Prospects for Charge Sensitive Amplifiers in Scaled CMOS[1]

P. O'Connor and G. De Geronimo
Brookhaven National Laboratory, Instrumentation Division, Bldg 535B, Upton, NY 11973, USA

Abstract

Due to its low cost and flexibility for custom design, monolithic CMOS technology is being increasingly employed in charge preamplifiers across a broad range of applications, including both scientific research and commercial products. The associated detectors have capacitances ranging from 50 fF to several hundred pF, and applications call for pulse shaping from tens of ns to tens of μs, and constrain the available power per channel from tens of μW to tens of mW. At the same time a new technology generation, with changed device parameters, appears every 2 years or so. The optimum design of the front end circuitry is examined taking into account submicron device characteristics, weak inversion operation, the reset system, and power supply scaling. Experimental results from recent prototypes will be presented.

We will also discuss the evolution of preamplifier topologies and anticipated performance limits as CMOS technology scales down to the 0.1 μm / 1.0 V generation in 2006.

I. Introduction

Charge sensitive amplifiers (CSAs) are used extensively in processing the signals from capacitive sensors such as photodetectors, pressure sensors, particle and X-ray detectors, and pyroelectric detectors. Increasingly, these amplifiers are being implemented in monolithic processes where there is a need for high-volume production (e.g. particle physics collider detectors), for interfacing to a dense array of sensors (pixel detectors), or whenever miniaturization and high functional integration are at a premium. The high input impedance of the MOSFET makes CMOS an attractive technology for fabricating such amplifiers, particularly if a high level of integration is desired. However, CMOS technology development is driven by the needs of digital VLSI, and the resulting rapid feature size scaling presents several challenges for high dynamic range CSAs:

- increase in MOSFET noise due to carrier heating in the channel, higher interface trap density, gate tunnelling current, and larger parasitic resistances;
- reduced power supply voltage which restricts the output swing, constrains the circuit topology, and increases the noise of current sources;
- increased application demand for mixed-signal circuits having digital switching activity occurring on the same substrate with highly sensitive CSAs;
- decreased availability of quality passive components for analog design;
- poor modeling of the DC, AC, and noise properties of the devices.

[1] Work supported by the US Department of Energy, Contract No. DE-AC02-98CH10886.

Many of these drawbacks have been discussed in the overall context of analog design in scaled CMOS [1,2]. In the following sections we consider only those aspects of CMOS scaling which impact the performance of CSAs. After a review of CSA operation and CMOS scaling in Sections II and III, Section IV covers the noise sources in MOSFETs and Section V examines the expected scaling of the input-referred noise charge and dynamic range. Section VI discusses sources of noise other than the input transistor and additional scaling effects.

II. CSA Operation

Charge measurement systems are characterized by system requirements which vary tremendously from application to application:

- dynamic range 6 to 20 bits;
- sensor capacitance 50 fF to 10 nF;
- speed of response 5 ns to 1 ms;
- power dissipation 10 μW to 100 mW.

A generic block diagram, shown in Figure 1, represents such a system. The sensor, with capacitance C_{det}, produces pulses of charge which are integrated on the feedback capacitor C_F. The amplifier is characterized by a series input voltage noise source e_n with a white component set by the input device transconductance g_m and a 1/f component that is inversely proportional to the device area WL.

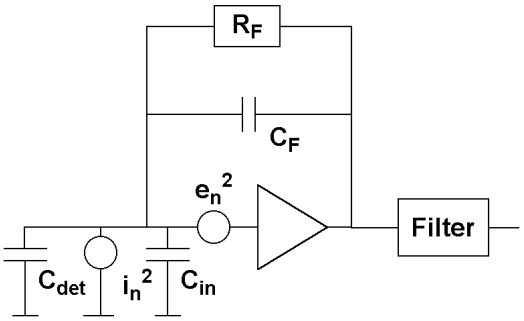

Figure 1: Generic CSA

In the case of CMOS the equivalent input noise current is negligible; however the secondary feedback element R_f which discharges the feedback capacitor contributes noise whose impact must be carefully considered. A filter amplifier following the preamplifier is necessary to eliminate out-of-band noise. The resolution of the system is expressed by the equivalent input noise charge (ENC) [3,4]:

$$ENC^2 = (C_{det} + C_{in})^2 [2kTa_1R_s/t_s + a_2K_f /C_{ox}WL] + 2kTa_3t_s/R_p$$

