3 The Front-End Electronics

Abstract. One of the biggest challenges in the construction of a pixel detector is the design of suited readout chips with several thousand electronic channels. These pixel chips contain a regular arrangement of pixel unit cells, each one serving one sensor pixel connected through a bump-bond connection and, in addition, on-chip circuitry by which hit information is transported to the bottom of each chip from where it can be transmitted out. The pixel unit cells must be as small as possible because their area dictates the sensor pixel size and therefore the spatial resolution. While dissipating typically well below 100 μW, every pixel circuit must provide a low noise amplification of the sensor charge, hit discrimination, and a readout architecture adapted to the application. A sufficient speed of the analog chain and the digital section, a well-defined threshold, and possibly a high-radiation hardness must be guaranteed. This chapter discusses the various requirements and presents solutions to the above-mentioned challenges. Various existing readout architectures for experiments in particle physics and biomedicine are described.

3.1 Introduction

The huge number of channels in pixel systems can only be addressed with highly integrated custom-designed electronics circuits. Only the access to suited chip manufacturing technologies has therefore made the development of pixel systems possible. Early designs had to use technologies with gate lengths of $\approx$3 μm so that only a few dozen transistors could fit into pixel cells of 330×330 μm^2 size [191]. With shrinking technology feature sizes, more complicated functions (in particular in the parameter tuning and in the data processing sections) could be integrated into pixels with significantly smaller area. The access to deep submicron technologies (DSM) with feature sizes of 0.25 μm and below has made possible pixel cells of 55×55 μm^2 containing several hundred transistors [193] each. The increasing design possibilities had to satisfy more and more demanding requirements. In the analog part, higher leakage currents and lower signal charges due to radiation damage in the sensor had to be coped with. The analog section had to become significantly faster due to the higher interaction rates in the new experiments (40 MHz at LHC). On the other hand, the power dissipation allowed per pixel decreased because the cooling of electronics and sensor becomes difficult in the low-mass support structures required in particle physics. The digital readout

architectures had to process much higher data volumes and so new concepts had to be implemented.

The first operational chips and systems were developed for experiments in particle physics using silicon as the sensor material [191,194]. In these applications possible hits in the pixels occur at precisely known moments, namely shortly after the collision of a particle beam with a target or with another beam of opposite direction. A trigger signal from other detector components is often available after a fixed time interval, the *latency*, to select interesting events for readout. The generation of the trigger signal typically requires a few microseconds which corresponds to ≈100 interactions at the LHC. Pixel chips for particle physics applications therefore perform an on-chip zero suppression ~~and then~~ buffer the hits until the trigger signal arrives and send only a fraction of the hits to the data acquisition (DAQ). Several different architectures which have been implemented to achieve this goal are presented in Sect. 3.4. Some recent chips [195,196] send out the zero-suppressed hit data immediately as fast as possible so that the pixel information can contribute to the trigger.

Au: Please spell out PET

Pixel detectors are also promising devices for medical applications (X-ray imaging, autoradiography, X-ray tomography, PET; see Sect. 5.3). Several research groups have built chips which count the number of hits in every pixel [165,197,198]. Dual threshold chips with two counters in every pixel [199] allow an energy discrimination for every single hit. This should lead to a contrast enhancement in the image.

3.1.1 Generic Pixel Chip

Although the existing pixel chips use different geometries, readout philosophies, and analog circuits, several building blocks and properties are common to the major part of the designs. As illustrated in Fig. 3.1, the chips can be divided into an *active area* which contains a repetitive matrix of nearly identical rectangular or square pixels and the *chip periphery* from where the active part is controlled, where data is buffered and global functions common to all pixels are located. The wire bond pads for the connection are located only at the lower edge so that the chips can be abutted side by side on a module as indicated in Fig. 3.1. The gap between the chips is made as small as possible and so the size of the underlying sensor pixels must be increased only slightly (see Sect. 5.2.1). The minimum gap size is dictated by the tolerances in chip dicing and by the space required during the flipping procedure.

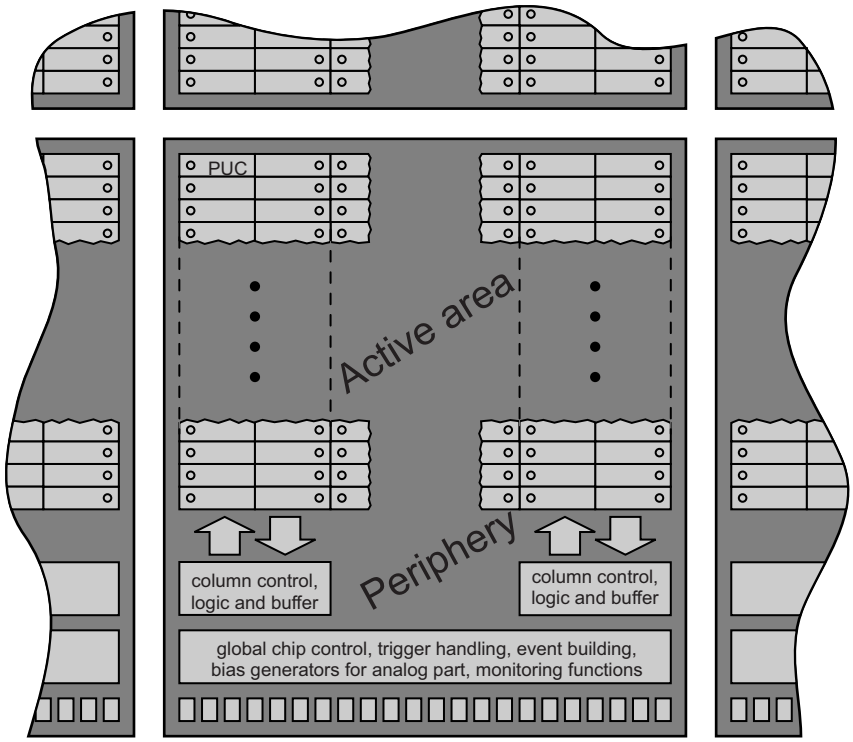


Fig. 3.1. Geometry of a generic pixel chip. The arrangement of several adjacent chips on a module is indicated

3.1.1.1 Active Area

The pixel unit cells (commonly denoted as PUCs) in the active area usually have the same area as the corresponding sensor pixel.[1] They are often grouped in columns; i.e., power, bias, and control signals and the output data flow are routed vertically and only very few signals run horizontally. This column-based approach is advantageous, in particular for rectangular pixels because less area in the PUC is occupied by the bus signals. This is important when only few metal layers are available for routing. Two columns are often grouped together to a "column pair" in order to share circuitry between pixels and to reduce cross talk between the digital and the analog sections. If, for instance, large fractions of the readout section are used by the pixels in two adjacent columns, a mirrored arrangement with the analog parts and the bump pad at the outside and the common readout in the middle of the column pair (see

[1] Size and geometry of the pixels on the sensor and the chip can be different if a routing layer is introduced in between, as for instance in the MCM-D approach [200] described in Sect. 6.2.1

Fig. 3.1) is advantageous. The column-based layout with all signals coming from the bottom of the chip requires no circuitry at the side or above the active area and so the distance between the active area and the chip cut edge can be as small as permitted by the design rules of the technology used (typically 50–100 μm). This is important to allow a denser placement of the chips on the module.

