. This arises because we always increase Iout-max in the same proportion as any increase in sampling frequency, to fully utilize the maximum output swing Vfs.

It is desirable to maximize the use of the full output swing (i.e. maximize Iout-max) not only to get the fastest settling, but to also provide maximum immunity to voltage noise at the output, since voltage noise will then be a smaller fraction of full scale. In physical terms, this maximizes the amount of charge stored on the capacitor, so any stray charges injected by noise then show up as a smaller fraction of the total. We shall see later that improving this ratio of Vfs / Vd-max will turn out to be highly important in increasing our precision.

Our current ratio of Vfs / Vd-max gives us 4.3 time constants, and indicates a settling to better than 1 part in 72, which is sufficient for 6 bits.

5.3 Noise-canceling benefits

The continuous auto-zeroing process also endows a huge win in terms of reducing our low frequency noise (particularly, 1/f noise).

In section 3.2, we already saw how an integrating stage attenuates high frequency noise outside our band of interest, by convolution with a box and then sampling by multiplication with an impulse train. High frequency noise is attenuated because the spectrum gets multiplied by a sinc before sampling.

Auto-zeroing performs the exactly same operations, just in the reverse order: we first sample the offset by multiplying with an impulse, and then hold that sample by convolving with a box. This "sample and hold" effect also multiplies the spectrum of the offset by a sinc, thus preserving all the low frequency information in the offset.

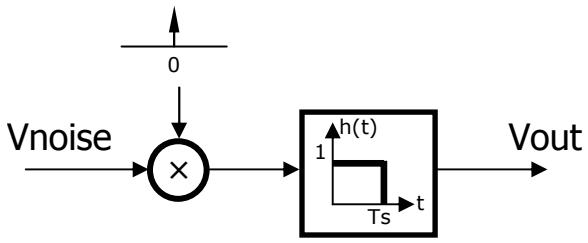


Fig. 21. Sampling of the offset and noise onto Vout, and held over one conversion cycle (Ts).

The sample and hold will therefore create a staircase series of voltages on Vout, if we perform this operation every cycle. However we do not want to consider the spectrum of this staircase waveform as each edge contains large amounts of high frequency noise that we do not care about when Vout is being reset. As we care about the noise cancellation only within one hold cycle, after which the sample is reset and the process starts over, we can consider just a single cycle starting at t=0 without loss of generality.

After the offset is sampled, it is placed in series with our input signal, but opposite in polarity to the original noise source, as follows:

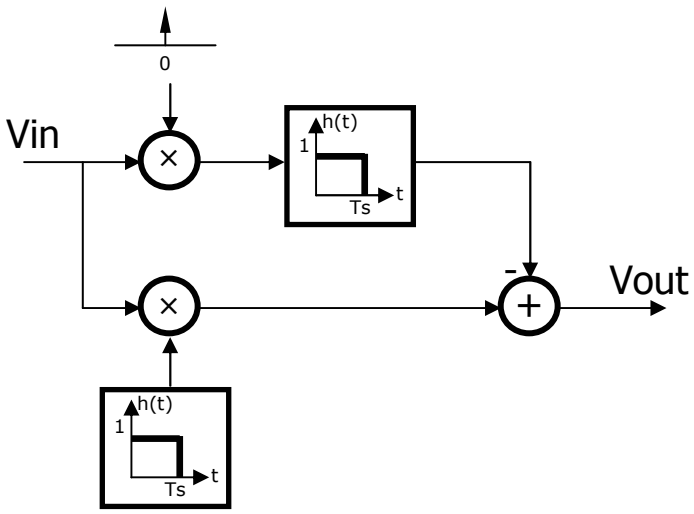


Fig. 22. Subtraction of the offset during normal operation over one conversion cycle (Ts).

In the frequency domain, this amounts to subtracting the sinc of the input noise from the noise convolved with a sinc (which is close to the original noise itself), thus showing that low frequency noise below the cutoff of $2\pi/t_{int}$ is indeed taken out by autozeroing. This is particularly useful since that is exactly our signal's frequency band of interest!

6 The dual-slope converter in full operation

6.1 Details of a conversion cycle

The circuit which implements the full-fledged dual-slope converter is shown below.

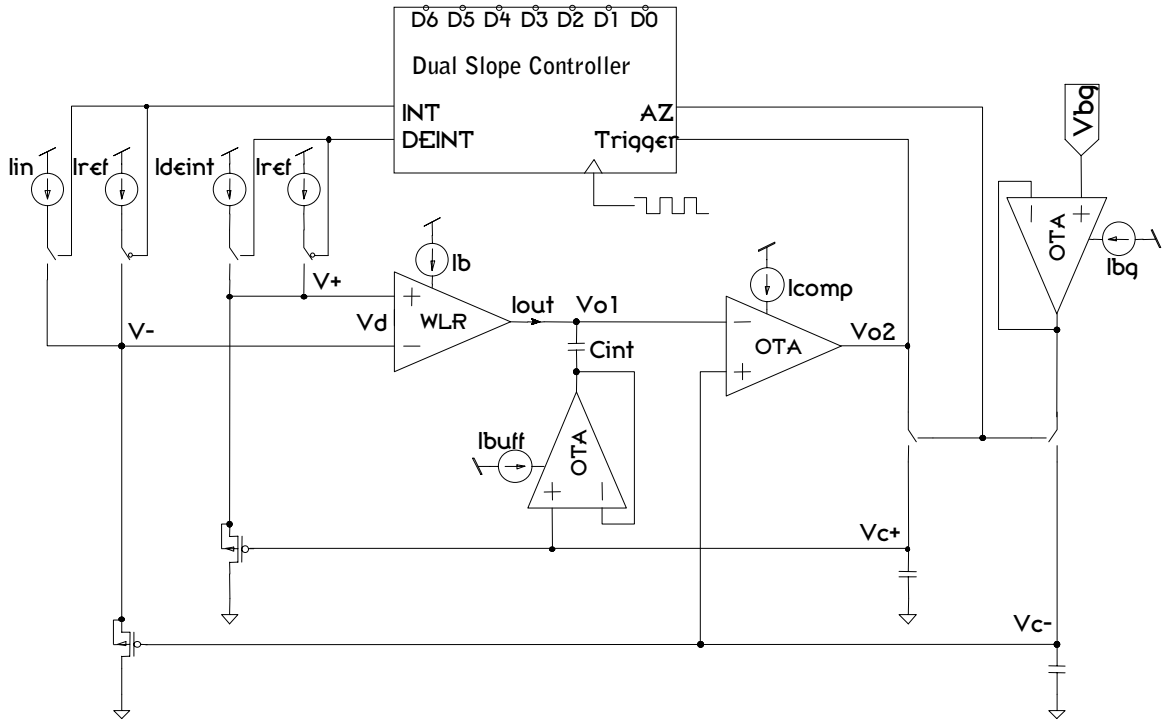


Fig. 23. The complete analog section is diagrammed here. It differs from Fig. 17 only by showing the logic switches that move us between the integrate, deintegrate and autozero phases.

We can summarize its operation as follows:

I) in the autozero phase, the controller raises digital signal AZ, which closes the switches to the two capacitors Vc+ and Vc-. Digital signals INT and DEINT are both low, so Iref is put into both PMOS source-followers. The miller compensated autozero loop then servos Vc+ to a differential offset away from Vc-, which is set up at some midrail DC level by Vbg.

II) in the integrate phase, AZ goes low and the feedback loop is broken, leaving the nodes Vc+ and Vc- to float. As long as equal amounts of charge are injected, the offset between them is preserved, and any common mode change is rejected by the CMRR of the WLR. Hence the voltage reference is laid out to look just like the comparator (in unity feedback) to match the charge injection to both sides. After the AZ switches open, INT goes high so the input current Iin is steered onto V-. DEINT is still low, so Iref is steered onto V+. This causes Vo1 to ramp down as the WLR sucks GmVd off Cint. The controller counts off a time t_{int} that is equal to ¼ Ts.

III) in the deintegrate phase, AZ stays low and INT goes low, while DEINT goes high, putting Iref into V- and I_{deint} into V+. This reverses the polarity of Vd so the WLR pumps a current = GmVd back onto Cint, while counting off a time t_{deint} that is latched when the comparator sends a falling edge trigger back into

the controller. I_{deint} is set up so a full scale count of t_{deint} occupies exactly ½ Ts, leaving a minimum of ¼ Ts left for the autozero phase.

A screen shot of a typical dual-slope waveform on Vo1 is shown below, together with the corresponding comparator output on Vo2:

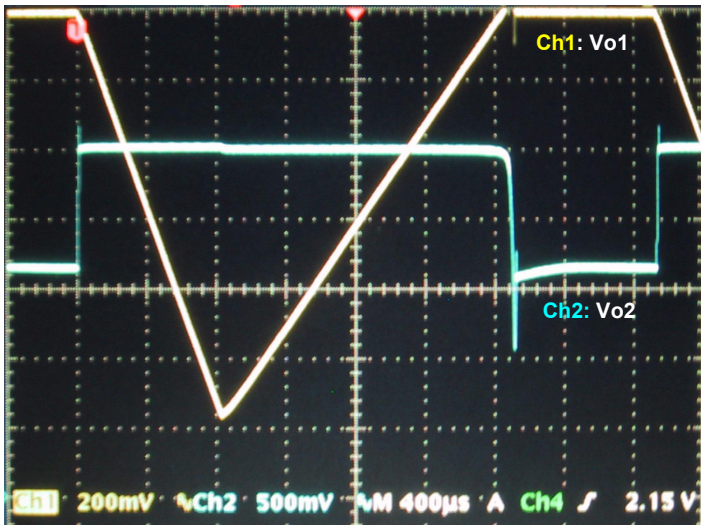


Fig. 24. The dual-slope ramps on Vo1 are displayed on Ch1, set to 200mV/div. The comparator output on Vo2 is displayed on Ch2, set to 500mV/div. Each horizontal division is 400 µs. This shows a full scale input of Vd-max being applied, hence Vo1 ramps down to a full scale V_f of ~1.2V. Note that t_{int} lasts for 833µs, which is ¼ Ts for a 300Hz sampling rate. t_{deint} lasts for ~1667µs before crossover, and thereafter leaves another ¼ Ts for autozeroing before the start of the next integration.

6.2 Analog power consumption

Referring to Fig. 23, the maximum current biases in the analog section are listed below:

Bias	Current (nA)	Mirrored Copies	Total Current (nA)	Total Power (nW)	Biasing Mirrors	Biasing Current (nA)	Biasing Power (nW)
Ib	15	1	30	90	1	15	45
Ibuff	100	0	100	300	1	100	300
Icomp	100	0	100	300	1	100	300
Ibg	100	0	100	300	1	100	300
Iref	0.4	1	0.8	2.4	1	0.4	1.2
Ideint	90	0	90	270	1	90	270
Iin (max)	400	0	400	1200	1	400	1200
TOTAL	805.4		820.4	2462.4		805.4	2416.2

Fig. 25. Table showing the power consumption of the converter. The converter is run off a +3V supply, for ease of using D cells to power.

The current consumption of the analog section is 820nA, but is roughly doubled when you include current mirror biasing. In the future we will include a single on-chip bias that should eliminate the need for double the power in biasing.

The converter currently consumes 2.5 μW when operating with maximum input current, but we will note that the input current takes up ½ the total power. If our average input current is the geometric mean of Iin and Iref (the average on a log scale), then <Iin> = 12.6nA and we should consume only an average of 1.2 μW.

6.3 Digital power consumption

Digital power goes as $f \cdot Ctot \cdot Vdd^2$, so we can reduce the digital power by trying to lower the clock frequency or the gate capacitance. I implemented 3 techniques to lower both Ctot and f:

1) The digital controller was designed to run asynchronous from the clock when responding to the trigger and latching the bits. Hence I could reduce the total capacitance (Ctot) by a factor of 5 from a synchronous design.

2) The clock itself doubles up as the LSB (using the low time to count a '0' and the high time to count a '1') thus one counter stage can be removed to run a factor of 2 slower than standard counters.

3) If we choose t_{int} = ½ Ts and t_{deint} = ¼ Ts, we need to run the clock frequency 4× as fast as $f_s \times$ the number of bits. This is because we need to count the full number of bits in the DEINT cycle, but the DEINT cycle is only ¼ of Ts. However if we give t_{deint} a bigger fraction of Ts by choosing t_{int} = ¼ Ts and t_{deint} = ½ Ts, the clock frequency only needs to be $2 \times f_s \times$ the number of bits, allowing us to run at ½ the speed of the previous case[8].

The total power savings come up to a factor of 5×2×2 = 20× better than an initial design, and has proven to lower the digital power from 19 μA to <1 μA. We can further improve the digital power draw by lowering DVdd. The circuit has been shown to work down to 2.5V before the transmission gates are not turned on hard enough.

Plots of the digital current draw and digital power dissipation against DVdd are shown below:

[8] This reduction in t_{int} also provides more accurate sampling (less frequency-dependent attenuation) at the cost of admitting more high frequency noise, as derived in section 3.2.

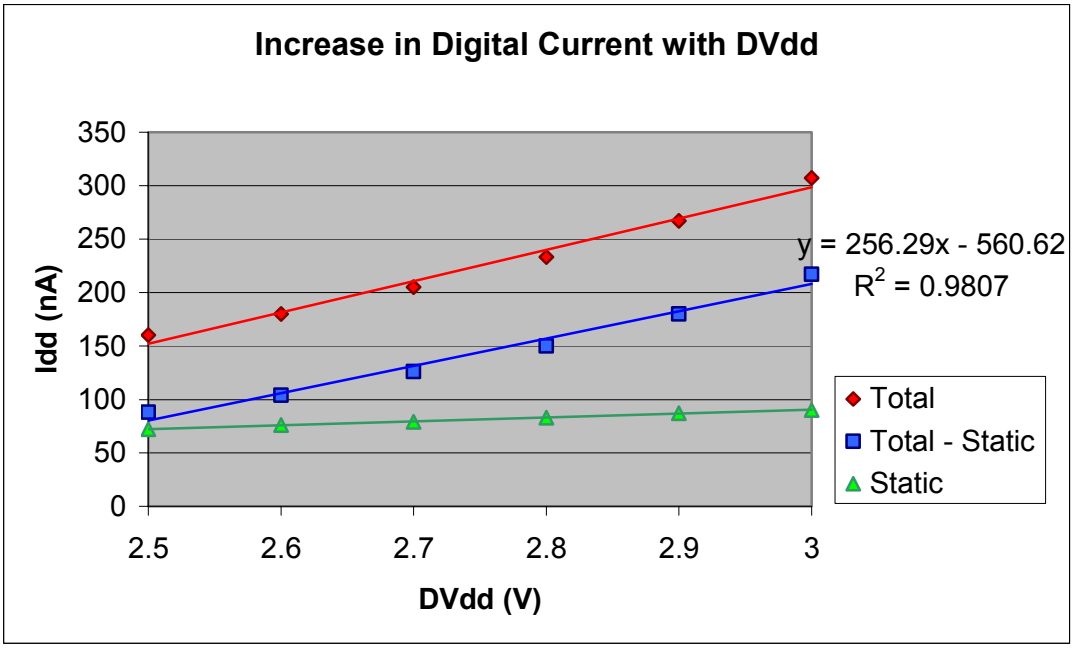


Fig. 26. Digital Current drawn by the Dual-Slope Controller. The static Idd was measured when the loop is reset to stay in the auto-zeroing cycle, and total Idd measured with a maximum Iin = 230nA. As expected, the current rises with DVdd, since Idd · DVdd = f · Ctot · $DVdd^2$.

The slope of the Total – Static Idd in Fig. 27 equal f · Ctot, but as we can see, the increase is not very linear even when the static power is subtracted. However, a linear plot of Idd vs. DVdd assumes that the power reaches a minimum static floor only when DVdd = 0. We may suspect otherwise, i.e. power dissipation may reach a static minimum at some DVdd > 0 due to leakage.

Hence if we plot the power dissipation of Total Idd · DVdd against DVdd instead, we can fit the curve to a quadratic that will take the leakage offset into account (by allowing another degree of freedom to place the minimum point in the quadratic). This curve is plot below, and indeed has an excellent fit:

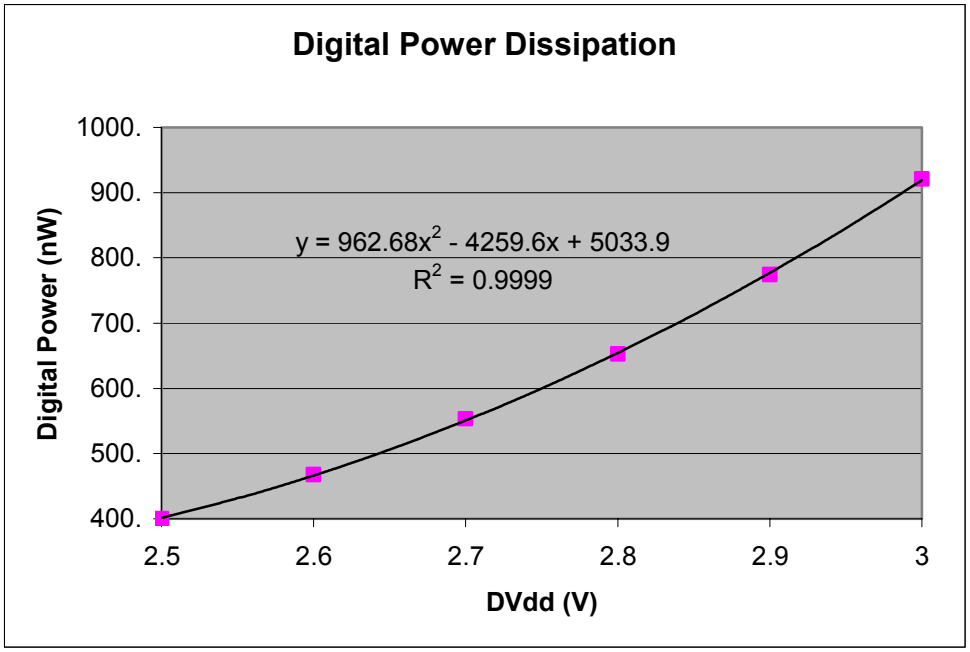


Fig. 27. Digital Power Dissipation in the Dual-Slope Controller. As expected, the power rises quadratically with DVdd.

If we complete the square for the quadratic fit above, we get $y = (31x - 68.6)^2 + 322$. This says that the quadratic should level off to a minimum when DVdd = 68.6/31 = 2.2V, and that the static floor is 322 nW independent of DVdd.

Furthermore we can predict our combined total capacitance, Ctot to be = 962.68n / f = 962.68n / (64 × 300) Hz = 50 pF in the entire digital section, that is charged and discharged every period.

Although the digital section consumes almost an additional µW of power, this should scale with technology. In a smaller process we can reduce this value of Ctot by the square of the feature size, and since we are running very slowly we can design custom cells with smaller gates (as opposed to using standard cells with huge gate drives). At that point, the digital power can be driven even further below a µW.

6.4 How many bits of precision?

When observing the dual-slope waveform on the scope, the A/D shows around 6 bits of precision both in measurements of the timing jitter and in the error rate of the digital output.

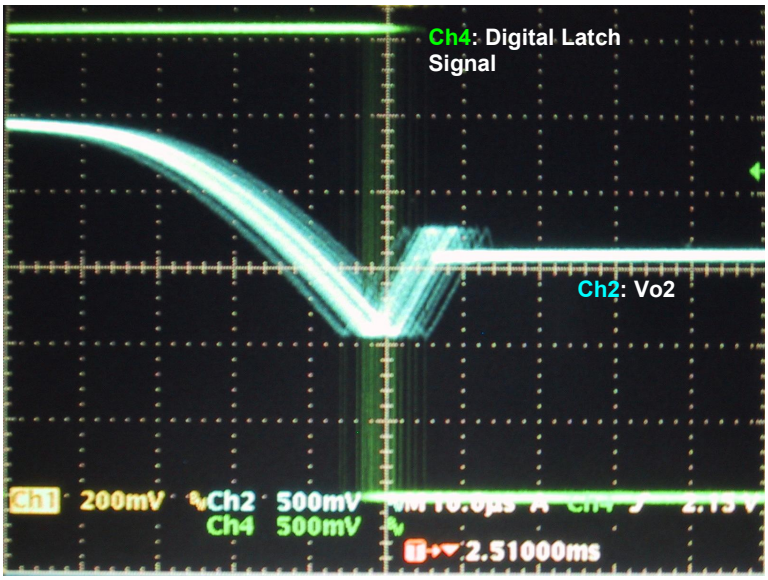


Fig. 28. A snapshot of the jitter at the end of deintegration is shown above. The comparator output Vo2 falls, and when it hits the inverter threshold of the digital latch, the digital latch sends a falling edge that latches the counter bits. We can observe here a jitter of ~12 µs, and ~20 µs of jitter is the maximum that is ever observed.

To quantify this more precisely, let us take some measurements. For a 300 Hz sampling frequency, with a full scale t_{deint} = 1.67ms, the maximum peak-to-peak jitter observed on a scope was 20 µs, which is about 1.2% or 1 part in 83.5. According to the standard definition that the number of bits (n) in an A/D can be found from its SNR as 6.02n + 1.76 (in dB), this simply says that:

$$SNR = \frac{t_{\text{full scale}}\ (\text{rms})}{t_{\text{noise}}\ (\text{rms})} \triangleq \frac{2^n \cdot LSB/2\sqrt{2}}{LSB/\sqrt{12}} = 2^n \cdot \sqrt{1.5} \qquad (10)$$

Since our SNR is measured to be about 83.5, by this definition the converter has exactly 6.09 bits of precision.

6.5 Static characterization of the A/D

For this measurement, the input current was swept over the 60dB input range we expect to see, by increasing the gate voltage on current source Iin. The digital output was then read by a logic analyzer, and plot against Iin. For each input value the logic analyzer would read 4096 samples,

and as expected, there is always bit error whenever the noise window straddles a bit transition edge. So I plot the mean and the median of the 4096 samples, together with the error rate on a separate chart.

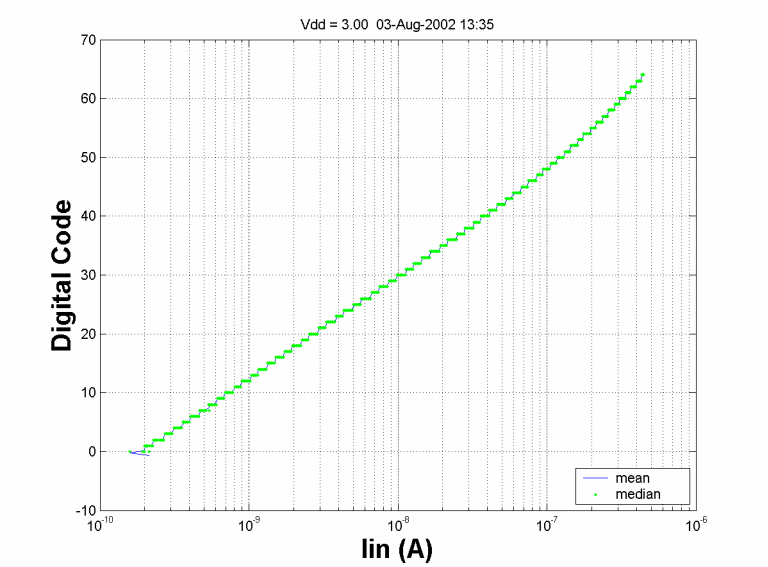


Fig. 29. Iin is increased on a logarithmic scale and the digital output is read by a logic analyzer.

