

Analog CMOS Circuit Design

Paul O'Connor, BNL

Selected slide captions

Slide 16

Recall that most radiation detectors behave as high-impedance capacitive sources. The ideal integrator is composed of an inverting voltage amplifier and capacitor in feedback. To maintain the input node at virtual ground the output will slew to a voltage $V_o = -Q_i/C_f$ in response to a current pulse at the input. If the loop gain of the amplifier is sufficiently high the integrator gain V_o/Q_i will be $1/C_f$, independent of detector and amplifier properties.

Slide 17

Many interrelated factors enter in to the design of an MOS charge amplifier. The relevant detector parameters are the capacitance, leakage current, and charge collection time. The experiment dictates the time and amplitude distribution of events. Typically natural sources have Poisson-distributed arrival times characterized by an average rate, while accelerator sources may have a well-defined time structure. The experiment also sets the acceptable limits for noise, linearity, and maximum signal. A very important, but frequently overlooked system constraint is the power budget for the front-end ASIC.

The designer must choose an appropriate MOS technology from the many varieties offered by today's foundries. The input device characteristics (NMOS/PMOS, gate dimensions, and bias condition) play the largest role in determining the ENC, and this is where the most careful optimization must take place. The relative allocation of the power budget to the input device and the rest of the circuit is also involved in this decision. As we will see, there are many options for the reset system (low-frequency circuit to discharge the feedback capacitor). Finally the shaper impulse response sets the weighting

function for noise and can be tailored to handle the anticipated event rate of the experiment.

Slide 18

As shown in the first lecture, the noise sources can be divided into serial (red), parallel (green), and 1/f (blue) components having $t_m^{-1/2}$, t_m^0 , and $t_m^{1/2}$ dependence on measurement time respectively. If shaping time is a free parameter, there is an optimum where series and parallel noise are equal:

$$t_{opt} = C_{det} \sqrt{R_P R_S}$$

where R_P and R_S are the equivalent parallel and white series noise resistances,

$$R_S = e_n^2 / 4kT$$

$$R_P = 4kT / i_n^2$$

If power is unconstrained, the equivalent series noise resistance R_S can be reduced by increasing the bias current in the input FET. Parallel noise is not inherent to the amplification process and can be minimized or eliminated in many cases. The ultimate limit to the noise performance of an MOS charge amplifier is the 1/f noise of the technology.

Normally the measurement time is not a free parameter but is constrained by pileup, timing precision requirement, or ballistic deficit due to finite charge collection time.

Slide 19

Considering for the moment only the series white noise component, it is clear that the input MOSFET width has an optimum. This is because both the transconductance and gate capacitance of the FET increase as width is increased. The former leads to a reduction in noise spectral density of the FET, while the latter increases the effective input capacitance of the detector + front end.

Both C_{gs} and g_m depend on the technology (gate oxide thickness), physical gate dimensions (width and length) and the bias conditions (drain current density, degree of inversion). In discrete designs the effective width of the FET is normally changed by introducing one or more devices in parallel, keeping each one at the same drain current. In that case the white series noise is minimized when the combined transistor gate capacitance is equal to the detector capacitance. In custom integrated circuits with power constrained, the drain current budget is fixed and the device width is changed directly. This case leads to an optimum at $C_{gs} = C_{det}/3$ when the device is operated in strong inversion.

The $1/f$ noise coefficient K_F is more or less independent of bias condition. Therefore the optimum device size for $1/f$ noise is easily found to be $C_{gs} = C_{det}$.

Slide 20

When we combine the white and $1/f$ noise components we can't get a clean analytic solution to the optimization. We find the solution numerically by summing in quadrature the white, $1/f$, and parallel noise contributions to the ENC. Here is an example
The $1/f$ match at $C_{gs}/C_{det}=1$ as always. The white match is at about $C_{gs}/C_{det}=0.15$, which means the device is just at the weak-strong inversion boundary. The overall optimum is about 0.3. Note also that the minima are fairly broad. An error by a factor of about 3 either way would result in about a 40% increase in noise.