3.1.1.2 Chip Periphery

The bottom part of the chip is usually divided into repetitive blocks to interface to the columns, a global control and bias section, and the wire bond pads. An analog test pulse generator is often included on chip to inject known charges into the pixels. The interfaces to the columns distribute bias signals for the analog sections in the PUCs, they provide buffered digital control signals, and they contain receivers and possibly buffer memory for the data sent down from the pixels. The circuitry to buffer data until reception of a trigger signal can be fairly complicated and space-consuming. The global control part is responsible for the communication with the outside world. In order to save wire bond pads, a serial protocol is very often used for downloading configuration data to the chip and to send hit information from the chip to the DAQ. The configuration data may include the following:

- Bias settings for the analog part which are often generated with on-chip digital-to-analog converters (DACs) as currents or voltages
- A global threshold value and threshold trim values which are written into every PUC
- Mask patterns to disable defective pixels or to switch off complete columns
- Bits to switch between different readout modes, to enable test features, to inject digital test patterns, etc.

The signals going to and coming from the chip which are active during data taking might cross couple into the very sensitive amplifier inputs. Low swing differential signals are therefore often used, so that the number of required wire bond pads is increased. Independent multiple pads are usually used for analog and digital supply voltages to decrease ohmic and inductive losses, and to add redundancy.

3.1.1.3 Examples of Chip Geometries

The size and number of PUCs and the total active area of some existing chips are collected in Table 3.1.[2] The total chip size is a compromise between a good fill factor (requiring large chips) and the production yield (penalizing large chips). The more recent chips use technologies with a gate length of

[2] According to [201], the 3 μm self-aligned CMOS (SACMOS) process used for several designs has the density of a typical 1.5 μm process

Table 3.1. Geometrical properties of some pixel chips

Experiment or name of chip	PUC (μm^2)	Colums × rows	Active area (mm^2)	Technology (μm)	Ref.
OMEGA1	200 × 200	12 × 9	2.4 × 1.8	3 (SACMOS)	[26]
OMEGA2	500 × 75	16 × 64	8 × 4.8	3 (SACMOS)	[201]
DELPHI	330 × 330	24 × 24	7.9 × 7.9	3 (SACMOS)	[191]
LHC1/OMEGA3	500 × 50	16 × 128	8 × 6.4	1 (SACMOS)	[202]
ATLAS	400 × 50	18 × 160	7.2 × 8	0.25	[203]
CMS 0.8 μm	150 × 150	52 × 53	7.8 × 7.9	0.8	[204]
CMS 0.25 μm	150 × 100	52 × 80	7.8 × 8	0.25	[205]
BTeV, FPIX1	400 × 50	18 × 160	7.2 × 8	0.5	[195]
ALICE/LHCB	425 × 50	32 × 256	13.6 × 12.8	0.25	[206]
MPEC1.0	433 × 50	12 × 63	5.2 × 3.1	0.8	[165]
MPEC2.1	200 × 200	32 × 32	6.4 × 6.4	0.8	[207]
XPAD	330 × 330	24 × 25	7.9 × 8.2	0.8	[208]
PILATUS	217 × 217	48 × 85	9.6 × 17	0.8	[198]
MEDIPIX1	170 × 170	64 × 64	10.9 × 10.9	1 (SACMOS)	[197]
MEDIPIX2	55 × 55	256 × 256	14.1 × 14.1	0.25	[193]

0.25 μm and up to six metal layers for routing so that the pixels can be made smaller. These technologies should still provide a good yield for chips with an area of more than 1 cm^2, which is a common size for commercial chips fabricated in such technologies. Figure 3.2 shows a photograph of a single pixel chip mounted on a printed circuit board for testing.

3.1.2 Simple Sensor Model

A sensor pixel can be modeled by the capacitances to the backside (C_{back}) and to the first (or more) neighbors (C_{interpix}) as illustrated in Fig. 3.3a (see also Sect. 2.3.4). Assuming that the neighbor pixels are held at constant potential by the connected amplifiers (which is not perfectly true), the capacitances can be summed up leading to the effective detector capacitance C_{det}, which is a crucial quantity for circuit analysis. Cdet is the central element of the simple sensor circuit model shown in Fig. 3.3b. A constant current source is added to model the fraction of the sensor leakage current flowing into the central pixel. Electron–hole pairs created in the sensor volume are drifting to the electrodes under the influence of the electric field generated by the

AU: Capacitances to the backside is denoted by C_{back} in the text and by C_{backside}in the artwork. Kindly check this discrepancy and modify as appropriate.

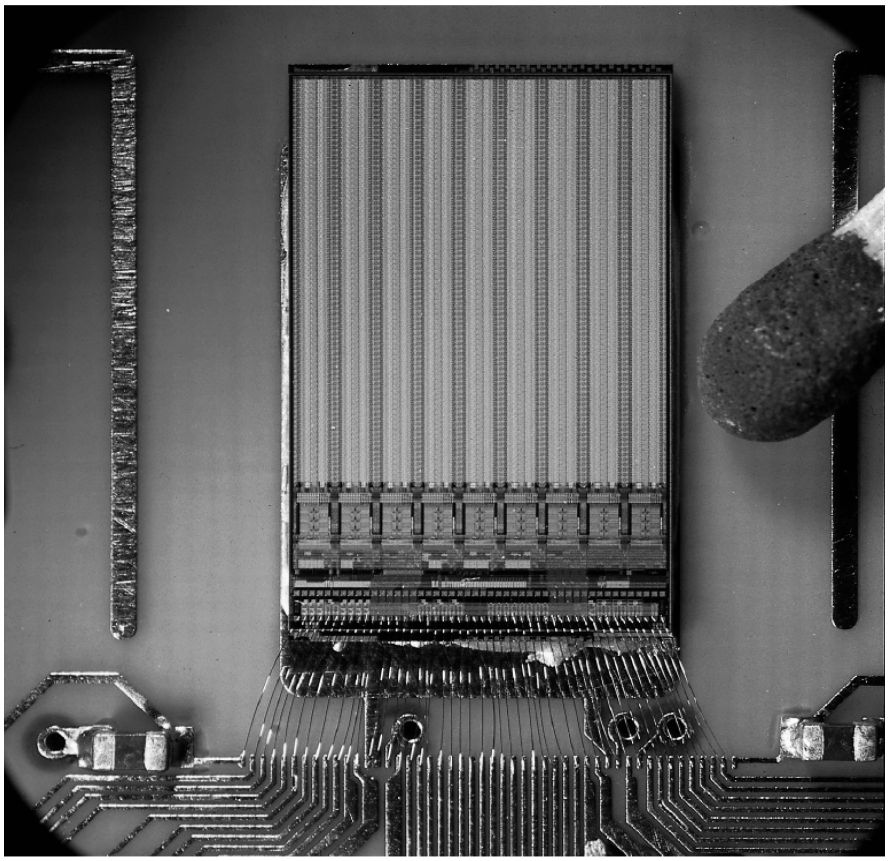


Fig. 3.2. Photograph of a pixel chip with 18×160 pixels. The regular active area in a column-based structure, the periphery with data buffers at the bottom of column pairs, and the global control circuitry with the wire bond pads at the bottom can be clearly distinguished. A match is shown for size comparison