There is a small offset error since Iref was set to be 400pA rather than 200pA (which is where the zero-crossing occurs). However the transfer characteristic is generally linear and definitely monotonic. At the minimum input, many bit errors start to occur, so ideally we want to incorporate an intentional offset to avoid operating the dual slope with a slope of zero. The curve starts to slope upwards at the high current levels because the PMOS transistors start to enter moderate inversion, and are no longer exponential devices. Hence they drop a larger voltage than expected, resulting in a larger digital code being reported. This non-linearity can be confirmed by plotting the digital code against the voltage input used to bias Iin:

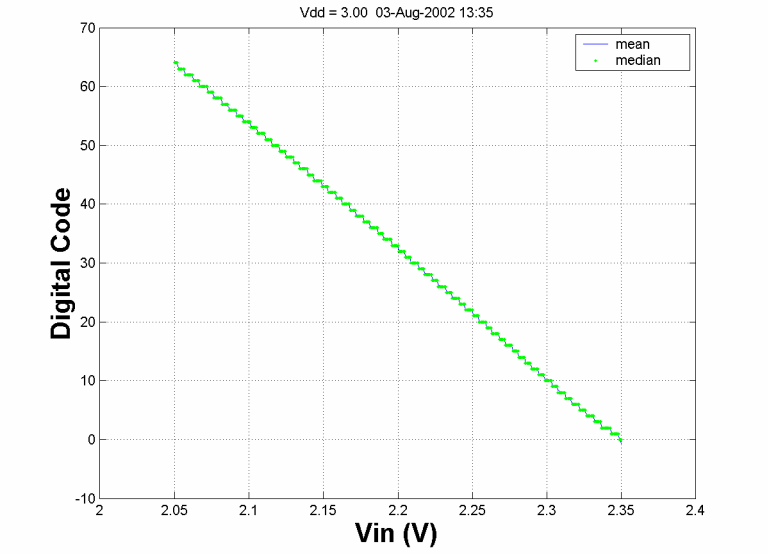


Fig. 30. Here we see that the converter is indeed linear in voltage, even at the high end of current, confirming that the non-linearity arises from the PMOS diodes leaving the sub-threshold region.

The bit errors for each sample point are shown below:

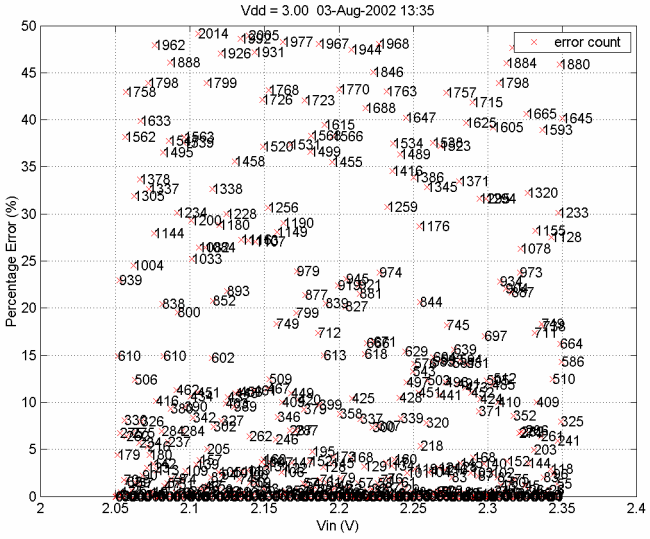


Fig. 31. The bit errors for the two transfer curves above. The numbers represent the number of errors (out of 4096 samples) that were read, taking the median value to be correct. As we expect, there are large numbers of bit errors due to bit transitions being crossed, but the errors are bounded by a maximum of 50% since we are never off by more than 1 LSB. In addition, the presence of a large number of perfect readings (0% error) indicates that the noise window is indeed smaller than 1 LSB, confirming that we indeed have ≈ 6 bits.

6.6 Linearity of the A/D

We will now quantify the exact linearity of the A/D using the following standard definitions:

Integral Non-Linearity (INL): INL is defined as the largest vertical difference (expressed in LSBs) between the code center points of the actual characteristic curve and the line connecting the endpoints on the curve [15].

Differential Non-Linearity (DNL): DNL (expressed in LSBs) is defined as the largest deviation in the width between two adjacent thresholds from the ideal width of 1 LSB on the analog (horizontal) scale [15].

Taking a linearity plot over 60dB from a minimum current of 210pA to 210nA, we find that the DNL is acceptably small (< ½ LSB) but the INL is rather high:

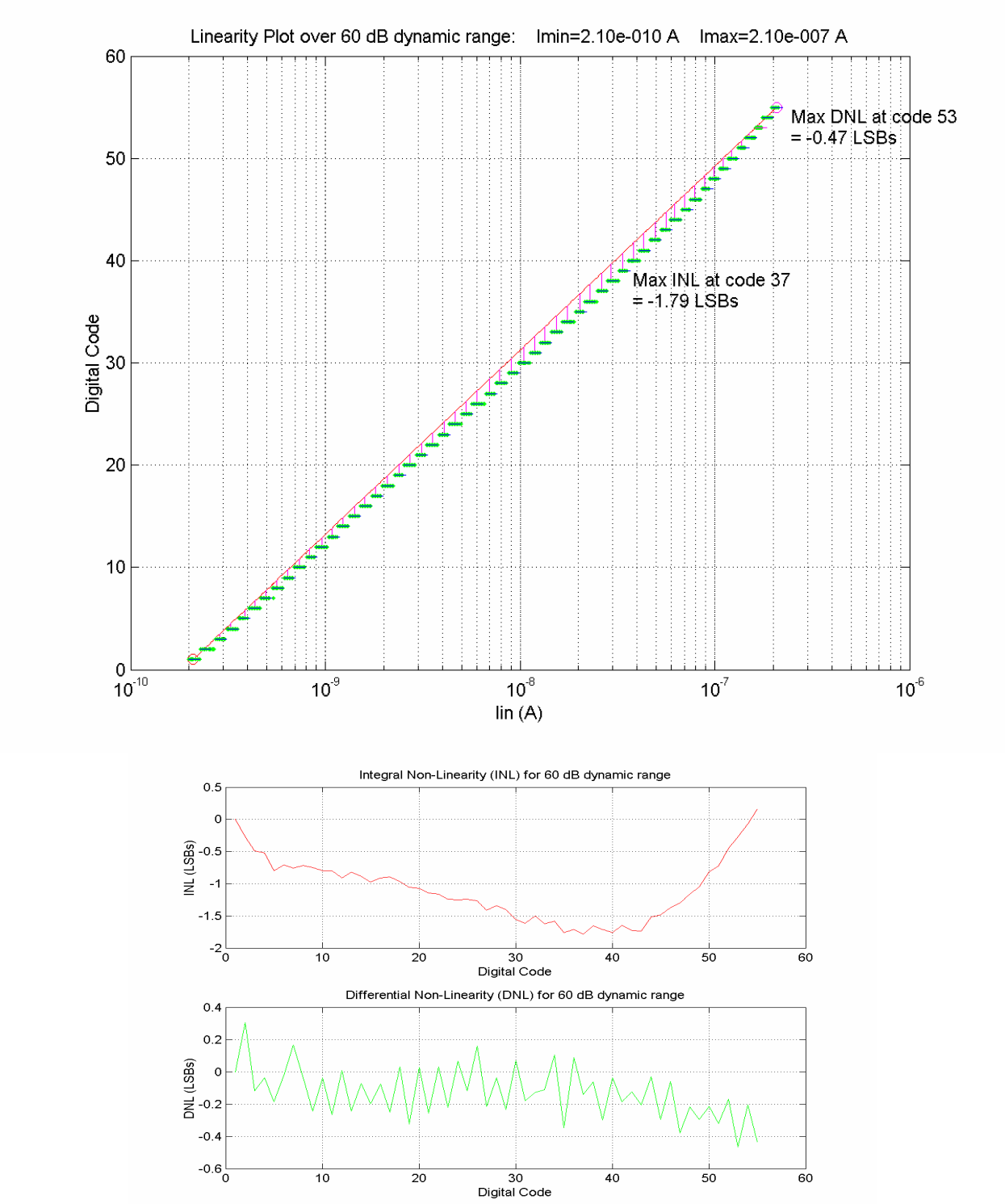


Fig. 32. Linearity plot over 60dB of dynamic range, showing the INL and DNL. The vertical lines in the linearity plot show the measured INL, while the horizontal lines represent one ideal LSB, for reference in measuring the DNL. Although the curve looks fairly linear, the transition to moderate inversion at 200nA induces almost 2 LSBs of INL.

To confirm that we indeed have better linearity if we avoid moderate inversion, a plot over a 54dB of dynamic range is shown below. The maximum current was reduced by only a factor of 2, from 210nA to 100nA:

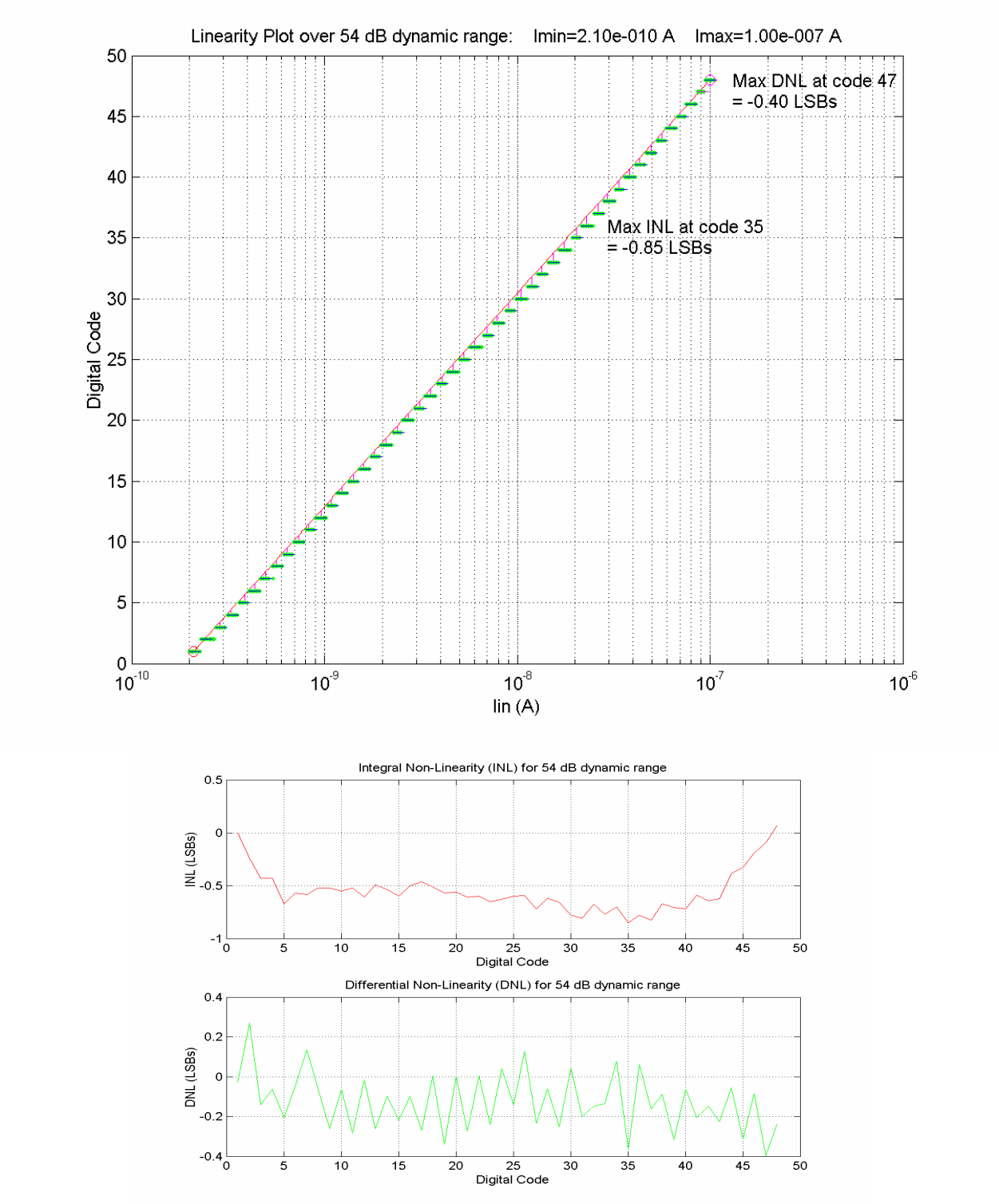


Fig. 33. Linearity plot over 54dB of dynamic range, showing the INL and DNL. Indeed if we just avoid the 100-200nA region of moderate inversion, our INL is capped below 1 LSB. The DNL remains low below ½ LSB.

This shows that moderate inversion of the MOS diodes does adversely affect our linearity. However we can conclude from Fig. 33 that our worst case INL of 2 LSBs only arises when the signal approaches the uppermost 6dB of operation. And in most cases, we can expect that INL is less important than DNL, because ascertaining the precise loudness of sounds is not something we are usually concerned about[9]. However DNL will be important to deaf patients because it allows them to accurately discriminate between two sounds when one is only slightly louder than another. With < ½ LSB of DNL, we therefore have better than 1% linearity for the purposes of discrimination.

7 Dynamic characterization of the A/D

7.1 Direct sampling of sinusoidal inputs

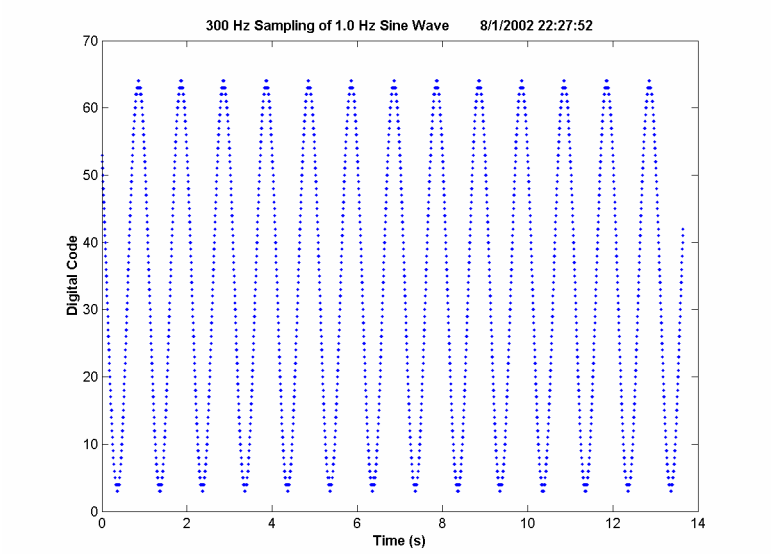


Fig. 34. Sampling of a 1 Hz input sinusoid. Note that the period is precisely 1s, and there are 300 points along 1 cycle since the sampling period is 3.3ms.

[9] For example, an integral non-linearity of 2 LSBs would mean a sound may appear 5× louder than another when it should really be only 4.8× louder.

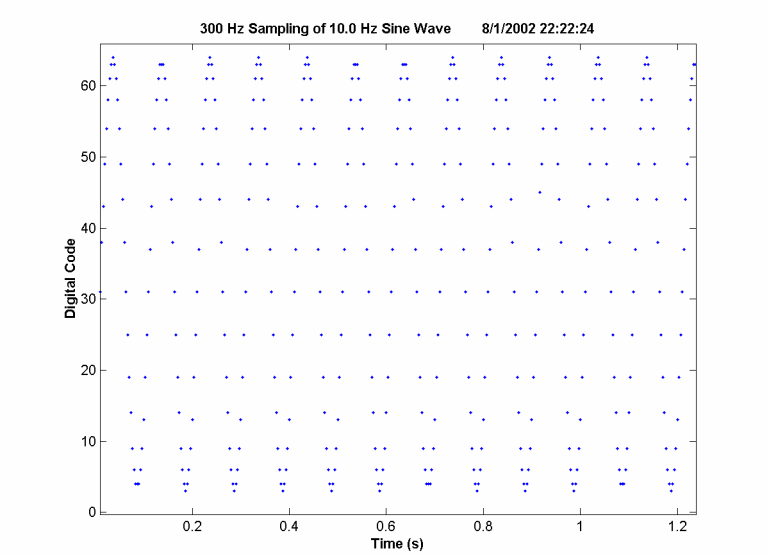


Fig. 35. Sampling of a 10 Hz input sinusoid. Note that the period is precisely 100ms, and there are 30 sample points per cycle.

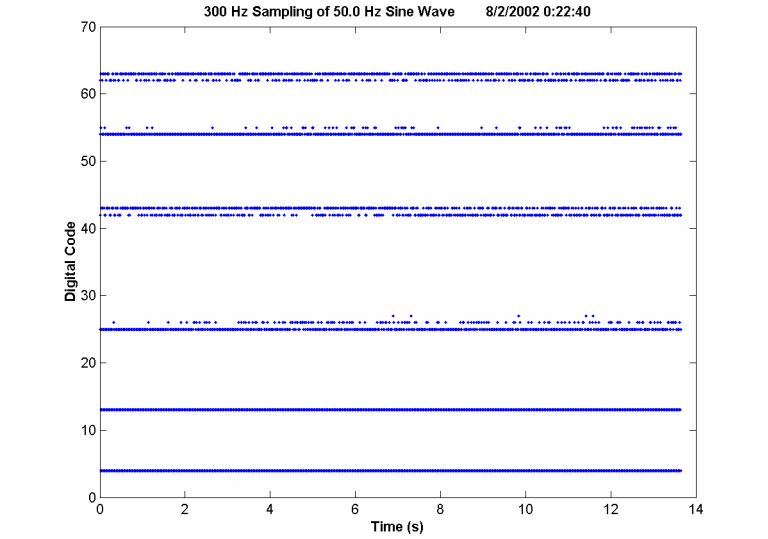


Fig. 36. Sampling of a 50 Hz input sinusoid. We now only have 6 points per cycle, which are almost exactly in phase from cycle to cycle, thus creating a striation of only 6 levels over the whole 14s.

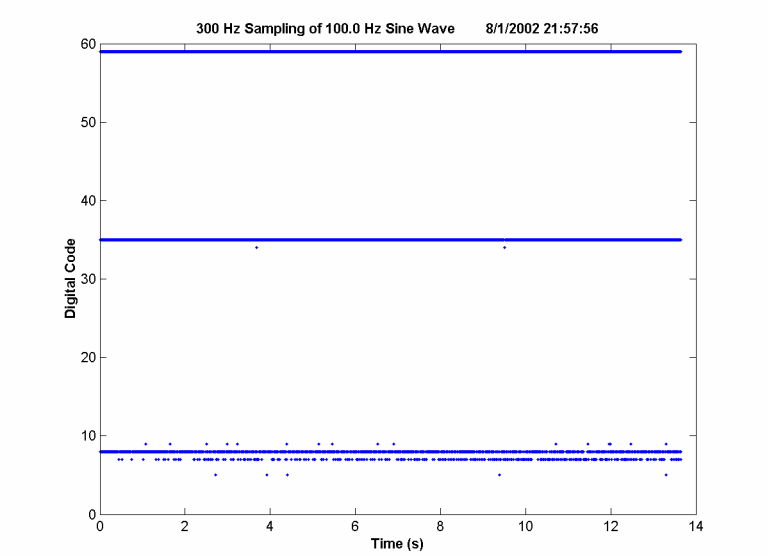


Fig. 37. Sampling of a 100 Hz input sinusoid. There are now only 3 levels since a 300 Hz sampling rate will pick off only 3 points per cycle.

7.2 Strobed sampling of sinusoidal inputs

We can get around the inability to see a completely sampled waveform if we apply an input frequency that is slightly faster than some integer fraction of the sampling frequency, by exploiting the stroboscopic effect to "walk" down the phase of a waveform.

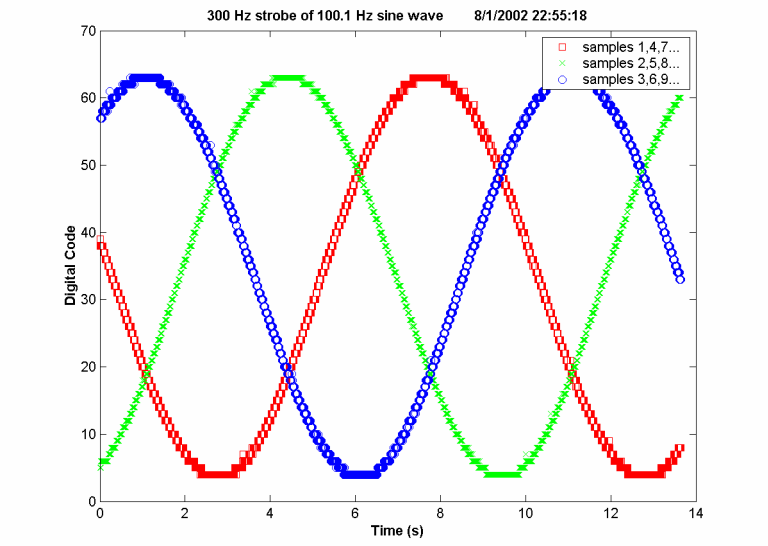


Fig. 38. Stroboscopic walk of a 100 Hz input sinusoid. The frequency of the walk is 0.1Hz, showing 1 complete cycle in 10s, which is easily visible. There is minimal distortion, confirming that the bandwidth of our converter is sufficient.

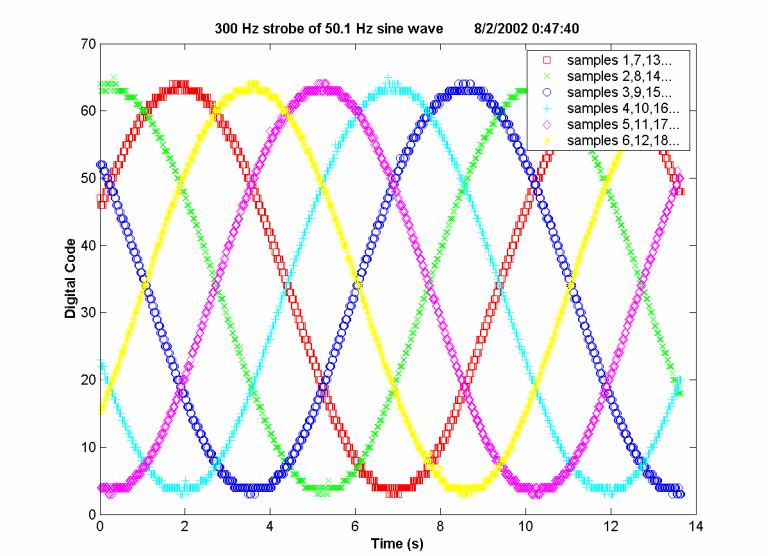


Fig. 39. Stroboscopic walk of a 50 Hz input sinusoid. The frequency of the walk is again 0.1Hz, but we have 6 overlapping waveforms since there are 6 points per cycle that get stroboscopically "walked."

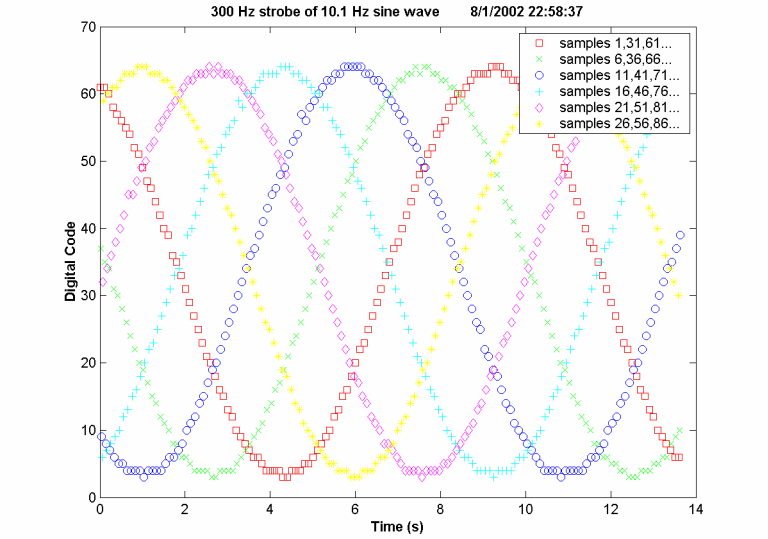


Fig. 40. Stroboscopic walk of a 10 Hz input sinusoid. There are now 30 points per cycle which get "walked" by the strobe, so we pick off only 6 for display.