Slide 21

The relationships between device width, drain current, transconductance, capacitance, and noise source spectral density depend on the region of operation. As seen in the second lecture the most advanced CMOS processes favor device operation in the moderate inversion region (low current density). Moderate inversion is even more prevalent in charge amplifier input transistors because they must match (within a factor of ~ 5) the capacitance of the detector. Detector capacitance is always at least 100 times

larger than the capacitance of a minimum-sized transistor in modern processes. The correspondingly large device width together with the low drain current allowed by power dissipation constraints results in a low current density, low inversion coefficient region of operation.

Slide 22

Analytic expressions for MOSFET behavior in moderate inversion have been lacking until the mid-90's when the "EKV" model was introduced. This new model, which is slowly making its way into engineering textbooks and circuit simulators, has simple expressions that are valid for the entire range of operating conditions from weak to strong inversion. Hence it is particularly useful for analog design in modern submicron processes, where the favored operating point often places the device in moderate inversion. The EKV model makes use of the inversion coefficient i , which is a measure of drain current normalized in such a way that $i \ll 1$, $i \gg 1$ represent weak and strong inversion respectively.

Using the EKV expressions for transconductance, capacitance, and noise spectral density allows the ENC to be optimized without resorting to full analog simulation and without need for a detailed transistor model from the foundry. Since the EKV parameters are simple and physics-based, it is also possible to predict noise performance for future scaled technologies.

Slide 23

This series of graphs shows the transconductance of NMOS and PMOS devices as a function of gate capacitance. As the device is made wider (higher C_G) at constant drain current it goes from strong to weak inversion. In strong inversion $g_m \sim C_G^{1/2}$, while in weak inversion g_m becomes independent of device size. The weak-strong transition occurs earlier for smaller feature-size technologies, and for lower currents. In strong inversion, the NMOS device has higher g_m/C_G ratio than PMOS due to the higher mobility of electrons compared to holes.

Slide 24

Series white noise has a minimum due to the competing effects of gate capacitance and transconductance as a function of gate width. Higher drain current (more power) and smaller technology feature size lead to lower noise levels. Strong inversion conditions (large I_D/C_G ratio) give rise to a minimum at $C_G \sim C_D/3$, while in moderate and weak inversion the minimum is at smaller values of gate width.

Slide 25

The next 2 slides show some noise and matching trends for the composite series noise. Slide 25 shows the fully optimized noise for 0.5 μm gate length NMOS and PMOS transistors as a function of detector capacitance. NMOS results are for a shaper peaking time of 50 ns, while PMOS is optimized for 5 microsecond peaking. Noise (solid lines) and capacitive match (dotted) are shown for three power levels in the input device. When detector capacitance is large and drain current is low, the optimized input device tends to operate in moderate inversion; under these conditions the optimized gate capacitance can be as small as 1% of the detector capacitance.

Slide 26

This figure shows the fully optimized noise as a function of power dissipation for 0.25micron CMOS technology for a 1 pF detector capacitance. Shaper peaking times from 10ns to 3 μs are shown.

At short peaking time and low power the series white noise dominates and the corresponding optimum ENC decreases roughly as $P^{-0.4}$. As power (and/or shaping time) increase, the 1/f noise component becomes increasingly important and the slope $dENC/dP$ flattens. In this example, an ENC of around 12 e- is the minimum possible for this combination of CMOS technology and detector capacitance.

Slide 27

In the preceding analysis the gate capacitance was modeled as a single value directly proportional to the active gate area. A more accurate model reflects the fact that the gate-source and gate-drain overlap capacitances contribute to the total input capacitance in parallel with C_{det} . C_{GSO} and C_{GDO} depend only on the device width (not area) and are bias-independent. However, these parasitic capacitances do not influence the 1/f noise. It is found empirically that for some technologies the 1/f noise coefficient K_F increases for the shortest gate lengths, and that the frequency dependence is $f^{-\alpha}$ where α is in the range 0.85 to 1.1. This behavior influences the choice of gate length and the peaking time dependence of the noise. See the reference for details:

G. De Geronimo, P. O'Connor, "MOSFET Optimization in Deep Submicron Technology for Charge Amplifiers", IEEE Trans. Nucl. Sci. 52(6), 3223-3232 (Dec. 2005).