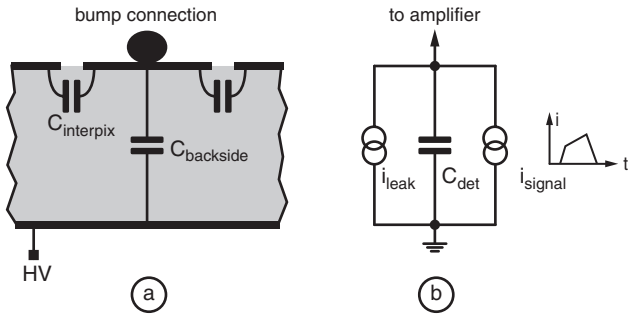


Fig. 3.3. Capacitances in a (**a**) sensor and (**b**) simple equivalent circuit with leakage current source and signal source

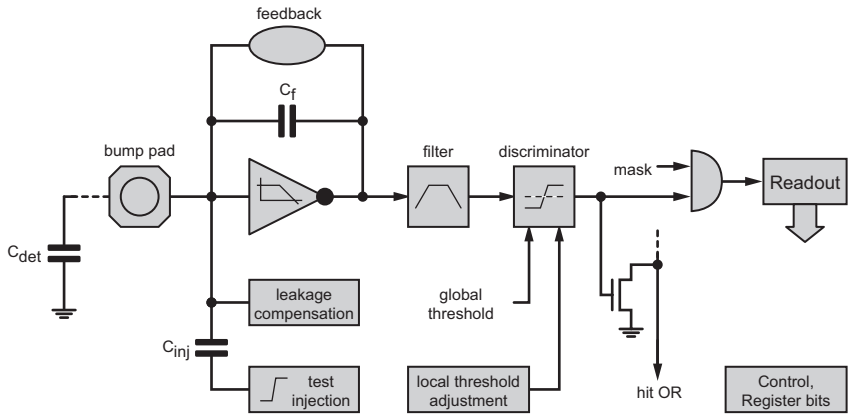


Fig. 3.4. Components of a generic PUC

bias voltage. The signal induced on the pixel and its neighbors already *during* the charge motion is modeled by a time-dependent current source. The exact temporal shape of this current signal depends on many factors like the position of the charge deposition, the sensor material properties (mobilities, trapping), the bias voltage, and the pixel geometry (small pixel effect; see Sect. 2.2.3). A total drift time of ≈10 ns is often used to model a 300-μm-thick silicon sensors in which a minimum ionizing particle (see Sect. 2.2.2.1) deposits a total charge of ≈4 fC. The polarity of the signal is determined by the type of charge carriers collected on the pixel. For sensor materials with a diode junction like silicon, the voltages must be chosen such that the diode is reversely biased. For other "ohmic" materials like CdTe or diamond, negative input signals are more common due to the higher mobility of electrons. This requires a ~~positive~~ backside bias voltage with respect to the pixels. Typical pixels have capacitances in the order of 100 fF and leakage currents of the order of 10 pA before irradiation.

3.1.3 Generic PUC

Several circuit elements are nearly always present in the elementary PUCs in the active part of the chip. The common circuit blocks shown in Fig. 3.4 are therefore briefly discussed in the following sections. Figure 3.5 shows part of the layout of a PUC as an example.

3.1.3.1 The Bump Pad

A small square or octagonal bump pad is used for the connection to the sensor pixel. The size of the opening in the chip passivation layer depends on the available space and on the bump technology used and ranges from

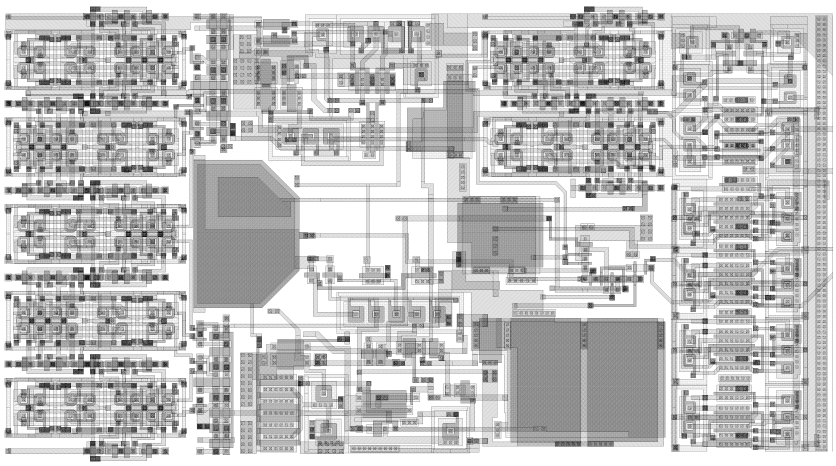


Fig. 3.5. Layout of analog part and control section of the PUC of the ATLAS FEI Chip in 0.25 μm technology. The height of the shown part of the PUC is 50 μm, and the width is ≈100 μm. Only the lowermost two metal layers, the polysilicon layer, and contacts are shown for clarity

12 μm (ATLAS) to 50 μm (MPEC2). The passivation layer must not be too thick for some bump-bonding techniques. The bump pad being connected to the input of the amplifier is a very sensitive node and care must be taken in the layout to shield it toward the underlying circuitry and the substrate. An elegant possibility is to use the capacitance between the pad and the metal shield underneath to implement the feedback and the injection capacitors.