7.3 Sampling of a typical attack and release envelope

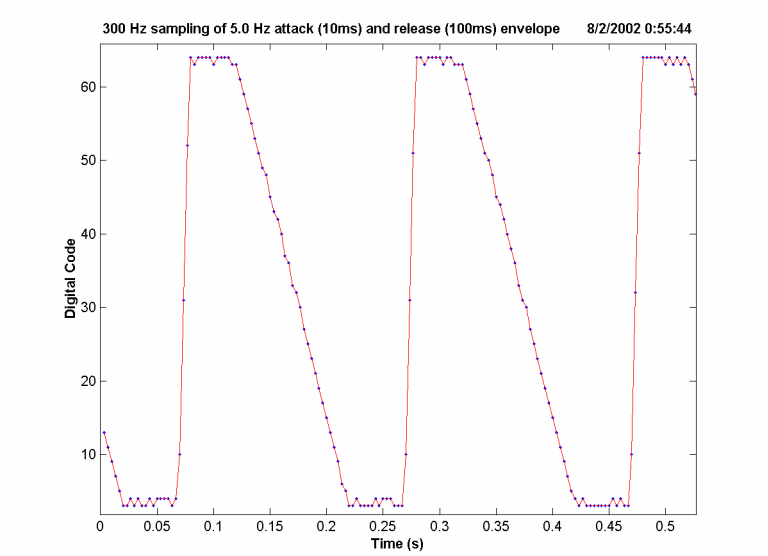


Fig. 41. An piecewise waveform was put in with a 10ms rise time and a 100ms fall time, as we expect from the attack and release time constants of the envelope detector. The logmap samples the envelope as expected, with 3 samples along the rising edge and 30 samples on the falling edge.

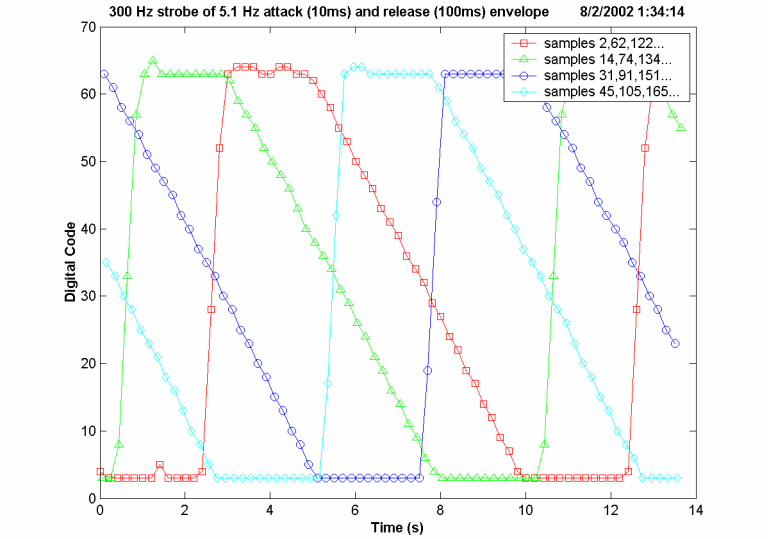


Fig. 42. Strobing the envelope to pick off different points along the rising edge, we confirm that the logmap has enough bandwidth to handle the speed of the envelope.

8 Characterization of the WLR and OTA

One of the most important questions to ask is: what limits our precision to 6 bits?

To understand the noise inherent in our system, the first piece of information we need is a complete characterization of both the WLR and OTA (used as the comparator and buffer). In this section I present data that confirms each of their effective V_L's and noise power spectral

densities (PSD). I would then be able to predict the noise contributions from the WLR and OTA in the converter.

8.1 Finding the BW, and hence V_L of the WLR

I had a WLR hooked up in a Gm-C buffer, so I used the spectrum analyzer to take a bode plot and fit the measured data with both theoretical values and simulation data. Simulation turned out to line up exactly with measurements if I added a 355fF capacitance to the well input, due to the unexpectedly large junction capacitance of the parasitic well-to-substrate diode.

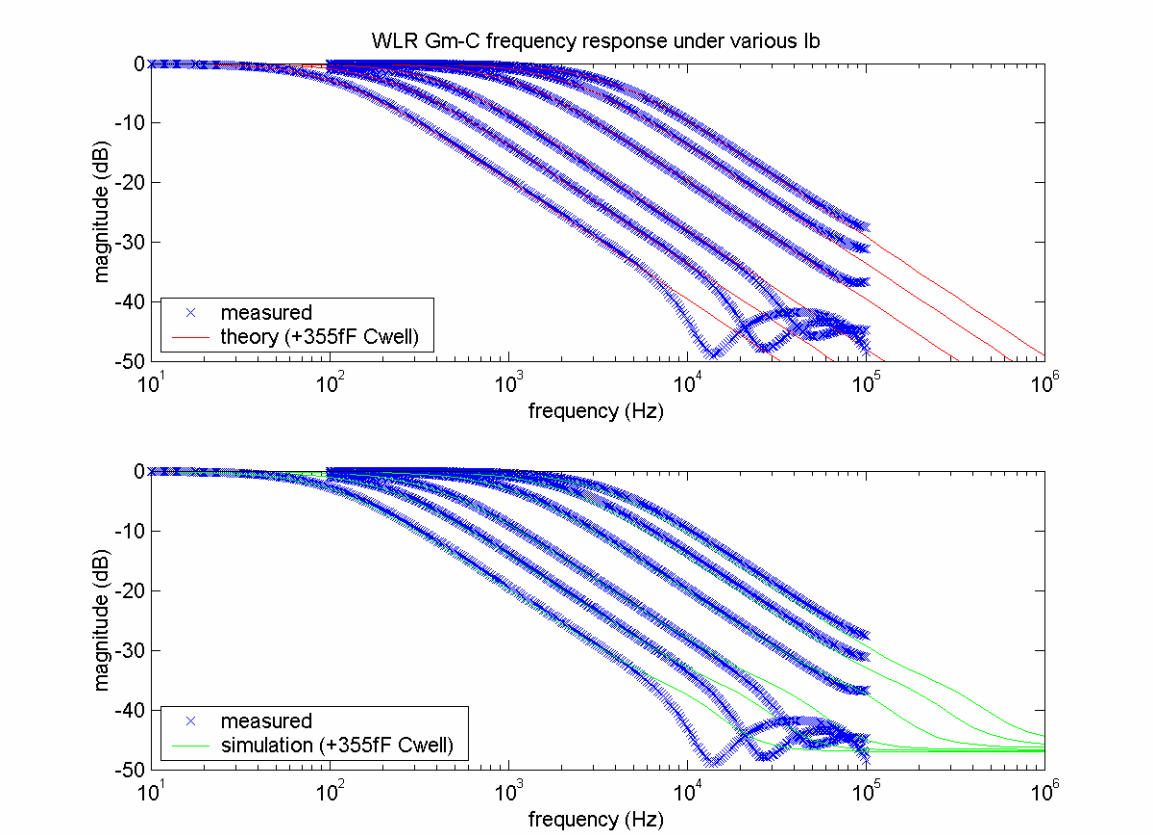


Fig. 43. Bandwidth of the WLR Gm-C filter at various bias currents. C=1.5pF, as is the case in the actual logmap. Simulation lines up extremely well with a value of 1.2V for V_L. Both theory and simulation confirm that there is an additional 355 fF on the well input that needs to be taken into account.

8.2 Finding the noise PSD, and confirming N (# noise sources) in the WLR

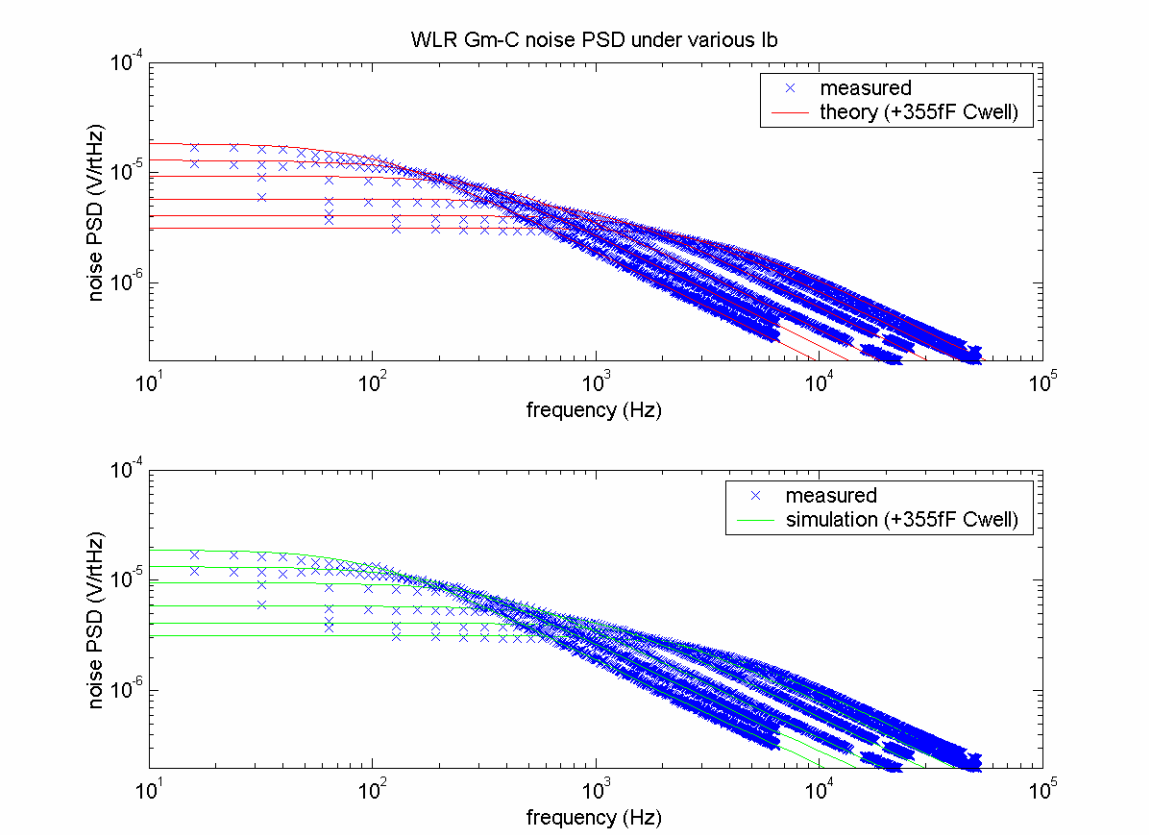


Fig. 44. Noise PSD for the WLR Gm-C filter at various bias currents. Both simulation and theory line up well, and confirms that the number of noise sources, N = 3.5.

The NoisePSD^2 predicted by theory is:

$$NoisePSD^2 = \frac{N \cdot 2q \cdot (0.3Ib)}{Gm^2} = \frac{Nq\,(0.6)\,V_L{}^2}{Ib} \qquad (11)$$

The theoretical number of noise sources in the WLR is ~3.5 [derived in appendix]. V_L is confirmed to be ~1.2V, and the value of Ib is well determined, so the only variable left to affect our NoisePSD is the factor by which the bump linearization reduces the current in each transistor. We shall call this the "bump current reduction," or BCR. Simulation shows that the current does not split evenly between left, bump and right transistors (see Fig. 13), as 1:1:1, but

rather divides 3:4:3 (consistently), over the entire range of bias currents. Hence I used a BCR of 0.3 in Eq. (11), rather than 0.33.

8.3 Integrating the total noise in the WLR

With values for both the NoisePSD and bandwidth, we can confirm the amount of noise we expect by integrating the curves in Fig. 44:

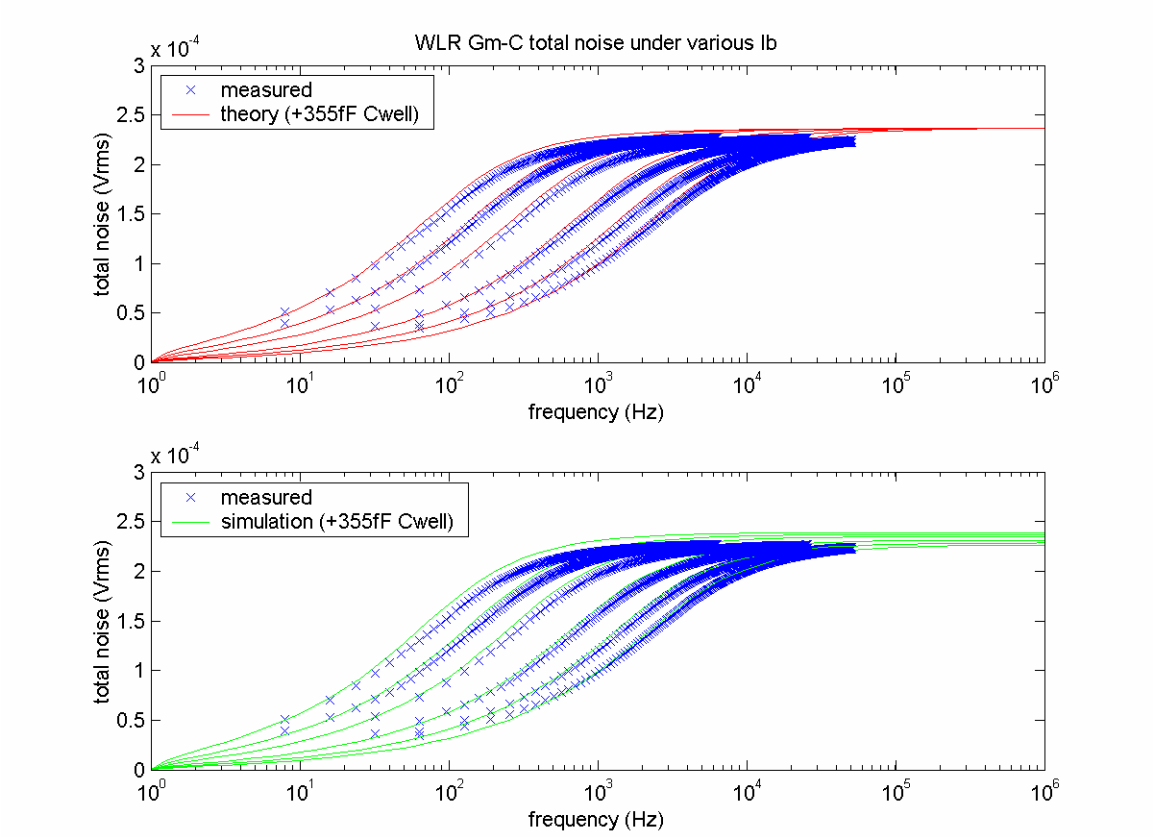


Fig. 45. Total noise of the WLR Gm-C filter, calculated by integrating the PSD at various bias currents. Again, measurements line up well with simulation and theory, and come up to about 240 µVrms.

The total noise is independent of bias level, since it is fundamentally a N·kT/C result, and is borne out by all our measurements.

$$
\begin{aligned}
TotalNoise^2 &= NoisePSD^2 \cdot BW \cdot \frac{\pi}{2} \\
&= \frac{N \cdot 2q \cdot (0.3Ib)}{Gm^2} \cdot \frac{Gm}{2\pi C} \cdot \frac{\pi}{2} \\
&= \frac{Nq\,(0.6)\,V_L}{4C}
\end{aligned}
\tag{12}
$$

Eq. (12) confirms V_L must be correct, as $NoisePSD^2$ goes as V_L^2 while $TotalNoise^2$ goes as V_L. And since both $TotalNoise^2$ and $NoisePSD^2$ agree with measurement and simulation, the value of V_L we are using must be yielding both the correct noise PSD and the correct bandwidth.

Using theoretical levels as a conservative characterization of the WLR:

Ib (nA)	**BW (Hz)**	**PSD (µVrms/√Hz)**	**Noise (µVrms)**
1.5	105	18.3	236
3	210	13	236
5.75	402	9.37	236
15	1050	5.8	236
30	2100	4.1	236
50	3500	3.18	236

8.4 Finding the BW, and hence V_L of the OTA

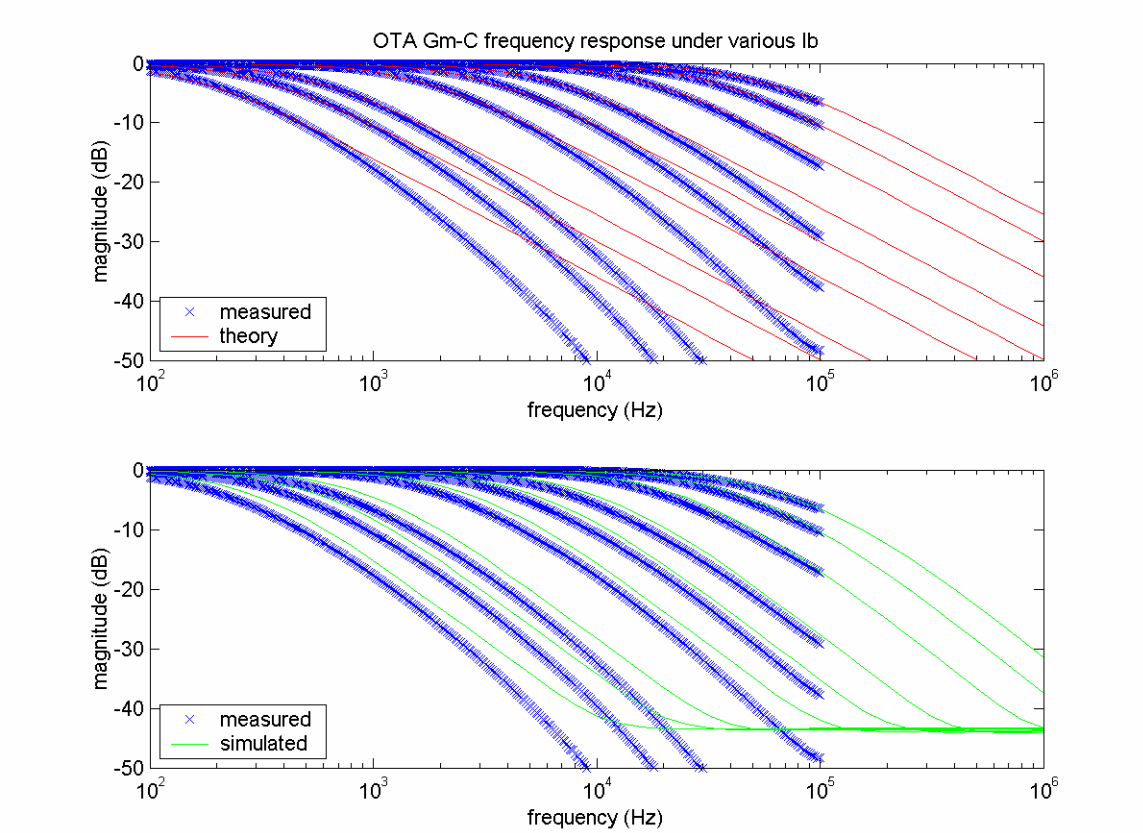


Fig. 46. Bandwidth of the OTA Gm-C filter at various bias currents. C=2pF, as is the case in the actual logmap. Theoretical calculations use a value of 75mV for V_L and model a corner frequency that agrees better than simulation, but does not predict the higher-order rolloff from parasitic poles.

8.5 Finding the noise PSD, and confirming N (# noise sources) in the OTA

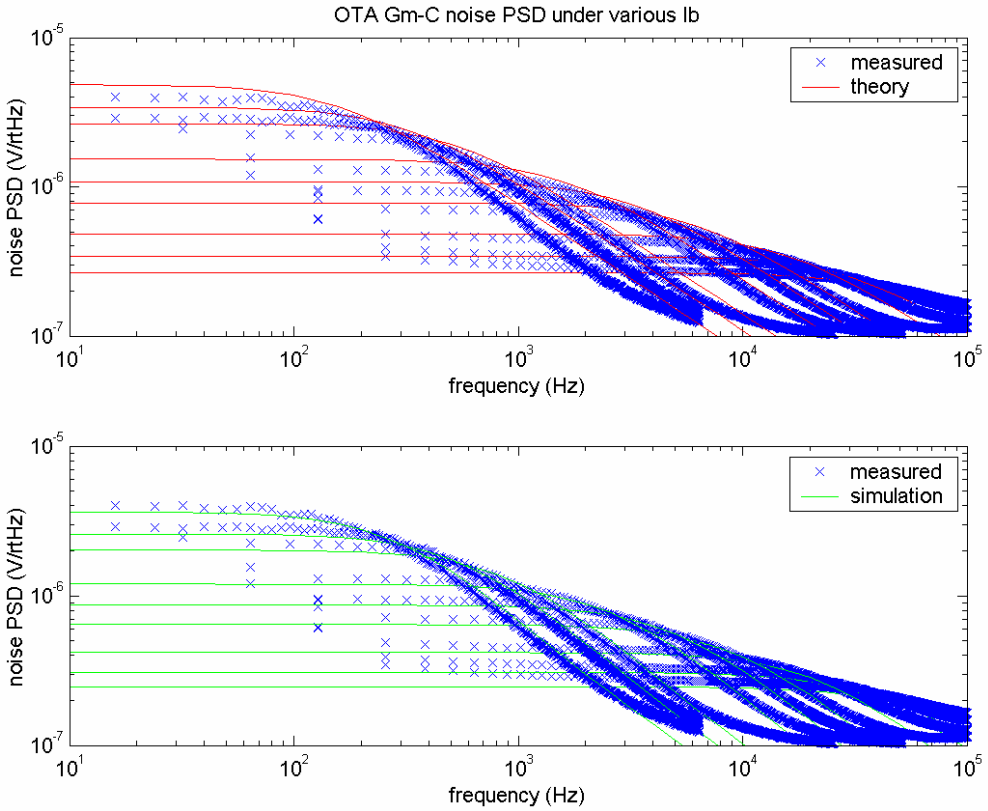


Fig. 47. Noise PSD for the OTA Gm-C filter at various bias currents. Theory shows a PSD which is slightly higher than the measured data, while simulation shows a PSD which is slightly lower. Note there is a tilt upwards in the PSD at higher bias currents due to the suspected presence of 1/f noise. N, the number of theoretical noise sources in this OTA = 4, and is well borne out by these measurements.

Measured noise could be uniformly lower than theory because of higher-order poles that lower the effective bandwidth. The higher-order rolloffs are better modeled in simulation, as shown in Fig. 46. However simulation may then have predicted a noise level which is too low, because it does not model the 1/f noise seen in Fig. 47.