Slide 28

The choice of NMOS/PMOS input device is influenced not only by reaching the minimum ENC, but also by practical concerns of signal swing and voltage headroom, signal return path, and off-chip interfacing. NMOS devices will have lower noise only for operating points far into strong inversion and at short peaking times where the white series noise is strongly dominant over 1/f. For the lowest-noise configurations, e.g. small C_{det} , high power budget, and long shaping time, PMOS will have an advantage over NMOS due to its lower 1/f noise coefficient.

Slide 29

These expressions are same as those given in Fig. 21 of the first lecture save for numerical constants which depend only on weighting function integrals.

Using these expressions one can crudely determine if the noise will be dominated by the white or 1/f series contribution.

Recall kT/K_F typically ~ 3000 for NMOS, $10,000 - 30,000$ for PMOS. When $t_m < 1000 t_{el}$ the amplifier noise is white series noise limited.

These rules of thumb can also help with the selection of NMOS/PMOS.

Summary – charge amplifier input transistor optimization

From slides 17-39 it should be apparent that many interrelated factors influence the design of an optimized charge amplifier input transistor. In particular, recognize that every combination of detector, shaping time, power dissipation limit, and CMOS technology requires re-optimization to achieve the lowest noise. For this reason it is usually the case that ASIC charge amplifiers are custom-designed for every experiment. The quick pace of CMOS technology scaling also means that every few years another technology family becomes obsolete. Since amplifier designs from older technologies are not readily portable to new generations, the initial fabrication run of parts must be sized to cover the lifetime needs for a particular experiment.

Slide 30

R_{eq} expressions can be derived by breaking the long gate finger up into infinitesimal transistors in parallel with their drain currents summed.

It doesn't take into account any time-varying transmission of signal down the gate, which is important for microwave applications.

Slide 31

Applies to any noise source originating in the bulk -- like coupled digital noise. Noise is flat with frequency up to pole from $C_{gate-channel} * R_{sub}$ (in reality, the bulk resistor and gate-channel capacitor form a distributed RC line). Therefore this noise is superimposed with white channel thermal noise and can masquerade as high gamma. It is more significant at shorter gate length, since it depends on g_{mb}^2 whereas the channel thermal noise depends only on g_m .

Slide 32

It is important during layout of the input device to prevent the introduction of parasitic resistances in the gate polysilicon and from the bulk. One typically computes the maximum allowable gate finger width and the maximum distance from active transistor area to nearest substrate contact. Then the input transistor is subdivided into cells which obey the resistance constraints, tied together in parallel with low-resistance metal wiring. To prevent pickup of induced currents flowing in the substrate, one or multiple guard rings surrounding the input transistor are advisable.

Slide 33

Practical integrators need some method to discharge the feedback capacitor between events. In classical discrete designs a high-value resistor performed this function, giving excellent linearity and low parallel noise contribution. However, another solution must be found in monolithic circuits where the highest generally-available materials have sheet resistances lower than 200 Ohms/square, making resistors with $R > 100\text{k}\Omega$ impractical. For high-precision circuits the reset system must be linear and low-noise over a wide signal range. It should also be insensitive to process/temperature/supply voltage variation, and should not add parasitic capacitance to the input node.

In addition to resetting the feedback capacitor, it is an important advantage if the reset circuit can supply (or sink) the DC leakage current of the detector, thereby allowing direct coupling of detector to preamp without bulky DC-blocking capacitors.