3.1.3.2 The Charge-Sensitive Preamplifier

An inverting amplifier with a feedback capacitor C_f converts an input charge Q_in to a voltage. An infinite gain would keep the input at a perfect virtual ground and the output voltage step in this ideal case is $\Delta U_\mathrm{out} = -Q_\mathrm{in}/C_\mathrm{f}$. If the gain $-v_0$ is finite, however, a small residual voltage

$$\Delta U_\mathrm{in} = \frac{Q_\mathrm{in}}{C_\mathrm{in} + (1 + v_0)\, C_\mathrm{f}} \tag{3.1}$$

remains at the input. The effective input capacitance

$$C_\mathrm{eff} = (1 + v_0)\, C_\mathrm{f} \approx v_0 C_\mathrm{f} \tag{3.2}$$

of the charge-sensitive configuration acts in parallel to C_in, which is the sum of the detector capacitance C_det, the preamplifier input capacitance C_amp, and parasitic contributions C_stray. The output voltage step for an amplifier with a finite gain $-v_0$ hence is

$$\Delta U_{\text{out}} = -\frac{Q_{\text{in}}}{C_{\text{f}}} \times \frac{1}{1 + \frac{1}{v_0} + \frac{C_{\text{in}}}{v_0 C_{\text{f}}}} \tag{3.3}$$

and so a significant fraction of the amplitude is lost when C_{in} approaches $v_0 C_{\text{f}}$. It is therefore usually required that the effective input capacitance be significantly larger than the detector capacitance, i.e.

$$C_{\text{eff}} \approx v_0 C_{\text{f}} \gg C_{\text{in}} = C_{\text{det}} + C_{\text{amp}} + C_{\text{stray}}. \tag{3.4}$$

This condition also keeps the voltage step at the input ΔU_{in} small so that cross coupling to neighboring pixels via the interpixel capacitance C_{interpix} is reduced.

The rise time of the preamplifier output signal for an instantaneous input charge depends on its gain–bandwidth product and on the closed loop gain, which is roughly set by the ratio of the detector and the feedback capacitor. A large detector capacitance or a small feedback capacitance leads to a slower rise time. The rise time is of course limited by the duration and the exact shape of the signal induced during charge collection, in particular if slow sensor materials or low bias voltages are used.

The choice of the value of C_{f} is therefore a compromise between charge gain, input impedance, speed, stability, and matching (gain uniformity between channels). Values of 4–30 fF are used in most pixel front-end chips.[3] A typical inverting gain of a few hundred leads to an effective input capacitance of a few picoFarads which is an order of magnitude larger than typical sensor capacitances.

The preamplifier is one of the most crucial parts in the PUC. It must provide an inverting gain of well above 100 with a sufficient bandwidth. Its power consumption must be kept very low in most applications (typically 50 μW in particle physics) in order to limit the heat dissipated in the active area. As the noise of the preamplifier is a crucial factor for the performance of the PUC, both the frequency-independent white noise as well as the $1/f$-noise must be minimized for a given detector capacitance and shaper characteristics. The input transistor is usually the dominant noise contribution and so its type (NMOS, PMOS, bipolar transistor), its size, and the biasing condition must be chosen carefully (see Sect. 3.3.6). Another design goal is a good immunity of the amplifier to fluctuations in the supply voltage which could be generated by changes in chip activity and cross coupling from the digital to the analog part. A high power supply rejection ratio in the frequency range of interest is therefore desirable. It can be achieved, for instance, using differential topologies.

[3] Chips for the readout of strip detectors must use much higher values of C_{f} because of the larger detector capacitance

3.1.3.3 The Feedback Circuit

A feedback circuit is required to define the DC-operation point of the charge-sensitive preamplifier and to remove signal charges from the input node (or from C_f after the dynamic response of the amplifier) so that the preamplifier output voltage returns to its initial value. The discharge should be slow if further filtering is used as shown in Fig. 3.4, otherwise the pulse shape after the filter is degraded by the falling edge of the preamplifier output signal (see Fig. 3.7). The discharge must be fast enough, on the other hand, to avoid saturation or nonlinearities of the preamplifier at the maximum allowed hit rate. If the discriminator is DC-coupled to the preamplifier output, a "pileup" of pulses leads to unwanted threshold shifts.

The discharge of a feedback capacitor of $C_\mathrm{f} = 10$ fF in $\tau = 1$ μs with a resistive element requires the very large resistance of $R_\mathrm{f} = \tau/C_\mathrm{f} = 100$ MΩ which is difficult to implement on a chip. MOS transistors operated in the linear region for instance, which are commonly used in chips for the readout of strip detectors, cannot be used easily because the drain saturation voltage becomes very small at the required very low gate voltages. Possible solutions to this problem are presented in Sect. 3.3.2.

In a circuit where no separate filter is present and the preamplifier output is directly used for hit discrimination, the discharge must be completed before the next signal arrives. A fast discharge is required in this case. This can lead to a reduction in the peak amplitude if the discharge starts before the signal has reached its maximum due to limited rise time of the amplifier. This "shaping loss" is illustrated in Fig. 3.6a, b.

Several chips use a constant current discharge which can be implemented very easily. This leads to saw-tooth-shaped output signals (Fig. 3.6b, d) with a width proportional to the input charge. This property can be used to determine the deposited charge by measuring the width of the discriminator output pulse, the "time over threshold" (ToT). A constant current feedback is less suited for high-rate applications due to the increasing dead time for large signal amplitudes.

In some rare cases large charges deposited in the sensor by heavy particles, curled tracks, or pulse height fluctuations can saturate the amplifier for a relatively long time, in particular if constant current feedback is used. Some designs therefore introduce an additional nonlinear element (e.g. a diode) in the feedback which provides a high discharge current for anomalously high output voltages.

3.1.3.4 The Leakage Compensation Circuit

The sensor pixels are usually DC-coupled to the preamplifier inputs (this eliminates the need for biasing structures and AC-coupling capacitors on every sensor pixel) so that the leakage current I_leak of up to several 10 nA after irradiation must be sunk/sourced by the pixel circuit. Without any

Does "several 10 nA" mean "several tens of nanoamperes" ar "some 10 nA"? check thro'out

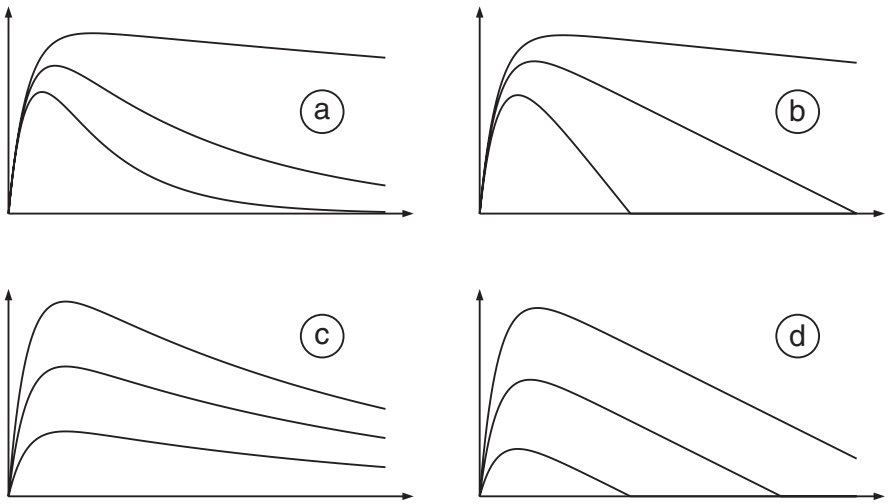


Fig. 3.6. Preamplifier signals (amplitude vs. time) for resistive feedback *(left)* and for constant current feedback *(right)*. The upper graphs (**a**) and (**b**) show a variation in the feedback time constant, and the lower graphs (**c**) and (**d**) show different input charges

further measure, this current would have to flow through the feedback circuit. It would introduce a DC-offset at the preamplifier output of $R_{\mathrm{F}} I_{\mathrm{leak}}$ for the case of a purely resistive feedback. For $I_{\mathrm{leak}} = 10\,\mathrm{nA}$ and $R_{\mathrm{F}} = 100\,\mathrm{M\Omega}$ this leads to $\Delta V \approx 1\,\mathrm{V}$, a value which would bring typical circuits out of their operation regime. Even much smaller offsets would lead to significant threshold shifts in schemes where the discriminator is DC-coupled to the preamplifier.