8.6 Integrating the total noise in the OTA

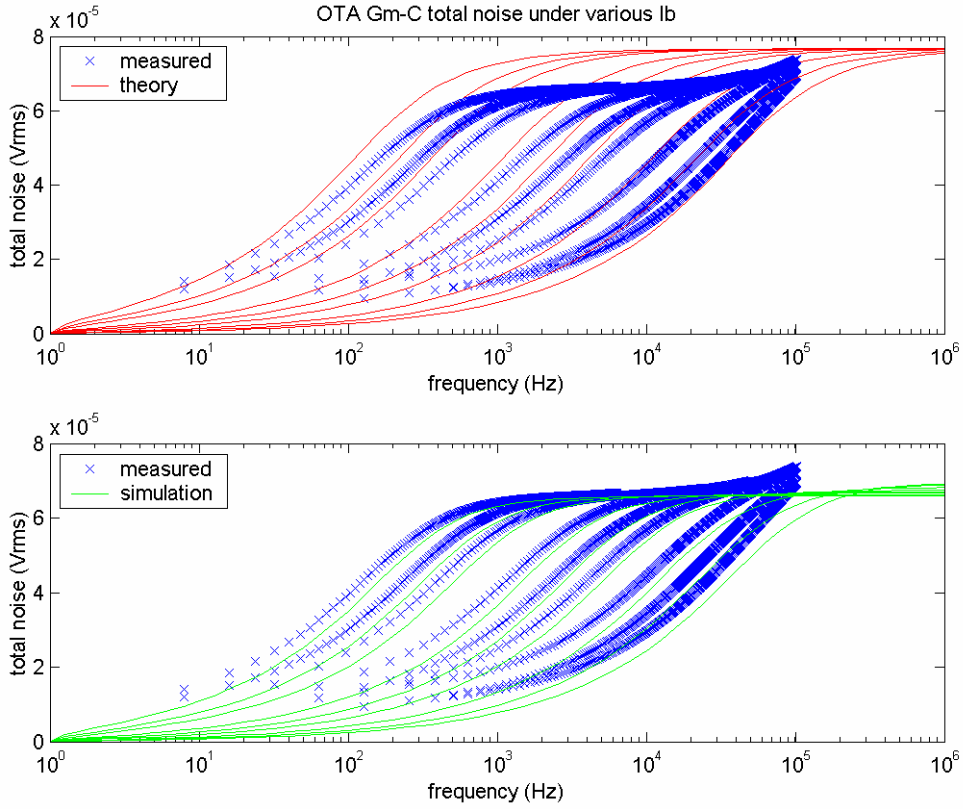


Fig. 48. Total noise of the OTA Gm-C filter, calculated by integrating the PSD at various bias currents. Although higher order poles cause the noise to roll off faster than theory, the total noise still converges to the same fundamental level. Thus measurements again are close to what we predict from simulation and theory, and come up to about 70 µVrms.

Using theoretical levels as a conservative characterization of the OTA:

Ib (nA)	BW (kHz)	PSD (µVrms/√Hz)	Noise (µVrms)
0.15	159	4.84	76.5
0.3	318	3.42	76.5
0.5	531	2.65	76.5
1.5	1.59	1.53	76.5
3	3.18	1.08	76.5
5.75	6.1	0.781	76.5
15	15.9	0.484	76.5
30	31.8	0.342	76.5
50	53.1	0.265	76.5

9 Theoretical noise predictions seen during auto-zeroing (unity feedback operation)

9.1 Motivation for scrutinizing the auto-zeroing phase

If we observe the start and end of the integrating slope, it becomes clear that some random error causes the integrating slope to diverge into a "ray":

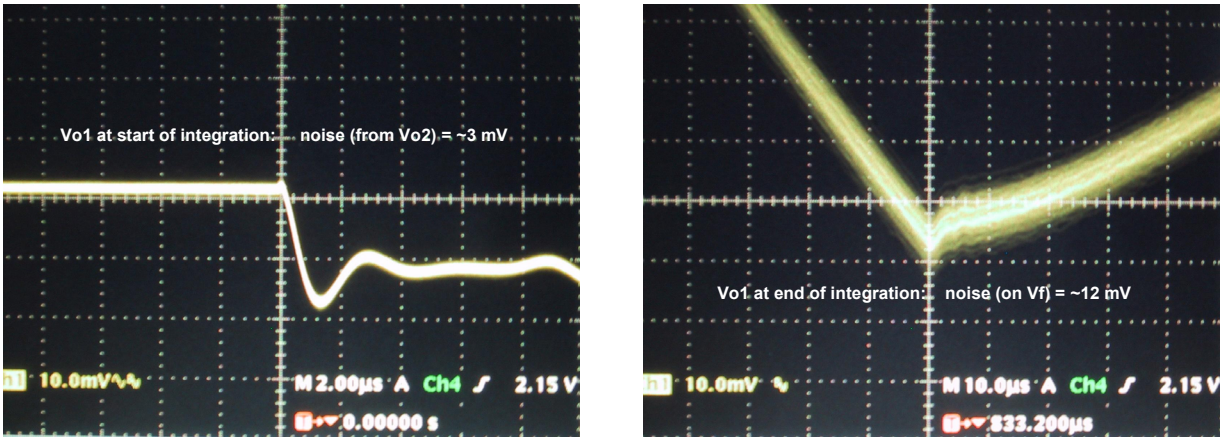


Fig. 49. Captured waveforms showing divergence of the integrating slope. The voltage noise has increased from ~3mV to ~12mV, suggesting that the gain of the WLR (Vfs / Vd-max ≈ 4×) has been applied to an input noise.

Furthermore, the noise jitter in voltage V_f was observed to translate directly to a corresponding jitter in time t_{deint}, as given by Eq. (2). The timing jitter then matched the precision observed on our digital output bits. The voltage jitter in V_f thus accounts almost entirely for our observed precision.

This observation suggests that the dominant source of noise must be low-frequency in nature (at least slower than one conversion period Ts), and that it must exist primarily at the input of the WLR, to alter the nominal slope from cycle to cycle. The auto-zero loop is then most suspect, since kT/C noise sampled onto the storage capacitors will show up directly at the WLR inputs via the input stage source followers.

We can infer from our SNR of 83.5 that there must be an input-referred noise of about 3.3mV (3.3mV / 280mV ≈ 1 part in 83.5), which gives us 6 bits of precision. When I examined Vo2 on a scope, which is the output node that gets sampled (see Fig. 23), I was surprised to actually *see* 3mV of peak-to-peak noise:

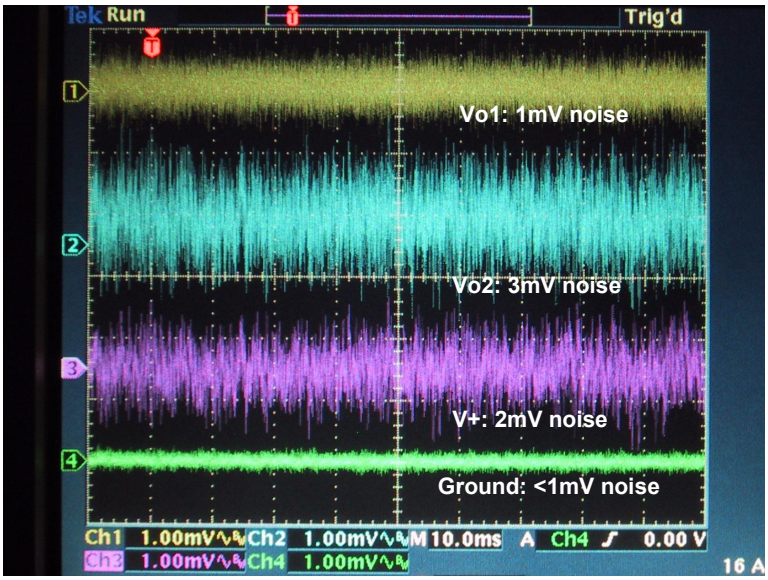


Fig. 50. Noise levels seen during auto-zeroing. Vo2 can be seen to exhibit a maximum peak-to-peak noise of 3mV.

To confirm that this noise is fundamental, I decided to form theoretical predictions of the noise I should see on three important nodes in the loop (which I had access to via analog widepad buffers). These nodes are Vo1, Vo2 and V+ as labelled in Fig. 23. If I could verify theoretical predictions with measured data, then I could be confident the noise observed on the scope was real. And it would assure us that sampled kT/C noise is indeed responsible for our loss of precision.

This section therefore details the spectral density and total noise that I expect to see on each of the nodes Vo1, Vo2 and V+.

9.2 Closing the major loop in unity feedback

In the auto-zeroing phase, the major loop is closed by a source-follower biased at Iref. Hence we can add it into the block diagram of Fig. 19 as follows:

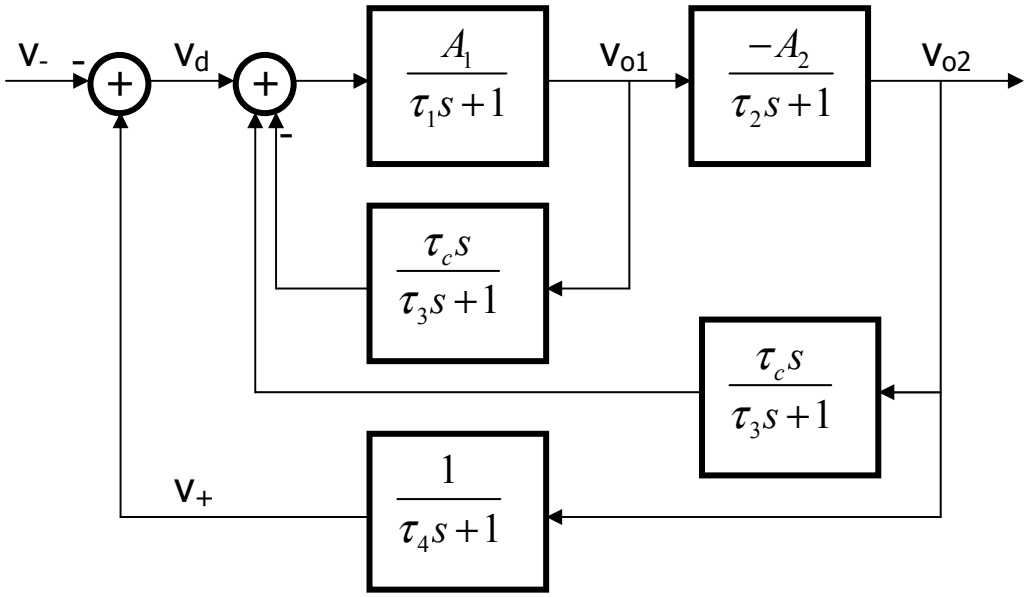


Fig. 51. Autozero loop with major loop feedback, closed using a source follower with time constant τ_4. τ_4 is given by $C+/G_{m4}$, C+ being the total parasitic capacitance seen at input V+. We can predict that it is at least 355fF due to the junction capacitance of the well, and that $G_{m4} = \kappa \cdot Iref / \phi_T$.

Now that we have the complete loop, we can add in the noise sources and perform some quick simplification. From the analysis in section 5.1 and from the forward transmission of |Vo2/Vd| as plot in Fig. 20, we know that the forward transmission is going to be split-poled by the feedback from Vo1. The two poles at τ_1 and τ_3 are hence split to become τ_s (for slow) and τ_f (for fast) respectively:

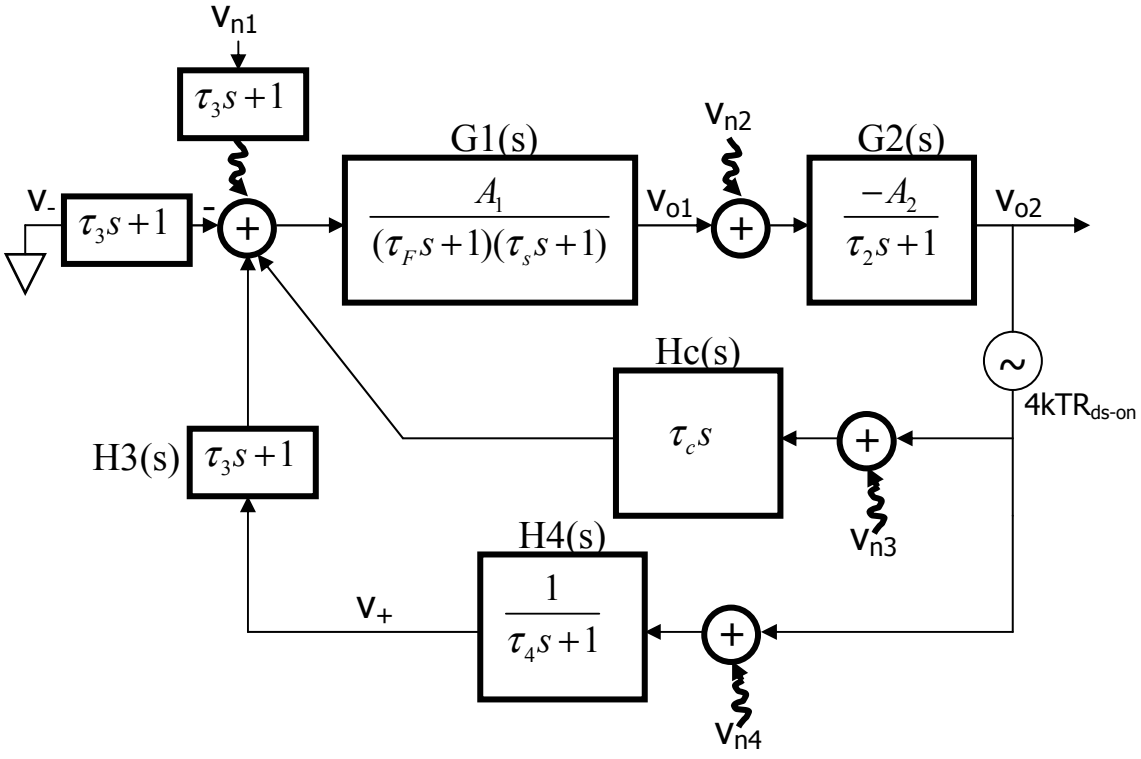


Fig. 52. Autozero loop simplified, adding in all the noise sources. We have moved H3(s) = $\tau_3 s + 1$ through the summing junction to simplify the minor loop of G1, G2 and Hc since H3 creates a pole-zero cancellation in G1 and Hc. The 4kTR term comes from the resistance of the switch connecting Vo2 to the storage capacitor.

It will also help greatly if we combine all the H(s) feedback blocks into a single one for the calculation of the transfer functions to Vo1 and Vo2:

$$\begin{aligned} Hc34(s) &= \text{Hc(s)} + \text{H3(s)H4(s)} \\ &= \tau_c s + \frac{\tau_3 s+1}{\tau_4 s+1} = \frac{\tau_c \tau_4 s^2 + (\tau_c + \tau_3)s+1}{\tau_4 s+1} \\ &= \frac{(\tau_c^{+} s+1)(\tau_c^{*} s+1)}{\tau_4 s+1} \end{aligned} \tag{13}$$

We will note that since $\tau_3 \ll \tau_4$, the zeroes of Hc34(s) will become complex and also shift down to a lower frequency, since the 1st-order term in the quadratic ($\tau_c + \tau_3$) is much smaller than the sum of ($\tau_c + \tau_4$). Hence we will denote the two zeroes as τ_c^{+} and τ_c^{*} to indicate they are both bigger in magnitude than the original τ_c and are also conjugate pairs.

We will also lump the 3 noise sources Vn3, Vn4 and $4kTR_{ds\text{-}on}$ into one, and still call it Vn3:

$$\begin{aligned} Vn3 &= \sqrt{\text{Vn3}^2 + \text{Vn4}^2 + 4\text{kTR}_{\text{ds-on}}} \\ &= \sqrt{\frac{4q(I_{buff}/2)}{G_{m3}^{\ 2}} + \frac{2qI_{ref}}{G_{m4}^{\ 2}} + 4\text{kTR}_{\text{ds-on}}} \\ &= \sqrt{\frac{4q(50n)}{(1.25\mu)^2} + \frac{2q(400p)}{(10n)^2} + 4(26m)(1.6\text{e-}19)(2.46k)} \\ &= \sqrt{2.048e-14 + 1.28e\text{-}12 + 4.1e-17} \\ &= 1.14\ \text{V}/\sqrt{\text{Hz}} \end{aligned} \tag{14}$$

Hereafter when we refer to Vn3, we will actually be referring to the combined sum of all 3 noise sources. Also note that by taking the "vector magnitude" of the 3 sources, we can see immediately that Vn4 emerges as the dominant noise source acting at this node.

If we go ahead and make some reasonable assumptions about A_1, A_2, C_1 and C_4, we can go ahead and calculate all the time constants and NoisePSD's used in our theoretical calculations:

Time Constant	**value (est./meas.)**	**NoisePSD**	**value (est./meas.)**
τ_1	2.5 ms	Vn1	5.8 μV/√Hz
τ_s	15 ms	Vn2	0.25 μV/√Hz
τ_2	160 μs	Vn3	1.27 μV/√Hz
τ_c	120 μs		
τ_4	50 μs	**Gain**	**value (simulated)**
τ_3	1.2 μs	A_1	100
τ_f	0.2 μs	A_2	100

From simulations, C_1 was estimated to have a maximum of 300fF, and C_4 was estimated to be 500fF since we know there is at least 350fF due to the well input.

We are finally ready to tackle the task of generating transfer functions from each of the noise sources {Vn1, Vn2, Vn3} to each of the measurable nodes {Vo1, Vo2, V+}.

9.3 Transfer functions from {Vn1,Vn2,Vn3} to Vo2

According to the block definitions in Fig. 52, we can write out the three transfer functions:

$$\begin{aligned}
\frac{Vo2}{V_{n1}}(s) &= \mathrm{H3(s)}\cdot\frac{\mathrm{G1(s)}\cdot\mathrm{G2(s)}}{1-\mathrm{G1(s)}\cdot\mathrm{G2(s)}\cdot\mathrm{Hc34(s)}} \\
\frac{Vo2}{V_{n2}}(s) &= \frac{\mathrm{G2(s)}}{1-\mathrm{G1(s)}\cdot\mathrm{G2(s)}\cdot\mathrm{Hc34(s)}} \\
\frac{Vo2}{V_{n3}}(s) &= \frac{\mathrm{G1(s)}\cdot\mathrm{G2(s)}\cdot\mathrm{Hc34(s)}}{1-\mathrm{G1(s)}\cdot\mathrm{G2(s)}\cdot\mathrm{Hc34(s)}}
\end{aligned} \qquad (15)$$

If we plot these 3 transfer functions as $|H(f)|$ and $\int |H(f)|^2\,df$, we will later be able to simply multiply them by the in-band NoisePSD and NoisePSD^2 to get the PSD(f) and TotalNoise^2 values directly:

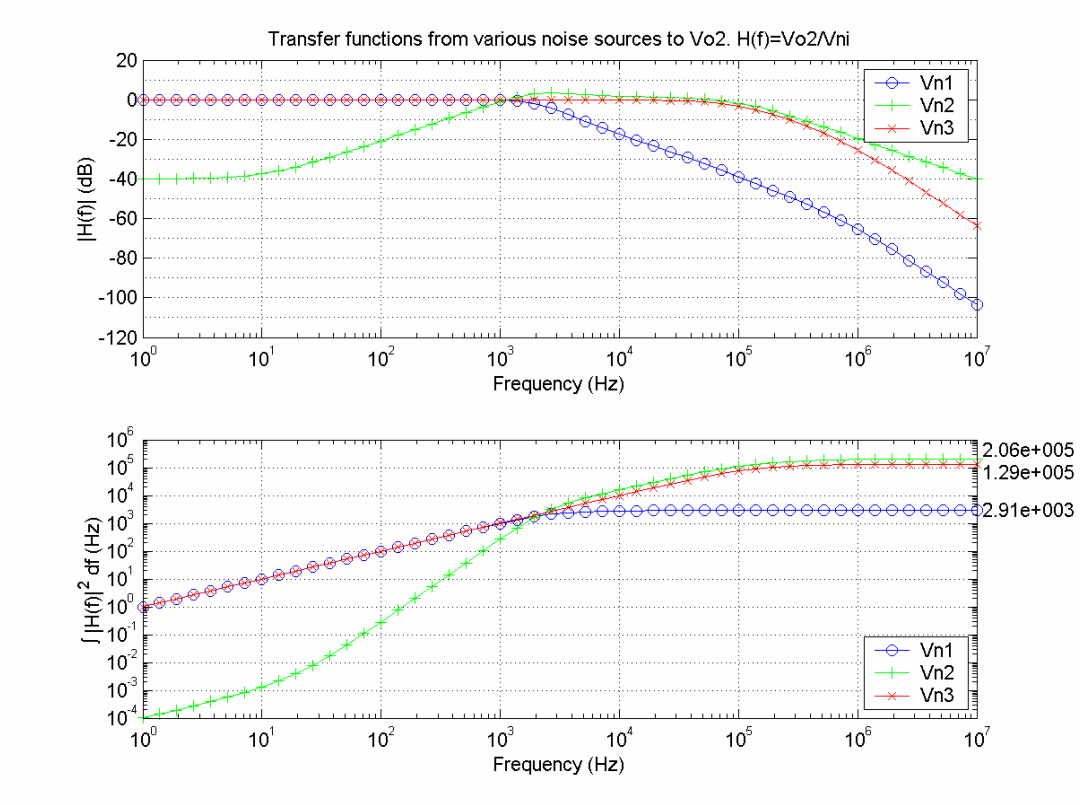


Fig. 53. Transfer functions from {Vn1, Vn2, Vn3} to Vo2.

We can make some notes about each transfer function.

From Vn1, we'll note that it looks almost like a unity feedback system that rolls off at around τ_c. This is because Hc34(s) has poles and zeroes all past τ_2, and hence are not seen before the forward transmission has already crossed over.

We might have expected the transfer function from Vn2 to be small, since it has the gain block G2(s) in the feedback path. However if we observe the plot of $\int |H(f)|^2$ df, we will see that G2(s) in feedback actually did not do any good by the time it reaches high frequencies. This is because the gain in G2(s) rolls off very early, at τ_2. The lesson here is that even though we have gain in the feedback path, we do not necessarily get any benefit of noise attenuation!

Lastly we expect the effect of Vn3 to be large because the transfer function is a purely unity feedback system, so all the noise up to the maximum bandwidth of the loop gets integrated in $\int |H(f)|^2$ df.

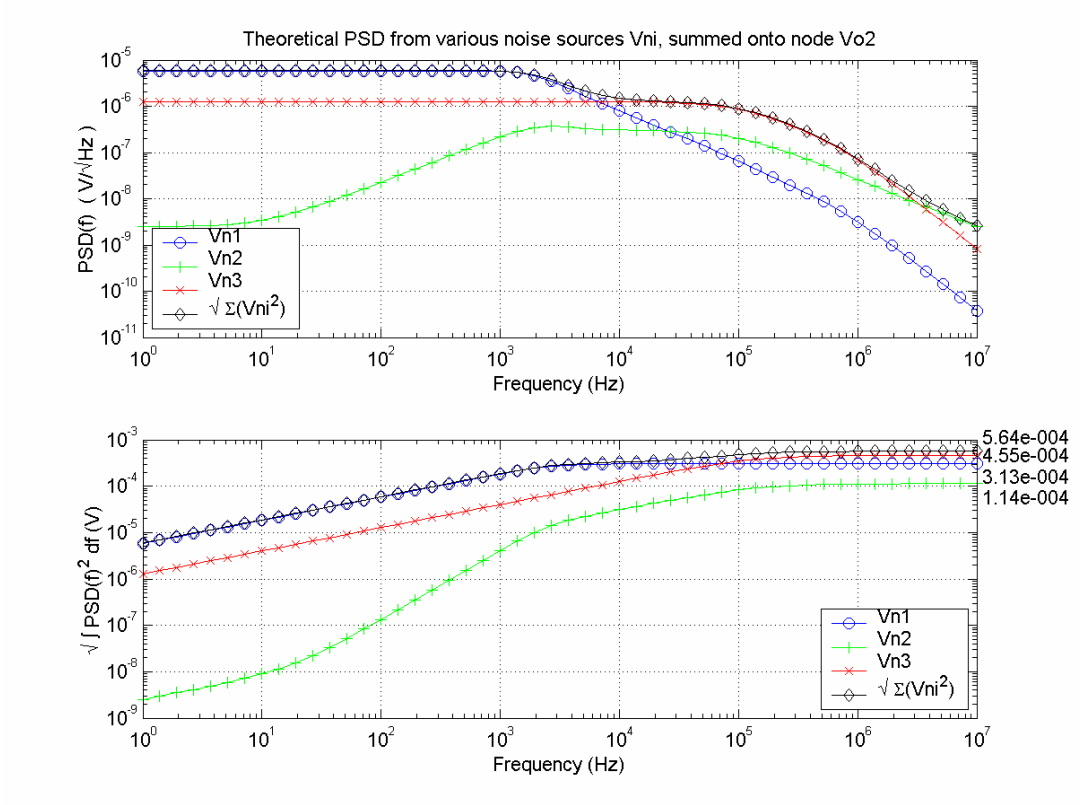


Fig. 54. Transfer functions from {Vn1, Vn2, Vn3} to Vo2, weighted by their NoisePSD.