Slide 34

Physical resistors in a monolithic process are polysilicon or metal meanders and form a distributed RC line with their associated parasitic capacitance to the conductive substrate. Although their DC resistance can be made arbitrarily high by increasing the line length, the noise properties are determined by the real part of the input impedance (frequency-

dependent). Hence their parallel noise contribution will be higher than the classical Johnson noise $4kT/R$ whenever the measurement time is less than the RC time constant of the line.

A MOS switch can be used to reset the feedback capacitor in some applications. In the OFF state the switch contributes negligibly to the parallel noise, as leakage currents are usually in the pA region at room temperature. For low event rates the switch may be maintained in the OFF state up to ~ 1 millisecond. Immediately after discharge, there will be a transient disturbance which must be allowed to settle before the amplifier can be sensitive again. Also, each reset will leave a noise charge kTC on the feedback capacitor, which can be removed by double sampling the output before and after an event arrives.

Slide 35

Many continuous reset circuits that use active elements have been developed. The first circuit shown in this figure relies on the excellent matching of monolithic devices to create a current amplifier by coupling an imperfect integrator stage with a voltage amplifier having a transfer function that exactly cancels the nonlinearity of the integrator. This circuit is discussed further in slides 37 – 39.

The second solution uses a low-frequency differential amplifier to maintain the output of the integrator at constant potential V_{ref} . A pulse of charge creates a step at the output of the integrator, and the voltage is returned to V_{ref} with a time constant given by the product of the transconductance of the low-frequency stage and C_f . Since a second feedback loop is present the circuit must be carefully stabilized to prevent oscillation. Thermal noise of the transistors in the low-frequency amplifier contributes to the ENC.

Slide 36

R-scaling circuits utilize an attenuating current mirror to create a circuit element that behaves like a resistor of value $R*N$, where R is a physical resistor and N is the attenuation ratio of the mirror. The voltage at the output of the integrator V_{out} is applied to a voltage-to-current converter producing a current V_{out}/R , which is then sent to the input

via a current mirror with ratio I/N . The effective resistance $R*N$ must be $>10^6\Omega$ to be effective as an integrator reset; this sometimes requires difficult mirror ratios to be designed.

The slew-rate limited configuration uses a constant current to discharge the feedback capacitor. This produces a linear, rather than exponential return to baseline and can be used with a discriminator to produce a pulse whose width is linearly proportional to the deposited charge.

Slide 37

The diagrams illustrate the similarity of the self-adaptive current-amplifying configuration to the classical pole-zero compensation circuit used in discrete preamplifier designs. In the frequency domain, a zero is created by R_C*C_C that cancels the pole formed by the integrator feedback elements. The monolithic version relies on matching of capacitors and transistors to create the cancellation. Moreover, the transistor's nonlinearities are canceled because their gates and sources are common, and their drains connect to virtual grounds at the same potential (inputs of A1 and A2). The overall circuit (A1, CF, MF, CC, and MC) injects a current into A2 which is an N-times scaled replica of the input current from the detector.

Note that the PMOS transistors shown (MF, MC) are appropriate for amplifiers which must respond to *negative* input charge. If the detector injects positive conventional current into the amplifier, the reset and compensation transistors must be NMOS.

Slide 38

Experimental results confirm that the nonlinear compensation is effective, and that the amplifier self-adapts to leakage current up to 70 nA with negligible change in gain.

Slide 39

A summary of the advantages and disadvantages of the self-adaptive, nonlinear compensated reset system.

Slide 40

In a properly designed charge amplifier the noise is dominated by the input transistor. However, in low-voltage CMOS processes it is difficult to degenerate the current sources which supply the input and cascode branches. Current source noise is minimized by using long-gate devices, but the lower the W/L ratio, the higher the drain-source voltages needed to operate in saturation.

It is also important to model the response of the charge amplifier to large signals. In some experiments, charge in excess of 10^3 the normal signal charge can be generated in the detector. The preamp's recovery from such overload can be extremely slow. If large overloads are expected the preamp must be designed to recover quickly, for instance by re-dimensioning the reset transistor or by providing a fast reset path that is triggered by saturating signals.