A compensation circuit to sink all or a significant fraction of the leakage current is therefore often implemented. A simple solution is the subtraction of a fixed current which is determined, for instance, by measuring the leakage in one or more "dummy" pixels [194]. More sophisticated designs use a current source in every pixel which is regulated such that the preamplifier output reaches a certain average DC-level [209]. Because signal current and leakage current can only be distinguished by their time structure, a very long time constant is involved in this solution. A different, very simple solution is provided by a constant current feedback which is able, due to its nonlinear nature, to provide leakage compensation and the discharge of the feedback capacitor simultaneously [210]. Various feedback circuits are presented in Sect. 3.3.2.

Most leakage compensation circuits can only sink (or source) current and so the polarity of the leakage current, and hence of the signal, is fixed. These pixel chips can then be used for one polarity type of readout only.

3.1.3.5 The Shaper

A band-pass filter (commonly referred to as *shaper*) is often included to explicitly limit the bandwidth of the preamplifier output signal. This is beneficial for a reduction of high- and low-frequency noise contributions introduced in particular by the sensor leakage current and by the input device. Special filter functions can be used to achieve ultimate noise performance for a particular noise spectrum. These filters are difficult to implement with few device components, however, and so a sequence of N simple high-pass (RC) and M low-pass (CR) stages is often used. The noise reduction by such CR^N–RC^M-shapers is discussed in Sect. 3.3.6.

In the time domain, higher order filters lead to shorter pulses for a given peaking time (see Fig. 3.30 in Sect. 3.3.6.1). This makes them useful for high-rate applications where the baseline must be restored as quickly as possible in order to be ready for the next signal pulse. The shaper outputs a unipolar signal only when its input is a step function (with a finite rise time). The preamplifier output, however, has to return to its baseline after some time in order to avoid saturation. This signal decay leads to an undershoot in the shaper response as illustrated in Fig. 3.7c, d. This problem can be addressed by a technique often referred to as "pole-zero cancellation".

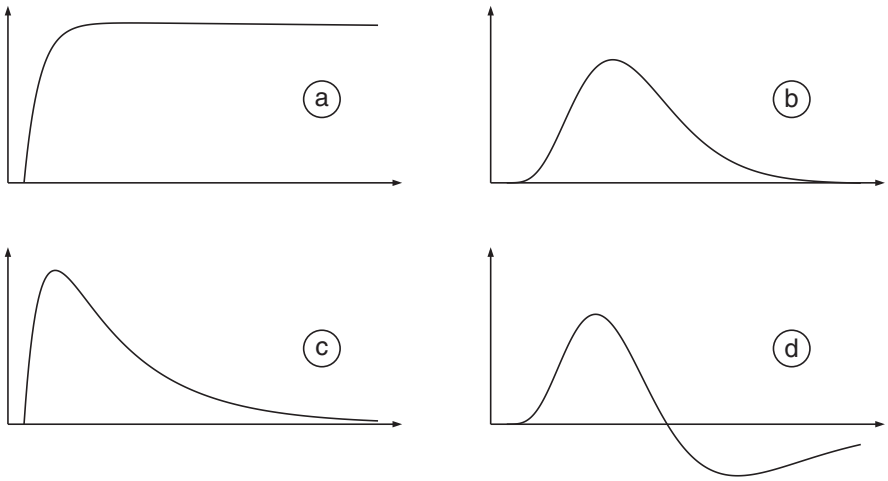


Fig. 3.7. Pulse shapes (amplitude vs. time) after a preamplifier with finite rise time *(left)* and after a CR–RC^4-shaper *(right)*. For a slow discharge of the feedback capacitor (**a**), the corresponding shaper output (**b**) is nearly unipolar. For fast discharge (**c**), however, the shaper output (**d**) has an undershoot

3.1.3.6 The Discriminator and Threshold Trim

Hits with a sufficiently large input charge are detected by a discriminator which compares the shaper output to a threshold value which is distributed globally to all pixels. The threshold is set as low as possible in order to maximize the detection efficiency but not too low, on the other hand, to keep the rate of noise hits at an acceptable level. Variations of the threshold of the individual pixels caused by transistor mismatch, voltage drops or preamplifier gain variations can lead to an increased noise hit rate or to a reduced sensitivity. It is therefore common practice to include a local threshold fine-tuning in every pixel to compensate for these variations. A well-defined threshold is particularly important in applications where the discriminator picks out hits from a continuous amplitude spectrum. In chips for X-ray detection for instance, two discriminators with different thresholds are used to distinguish between high- and low-energetic X-rays [199]. The possible contrast enhancement depends very much on the precision of the amplitude cuts provided by the discriminators and so a precise threshold trimming is required.

The local threshold adjustment is most often implemented by simple (DACs) with 3–70-bit resolution [196,211]. This solution requires several digital registers in every pixel to store the trim values. A more compact solution is the dynamic storage of a voltage on a capacitor [207]. This allows a continuous trimming over a wide range with high precision at the expense of the need for regular refresh operations.

The response time of the discriminator (in combination with the rise time of the preamplifier) is crucial in applications where the arrival time of a signal must be detected with high precision. This is for instance the case in particle physics experiments at the LHC where particle interactions occur every 25 ns. The "time walk" curve in Fig. 3.8 shows that hits with high amplitudes lead to a fast response. The discriminator needs much longer, however, to detect hits with amplitudes just above the threshold Q_{Thr}. A $\Delta T < 25\,\mathrm{ns}$ wide time window (some jitter must be allowed for other system components)

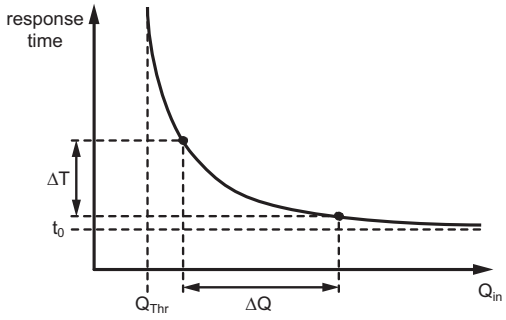


Fig. 3.8. Response time of a discriminator as a function of the input charge

can be selected on this curve by adjusting external system delays. This fixes an interval ΔQ of charges which are detected "in-time." The lowest amplitude which can still be associated with the correct interaction (the "in-time threshold") is higher than the threshold of the discriminator.