We can see the effect of the summation of all 3 sources by the points labeled by a black diamond, which traces out the maximum level of all 3 curves in PSD(f). This is consistent with the function sqrt(sum of squares) that takes the "vector magnitude" of all contributors.

We will note that Vn1 is our biggest noise contributor in magnitude, and accounts for the integration up to hundreds of μVs of rms noise. However: it looks like even if Vn1 was lower, Vn3 would push us up to the same level too. Vn3 prevents the noise from rolling off past the τ_c crossover, and adds around an extra 100 μV of noise at high frequencies.

9.4 Transfer functions from {Vn1,Vn2,Vn3} to Vo1

We will perform the same analysis for Vo1.

From Fig. 52, we can write out the three transfer functions:

$$\frac{Vo1}{V_{n1}}(s) = \mathrm{H3(s)} \cdot \frac{\mathrm{G1(s)}}{1 - \mathrm{G1(s)} \cdot \mathrm{G2(s)} \cdot \mathrm{Hc34(s)}}$$

$$\frac{Vo1}{V_{n2}}(s) = \frac{\mathrm{G1(s)} \cdot \mathrm{G2(s)} \cdot \mathrm{Hc34(s)}}{1 - \mathrm{G1(s)} \cdot \mathrm{G2(s)} \cdot \mathrm{Hc34(s)}} \qquad (16)$$

$$\frac{Vo1}{V_{n3}}(s) = \frac{\mathrm{G1(s)} \cdot \mathrm{Hc34(s)}}{1 - \mathrm{G1(s)} \cdot \mathrm{G2(s)} \cdot \mathrm{Hc34(s)}}$$

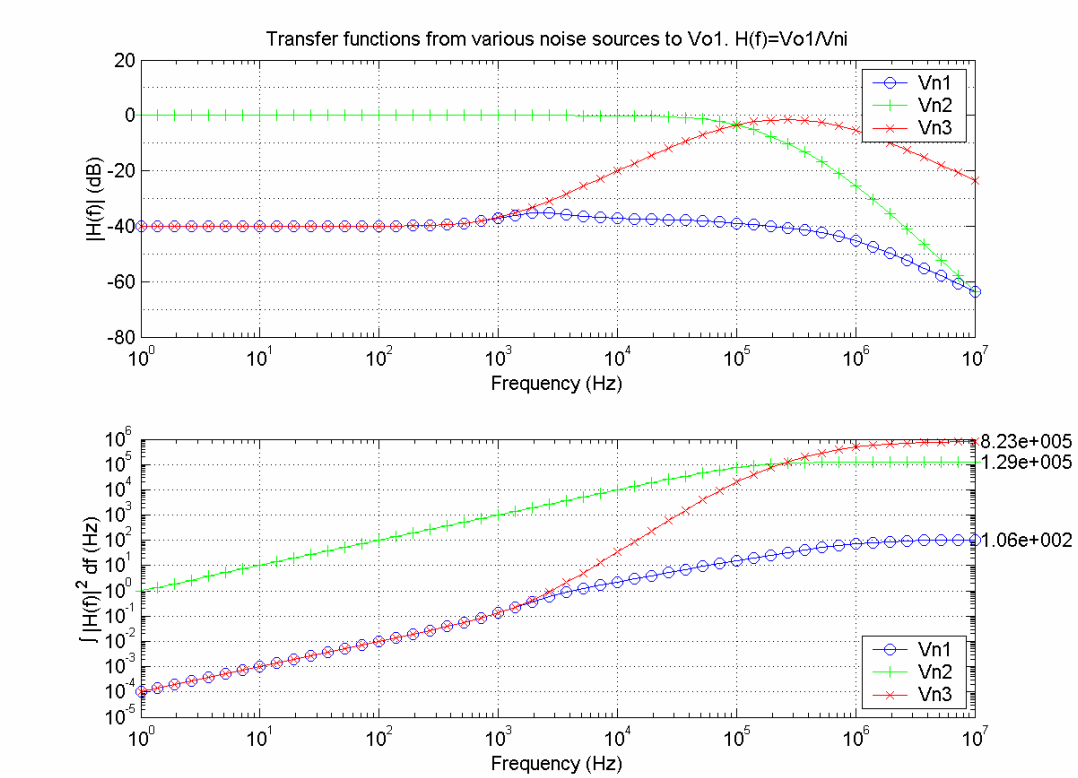


Fig. 55. Transfer functions from {Vn1, Vn2, Vn3} to Vo1.

Some comments on these transfer functions:

It leaps out at us that there is exceptional noise attenuation from Vn1 to Vo1. This is because there is gain in the feedback path of G2(s), but the additional block Hc34(s) with extra zeroes at high frequency that cancel the rolloff in gain of G2(s) help preserve the gain A2 to high frequencies. Thus the attenuation of -40dB lasts all the way past 1MHz.

This is in direct contrast with the transfer function Vo1/Vn3. Even though it starts out with the -40dB attenuation just like Vo1/Vn1, it does not have Hc34(s) in the feedback path to "preserve" the high gain at higher frequencies. Instead, Hc34(s) is in the forward transmission and actually hurts by creating a strong feedthrough of noise past the τ_c zeroes. Vn3 therefore is observed to

inject noise at frequencies even higher than the Vo1/Vn2 unity feedback transfer function, and is responsible for being the dominant noise source contributing to Vo1.

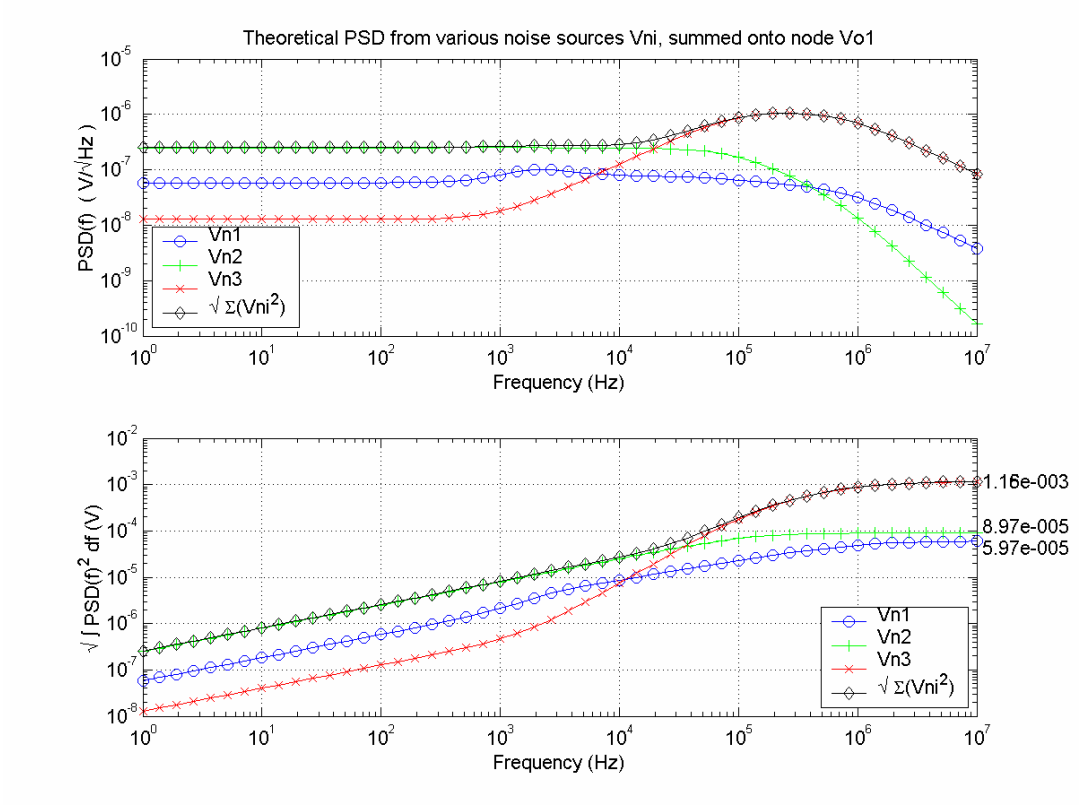


Fig. 56. Transfer functions from {Vn1, Vn2, Vn3} to Vo1, weighted by their NoisePSD.

As we predicted above, Vn3 bumps up our noise so much at high frequencies that we get 1mVrms total integrated noise when we integrate up to 10MHz. However this is not observed in lab, and will be accounted for in the next section.

For now, we will note that apart from the contribution of Vn3 at high frequencies (past 10kHz), the baseline in noise is actually quite low — about 300nV/√Hz up to 10kHz. This node is quiet because 2 of our 3 transfer functions have the gain block G2(s) in their feedback path.

9.5 Transfer functions from {Vn1,Vn2,Vn3} to V+

The analysis for V+ is more tricky, since we cannot use Hc34(s) anymore, and must split up the H(s) blocks individually. We have to start by first defining transfer functions for the minor loops:

$$
\begin{aligned}
P(s) &= \frac{\mathrm{G1(s)\cdot G2(s)}}{1-\mathrm{G1(s)\cdot G2(s)\cdot Hc(s)}} \\
Q(s) &= \frac{\mathrm{G2(s)}}{1-\mathrm{G1(s)\cdot G2(s)\cdot Hc(s)}} \\
R(s) &= \frac{1}{1-\mathrm{G1(s)\cdot G2(s)\cdot Hc(s)}}
\end{aligned}
\tag{17}
$$

And then close a major loop by incorporating H3(s) and H4(s) back in:

$$
\begin{aligned}
\frac{V+}{V_{n1}}(s) &= \mathrm{H3(s)}\cdot\frac{\mathrm{P(s)\cdot H4(s)}}{1-\mathrm{P(s)\cdot H4(s)\cdot H3(s)}} \\
\frac{V+}{V_{n2}}(s) &= \frac{\mathrm{Q(s)\cdot H4(s)}}{1-\mathrm{Q(s)\cdot H4(s)\cdot H3(s)\cdot G1(s)}} \\
\frac{V+}{V_{n3}}(s) &= \frac{\mathrm{R(s)\cdot H4(s)}}{1-\mathrm{R(s)\cdot H4(s)\cdot H3(s)\cdot G2(s)\cdot G1(s)}}
\end{aligned}
\tag{18}
$$

We can then plot these transfer functions as before:

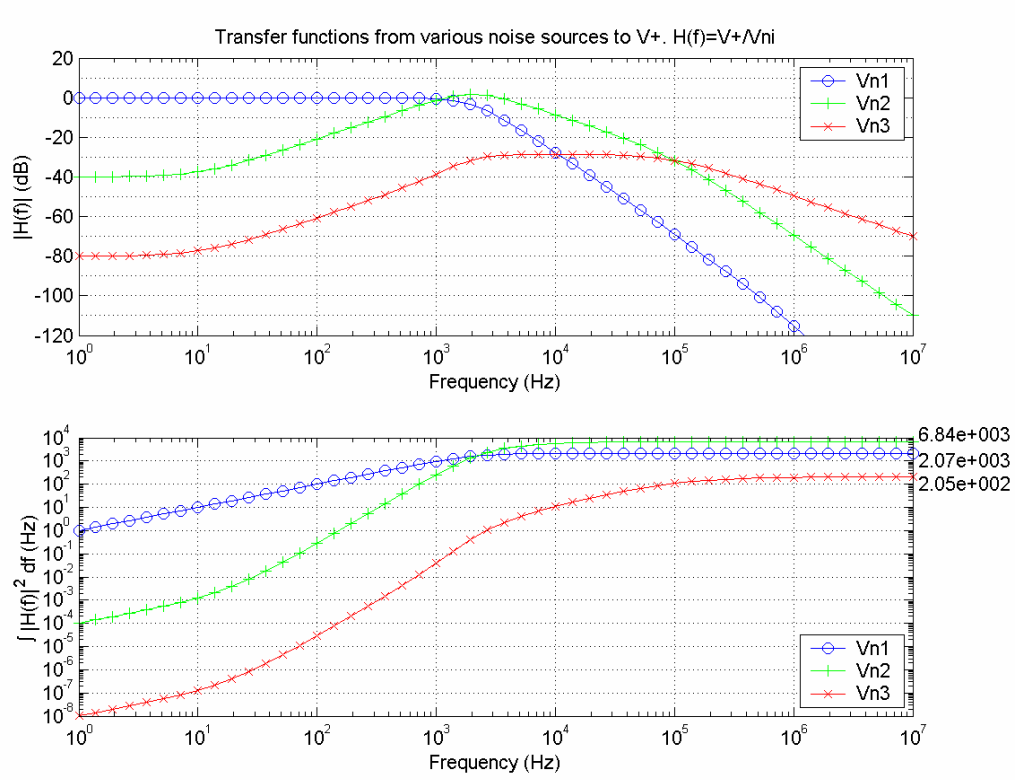


Fig. 57. Transfer functions from {Vn1, Vn2, Vn3} to V+

Interpreting the transfer functions in Fig. 57 is simpler if we break them down into forward transmission and feedback components. They have been plot here in the following chart:

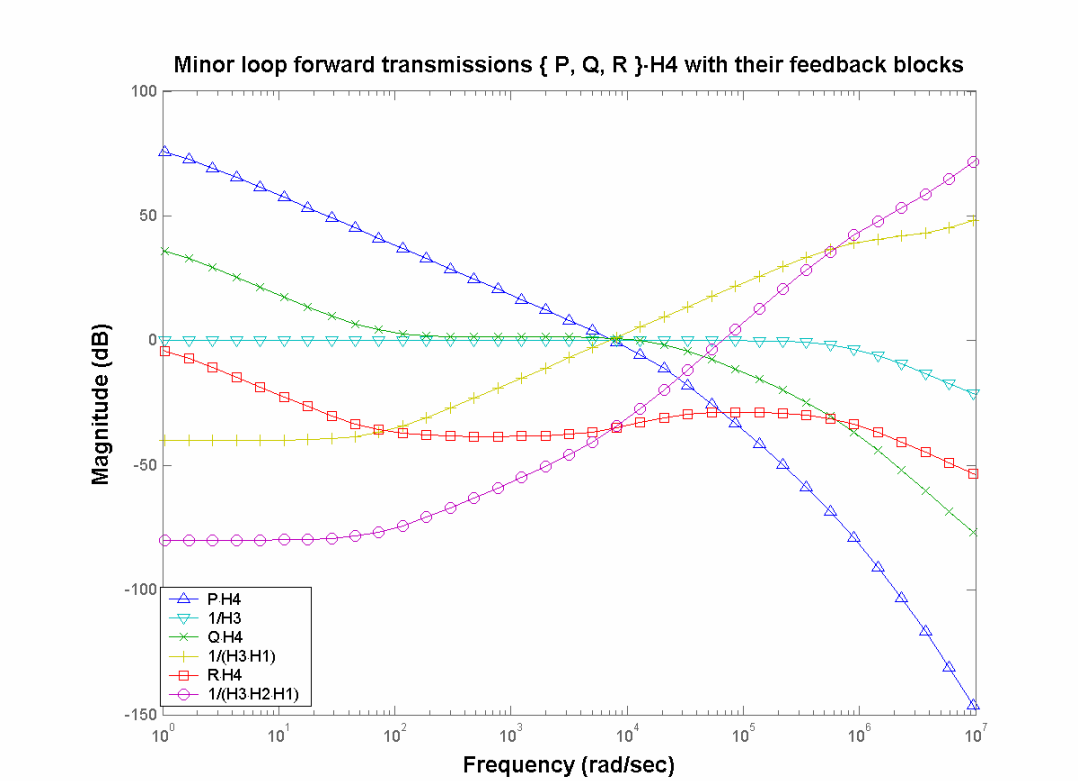


Fig. 58. Breakdown of |H(f)| in Fig. 57 into forward and feedback transmission components. |H(f)| is formed by taking the lower of the forward block and the 1/feedback block, observed from the expression G / (1+GH) = G(1/H) / (G + 1/H), which is a parallel combination of G and 1/H.

We can observe that the forward transmissions {P,Q,R}·H4 all have Hc in feedback and therefore dominant pole rolloffs, leading to large gains of (A1·A2) at DC. When put in feedback, the large gain in the forward transmission allows the feedback blocks to completely determine the transfer function at low frequencies as 1/feedback(s).

However, the gain in the forward transmissions start to roll off at higher frequencies, and the 1/feedback blocks also roll "up" around the same point, hence at high frequencies it is the forward transmissions that determine the transfer function.

In summary, it can be observed that the feedback transmissions completely determine the transfer function at low frequencies, but at high frequencies the forward transmissions take over.

By breaking the transfer function into low frequency / high frequency regimes, it is now straightforward to see what happens:

Vn1 does not see any attenuation at low frequency since there is no gain but rather just a high frequency zero (τ_3) in the feedback path that has no effect. V+/Vn1 therefore rolls off at high

frequencies along with the forward transmission P·H4, with a slope of -3 (-2 from P, -1 from H4).

Vn2 sees the attenuation of G1 at low frequencies, and thus starts off at -40dB. However G1's gain quickly rolls off, causing the attenuation to vanish, but then the forward transmission Q·H4 rolls off as well, around τ_c, and thereafter drops with a slope of -2 (-1 from Q, -1 from H4).

Vn3 sees the attenuation of both G1 and G2, and thus starts off at -80dB. Just like V+/Vn2, the gain of G1 and G2 quickly roll off at higher frequencies, and intersect the forward transmission R·H4. R·H4 rolls off with only a slope of -1 that comes from H4.

We are now ready to look at the transfer functions weighted by their thermal PSDs:

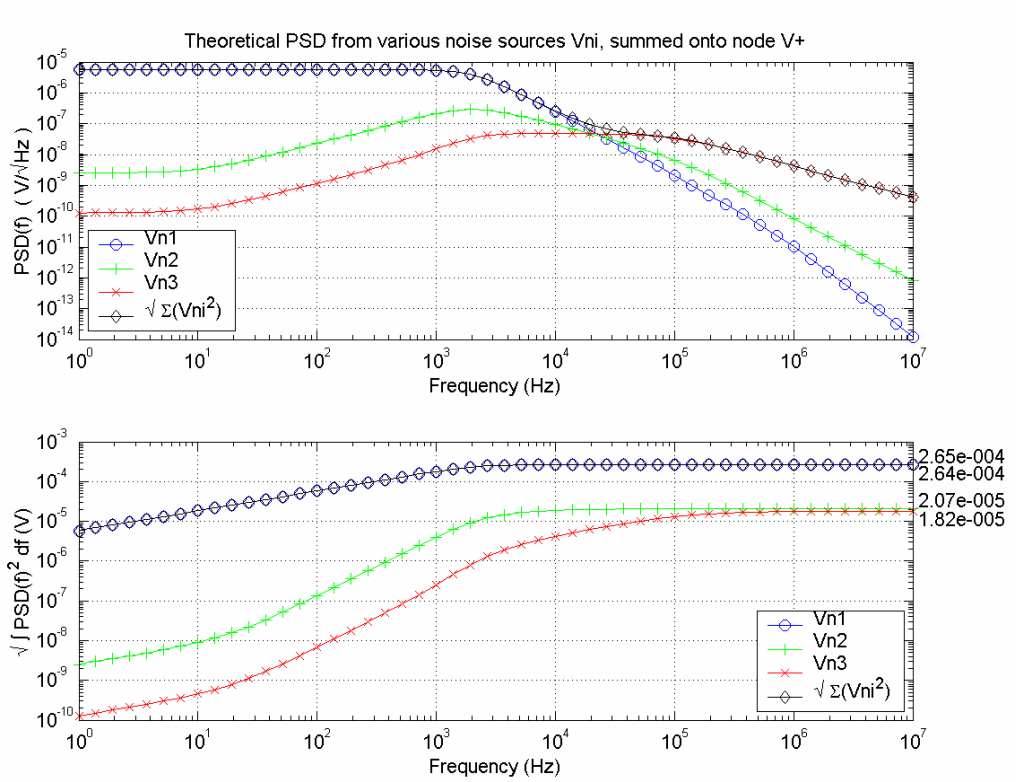


Fig. 59. Transfer functions from {Vn1, Vn2, Vn3} to V+, weighted by their NoisePSD.

It is clear that the transfer from Vn1, which is unattenuated by any gain in the feedback path, ends up dominating the system to produce 265 µVrms of noise (almost 1mVpp).

Vn2 and Vn3 noise sources end up hardly mattering, and fall completely below the integrated noise of Vn1.

10 Measurement of noise seen at Vo2, Vo1 and V+

We will see in this section how well our predictions line up with measured results. These measurements were made on an SRS785 spectrum analyzer in a Faraday cage.

10.1 Measured noise spectrum of node Vo2

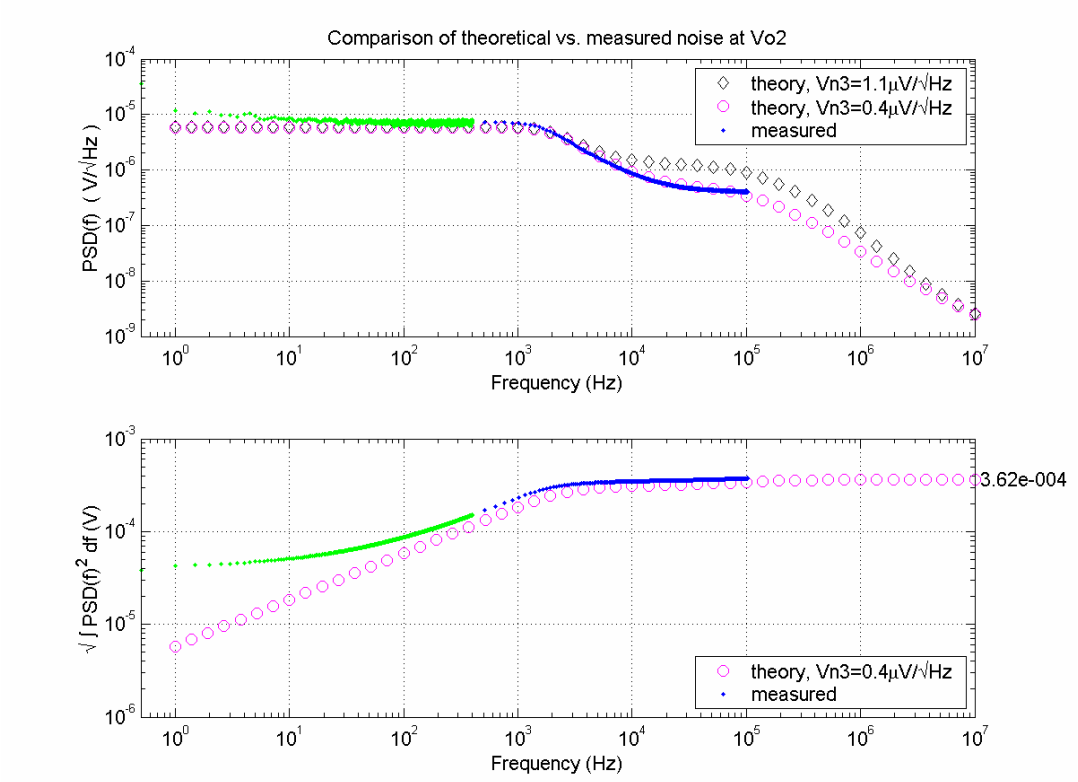


Fig. 60. Measured noise data at Vo2, plot against predictions from theory. One adjustment to the theory has been made to better fit the data: Vn3's PSD was reduced to 0.4 μV/√Hz.