where C_{in} is the capacitance of the input transistor, R_s is its equivalent series noise resistance, R_p is the effective noise resistance of the feedback element R_f, t_s is the characteristic time constant of the postfilter, K_f is the 1/f noise coefficient of the input transistor, and a_1, a_2, and a_3 are form factors (of order 1) related to the series white, 1/f, and white parallel noise respectively. Series white noise dominates for short shaping times, R_p for long times, and 1/f noise in the intermediate range. High capacitance detectors exacerbate the series noise terms. While both series white noise and parallel noise can be combated by circuit techniques, expenditure of more power, etc., the effect of 1/f noise on CSA performance is fundamentally limiting. The lowest-noise CSAs are always dominated by the 1/f noise properties of the input transistor.

An ideal CSA technology would have a high g_m/C_{gs} ratio (f_T) at low current, as well as low γ ($\gamma = R_s \times g_m$) for minimizing the white series noise; a low 1/f noise coefficient K_F; and controllable sub-nA current sources for low parallel noise. A high quality floating capacitor is needed for the preamp integrating capacitor and for the postfilter. Other desireable features for CSAs are:

- high g_m/g_d for amplifier gain;
- excellent AC isolation for integration with digital circuits;
- high supply voltage for driving subsequent stages and for cascodes;
- ESD – tolerant;
- radiation tolerant.

III. Digital CMOS Scaling

CMOS technology scaling takes place by a series of well-defined process generations, coordinated among foundries and equipment manufacturers. In each process generation, integration density doubles and speed increases by a 50%. Constant-voltage scaling, which had been followed up to the 0.8 μm generation, has given way to quasi-constant field scaling as oxide and junction breakdown limits have been reached. Here all dimensions are reduced by the scale factor λ. doping density is increased, and supply voltage scaled down by the same factor. In submicron device scaling below 0.5 μm the major challenges are to minimize undesireable short-channel effects, control power dissipation, and to ensure reliability to the level of one failure in 10^7 chip-hours of operation.

The roadmap for the next 6 technology generations is shown in Table I [5].

Table 1
CMOS Technology Roadmap

Year	1997	1999	2001	2003	2006	2009
Feature size (μm)	.25	.18	.15	.13	.10	.07
Supply (V)	2.5	1.8	1.6	1.5	1.2	0.9
Tox (nm)	5.0	4.0	3.3	2.8	2.2	2.0
Vth (mV)	500	470	440	420	400	370
Nsub (10^{16}/cm^3)	3.4	5	6	7	10	20
Xj (nm)	100	70	50	<50	<50	<50
10^6 FET/cm^2	8	14	16	24	40	64
Interconnect (km/chip)	.82	1.5	2.2	2.8	5.1	10

IV. Device Noise Trends

A. White series noise

For long channel MOSFETs the proportionality constant γ relating noise to transconductance ($R_s = \gamma/g_m$) has the value 2/3 in strong inversion, 1 in the linear region, and 1/2 in weak inversion [6] (here the small contribution from the bulk transconductance g_{mb} is neglected). In short channel devices the carriers can acquire enough energy from the electric field in the channel to raise their effective temperature above that of the lattice. Models of this heating effect [7,8] lead to an increase in γ above the long channel values. However, the predictions of different models are inconsistent, or are only supported by experimental data over a limited range of bias conditions. Predictions are particularly lacking for the normal bias point for a device used as the input transistor of a CSA, namely low V_{DS} (just above $V_{DS\text{-}sat}$) and low current density I_D/W (moderate inversion). Values of γ as high as 2 – 4 have been reported [9,10], but these results are from experimental short channel devices that do not follow the same scaling laws as current commercial devices (e.g., L_g/t_{ox} ratio.) Also high γ values are usually reported on devices at high V_{DS} and high I_D/W. Hence, we have measured noise on MOS transistors fabricated in representative commercial submicron processes, for devices in a typical CSA bias condition. These results (along with recently reported results from similar devices [11-13]) are shown in Figure 2, for four technologies with a minimum gate length from 0.7 to 0.25 microns. It can be seen that for devices from the same process there is a modest increase in γ as the gate length decreases. However, there is no trend towards higher γ for the smallest-L_g devices in each technology.