3.1.3.7 Test Charge Injection

In order to verify the correct operation of the PUCs already on the wafer level (when no sensor is connected), the controlled injection of known charges into the preamplifier inputs is crucial. This is easily accomplished by applying a known voltage step to a well-defined calibration capacitor C_{inj}. The relative matching of medium-size metal–metal capacitors with a capacitance of a few femtoFarads across a chip can be on the percent level so that the charge error can be made small. The voltage step can be supplied externally to the chip or generated in a "chopper" circuit in the chip periphery. It must be distributed to all pixels with a low-resistive bus so that the rise time degradation due to RC-effects is small. In order to inject only into a selected set of pixels (most readout concepts cannot cope with too many simultaneous hits) a switch in the PUC can disconnect the injection capacitor from the injection bus. The switch should be designed carefully because the parasitic drain–source capacitance in the open state can still lead to spurious charge injection. An alternative to a global voltage step (requiring a fairly powerful chopper circuit) is the local generation of the step in every pixel starting from static voltage levels distributed to the PUCs.

A drawback of the capacitive injection is the difficulty to generate several consecutive charge signals. This would require staircase signals which are more difficult to generate and which are limited in amplitude. Another approach is therefore to steer a known current I_{inj} into the preamplifier input during a known time interval T_{inj}. The injected charge is $Q_{\mathrm{inj}} = I_{\mathrm{inj}} T_{\mathrm{inj}}$, and so a current of 400 nA is required to accumulate a charge of 4 fC in a time interval of 10 ns. The injection can be accomplished easily with a simple differential pair. The transistors used to generate I_{inj} must be sized carefully to guarantee an adequate matching between the different channels.

3.1.3.8 Control and Test Circuitry

Most chips generate a fast hit-OR from the PUCs in one column. This signal can be used to test the analog part independently from the readout or to provide a fast timing signal. Some column readouts are started when a hit in the column is flagged by the hit-OR. Chips requiring a trigger signal can use the hit-OR to automatically produce a trigger whenever a signal is detected somewhere on the chip. This "self-triggered" mode is very useful for module testing with a radioactive source where the events occur at unknown times.

Depending on the area available in the layout, various control bits are used to disable ("mask") the readout of individual PUCs (for instance because a

broken front-end or a bad sensor pixel generates an excessive noise hit rate), to disable the fast OR, to turn on a digital test mode, or to switch off the preamplifier completely.

The control bits, the bit for test injection, and those for the threshold trim circuitry are stored in static registers in the pixels. They must be loaded with data during the chip initialization phase. x–y-Decoding schemes or shift registers to address the pixels have been used. A read back feature is useful to increase the testability. As with all registers on the chip, single event upset (i.e. the flipping of bits due to charges deposited on the storage node by ionizing particles) must be addressed for instance with the DICE cell discussed in Sect. 3.2.2.

3.1.3.9 The Readout

Various concepts for the further processing of the hit information have been proposed. The details of the readout architecture depend strongly on the target application. Two important classes are:

- Chips which buffer all incoming hits for a short time interval until a trigger signal selects a subset for readout and
- Chips which count the number of hits in every pixel

They are discussed in some detail in Sect. 3.4.

3.1.3.10 Extensions to the PUC

The existing pixel chip implementations add application-specific blocks to the generic PUC. These can be a second discriminator to sort hits according to their amplitude [193, 199], multiple discriminators to implement a low resolution ADC [195], an individual gain correction [211] or an analog sample and hold for a high-resolution readout of the signal amplitude [204].

3.1.4 Module Controller Chips

The hit information of the individual front-end chips must be sent to the data acquisition (DAQ) for further processing and storage. Serial links carrying digital or analog [196] signals are normally used in order to reduce the number of cables. If the bandwidth of the links is not exhausted by the data rate of a single front-end chip, the number of cables can be further decreased by merging the hit information already on the pixel module. This task is often accomplished by a separate "module control chip," the MCC [212]. As illustrated in Fig. 3.9, the MCC receives hit information from several pixel chips. A star topology is often preferred for its higher bandwidth and fault tolerance. The MCC buffers the hits until all pixel chips have delivered the data belonging to a given event and outputs a compact data block to the

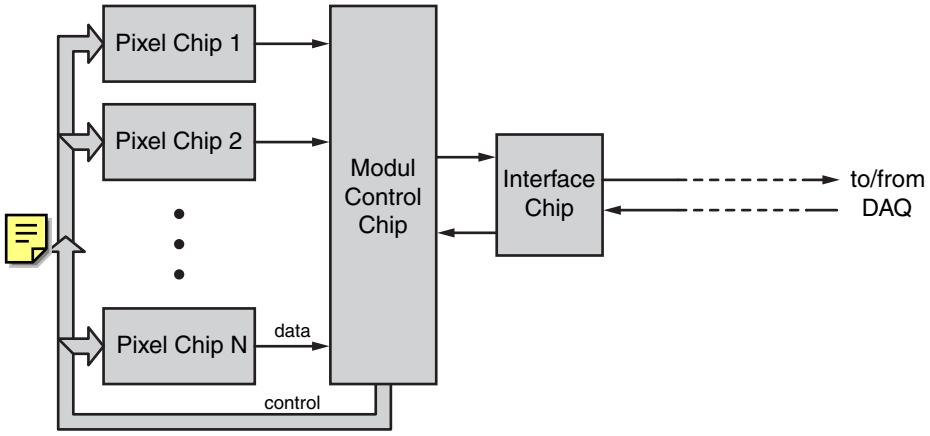


Fig. 3.9. Typical components of a pixel module with an MCC. The MCC receives data from the front-end chips, merges them, and sends them to the data acquisition. It is also used to configure the pixel chips and to provide timing and trigger information. Interface chips can be used, for instance, to serve an optical link

DAQ. It must provide means to handle various error conditions like defect pixel chips, buffer overflows, bad check sums, etc. The MCC can also be used to distribute timing and control signals and configuration data to the pixel chips. The token bit manager chip [213] of the CMS pixel system provides, among other functions, a start signal for the daisy chain readout of the front-tend chip.

The short local connections between pixel chips and MCC can use single-ended CMOS, differential CMOS, or differential low swing signals (LVDS) depending on how fast signals are and on how sensitive the environment is to cross coupling. LVDS is often preferred at the expense of twice the number of signal traces and more power dissipation in the LVDS drivers due to the required termination resistors. The long distance between the module and the DAQ often uses optical links which provide high speed, low cross talk and no risk of ground loops. The steering of the laser diode or LED for the outgoing data and the detection of the incoming timing and control signals are sometimes done with separate interface chips.

The architecture of an MCC is very experiment-specific. MCCs are therefore not further discussed in this book.