Indeed there is a good fit of the measured data to our theoretical predictions. Going back to Fig. 54, we can see that we would have integrated 560 μVrms of noise instead of 360 μVrms as is measured, if we did not make the correction to Vn3's PSD.

Note that at low frequencies, the amount of integrated noise tapers to some floor set by 1/f noise, which we did not include in the analysis in section 9. Our system is thermal noise limited, however, as the 1/f noise floor is only around 40 μV while we integrate up to at least 360 μV due to thermal noise.

From our understanding of the transfer characteristic that produces Vo1's noise spectrum (analyzed in section 9.4), we know that Vn3 dominates Vo1's total noise when we integrate to

high frequencies. Hence by measuring the noise spectrum of Vo1, we can confirm whether the value we used for Vn3 was indeed an overestimate.

10.2 Measured noise spectrum of node Vo1

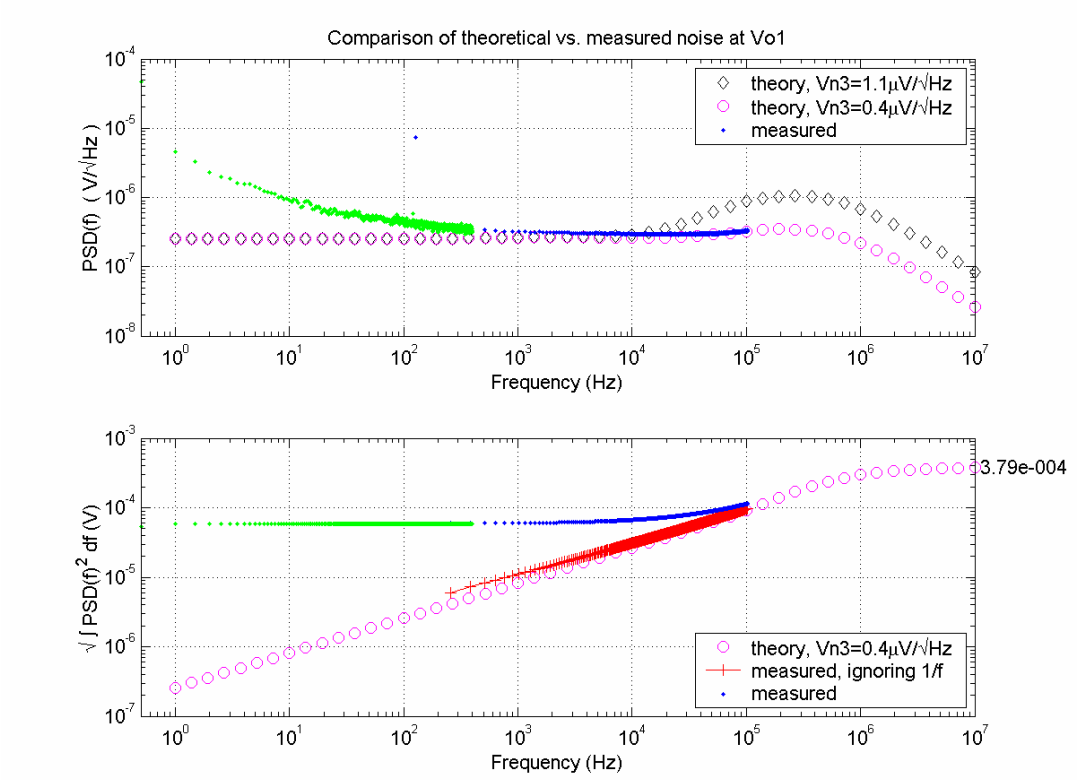


Fig. 61. Measured noise data at Vo1, plot against predictions from theory. The same adjustment we made in Fig. 60 (the reduction of Vn3 to 0.4 μV/√Hz) indeed gives us a better fit to the data.

If Vn3 was indeed as high as 1.1 μV/√Hz, we should see the noise spectrum lift up at about 10 kHz. The fact that it does not supports our guess that Vn3 is actually lower. Note the 1/f noise floor of about 50 μVrms that only gets exceeded by thermal noise when we start integrating past 100 kHz.

As we noted in section 9.4, the presence of the two complex zeroes in Hc34(s) in the forward transmission, nullifying the effect of the higher order poles, causes |Vo1/Vn3| to extend unattenuated to very high frequencies. This is the reason for Vn3 having such an adverse effect up to and even past 100 kHz.

Unfortunately our spectrum analyzer is unable to show us anything beyond that. We can only conclude that we should integrate up to 380 μVrms by 10 MHz, if our theory is correct.

I also included a plot of integrated noise if we ignore the first few data points which contain all the 1/f noise power. This allows us to confirm that the increase in white noise is indeed following

along the line predicted by theory, and that it eventually peeks over the 1/f floor at about 100 kHz.

10.3 Measured noise spectrum of node V+

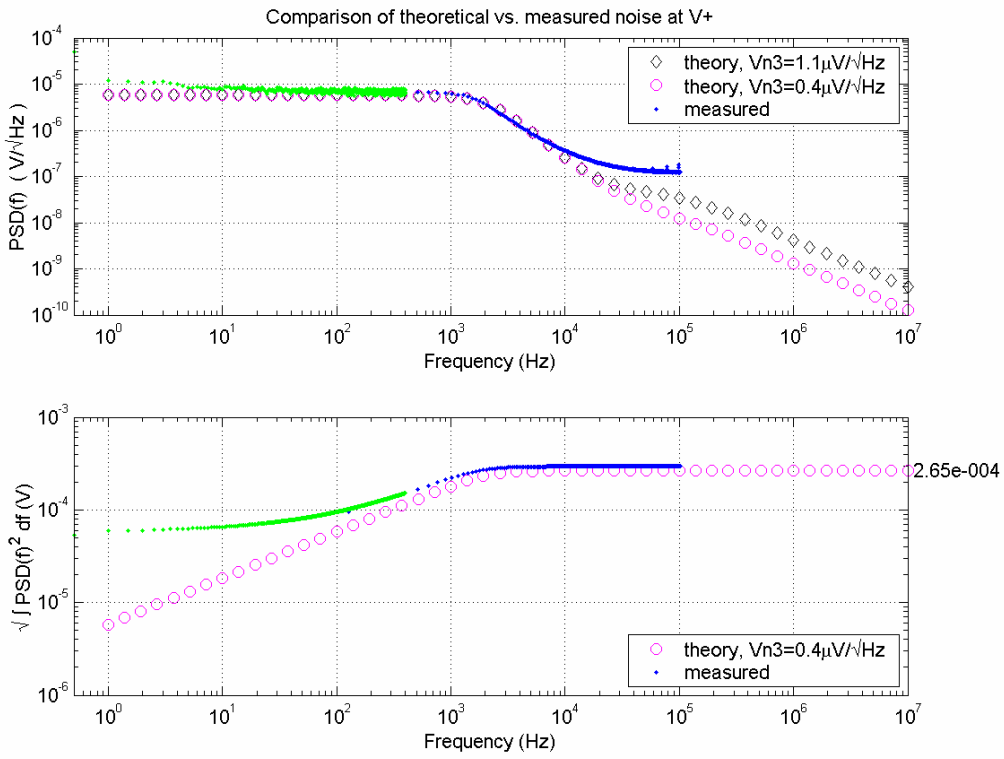


Fig. 62. Measured noise data at V+, plot against predictions from theory. The reduction of Vn3 to 0.4 μV/√Hz is included, but has very little bearing on V+.

As we noted in 9.5, the noise spectrum of V+ is completely dominated by Vn1. Even when the Vn3 curve rises above Vn1 near 10 kHz, the total noise is already completely determined by Vn1, as the effect of Vn3 is 2 orders of magnitude smaller than Vn1.

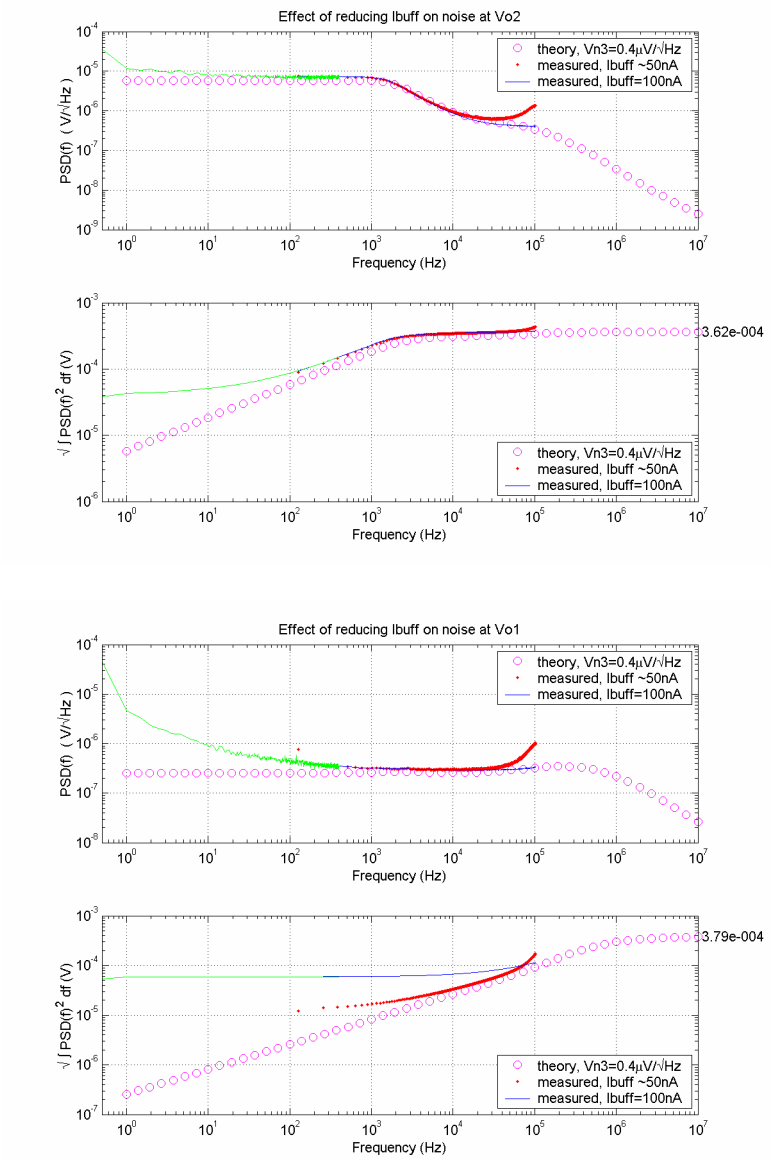
We observe that the spectrum flattens out at 100nV/√Hz around 50 kHz. This could happen if our net gain ($A_1 \cdot A_2$) was not 80 dB but just 70 or 60 dB, since it would result in Vn3 being attenuated by 10 or 20 dB less. However this discrepancy in gain, if it exists, is likely due to A_1 and not A_2, because the reduction in Vn3 is not seen in Vo2 (which is unattenuated) or in Vo1 (which is attenuated by A_2).

Alternatively, the 100 nV/√Hz that we see could possibly be spurious, as it shows a few significant noise spikes that definitely do not come from our system! All our measurements were taken in a Faraday cage, so it is likely that the noise spikes are instrument noise being picked up by the circuit.

What is important is that the total noise seen at V+ is still completely dominated by Vn1, which is unattenuated by any of the gain blocks G1 or G2, and hence integrates up to about 265 µVrms, exactly as predicted by theory.

11 Effect of reducing Ibuff on the noise spectrum at Vo2, Vo1 and V+

To confirm the effects of Vn3, I reduced Ibuff from 100nA to ~50nA and took three new sets of noise spectra.



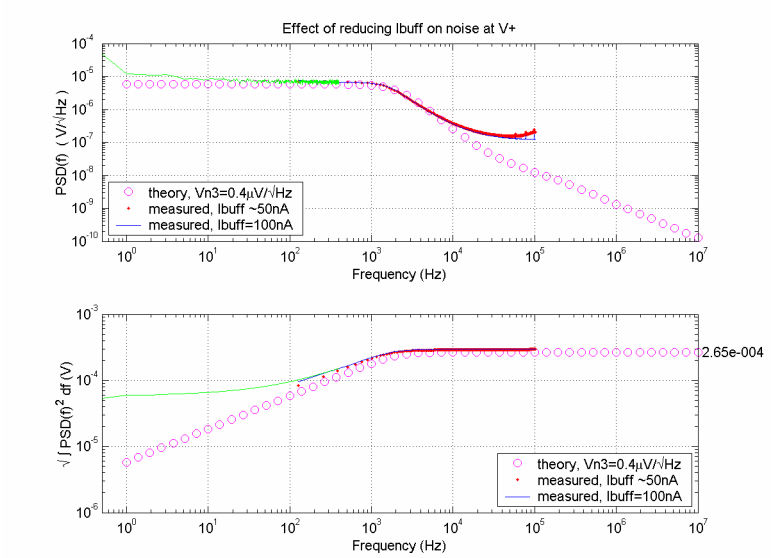


Fig. 63. Effect of reducing Ibuff on the noise spectra of all 3 nodes. Note that the curve lifts up in all 3 cases at the edge of 100 kHz.

From these plots, we know we are indeed seeing the spectrum of Vn3 at the edge of our measurable frequency range (100 kHz). Vn3 is the least filtered spectrum at those high frequencies, and the lifting of the spectrum suggests that Vn3 indeed extends past 100 kHz quite a bit before rolling off. In addition, this strongly suggests that there is more high frequency noise in the system that cannot be seen by the spectrum analyzer.

If this is so, then it is possible that some of the noise I see with the scope may be high frequency in nature, as the scope's bandwidth is much larger than 100 kHz. Also, the fact that I can discriminate around 0.2 mVpp on the ground (as evidenced by Fig. 50) suggests that the input referred noise of the scope is low enough for its measurements to be valid.

12 kT/C error analysis

The fundamental kT/C noise predicted and measured in the previous two sections gives rise to three independent errors, which I will describe below.

12.1 Gain error: a level shift at the WLR input

By definition, an amplifier is auto-zeroed by level-shifting one or both its inputs to take out the offset. Unfortunately, this level-shift will always have kT/C noise and charge injection riding on top of it.

Even in the absence of high frequency noise above 100kHz, and ignoring charge injection, we have concluded that there must be *at least* 400 μVrms of kT/C noise on our critical node, Vo2.

Furthermore, this value reported by the spectrum analyzer is just 1σ, while we need to take 6σ (±3σ) to account for a 99.8% spread of a (roughly) Gaussian distribution. As the scope will show us a 6σ spread in peak-to-peak variation, which then corresponds to 2.4 mVpp, the fact that we observe 3 mVpp is now very reasonable. If we capture the timing waveforms of the noise on Vo1, Vo2 and V+ (as shown in Fig. 50) and then plot their probability distributions, we indeed confirm that there is a 3mV spread in Vo2 when the standard deviation is only 415 μVrms:

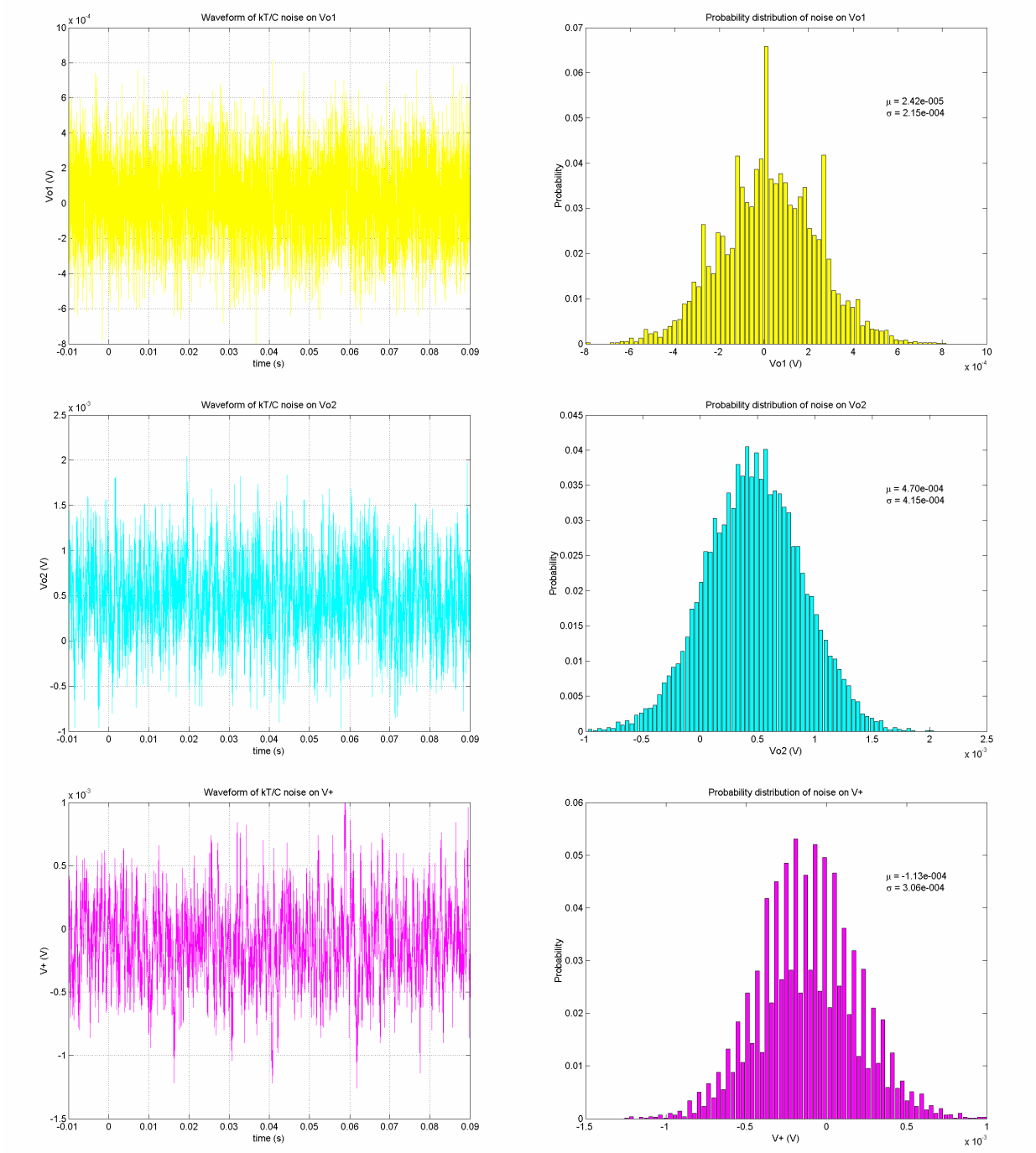


Fig. 64. Scope waveforms of the noise on Vo1, Vo2 and V+, and accompanying histograms showing their probability distribution. MATLAB was used to compute their mean (μ) and standard deviation (σ).

As the noise on Vo2 appears directly at the input, it gets multiplied by the effective WLR gain and gives rise to a gain error.

To understand the effective WLR gain, we can first observe from Fig. 49 that a 3mV noise in the initial condition is gained up to ~12mV by the end of integration, which is a factor of ~4×. This is predicted quantitatively from Eqs. (2) and (6):

$$\begin{aligned} \text{Since } V_f &= \frac{Gm1 \cdot Vd}{Cint} \times t_{int} \\ \text{WLR gain } \frac{V_f}{Vd} &= \frac{Gm1}{Cint} \times t_{int} \\ \text{which must be set to} &= \frac{Vfs}{Vd\text{-}max} \simeq \frac{1.2V}{280mV} = 4.3 \end{aligned} \tag{19}$$

Any error in the input Vd therefore gets gained up by a factor of 4.3 to become a larger error in the output V_f. It is helpful to think of the 3 mV noise at the input as an error of 3 parts out of 280, since the full scale voltage at the input is 280 mV. Equivalently, this error is also 4.3 × 3 mV = 13 mV at the WLR output, or an error of 13 parts out of 1200, since the full scale voltage at the output is 1.2V. Therefore the gain error alone can be seen to limit our precision to 6 bits.

12.2 Offset (initial condition) error: a level shift in the comparator threshold

Referring back to Fig. 23, the comparator threshold is set by the voltage sampled onto one cap (Vc-), while the mid-rail "ground" for the integrator is set by the sampled voltage on another (Vc+). Thus any difference between the two sampled noises translates into an effective shift in the comparator threshold.

As Vo2 gets sampled to become Vc+, the same 3 mV of kT/C noise that produced a gain error at the WLR input will also produce an offset error at the WLR output.

Fortunately, this error is less significant than the gain error, because it gets divided by the gain of the WLR when referred to the input. Put differently, since the offset error appears at the WLR output which has a full scale voltage of 1.2V, the 3 mV noise constitutes an error of only 3 parts in 1200.

12.3 Differential error: a sum of two independent variances

Both gain and offset errors described in 12.1 and 12.2 are actually derived from a differential voltage between sampled nodes Vc+ and Vc-. Both these nodes have independent levels of kT/C noise, that then create a variance which is the sum of their squares. Fortunately the noise source

on Vc- is much lower than Vc+, because Vc- derives from a voltage reference and picks up only 70 μVrms by going through a buffer. Therefore this effect is almost negligible, since if we look at the net effect on the input due to Vc- and Vc+, we will note that 70 μV^2 + 360 μV^2 is only 366 μV^2.

12.4 Summary and Conclusion

Type of Error	Noise / Full Scale voltage	Fractional Error	Percentage Error
Gain Error	13/1200	1/92	1.08%
Offset Error	3/1200	1/400	0.25%
Differential Error	negligible	-	-
Total Error	16/1200	1/75	1.33%

Fig. 65. Summary of the approximate errors produced by sampled kT/C noise. Our experimental measurement of a 1/83.5 timing jitter shows that our fractional error is not as small as a pure gain error of 1/92, but not as large as the total predicted error of 1/75.

We have seen that sampled kT/C noise leads to 3 sources of error, but the gain error presented in 12.1 is by far the most significant. Therefore future work will be targeted at reducing the kT/C noise, and focused in particular on the gain error.

13 Future Directions

13.1 One easy improvement: burn more power

Right now the WLR is set to consume only 15nA. I can conceivably increase the size of capacitor Cint by 10× from 1.5pF to 15pF and burn 150nA in the WLR instead, which should reduce the noise seen due to Vn1 by a factor of √10.

We should therefore be concerned about what happens when the WLR leaves the sub-threshold region:

Linearity: The onset of moderate inversion has the effect of lowering the Gm per unit Ib, or raising the effective V_L. As V_L is a measure of linearity, we can expect the tanh-shaped transconductance to flatten and in fact improve the converter's linearity.

Saturation: The value of Vds-sat rises from ~ $4 \cdot \phi_T$ to κ(Vgs-Vth), so if Vds-sat gets too large, we may have to size the transistors carefully to avoid saturating the WLR mirrors.

Noise: We know that thermal noise goes as:

$$\text{NoisePSD}^2 \cdot BW \propto \begin{cases} \dfrac{\text{Ib}}{\text{Gm}^2} \cdot \dfrac{Gm}{C} & \text{(weak inversion)} \\ \dfrac{2/3\,\text{Gm}}{\text{Gm}^2} \cdot \dfrac{Gm}{C} & \text{(strong inversion)} \end{cases} \tag{20}$$

hence the fundamental kT/C noise is independent of Ib and scales only with C.

However, if we raise C enough, we may have to drive the WLR into strong inversion to keep the WLR gain $(Gm1/Cint)\cdot t_{int}$ constant at Vfs / Vd-max, as given by Eq. (19). In other words, we have to increase the current to recover the original Gm/C bandwidth. But since Gm goes with √Ib in strong inversion, it means Ib is less and less effective at recovering the bandwidth lost to an increase in C. Stated differently, if we increase C by a factor of 10×, we may need to increase Ib by *more* than 10×. This is another reason to avoid the less power-efficient regime of strong inversion as far as possible.