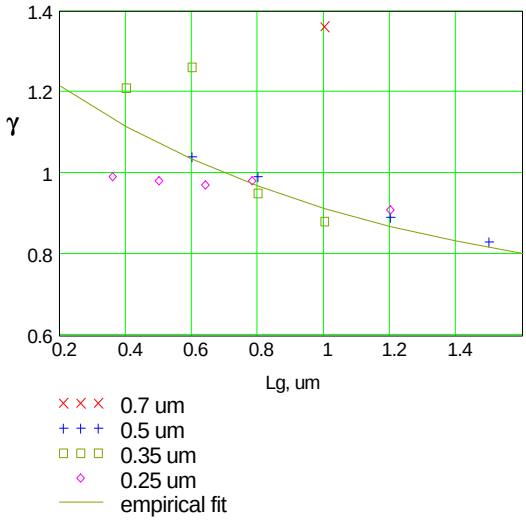

Figure 2: Thermal noise coefficient γ for devices from several commercial submicron processes

Another source of series thermal noise is the parasitic source and drain resistances (R_S, R_D) of the MOSFET. Scaled CMOS requires shallow junctions whose high resistance has been seen as a possible cause of increased noise. But unrestricted growth of R_S and R_D would be detrimental to digital performance as well. It can be expected that these parasitic resistances will be held to values much less than $1/g_m$ through silicidation and heavy doping of the contacts in the interest of maintaining logic speed; this will assure a minimal contribution to white series noise.

B. 1/f noise

High and variable 1/f noise has always been characteristic of MOSFETs. It is known to be strongly dependent on interface quality and gate processing. In long channel FETS, PMOS devices have 3 – 30 times less 1/f noise than equally-sized NMOS because of their buried channel conduction.

For deep submicron processes, the PMOS will be formed using p+ poly gates and retrograde well doping, causing the inversion layer centroid to be located closer to the Si-SiO_2 interface. The change of the PMOS from a buried channel to a surface channel device is predicted to lead to an increase in K_F to a value near that of NMOS. The relative advantage of using PMOS in 1/f noise-sensitive applications would then disappear. Although evidence for this effect has been reported for 0.25 and 0.18 μm devices measured at very low frequencies [2,14], this is not confirmed by recent measurements on amplifier devices down to the 0.25 μm generation. As shown in Figure 3, the scaled devices exhibit no significant increase in K_F for either NMOS or PMOS. The PMOS retains its 10X advantage over NMOS in these particular processes.

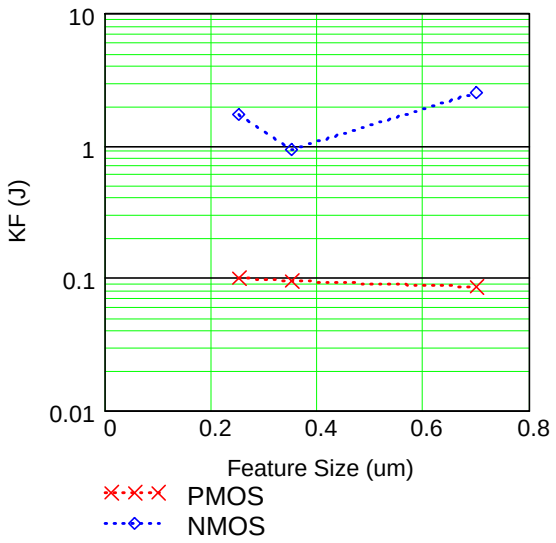

Figure 3 : 1/f noise coefficient K_F [10^{-24} J] versus minimum feature size for NMOS and PMOS transistors from 3 submicron foundries

The shallow junctions required for scaled devices can only be preserved by limiting the thermal budget -- hence gate processes in deep submicron devices will have reduced post-oxidation anneal and higher trap density. For ultrathin gate dielectrics, new materials with higher trap densities than SiO_2 will be used (nitrided, halogenated, H_2 annealed). These alternative methods of gate oxide formation have been shown to increase the 1/f noise by more than one order of magnitude [14].

Hot carrier stress is another mechanism that introduces noise-producing trap states in submicron MOSFETs [14,15]. The scaled MOSFET is engineered to have acceptable degradation of DC characteristics (gm, Vth) over the expected lifetime. However, 1/f noise is found to be far more sensitive to stress than the DC parameters. For example, a device in which gm degraded 10% during a 7-hour over-voltage stress exhibited a 400% increase in low frequency noise[15]. The effect is found to be worse for short channel devices, and depends strongly on bias condition during stress. Submicron MOSFETs experience some level of hot carrier stress during normal operation at the permitted supply voltage. Hence there is a possibility that devices that have been engineered for "acceptable" degradation of DC parameters may experience an unacceptable increase in 1/f noise.