3.2 Design Aspects

This chapter discusses various aspects of the design of pixel chips. Section 3.2.1 lists some typical specifications which guide the design. The radiation hardness of the circuits (i.e. the long-term stability when ionizing

and nonionizing particles irradiate the device) and the ability to cope with local charge depositions which are large enough to flip storage nodes (single event upset, "SEU") are discussed in Sect. 3.2.2. The digital activity on the chip can easily inject spurious signals into the analog section. This cross talk can significantly degrade the analog performance (noise, lowest possible threshold). It is addressed in Sect. 3.2.3.

3.2.1 Typical Specifications

Several partially conflicting requirements must be met by the readout electronics. Most of them concern the analog pixel part which consists of a low-noise charge-sensitive preamplifier, a discriminator with adjustable threshold (two discriminators in a dual discriminator scheme) and very often a test charge injection circuit. The most important requirements are discussed in this section.

3.2.1.1 Pixel Size and Geometry, and Spatial Resolution

A small pixel size is the prime requirement to achieve a high spatial resolution. While the effective pixel *pitch* must be identical for sensor and electronics in both the x- and the y-direction in order to achieve a one-to-one connectivity through the bump bonds, the exact geometries can be different. It is for instance possible to connect an array of hexagonal sensor pixels to a chip with rectangular pixels with a side ratio of $x/y = 2\sqrt{3}/3 \approx 1.155$ as shown in Fig. 3.10. While rectangular sensor pixels seem to be the most natural choice, a hexagonal geometry has the advantage that at most three pixels touch in the corners. The deposited signal charge is therefore shared between three pixels only and so more charge is left per pixel as compared to rectangles. The same advantage is obtained if square pixels are "bricked"; i.e., every second row is shifted by half a pitch (see Sect. 2.3.6). It can also be

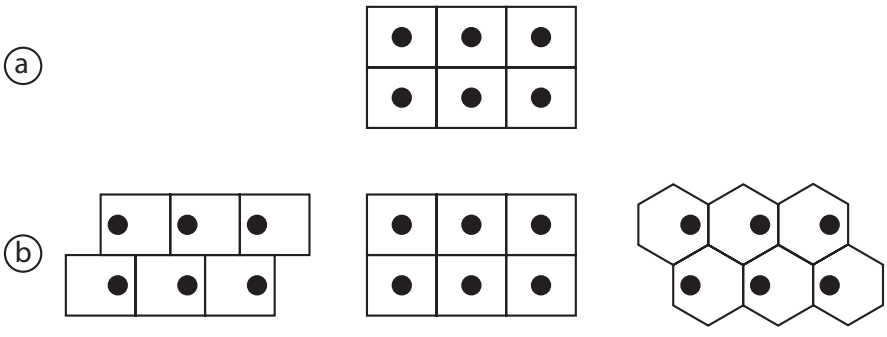


Fig. 3.10. A readout chip with rectangular pixels with side ratio $x/y = 2\sqrt{3}/3$ (**a**) can be mated with a sensor with bricked, rectangular, or hexagonal pixels (**b**)

shown easily that the theoretical resolution of such geometries with a purely binary readout is slightly better than for chessboard square pixels. Hexagonal pixels have probably not been used widely so far because of the slightly more complicated implementation and coordinate reconstruction. Table 3.1 lists the geometries of some existing pixel readout chips. Rectangular pixels are used in applications where asymmetric resolutions in x and y are required, for instance in particle physics where tracks are bent in a magnetic field. The measurement of the curvature of the tracks in the plane perpendicular to the field is particularly important. From the chip layout point of view rectangular pixels simplify the distribution of signals if the busses run in the "short" direction. If technologies with few metal layers are used, this can be a significant advantage over square pixels of the same area. The rectangular geometry has the drawback that the capacitance between adjacent pixels is relatively large which can lead to a noise increase and to increased cross talk between pixels.

The exact bump pad location within the pixel is sometimes chosen differently for odd and even rows or columns, the advantage being for instance that the analog sections of adjacent pixels are located next to each other on the readout chip (see Fig. 3.1). Power and control signals can be shared between the pixels and the cross coupling between analog and digital parts is reduced. This can be important if few routing and shielding possibilities are offered by the technology.

The spatial resolution for point-like hits (i.e. when the charge clouds by electron and hole diffusion in the sensor are much smaller than the pixel pitch) is $a/\sqrt{12}$, the rms of a box distribution of width a, a being the pixel pitch (see Sect. 2.3.6). Hits close to the pixel edge have the largest errors in this case. When charge diffusion spreads the hits over two adjacent pixels, the "double hits" occurring at the pixel edges have small reconstruction errors. The resolution is further increased if a pulse height information, i.e. the amount of charge deposited in the pixels of a hit cluster, is available. This general behavior is regularly observed in test beam measurements where the position reconstructed from the pixel data is compared to the true impact position measured by a high-resolution reference detector, for instance a silicon strip detector. Examples of such measurements are presented in Sect. 2.3.6.

3.2.1.2 Threshold and Threshold Variations

The thresholds in the many pixel elements on one chip are not perfectly identical. They exhibit random variations due to component mismatch (production fluctuations of doping concentrations, oxide thickness, geometrical size, etc.) and possibly systematic fluctuations due to voltage drops along the column or component mismatch from mirrored layouts. Furthermore, the effective threshold varies as a function of the capacitance connected to the preamplifier input. This is due to the finite gain of the preamplifier which leads to an amplitude loss for increasing input capacitance (see (3.3)). This

situation typically occurs at the side and at the top of the chip where the sensor pixels are made larger to fill the area between adjacent chips (refer to Sect. 5.2.1).

The local threshold trim possibility described in Sect. 3.1.3 has two main goals. On one hand, few pixels with particularly low or high thresholds should be brought close to the nominal value. This requires a large trim step size in order to extend the reachable range. On the other hand, the width of the threshold distribution of the majority of the pixels should be narrowed. This requires a small trim step. The optimal trim step is a compromise between these conflicting goals. It is therefore very much desirable to be able to adjust the trim step size dynamically (at least on the chip level) during the threshold trimming procedure.

The thresholds have an rms spread of σ_{thr} after trimming. Pixels with particularly low thresholds are susceptible to noise hits, while pixels with high thresholds may be less efficient. As an example for the estimation of the minimum required threshold, a pixel system at LHC with 250-µm-thick silicon sensors is considered. The signal charge collected from a minimum ionizing particle after several years of operation and corresponding irradiation degrades to $\approx 12,000$ e^- due to a reduction of the depletion depth and to charge trapping (see Sect. 2.4.1). This charge is often spread over two adjacent pixels so that only $\approx 6,000$ e^- may be seen per pixel. Taking into account statistical fluctuations of the signal charge (see Sect. 2.2.2), a threshold of $Q_{\mathrm{thr}} \leq 3,000$ e^- is typically required for good efficiency after degradation of the sensor due to irradiation.