Fortunately, the WLR is likely to remain in sub-threshold operation even with 150nA because the bump linearization transistors steal a third of Ib for most of the time, so each device will only carry 50nA when balanced. Furthermore, at a full scale input differential, Iout as a fraction of Ib is only 280mV / 1.2V = at most 25%. Hence the currents in each leg differ only by a ratio of 5:4, which in absolute currents is 83nA : 67nA.

In conclusion, this improvement is promising because the total integrated noise is dominated by Vo2's PSD · bandwidth product. Increasing Cint while burning more current to preserve the bandwidth will reduce our PSD and hence our noise.

13.2 A proposed solution: avoid global feedback with multi-stage compensation

The main problem with this topology is the large amount of kT/C noise seen directly by the input, where we are most sensitive to noise. And if we want to auto-zero the WLR, we must inevitably sample a large kT/C error, since the WLR has many devices and is designed to have a large V_L.

Although we cannot avoid sampling the kT/C noise of the WLR, we can avoid all other kT/C contributions if we store the WLR and comparator offsets sequentially, rather than together in global feedback. This can be implemented similar to a multi-stage comparator as described in [16] and is shown below:

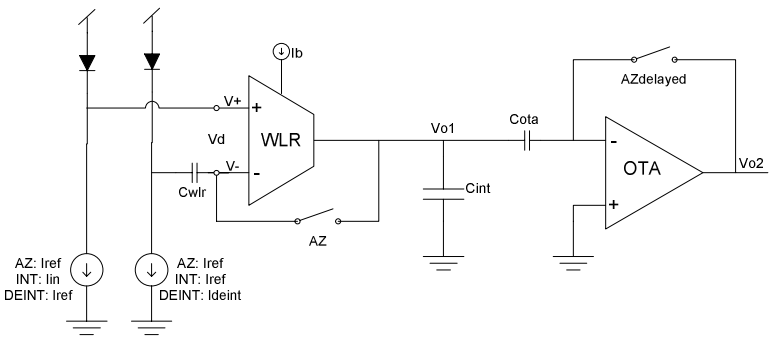


Fig. 66. A multi-stage offset compensation strategy. At the end of auto-zeroing, switch AZ is opened before AZdelayed.

If we keep the same Gm/C ratio, the total noise we expect is now 240 μVrms (from measurements of a Gm-C topology in Fig. 45) rather than 360 μVrms that we see currently. This is because the number of noise sources dumping kT/C noise at the input is reduced. Noise from any device beyond the WLR is thereby prevented from affecting the input.

13.3 Secondary benefits of multi-stage compensation

The multi-stage topology provides a host of secondary benefits, if implemented carefully.

1) **More power is freed up**: by removing the need for two OTA buffers, the decrease in power consumption of 200nA can go directly to decrease the WLR's noise PSD instead, as proposed in section 13.1.

2) **Elimination of WLR's contribution to the offset error**: just as the OTA's kT/C noise no longer affects the gain error (section 13.2), the WLR's kT/C noise has also been decoupled from the offset error. All charge injection and kT/C noise stored on capacitor Cwlr will be ignored by the OTA because the OTA is still being reset during that time. Thus only ~70 μVrms of noise stored on capacitor Cota will level-shift the comparator threshold, which is much improved over 400 μVrms because the V_L of the OTA is 20× smaller than the WLR!

   Furthermore, this offset error is divided by the effective gain of the WLR when input-referred; as before it is out of a large 1.2V full scale at the output rather than out of a 280 mV full scale at the input. Hence the offset error will be virtually eliminated.

3) **Elimination of any differential error**: by moving to a single-ended offset storage mechanism, we avoid summing the effect of two kT/C noise sources.

4) **Elimination of 70 μVrms noise from voltage reference**: our comparator threshold is now externally referenced (not sampled), and does not need to pick up additional thermal noise by going through a buffer. Any input-referred noise due to the voltage reference setting up the comparator threshold will therefore be smaller by 70 μVrms.

13.4 Possible tradeoffs to improve performance

We now examine some changes that can be made, which address various shortcomings in the current converter:

1) **Obtaining linearity with moderate inversion, rather than gate degeneration**: Gate degeneration improves our linearity, but the increase in V_L comes at the cost of increased noise. However, as noted in section 13.1, with moderate inversion we can obtain higher linearity with a *decrease* in noise, by increasing Ib together with Cint. Hence we may want to remove gate degeneration and lower V_L by a factor of $(1+\kappa_p/\kappa_n)$ as given by Eq. (7). The loss of linearity can then be recovered by increasing the current Ib.

2) **Improve linearity (and lower Vd-max) by using bipolars**: instead of MOS diodes, we can use bipolars as our log I-to-V converters (indicated by the diode symbols in Fig. 66). Bipolar diodes have the advantage of remaining logarithmic over the entire 60dB of dynamic range. As this reduces the input voltage range (Vd-max) from 280mV to $3 \times 60\text{mV} = 180\text{mV}$, we can then lower V_L to reduce the output noise. Unfortunately, this reduction in noise is almost completely offset by an increased sensitivity to gain error at the input, caused by Vd-max being lower.

   For example, if we switch to bipolars and then lower V_L by removing gate degeneration, we effectively increase the effect of input noise by 280mV / 180mV = 1.56, but then decrease the actual output noise by $\sqrt{(1+\kappa_p/\kappa_n)} = \sqrt{2.42}$ which is also = 1.56. Hence our net precision is unaffected by the change[10].

   Lastly, as described in 5.2, the WLR gain = Vfs / Vd-max also determines the number of settling time constants available for auto-zeroing. So by decreasing Vd-max, what we are really doing is allowing for an increase the Gm/C bandwidth of the auto-zeroing loop, which we will then exploit[11]. Hence bipolars indirectly allow us to improve our settling time which is important if we want to achieve a precision greater than 6 bits.

3) **Raise Vfs by removing Wilson mirrors**: raising Vfs has a similar effect as reducing Vd-max with regards to settling time. However it also improves our attenuation of offset error, since any noise voltage referred to Vo1 then sees a larger full-scale voltage. Although the kT/C noise should be small, as argued in 13.2, we will also benefit from the attenuation of charge injection and other stray leakage onto Vo1. Physically, this comes about because any noise injected onto Vo1 is then a smaller fraction of the total charge stored at that node.

---

[10] The reduction in Vd-max implies that the WLR gain = Vfs / Vd-max = $Gm1/Cint \cdot t_{int}$ must be increased accordingly, to maximize Vfs. We have two options to do this: lower V_L by changing the topology, or increase Ib after the fact, during operation. If we do not lower V_L but instead increase Ib, we would not lower the fundamental kT/C noise but in fact degrade the precision by a factor of 1.56 by using bipolars.

[11] We should remember from section 5.2 that the Gm/C settling bandwidth is not affected by Ib or V_L but dictated by Vfs / Vd-max! We cannot increase the Gm/C ratio to affect the settling bandwidth unless we enlarge Vfs / Vd-max.

To raise Vfs, we need to remove the Wilson mirrors because they rob us of almost a volt of output swing. The Wilson mirrors are meant to provide higher output impedance and a lowered number of noise sources. But we have shown in the appendix that Wilson mirrors do not provide much improvement in output resistance when fed by a low impedance source, like our gate-degenerated diff pair. Thus replacing them with normal mirrors should not degrade the ideal performance of the integrator.

The second benefit of a Wilson mirror is the lowered number of noise sources from N=2 to N=0.67, when fed by the gate-degenerated diff pair. But if we lower V_L to reduce our noise by removing the gate-degeneration, we unfortunately bring N back up to 2 again [as shown in the appendix]. Hence the number of noise sources in the WLR will rise by ~2.7 from 3.5 to ~6.5. We may choose to keep the Wilson mirrors for their high output impedance, but as the benefit of increasing Vfs should be greater, the Wilson mirrors should be removed to improve the attenuation of offset errors. Vfs can likely be raised from 1.2V to 2V in this new topology.

A summary of these tradeoffs is presented below:

Proposed Change	Moderate Inversion	Removal of gate degeneration	Bipolar diodes	Replacement of Wilson mirrors	Total effect
Effect	↑Ib and ↑Cint (preserving Gm/C BW)	↓V_L by $1+\kappa_p/\kappa_n$ = 2.42×	↓Vd-max by 280/180 mV = 1.56×	↑N by 6.5/3.5 = 1.86× ↑Vfs by 2/1.2 = 1.67×	
Gain Error	↓ in kT/C by √Cint	↓ by √2.42 = 1.56× ↑ in Wilson mirrors by √1.86 = 1.36× (if not replaced)	Effectively ↑ by 1.56×	↑ by √1.86 = 1.36×	1.36×↓from √Cint
Offset Error				↓ by 1.67×	0.60
Linearity	↑ in strong inversion as ~ √2·(Vgs-Vth)	↓ by 2.42×	↑ by 1.56× & ideally logarithmic	-	0.64×↑from moderate inversion
Gm/C settling bandwidth	Preserved, but ↑Ib must exceed ↑Cint (diminishing returns)	Unchanged, as ↓V_L must be matched with ↓Ib to preserve Vfs/Vd-max	↑ by 1.56× (to match ↓Vd-max)	↑ by 1.67× (to match ↑Vfs)	2.6×

Fig. 67. Table summarizing the tradeoffs involved in 4 possible improvements, along with the estimated effect of each change.

Weighing the effect of each technique, we come to the following conclusions:

1) it is definitely worthwhile to bias the WLR in moderate inversion, to decrease gain error and improve linearity.

2) The removal of gate degeneration is also worthwhile on its own, as it attenuates the gain error by 1.36/1.56 at only the cost of some linearity (that can be regained).

3) The use of bipolar diodes is more of a mixed blessing. As the increase in gain error is substantial, we should be better off without them. It may also be possible to compensate for any square-law behavior in MOS diodes (as exhibited in section 6.5) by the flattening in WLR transconductance, to get away with as a low a V_L as possible.

4) Once gate degeneration is removed, there is no disadvantage to removing the Wilson mirrors, since the reduction in N has already been given up. The resultant increase in Gm/C settling bandwidth then gives us $4.3 \times 1.67 = \sim 7$ settling time constants, which yields a settling accuracy of better than 1 part in 1000, in addition to improved offset error attenuation.

13.5 Two strategies for getting 8-bit precision

In conclusion, the multi-stage topology is promising in many respects. However, we must still resort to burning more power in the WLR and increasing Cint + Cwlr, if we are to attenuate the kT/C gain error. Having fewer noise sources at the input already lowers the noise by a factor of 360 μV / 240 μV = 1.5 (from section 13.2), thus a 10× increase in capacitance which lowers the noise by a further factor of √10 can possibly improve the precision by up to 2 bits.

A variant on the multi-stage idea that we might try would be to store the offset on a second WLR that is biased at a much lower Gm, so the effect of any kT/C error is further reduced by the ratio of the two Gm's:

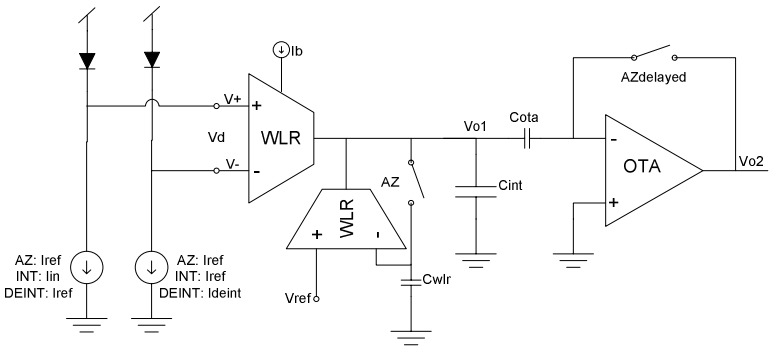


Fig. 68. A second WLR to perform offset compensation at the output node, Vo1. The second WLR must be biased with a smaller Gm to win by the ratio of the Gm's.

This requires just a very small bias, and should not hurt our power budget. However we must be careful not to introduce more noise due to the additional WLR. Now that we fully understand the problem, further work in this direction should soon enable us to achieve an 8-bit precision.

14 Bibliography

[1] **The all-electric ear**
Dettmer, R.
IEE Review , Volume: 34 Issue: 5 , 12 May 1988
Page(s): 195 -198

[2] **A Low-Power Wide-Dynamic-Range Analog VLSI Cochlea**
R. Sarpeshkar, R. Lyon, C. Mead
Analog Integrated Circuits and Signal Processing, Volume: 16(3), Aug 1998
Page(s): 245-274

[3] **Neuromorphic electronic systems**
Mead, C.
Proceedings of the IEEE , Volume: 78 Issue: 10 , Oct. 1990
Page(s): 1629 -1636

[4] **Biologically-based auditory signal processing in analog VLSI**
Lazzaro, J.
Signals, Systems and Computers, 1991. 1991 Conference Record of the Twenty-Fifth Asilomar Conference on , 1991
Page(s): 790 -794 vol.2

[5] **All About Decibels, Part I: What's Your dB IQ?**
Dumond, L.
Internet Column, Apr. 1999
http://www.prorec.com/prorec/articles.nsf/articles/EA68A9018C905AFB8625675400514576

[6] **An analog electronic cochlea**
Lyon, R.F.; Mead, C.
Acoustics, Speech and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on , Volume: 36 Issue: 7 , July 1988
Page(s): 1119 -1134

[7] **A comparison of different methods to assess phase-locking in auditory neurons**
Moissl, U.; Meyer-Base, U.
Engineering in Medicine and Biology Society, 2000. Proceedings of the 22nd Annual International Conference of the IEEE , Volume: 2 , 2000
Page(s): 840 -843 vol.2

[8] **Mimicking the human ear**
Loizou, P.C.
IEEE Signal Processing Magazine , Volume: 15 Issue: 5 , Sept. 1998
Page(s): 101 -130

[9] **Auditory prostheses research with multiple channel intracochlear stimulation in man**
Eddington, D.K.; Dobelle, W.H.; Brackmann, D.E.; Mladejovsky, M.G.; Parkin, J.L.

Ann. Otol. Rhinol. Laryngol., Volume: 87 (Suppl. 53), 1978
Page(s): 1-39

[10] **A Low-Power Analog Front-end Module for Cochlear Implants**
R. J. W. Wang, R. Sarpeshkar, M. Jabri, and C. Mead
presented at the XVI World Congress on Otorhinolaryngology, Sydney, March 1997

[11] **Analog-to-digital converter survey and analysis**
Walden, R.H.
Selected Areas in Communications, IEEE Journal on , Volume: 17 Issue: 4 , April 1999
Page(s): 539 -550

[12] **The Integrating A/D Converter (ICL7135)**
Intersil Application Note 017, 1999

[13] **Analog VLSI and Neural Systems**
C. A. Mead
Addison-Wesley Publishing Co., Reading, MA, 1989

[14] **A low-power wide-linear–range transconductance amplifier**
R. Sarpeshkar, R. F. Lyon, and C. A. Mead
Analog Integrated Circuits and Signal Processing 13(1/2), May/June 1997
Page(s): 123-151

[15] **Introduction to CMOS Op-Amps and Comparators**
R. Gregorian
John Wiley & Sons, Inc., 1999

[16] **Analog Integrated Circuit Design**
D. A. Johns, K. Martin
John Wiley & Sons, Inc., 1997

15 Appendix: Derivation of noise and impedances in the Wilson mirrors

15.1 Small signal circuit diagrams

A classic N-type Wilson mirror is shown below, taking as its input a current source Iin with input impedance R_1.

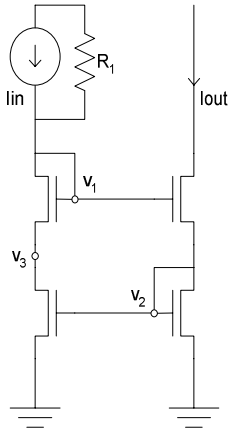


Fig. 69. A classic N-type Wilson mirror.

The small signal diagram for the circuit above can be drawn as follows:

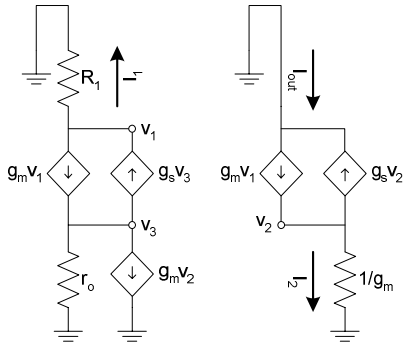


Fig. 70. Small signal diagram for the Wilson mirror in Fig. 69.

Where I_{out} is the short circuit output current. We can simplify this circuit model by replacing $g_s v_2$ with an equivalent resistor and replacing $g_m v_2$ with its effective function, which is to mirror the current I_2. Performing these simplifications and adding noise sources for each device, we get:

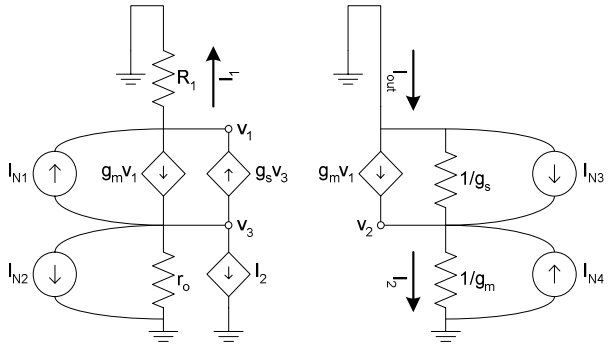


Fig. 71. Simplified small signal diagram including noise sources.

15.2 Block diagrams

We can therefore draw the block diagram from I_{N1} to I_2 as follows:

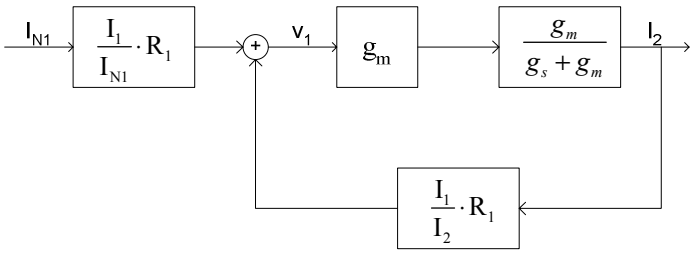


Fig. 72. Block diagram for transfer function I_2/I_{N1}.

From which we can intuitively interpret the loop gain in terms of 2 current dividers and a current gain:

1) **current division between g_m and g_s**: any current injected into node v_2 by g_m generator $g_m v_1$ will divide between resistors g_m and g_s to give current I_2;
2) **current division from I_2 to I_1**: I_2 then divides to give I_1 according to current divider ratio I_1/I_2;
3) **current gain by $g_m R_1$**: I_1 then gets multiplied by R_1 to give v_1, which in turn controls the g_m generator.

The current divider ratio in (1) is fixed at $\kappa/(1+\kappa)$, and does not have a big effect on the loop gain L, which is therefore $\kappa/(1+\kappa) \cdot I_1/I_2 \cdot g_m R_1$. However we can see that the value of R_1 is crucial in determining L, because it affects both the divider ratio $\mathbf{I_1/I_2}$, as well as the current gain $\mathbf{g_m R_1}$.

Let us therefore consider 2 likely cases for R_1:

	High Impedance	Low Impedance
R_1	$\approx r_o$	$\approx 1/g_m$
Divider ratio $\lvert I_1/I_2 \rvert$	$\approx ½$	≈ 1
Current gain $g_m R_1$	$g_m r_o = A$	≈ 1
Expected $\lvert L \rvert$	$\approx ½ A \cdot \kappa/(1+\kappa)$ ≈ 20	$\approx \kappa/(1+\kappa)$ ≈ 0.4

Fig. 73. Table summarizing the components of loop gain L. Values shown use an expected value of κ=0.6 and A=100.

First we will note that the I_1/I_2 divider ratio does not affect L by more than a factor of 2. L is therefore set primarily by current gain $g_m R_1$.

In the high impedance case when $R_1 \approx r_o$, and defining $A = g_m r_o$ to be a large gain » 1, L is then large and on the order of A.

However in the low impedance case when $R_1 \approx 1/g_m$, L is then just on the order of $\kappa/(1+\kappa)$, which gives a final value of $L < 1$.

Now we have a useful interpretation of the loop gain L and how it is affected by R_1, we can redraw the block diagram to include the remaining noise sources:

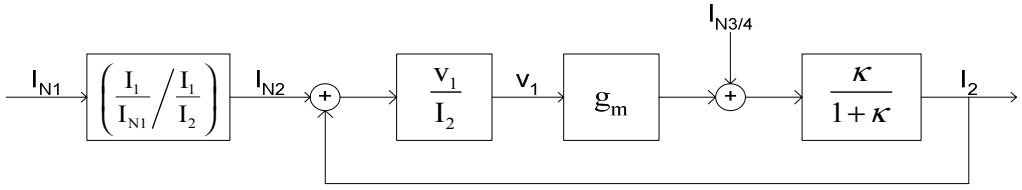


Fig. 74. Block diagram from each noise source to current I_2.

Firstly, we have combined the block $I_1/I_2 \cdot R_1$ as V_1/I_2 and moved it through the summing junction. Noting that I_{N2} acts just like I_2, I_{N2} therefore goes through the feedback loop in unity gain. To get I_{N2} from I_{N1}, we just have to multiply by the ratio of the current dividers $(I_1/I_{N1}):(I_1/I_2)$, or equivalently, $(v_1/I_{N1}):(v_1/I_2)$.

I_{N3} and I_{N4} both get injected into node V_2 in the same way as the g_m generator $g_m v_1$. For all noise sources except I_{N4}, $I_{out} = I_2$. I_{N4} is an exception because I_{N4} and I_{out} instead sum to form I_2. Hence $I_{out} = I_{N4} - I_2$, and dividing through by I_{N4}, we can obtain I_{out}/I_{N4} from I_{out}/I_{N3} simply as $1 - I_{out}/I_{N3}$. These transfer functions are summarized in the following equations:

$$\begin{aligned}
\frac{I_{out}}{I_{N1}} &= \left(\frac{v_1}{I_{N1}} \Big/ \frac{v_1}{I_2}\right) \cdot \frac{L}{1-L} \\
\frac{I_{out}}{I_{N2}} &= \frac{L}{1-L} \\
\frac{I_{out}}{I_{N3}} &= \left(\frac{\kappa}{1+\kappa}\right) \cdot \frac{1}{1-L} \\
\frac{I_{out}}{I_{N4}} &= 1 - \frac{I_{out}}{I_{N3}}
\end{aligned}
\quad \Bigg| \quad
\begin{aligned}
\text{where } L &\triangleq \frac{v_1}{I_2} \cdot g_m \cdot \left(\frac{\kappa}{1+\kappa}\right) \\
&= \frac{I_1}{I_2} \cdot g_m R_1 \cdot \left(\frac{\kappa}{1+\kappa}\right)
\end{aligned}
\qquad (21)$$

Fig. 75. Transfer functions from each noise source to I_{out}.

15.3 Calculation of current divider ratio I_1/I_2

What remains is for us to calculate the transfer functions v_1/I_2 and v_1/I_{N1}.

To deduce how I_2 divides as I_1/I_2, we must first find the Thevenin equivalent resistance looking up into node v_3. Redrawing that section of the small signal circuit as follows:

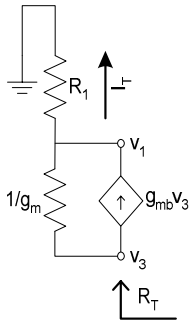


Fig. 76. Diagram to find small signal Thevenin resistance R_T.