C. Gate current

At a bias of 1.5V, gate current density increases by 10 orders of magnitude as the oxide thickness decreases from 3.6 to 1.5 nm [16]. This corresponds to the expected oxide thickness change from the 0.15μm to the 0.07μm generation shown in Table I. A gate current density of I_G = 1A/cm^2 is considered tolerable for digital circuits based on power dissipation considerations (total gate area per chip ~ 0.1 cm$^{2.}$) At this current density, a typical CSA input FET optimized for low series noise Swould have I_G of the order 10 - 100 nA. A parallel noise of 200 to 700 e- would result in a system with 1μs shaping. To optimize noise, the selection of the device geometry (see Section V) would have to consider the simultaneous minimization of series and parallel noise as in bipolar front ends.

V. ENC AND DYNAMIC RANGE SCALING

For many applications white series noise is the most important source. Its scaling properties are influenced not only by the coefficient γ but also by the device f_T and supply voltage. To find the optimum ENC it is first necessary to determine the optimum size of the input transistor.

A. Dimensioning the input device: generalized capacitive match condition

By Eq. 1, the ENC for series white noise is

$$ENC_s^2 = (C_{det} + C_{ox}WL)^2 \, 4kT\gamma a_1/g_m t_s$$

It is well known that this noise has a minimum with respect to W, since increasing W increases both g_m and the device capacitance. Here we generalize the earlier results taking into

account the behavior of g_m over a wide range of current density. For devices whose f_T is independent of size (such as FETs operated at constant current density) it is well-known that ENC is minimized when Cgs of the input transistor equals Cdet [3,4]. For CMOS in strong inversion under the constraint of constant ID, optimum Cgs is 1/3 of Cdet [17,18]. Figure 4 shows the dependence of g_m on I_D/W, showing three regions in which

$$g_m \sim (I_D/W)^{\alpha}$$

where α = 1, 1/2, and 0 in the weak inversion, strong inversion/square-law and velocity saturated regions respectively. Note that as CMOS scales, the width of the strong inversion region, usually considered the normal operating condition for an analog MOSFET, shrinks until by about the 0.13 μm generation it disappears altogether, as the FET goes from weak inversion directly into velocity saturation.

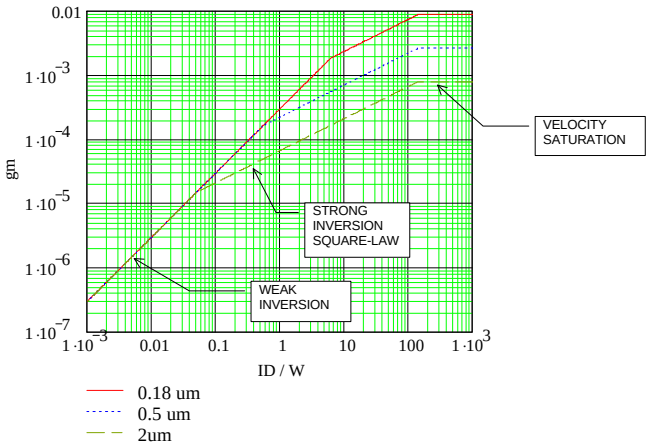

Figure 4: Transconductance as a function of current density for several CMOS generations

From the dependence of g_m on I_D/W and Eq. (1) it is possible to calculate the optimum gate width for any combination of detector capacitance and device drain current. The result depends only on the ratio C_{det}/I_D as follows:

Region I: $\frac{C_{\det}}{I_D} < \frac{6m}{v_{sat}^2}$ (velocity saturated)

$$C_{gs} = C_{det}$$

Region II: $\frac{6L^2}{m(nkT/q)^2} < \frac{C_{\det}}{I_D} < \frac{6m}{v_{sat}^2}$ (strong inversion)

$$C_{gs} = C_{det}/3$$

Region III: $\frac{C_{\det}}{I_D} > \frac{6L^2}{m(nkT/q)^2}$ (W.I. – S.I. boundary)

$$C_{gs} = \frac{2L^2 I_D}{C_{\det}(nkT/q)^2}$$

In the above expressions L is the (minimum) gate length, μ is the carrier mobility, and v_{sat} is the carrier saturation velocity. For high current densities the device is velocity-saturated and the cutoff frequency becomes independent of size; hence the optimum gate width is the one which gives Cgs = Cdet. For moderate current density the input FET is in the strong inversion – square law region and the optimum capacitance is Cdet/3. For lower current densities the device should be sized so that it is operating at the boundary of weak and strong inversion. This is because the transconductance becomes independent of device geometry in weak inversion, so the $g_m/(C_{gs} + C_{det})$ ratio can always be improved by decreasing the device width. The optimum input device capacitance, expressed as a fraction of the detector capacitance, is illustrated in Figure 5.