Slide 41

Slide 42

The shaping amplifier performs analog processing on the signal from the preamp to minimize the measurement error with respect to noise, and at high counting rates to minimize the effects of pulse overlap or pileup.

Shaping amplifiers whose system parameters do not change with time are referred to as time-invariant. Time-invariant systems are described completely by their impulse response. The impulse response function may be unipolar (always positive with respect to the baseline) or bipolar. Unipolar impulse response is characterized by its peaking time (typically from 1% above baseline to peak), width (1% -- 1%), slope at threshold crossing, and first and second moments. The symmetry of the pulse may be characterized by the ratio of rise to fall times, or peaking time to full 1% width. For bipolar functions

the time to zero crossing, time to negative peak, and ratio of positive to negative peaks are also important. Bipolar response generated by a linear system always has zero net area.

Slide 43

If the noise spectral density and input signal waveform are completely known, and if count rate is low enough to avoid pileup, then a matched filter can be constructed mathematically that will lead to minimum noise. However, these conditions are rarely met in practice and the ideal matched filter is usually difficult to realize with practical circuits. Selection of the appropriate shaping function is commonly constrained by rate, charge collection time of the detector, and acceptable level of circuit complexity. Some systems perform digital processing on a waveform sampled just after preamplification. This technique is more flexible in realizing an arbitrary shaping function, but must satisfy Nyquist sampling and quantization noise requirements, which often leads to unacceptable levels of power dissipation.

Slide 44

Simple shaper response functions can be realized by cascades of single-order low and high-pass filters. The higher the order of the filter, the more symmetrical the output waveform.

Slide 45

This slide gives expressions for the transfer function, impulse response, and peaking time of the classical semiGaussian ($CR-RC^n$, CR^2-RC^n) filters of Slide 44.

Slide 46

Ohkawa (NIM 1976) gives a formal derivation of the pole-zero constellations giving the closest approximation to a Gaussian pulse. As in the simple $CR-RC^n$ filters, higher-order filters give more symmetrical response, compared to the $CR-RC^n$ real-pole filters, the complex-pole shapers derived by this method have a superior symmetry for the same filter order.

Slide 47

The graph compares the impulse response of a 5th-order complex-pole shaper derived by the Ohkawa method with first- and fourth-order $CR-RC^n$ filters. The filters are compared at equal 1% widths, as appropriate for a pileup-limited signal chain. Since the more symmetric pulse has smaller derivative everywhere, its series noise weighting function is superior to the others.

Some circuits for realizing two poles per amplifier are shown in the lower part of the figure.

Slide 48

As in any amplifier the noise of the second and subsequent stages, when reported to the input, is attenuated by the gain of prior stages.

Slide 49

AC-coupled systems are subject to baseline wander with random pulse input, while DC coupling requires careful baseline stabilization against drift caused by detector leakage, temperature, or supply voltage variation. The low-frequency feedback system achieves good baseline stability and can also be used to equalize the baselines of many chips in a system. Note that this is not the same as the LF feedback loop around the preamp in the type of reset system shown in Fig. 35.

Slide 50

Bipolar pulse shaping automatically takes care of baseline wander and rejects low-frequency disturbances, but is noisier.

Slide 51

This implementation uses a slew-rate limited source follower and a very low-frequency low-pass filter in feedback to achieve good baseline stability.

Slide 53

With a limited power budget, it is critical to expend as large a fraction as possible on the preamplifier and shaper where it can contribute to reducing the noise. In typical applications, the input to the preamp/shaper has low duty cycle, i.e. low channel occupancy in time.

Off-chip analog I/O can be very costly in terms of power if the simplest output drivers, such as source followers, are used. To avoid wasting power the driver should use Class AB stage where the quiescent current can be much smaller than the maximum source or sink current that can be provided to a load.

System architecture choices can have a large effect on the need to drive analog signals over long lines and should be considered at the beginning of the experiment.