We can then write out R_T:

$$V_T = I_T \cdot R_1 + (I_T - g_{mb} V_T) \cdot \frac{1}{g_m}$$

$$R_T = \frac{R_1 + \frac{1}{g_m}}{1 + \frac{g_{mb}}{g_m}} = \kappa \left(R_1 + \frac{1}{g_m} \right) \tag{22}$$

Fig. 77. Calculation of R_T, taking body effect into account.

Which tells us that the body effect simply causes the effective resistance ($R_1 + 1/g_m$) to be degenerated by a factor of κ.

To see how the current I_2 divides between R_T and r_o, we can draw out the current divider with the equivalent block diagram:

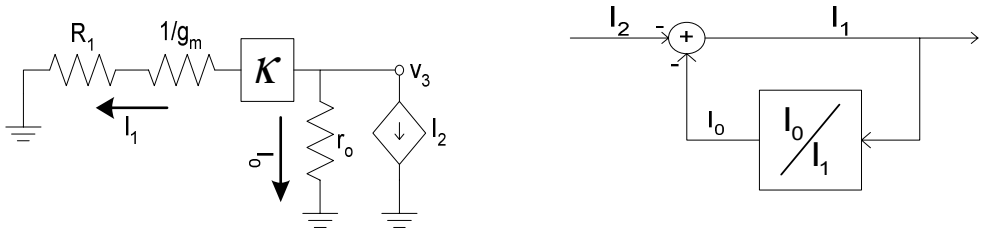


Fig. 78. Current division of I_2 between I_1 and I_0 according to ratio of resistances.

Since the ratio of the currents I_0/I_1 = ratio of the resistances R_T/r_o, the block diagram allows us to write out I_1/I_2 straightforwardly:

$$\frac{I_0}{I_1} = \frac{R_T}{r_o} = \frac{\kappa \left(\frac{1}{g_m} + R_1 \right)}{r_o} = \frac{\kappa (1 + g_m R_1)}{A}$$

$$\frac{I_1}{I_2} = \frac{-1}{1 + \frac{R_T}{r_o}} = \frac{-1}{1 + \kappa \left(\frac{1 + g_m R_1}{A} \right)} \tag{23}$$

Fig. 79. Current divider transfer function from I_2 to I_1.

As we anticipated in Fig. 73, when $R_1=r_o$, then $I_1/I_2 \approx -1/(1+\kappa) \approx -5/8$. But in the case where $R_1 \approx 1/g_m$, then $I_1/I_2 \approx -1$.

15.4 Calculation of current divider ratio I_1/I_{N1}

Now if we want to find I_1/I_{N1}, we can just write out a KCL equation taken at node V_1 (from Fig. 71):

$$I_1 + g_m v_1 = g_s v_3 + I_{N1}$$
$$I_1(1 + g_m R_1) = g_s(-I_1 r_o) + I_{N1}$$

$$\frac{I_1}{I_{N1}} = \frac{1}{1 + g_m R_1 + g_s r_o} = \frac{1}{1 + g_m R_1 + \frac{A}{\kappa}} \qquad (24)$$

$$= \frac{\kappa/A}{1 + \kappa\left(\frac{1 + g_m R_1}{A}\right)}$$

Fig. 80. Current divider transfer function from I_{N1} to I_1.

Hence the ratio of the current dividers $(v_1/I_{N1}):(v_1/I_2) = (I_1/I_{N1}):(I_1/I_2)$ is just κ/A, independent of resistance R_1. We can understand this from noting in Fig. 71 that I_{N1} will mainly circulate in the $1/g_m$ diode and not flow out through R_1, because R_1 is always in series with r_o.

15.5 Transfer functions for $R_1=r_o$

Substituting $(I_1/I_{N1}):(I_1/I_2) = \kappa/A$ into Eq. (21), we can then write out the approximate transfer functions for the case when $R_1=r_o$.

$$\begin{aligned}
\frac{I_{out}}{I_{N1}} &= \frac{\kappa}{A}\cdot\frac{L}{1-L} \approx -\frac{\kappa}{A} \\
\frac{I_{out}}{I_{N2}} &= \frac{L}{1-L} \approx -1 \\
\frac{I_{out}}{I_{N3}} &= \left(\frac{\kappa}{\kappa+1}\right)\cdot\frac{1}{1-L} \approx \frac{\kappa+1}{A} \\
\frac{I_{out}}{I_{N4}} &= 1-\frac{I_{out}}{I_{N3}} \approx 1
\end{aligned}$$

When $R_1=r_o$:

$$\begin{aligned}
\frac{I_1}{I_2} &\approx \frac{-1}{1+\kappa} \\
g_m R_1 &= A \\
L &= \frac{-A}{1+\kappa}\cdot\left(\frac{\kappa}{1+\kappa}\right)
\end{aligned} \qquad (25)$$

Fig. 81. Transfer functions from each noise source to I_{out}, given $R_1 = r_o$.

Substituting expected values of κ=0.6 and A=100, we can estimate the number of equivalent noise sources as:

Noise source	Transfer function	Noise contribution
I_{N1}	0.01	$0.01^2 \approx 0$
I_{N2}	0.96	$0.96^2 = 0.92$
I_{N3}	0.02	$0.02^2 \approx 0$
I_{N4}	0.98	$0.98^2 = 0.97$
Total		1.89

Fig. 82. Estimate of each source's noise contribution when $R_1 = r_o$.

Thus when $R_1=r_o$, the total number of noise sources is ≈ 2.

15.6 Transfer functions for $R_1=1/g_{mp}$

If we employed standard gate degeneration with a gate-to-drain connected diff pair as shown in below:

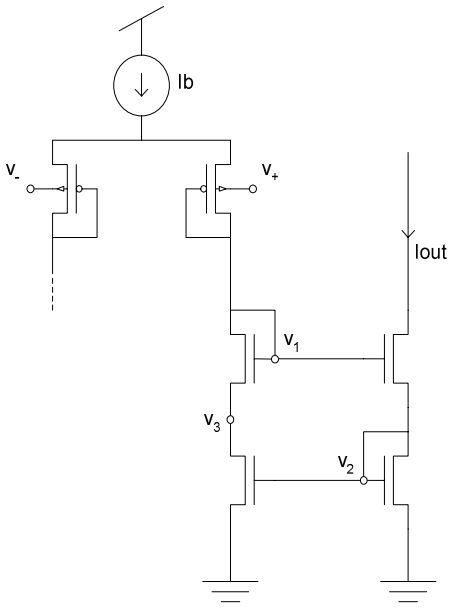


Fig. 83. Standard gate-degenerated diff pair; $R_1 = 1/g_{mp}$.

Then our value for R_1 would be $1/g_{mp}$ and our value of the current gain g_mR_1 would be = κ/κ_p, which is in fact not a gain but an attenuation!

Once again substituting $(I_1/I_{N1}):(I_1/I_2) = \kappa/A$ into Eq. (21), we can write out the approximate transfer functions for the case when $R_1 = 1/g_{mp}$:

$$\begin{aligned}
\frac{I_{out}}{I_{N1}} &= \frac{\kappa}{A}\cdot\frac{L}{1-L} \approx -\frac{\kappa}{A}\frac{1}{1+\frac{\kappa_p}{\kappa}\left(\frac{1+\kappa}{\kappa}\right)} \\
\frac{I_{out}}{I_{N2}} &= \frac{L}{1-L} \approx -\frac{1}{1+\frac{\kappa_p}{\kappa}\left(\frac{1+\kappa}{\kappa}\right)} \\
\frac{I_{out}}{I_{N3}} &= \left(\frac{\kappa}{1+\kappa}\right)\cdot\frac{1}{1-L} \approx \frac{\kappa_p/\kappa}{1+\frac{\kappa_p}{\kappa}\left(\frac{1+\kappa}{\kappa}\right)} \\
\frac{I_{out}}{I_{N4}} &= 1-\frac{I_{out}}{I_{N3}} \approx \frac{1+\frac{\kappa_p}{\kappa}\left(\frac{1}{\kappa}\right)}{1+\frac{\kappa_p}{\kappa}\left(\frac{1+\kappa}{\kappa}\right)}
\end{aligned}$$

$$\text{When } R_1 = \frac{1}{g_{mp}}:$$

$$\begin{aligned}
\frac{I_1}{I_2} &\approx -1 \\
g_mR_1 &= \frac{\kappa}{\kappa_p} \\
L &= -\frac{\kappa}{\kappa_p}\cdot\left(\frac{\kappa}{1+\kappa}\right)
\end{aligned} \qquad (26)$$

Fig. 84. Transfer functions from each noise source to I_{out}, given $R_1 = 1/g_{mp}$.

Substituting expected values for κ=0.6, κ_p=0.85 and A=100, we can estimate the number of equivalent noise sources as:

Noise source	Transfer function	Noise contribution
I_{N1}	0.001	$0.001^2 \approx 0$
I_{N2}	0.21	$0.21^2 = 0.04$
I_{N3}	0.30	$0.30^2 \approx 0.09$
I_{N4}	0.70	$0.70^2 = 0.49$
Total		0.62

Fig. 85. Estimate of each source's noise contribution when $R_1 = 1/g_{mp}$.

Thus when $R_1=1/g_{mp}$, the total number of noise sources is ≈ 0.6, which is much reduced! This is again to be anticipated from the intuition developed in section 15.2: the current gain g_mR_1 basically determines the loop gain L, and since L is <1, all the transfer functions turn out to attenuate the noise.

15.7 Single-diode loading using the Wilson mirror

The gate degenerated diff pair shown in Fig. 83 will see ~2 N-diodes as its effective load. This is the input impedance of a Wilson mirror:

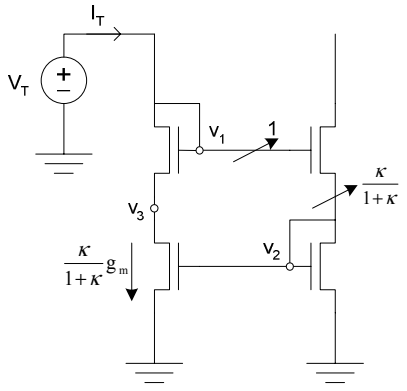


Fig. 86. Input impedance of a Wilson mirror: ≈ $(1+\kappa)/\kappa \cdot 1/g_m \approx 2.7/g_m$ using κ=0.6

The effective conductance is therefore I_2/v_1 which we could have also derived from the block diagram of Fig. 74 as $\kappa/(1+\kappa)\cdot g_m$ where v_1 is replaced with the test voltage source V_T.

However, our WLR was designed to be loaded with only 1 N-diode as follows:

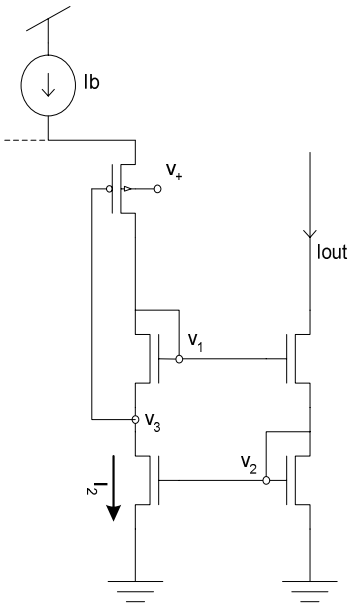


Fig. 87. Loading of gate degenerated diff pair with single diode drop from the Wilson mirror.

This gives rise to different transfer functions for v_1/I_2 and v_1/I_{N1} that should be used in Eq. (21).

We should first note that the transfer function v_1/I_2 should be very small, ≈ 0. We can understand this intuitively as follows: any current sunk by I_2 must be supplied from above by the PMOS and will therefore drop v_3 by I_2/g_{mp}. I_2 will also flow through the NMOS diode, and the Vgs drop v_1-v_3 must therefore be I_2/g_m. With the fall in v_3, this means that v_1 needs to move hardly at all, by $I_2/g_m + v_3 = I_2 \cdot (1/g_m - 1/g_{mp})$. This actually appears to be slightly positive at first glance, because $1/g_m > 1/g_{mp}$, and results in positive rather than negative feedback.

Fortunately, we are assisted by body effect which lowers the transimpedance v_1/I_2 because a fall in v_3 allows a smaller Vgs drop to support the same increase in current I_2. Hence v_1 need not be as large, and actually turns out to fall (rather than rise) with increasing I_2:

15.8 Small signal diagram of the single-diode load configuration

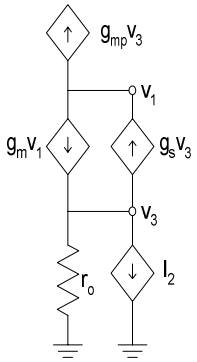


Fig. 88. Small signal diagram of the left hand transistors in Fig. 87.

We can calculate v_1/I_2 in 2 steps:

1) **v_3/I_2**: looking up into v_3, we see an impedance of $1/g_{mp}$ since any current that flows in response to an increase in v_3 is gated only by the PMOS.
   Thus $v_3/I_2 = -(1/g_{mp} // r_o) \approx -1/g_{mp}$.

2) **v_1/v_3**: Writing the node equation at v_1,

$$g_{mp} v_3 + g_m v_1 = g_s v_3$$

$$\frac{v_1}{v_3} = \frac{g_s - g_{mp}}{g_m} = \frac{1 - \kappa_p}{\kappa} \tag{27}$$

with $\kappa_P = 0.85$ and $\kappa = 0.6$, $v_1/v_3 = 0.15/0.6 = 0.25$.

Hence we can conclude: $v_1/I_2 = -(1-\kappa_p)/\kappa \cdot 1/g_{mp} \approx -0.25/g_{mp}$.

15.9 Finding v_1/I_{N1} for the single-diode load configuration

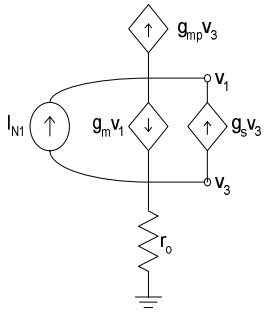


Fig. 89. Small signal diagram to find v_1/I_{N1}.

If we write down the two node equations at v_1 and v_3,

$$\begin{aligned} v_1 &= \frac{I_{N1} + g_s v_3 - g_{mp} v_3}{g_m} \\ v_3 &= \frac{g_m v_1 - I_{N1}}{g_s + \frac{1}{r_o}} \approx \frac{g_m v_1 - I_{N1}}{g_s} \end{aligned} \tag{28}$$

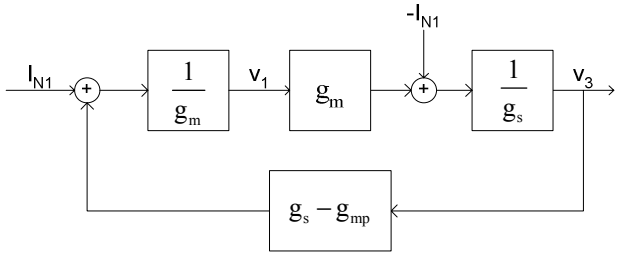


Fig. 90. Block diagram from which v_1/I_{N1} can be deduced.

We can then write down v_1/I_{N1} as:

$$\text{Loop gain, L} = \frac{g_s - g_{mp}}{g_s} = 1 - \kappa_p$$

$$\frac{v_1}{I_{N1}} = \frac{1/g_m}{1-L} - \frac{(1-\kappa_p)\cdot 1/g_m}{1-L} \qquad (29)$$

$$= \frac{\kappa_p \cdot 1/g_m}{\kappa_p} = 1/g_m$$

The ratio $(v_1/I_{N1}):(v_1/I_2)$ is therefore:

$$\frac{\frac{v_1}{I_{N1}}}{\frac{v_1}{I_2}} = \frac{\frac{1}{g_m}}{\frac{1-\kappa_p}{\kappa}\frac{1}{g_{mp}}} = \frac{\kappa_p}{1-\kappa_p} \qquad (30)$$

15.10 Transfer functions for single-diode load configuration

Substituting $(v_1/I_{N1}):(v_1/I_2) = \kappa_p/(1-\kappa_p)$ into Eq. (21), we can write out approximate transfer functions for single-diode load configuration:

$$\frac{I_{out}}{I_{N1}} = \frac{\kappa_p}{1-\kappa_p}\cdot\frac{L}{1-L} \approx -\frac{\kappa_p}{1-\kappa_p}\cdot\frac{1}{1+\frac{\kappa_p}{1-\kappa_p}\left(\frac{1+\kappa}{\kappa}\right)}$$

$$\frac{I_{out}}{I_{N2}} = \frac{L}{1-L} \approx -\frac{1}{1+\frac{\kappa_p}{1-\kappa_p}\left(\frac{1+\kappa}{\kappa}\right)}$$

$$\frac{I_{out}}{I_{N3}} = \left(\frac{\kappa}{1+\kappa}\right)\cdot\frac{1}{1-L} \approx \frac{\frac{\kappa_p}{1-\kappa_p}}{1+\frac{\kappa_p}{1-\kappa_p}\left(\frac{1+\kappa}{\kappa}\right)}$$

$$\frac{I_{out}}{I_{N4}} = 1-\frac{I_{out}}{I_{N3}} \approx \frac{1+\frac{\kappa_p}{1-\kappa_p}\left(\frac{1}{\kappa}\right)}{1+\frac{\kappa_p}{1-\kappa_p}\left(\frac{1+\kappa}{\kappa}\right)}$$

Single-diode load :

$$\frac{v_1}{I_2} \approx -\frac{1-\kappa_p}{\kappa}\frac{1}{g_{mp}} \qquad (31)$$

$$L = -\frac{1-\kappa_p}{\kappa_p}\cdot\left(\frac{\kappa}{1+\kappa}\right)$$

Fig. 91. Transfer functions from each noise source to I_{out} in the single-diode load configuration. Note that apart from the divider ratio (v_1/I_{N1}):(v_1/I_2), these transfer functions are identical to the case where $R_1 = 1/g_{mp}$ in Fig. 84, if we just replace all instances of $1-\kappa_p$ with κ.

Substituting expected values for κ=0.6 and κ_p=0.85, we can estimate the number of equivalent noise sources as:

Noise source	Transfer function	Noise contribution
I_{N1}	0.35	$0.35^2 = 0.12$
I_{N2}	0.06	$0.06^2 = 0.004$
I_{N3}	0.35	$0.35^2 \approx 0.12$
I_{N4}	0.65	$0.65^2 = 0.42$
Total		0.66

Fig. 92. Estimate of each source's noise contribution in the single-diode load configuration.

Thus the noise profile is significantly changed, but the total number of noise sources is still low, amounting to only ≈ 2/3 of a noise source.

15.11 Noise contribution from diff pair in WLR

The following small circuit diagram shows how noise from the diff pair will get to the output via I_2:

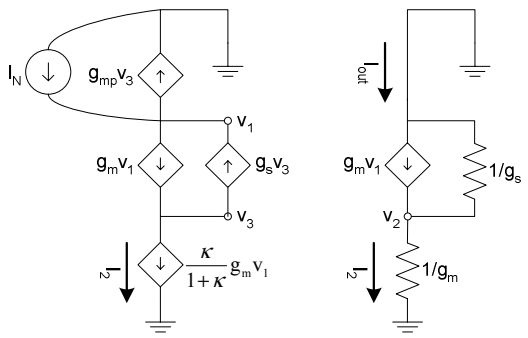


Fig. 93. Small signal diagram to show path from I_N to I_{out}.

I_N can be seen to divide between I_2 and $g_{mp}v_3$. All the current induced in I_2 appears at the output. So we just need to find how much current gets shunted away from I_2 by $g_{mp}v_3$.

We can simply observe that the difference in currents between $g_m v_1$ and $\kappa/(1+\kappa)\cdot g_m v_1$ must shunt back up through the $g_s v_3$ generator. Hence we know $g_s v_3 = 1/(1+\kappa) g_m v_1$, and therefore $v_1/v_3 = (1+\kappa)/\kappa$.

The current divider between I_2 and $g_{mp}v_3$ is therefore:

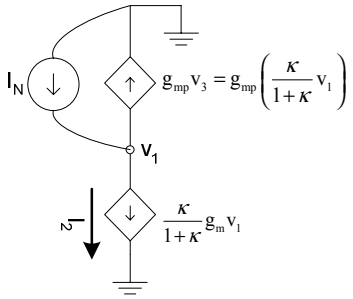


Fig. 94. Small signal diagram to show current division of I_N between I_2 and $g_{mp}v_3$.

Since the conductance looking down from v_1 is $\kappa/(1+\kappa)g_m$ (which was also observed in Fig. 86), and the conductance looking up from $v_1 = g_{mp}\cdot\kappa/(1+\kappa)$, the current divider I_2/I_N is just equal to the conductance divider ratio:

$$\frac{I_2}{I_N} = \frac{g_m}{g_m + g_{mp}} = \frac{\kappa}{\kappa + \kappa_p} \tag{32}$$

And this turns out to be identical to standard gate degeneration where the noise current must divide between a diode-connected PMOS and a diode-connected NMOS.

For expected values of κ=0.6 and κ_p=0.85, I_2/I_N = 0.41, thus the number of noise sources = 0.41^2 = 0.17 per device in the diff pair.

15.12 Noise contribution from P-Wilson mirror / Output impedance of N-Wilson mirror

The last piece of information we need is the noise contribution from the P-Wilson mirror. The P-mirror functions identically to the case in section 15.5 that we have already analyzed, where $R_1 \approx r_o$.

To get a more exact value for R_1, we need to find the output impedance of the N-Wilson mirrors that have gate-degenerated input sources.

The small signal diagram from which we can calculate R_1 is shown below:

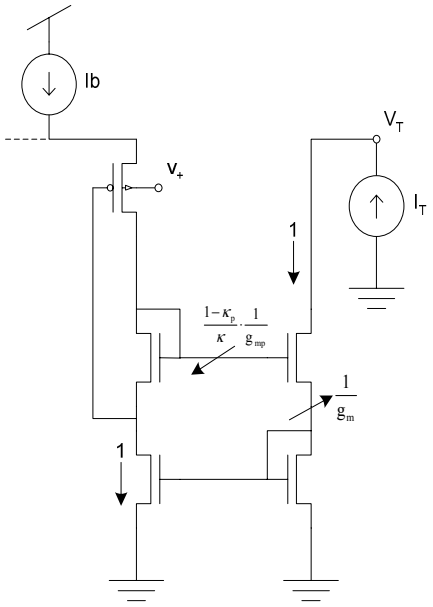


Fig. 95. Small signal diagram for calculation of our N-Wilson mirror's output impedance.

This allows us to find the output impedance R_T immediately:

Since $\frac{v_1}{I_2} = \frac{1-\kappa_p}{\kappa} \cdot \frac{1}{g_{mp}}$, we can write down:

$$R_T = \frac{V_T}{I_T} = \frac{1}{g_m} + \left(1 + g_m \frac{1-\kappa_p}{\kappa} \cdot \frac{1}{g_{mp}} + g_s \frac{1}{g_m}\right) r_o \qquad (33)$$

$$= \frac{1}{g_m} + \left(1 + \frac{1-\kappa_p}{\kappa_p} + \frac{1}{\kappa}\right) r_o$$

thus $R_T \approx (1+0.18+1.7)r_o = 2.9r_o$ using $\kappa_p=0.85$ and $\kappa=0.6$.

This is not as high as the gain of $g_m r_o/2$ that we are used to seeing in Wilson mirrors, because the effective current gain is very low at only $(1-\kappa_p)/\kappa_p$.

Nevertheless, the source impedance feeding the P-Wilson mirror is on the order of r_o, hence the number of noise sources in the P-Wilson ≈ 1.9 as given by Fig. 82.

15.13 Total number of noise sources in the WLR

The number of noise sources in the WLR can therefore be tabulated as follows:

Structure	# noise sources
Gate degenerated diff pair	2×0.17
N-Wilson mirrors	2×0.66
P-Wilson mirror	1.9
Total	3.56

Fig. 96. Summary of theoretical noise sources in the WLR of Fig. 13.