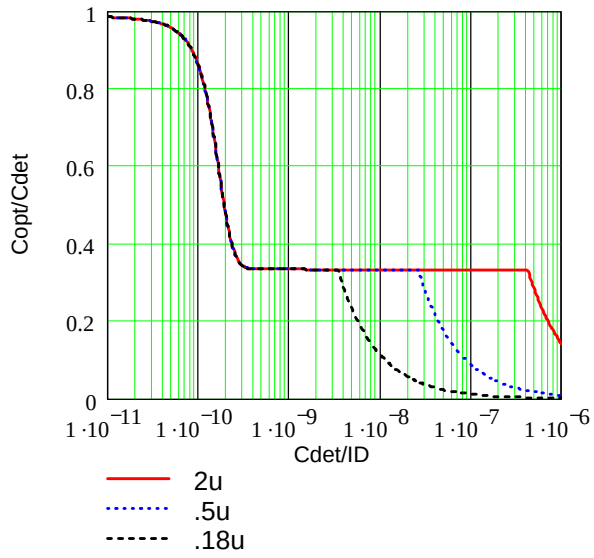

Figure 5: Optimum input capacitance as a function of the parameter C_{det}/I_D for three technology generations

If the excess noise factor γ is dependent on L, then the sizing becomes yet more complex because $L_{opt} \neq L_{min}$. The importance of accurate modeling is emphasized by these examples.

B. ENC scaling

At the capacitive match conditions given above the ENC is minimized. For devices in Region II or III (strong inversion the optimum ENC is

$$ENC^2 = x \cdot kTC_{\det} \cdot \frac{1}{t_m} \cdot L\sqrt{\frac{C_{\det}V_{DD}}{mP}}$$

where ξ is a numerical constant, t_m is the shaping time, and P = V_{DD}/I_D is the power dissipation of the input branch. The parameters that scale in the above expressions are

$$L' \rightarrow \lambda L$$

$$V_{DD}' \rightarrow \lambda V_{DD}$$

where λ is the feature size scaling factor. In Regions II and III therefore

$$ENC' \rightarrow \lambda^{3/4}\, ENC$$

This amounts to a 23% improvement in ENC per generation for the same detector capacitance and power dissipation.

Alternatively, the results of Subsection A can be solved for the power required to achieve a given ENC. In this case we get

$$P' \rightarrow \lambda^{3}\, P$$

and so a 60% decrease in power is expected per generation.

For the velocity saturated case (Region I) the optimum ENC is

$$ENC^2 = x' \cdot kTC_{\det} \cdot \frac{1}{t_m} \cdot \frac{L}{v_{sat}}$$

which scales as

$$ENC' \rightarrow \lambda^{1/2}\, ENC$$

or 16% per generation. In the velocity saturated case the power required to achieve a given ENC is independent of λ.

C. Dynamic range scaling

To calculate the maximum signal that can be processed, consider that the maximum output signal is $<V_{DD}$ and thus the maximum input charge is

$$Q_{in,max} = c\, C_{det}\, V_{DD}$$

where $c = C_F/C_{det} < 0.05$ in practical circuits. In this case the maximum signal to noise ratio can be expressed as

$$SNR = \frac{cm^{1/4} P^{1/4} C_{\det}^{1/4} V_{DD}^{1/4}}{\sqrt{10 a_1 kT / t_m} L^{1/2}}$$

for the case of strong inversion. Substituting scaled quantities for V_{DD}, L we find that

$$SNR' \rightarrow \lambda^{1/4}\, SNR$$

corresponding to a 10% decrease in SNR per generation since the decrease in supply voltage offsets the improvement in ENC found in the previous Section.

To compensate for this decline of SNR, we can

- further reduce the ENC by increasing the power in the front device;
- replace conventional amplifier stages with rail-to-rail stages, especially in the output stage;
- convert single ended signals to differential, recovering a factor of 2 in signal swing. Note that this is usually accompanied by an increase in power and noise.

The first option above requires power to scale as

$$P' \rightarrow \lambda^{-1}\, P,$$

or 43% increase per generation.

VI. Other Effects

A. Off-leakage

The MOSFET off-state leakage current is

$$I_{off} \sim \exp(q(V_G - V_{th})/nkT)$$

and it will increase by a factor of 10 for every 85 mV decrease in threshold voltage, or about 2.3X per generation. Other short-channel effects cause I_{off} to increase even more rapidly in scaled devices. This impacts the design of the feedback element of the preamplifier, which is sometimes required to have an equivalent resistance of hundreds of MΩ. To achieve such high equivalent resistance a much greater L/W ratio is needed, and the use of switches with minimum L must be curtailed.

B. Current source noise

A secondary source of noise is the current source which supplies the input transistor. The noise contribution of the current source is proportional to the ratio $(g_{m2}/g_{m1})^2$ where $g_{m1,2}$ is the transconductance of the (input device, current source). To minimize this ratio, we size the current source transistor to have the maximum L/W ratio while staying out of the linear region; this ratio must decrease as the supply voltage decreases. At the 0.13 μm generation, it will become impossible to make a current source that contributes less than 10% to the overall noise.

C. Crosstalk/coupling

Coupling of digital and high-swing analog nodes back into the sensitive amplifier inputs is already a problem for mixed-signal circuits incorporating CSAs. In future technology generations, there will be competing mechanisms effecting crosstalk and coupling. First, V_{DD} scaling will limit the capacitive coupling by reducing dV/dt transients on the aggressor nodes. At the same time the higher substrate doping will raise the source/drain junction capacitance of the aggressor nodes and also increase the back transconductance gmb of the sensitive circuits, i.e. coupling of devices to the substrate will increase. In the interconnect, the use of low-ε dielectrics will reduce internode coupling and the availability of many (5 – 10) metal layers will allow more effective shielding.

D. Passive Components

The most advanced CMOS processes are optimized for digital circuits and few foundries offer high quality passive components. We presented a new CSA structure that allows the resistor R_F in Fig. 1 to be replaced by a MOSFET in the triode region; a scaled copy of the R_F MOSFET provides pole-zero cancellation of all nonlinearities and has been experimentally verified [19-22]. A further modification to the classic CSA structure is shown in Figure 6. Here both the feedback resistor R_F and the feedback capacitor C_F are replaced by a MOS transistors. The device M1 with source and drain shorted acts is biased in strong inversion and uses the gate to channel capacitance. This nonlinear capacitance is compensated by the series element n*M1. The pole formed by M1-M2 is cancelled by the zero from the series transistors, and all transient nonlinearities are likewise cancelled. The overall circuit injects a current n*Iin into the second stage amplifier. In Figure 6 both amplifiers A1,A2 must have virtual ground inputs at the same DC potential. For the second stage feedback resistor, a line-

arized MOS resistor can be used for current to voltage conversion. The resulting circuit can perform amplification and filtering of low charge signals without relying on high quality passive components on chip.

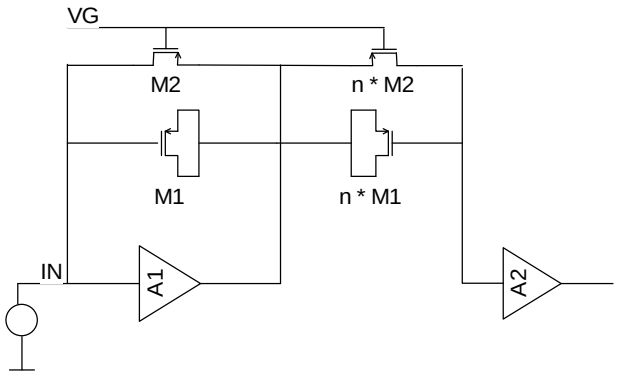

Figure 6. Method of charge amplification using MOSFETs only.

VII. ACKNOWLEDGEMENTS

The authors would like to acknowledge G. Anelli, A. Rivetti, M. Campbell, and W. Snoeys (CERN) for sharing prepublication data on 0.25μm devices; A. Kandasamy, D. Zhao, and J. Triolo (BNL) for helping with the noise measurements in Section IV; and V. Radeka (BNL) for help and encouragement with this project.

VIII. REFERENCES

[1] E. A. Vittoz, "The design of high-performance analog circuits on digital CMOS chips," IEEE J. Solid-State Circuits, vol. SC-20, no. 3, pp. 657-665, 1985

[2] Pelgrom, Marcel J.M., Vertregt, Maarten, "CMOS technology for mixed signal ICs", Nuclear Instruments & Methods A395, 298-305, 1997

[3] V. Radeka, "Low-Noise Techniques in Detectors", Ann. Rev. Nucl. Part. Sci.38,.217-277, 1988

[4] E. Gatti and P. F. Manfredi, "Processing the signals from solid-state detectors in elementary-particle physics," *Il Nuovo Cimento*, vol. 9, pp.1-146, 1986.

[5] Semiconductor Industry Association, *The National Technology Roadmap for Semiconductors*.San Jose, CA: SIA, 1994; 1997

[6] P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*.New York: Wiley, 1984; A. van der Ziel, "Thermal noise in field-effect transistors," *Proc. of the IRE* pp.1808-1812, 1962.

[7] D. Triantos, A. Birbas, "Thermal Noise Modeling for Short-Channel MOSFETs", IEEE Trans, Electron Devices 43, 1950 (1996); D. Triantis, A> Birbas, "Optimal Current for Minimum Thermal Noise Operation of Submicrometer MOS Gransistors", IEEE Trans. Electron Devices 44, 1990 (1997); A. Birbas et. al., "Input Capacitance Scaling Related to Short-Channel Noise Phenomena in MOSFETs", IEEE Trans. Electron Devices 46 (1999) 1253-1257.

[8] P. Klein, "An Analytical Thermal Noise Model of Deep Submicron MOSFETs", IEEE Electron Device Lett. 20 (1999) 399.

[9] R. Jindal, IEEE Trans. ED-33(1986) 1395

[10] A. Abidi, IEEE Trans. ED-33(1986) 1801

[11] G. Anelli et. al., "Radiation Tolerant VLSI Circuits in Standard Deep Submicron CMOS Technologies for the LHC Ecperiments: Practical Design Aspects", Proceedings 1999 NSREC, to be published in IEEE Trans. Nuclear Sciences

[12] A. Rivetti et. al., "Analog Design in Deep Submicron CMOS Processes for LHC", Proceedings 1999 LEB, Snowmass, CO Sept. 25-28, 1999

[13] J.C. Santiard, F. Faccio, "Noise and speed characteristics of test transistors and charge amplifiers designed using a submicron CMOS technology", Nuclear Instruments & Methods A380, 350-352, 1996

[14] E. Simoen, C. Claeys, Sol St Elect. 43 (1999) 865.

[15] E. Sheehan, Microelect. Rel. 38 (1998) 931.

[16] H-S. Wong,"Technology and Device Scaling Considerations for CMOS Imagers", *IEEE Trans. Electron Devices* 43 (1996) 2131-2142.

[17] Chang, Z.Y. and Sansen, W., *Low-Noise Wide-Band Amplifiers in Bipolar and CMOS Technologies*, Kluwer Academic Publishers, 19

[18] W. Sansen, Z.Y. Chang, "Limits of Low Noise Performance of Detector Readout Front Ends in CMOS Technology", IEEE Trans. Circuits and Systems 37 (1990) 1375-1382

[19] P. O'Connor, G. Gramegna, P. Rehak, F. Corsi, and C. Marzocca; CMOS Preamplifier with High Linearity and Ultra Low Noise for X-ray Spectroscopy; BNL 64027 ; IEEE Trans. Nucl. Sci. 44 (3) (1997) 318-325.

[20] G. Gramegna, P. O'Connor, P. Rehak, and S. Hart; CMOS Preamplifier for Low-Capacitance Detectors; BNL 64026 ; Nucl. Instrum. & Meth. A390 (1997) 241-250.

[21] G. Gramegna, P. O'Connor, P. Rehak, and S. Hart; Low-Noise CMOS Preamplifier-Shaper for Silicon Drift Detectors; BNL 64028 ; IEEE Trans. Nucl. Sci. 44 (3) (1997) 385-388

[22] G. De Geronimo and P. O' Connor, "A CMOS detector leakage current self-adaptable continuous reset system: theoretical analysis," *Nuclear Instruments Methods*, vol. A421, pp.322-333, 1999