The lower limit for the average threshold is set by the fluctuations σ_{thr} and by the pixel noise σ_{noise}. The noise hit rate is kept negligible if

$$Q_{\mathrm{thr}} \gtrsim 5 - 6 \times \sqrt{\sigma_{\mathrm{thr}}^2 + \sigma_{\mathrm{noise}}^2} \tag{3.5}$$

The threshold dispersion should therefore be comparable to the noise. A typical conservative design goal is $\sigma_{\mathrm{thr}} \approx \sigma_{\mathrm{noise}} \approx 200$ e^-.

A specification of the threshold dispersion after trimming can be necessary, for instance, in systems which select certain signal amplitudes (see Sect. 3.4.3), but very often it is only required that the noise hit rate be kept low while setting the lowest feasible threshold. This goal could be achieved with a threshold trimming mechanism which observes the noise hit rate and increases the threshold until the noise hit rate is sufficiently low. This approach would lead to minimal thresholds albeit at unknown absolute values.

3.2.1.3 Noise

The noise of a channel can be defined as the root mean square (rms) of the voltage fluctuation at the end of the analog processing chain divided by the gain (in volts per coulomb). The *equivalent noise charge* at the input,

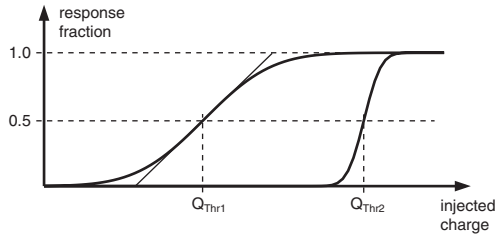


Fig. 3.11. Response of a discriminator (fraction of injections of a given charge firing the discriminator) for a channel with a high noise at a low threshold and for a channel with a low noise at a higher threshold

$$\text{ENC} := \frac{\text{noise output voltage (rms)}}{\text{signal output voltage for an input charge of one electron}}\,, \quad (3.6)$$

is commonly quoted in units of electrons as the figure of merit. In binary systems, the analog signal cannot be measured directly and so the ENC is determined from the response of the discriminator to multiple charge injections with increasing amplitude. Figure 3.11 illustrates this kind of measurement for noise with an assumed Gaussian distribution. The threshold is defined as the charge where 50% of the injections fire the discriminator. The slope s at this point can be used to determine the noise:

$$\text{ENC} = \frac{1}{\sqrt{2\pi}}\frac{1}{s}. \quad (3.7)$$

For quick estimations one can determine the interval from ≈5 to ≈95% which corresponds to ≈3.29ENC.

The noise is mainly determined by the total input capacitance and by the speed of the amplifier. It can often be reduced by increasing the transconductance of the input device (see Sect. 3.3.6) if the bandwidth of the system is kept constant.

As described in the previous section, the noise must be kept significantly lower than the hit threshold to limit the noise hit rate to an acceptable level. This is particularly difficult for the edge and corner pixels where the noise is higher due to the increased capacitance. A noise in the order or below the threshold dispersion level of $\sigma_{\text{Thr}} \approx 200\ \text{e}^-$ is usually required. In systems with an analog readout of the hit amplitude an even lower noise figure may be beneficial to increase the spectroscopic resolution or to improve the spatial resolution obtained by interpolation.

The calculations in Sect. 3.3.6 show that such a noise can be achieved fairly easily with an idealized amplifier. Several other sources of noise, in particular cross talk within the chip and noise on the supply lines and on the substrate, voltage drops, etc., make it a design challenge, however, to reach the theoretical value on a large chip when many pixels are active simultaneously.

3.2.1.4 Input Impedance and Cross Talk

A charge Q deposited on a single pixel can induce parasitic signals on neighboring pixels due to the interpixel capacitance. This "cross talk" can lead to an increased fraction of double and triple hits and may have to be taken into account in the position reconstruction algorithms. The approximate magnitude of this effect is calculated in this section for cross coupling along one dimension only as it is the case for elongated rectangular pixels where the interpixel capacitance on the long edge dominates. Figure 3.12a shows a simplified cross section of an (infinitely wide) sensor with backside capacitances C_{back} and interpixel capacitances C_{C}. Next-to-next neighbor capacitances etc. are not considered in this approximation. Every pixel is connected to a charge-sensitive preamplifier with an effective input capacitance C_{eff} (see (3.4)) which is usually designed to be much larger than the total pixel capacitance in order to "pull" all the deposited charge into the amplifier. The capacitance C_{G} of every node to ground is therefore much larger than the pixel capacitances:

$$C_{\mathrm{G}} \approx C_{\mathrm{eff}}, \qquad C_{\mathrm{G}} \gg C_{\mathrm{back}}, \qquad C_{\mathrm{G}} \gg C_{\mathrm{C}} \tag{3.8}$$

The deposition of a charge Q on pixel 0 in the lumped arrangement of Fig. 3.12b is analyzed. The infinite capacitor chain to the right consisting of C_{C}, C_{G}, $C_{\mathrm{C}}, \ldots$ can be replaced by a single equivalent capacitor C_{X} as illustrated in Fig. 3.12c. Its value can be determined from the observation that a C_{C}–C_{G} pair followed by C_{X} must have the effective capacitance C_{X} due to the infinite arrangement. This is illustrated in Fig. 3.13. C_{X} must therefore be equal to the series connection of C_{C} and $C_{\mathrm{X}} + C_{\mathrm{G}}$. This leads to

$$C_{\mathrm{X}} = \frac{C_{\mathrm{G}}}{2}(W-1) \quad \text{with} \quad W = \sqrt{1 + 4\frac{C_{\mathrm{C}}}{C_{\mathrm{G}}}}. \tag{3.9}$$

The charge Q deposited on pixel 0 with the total capacitance $C_{\mathrm{G}} + 2C_{\mathrm{X}}$ leads to a voltage $U_0 = Q/(C_{\mathrm{G}} + 2C_{\mathrm{X}})$ at that node. The charge Q_0 on C_{G} of this

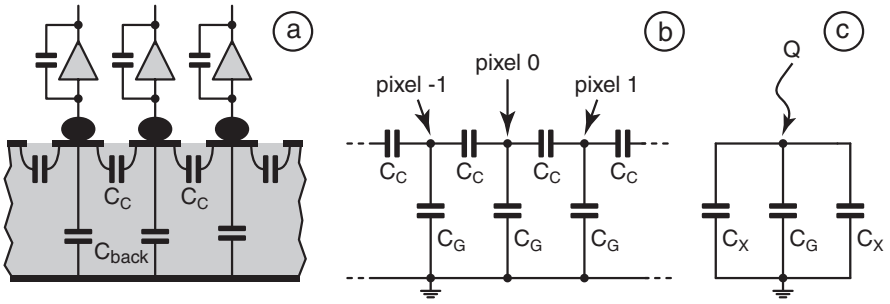


Fig. 3.12. (**a**) Model used for the estimation of cross talk to neighboring pixels, (**b**) lumped model of capacitances (**b**), and (**c**) left and right capacitances merged to C_{X}

pixel is $U_0 C_\mathrm{G}$, and so the corresponding charge fraction q_0 is

$$q_0 := \frac{U_0 C_\mathrm{G}}{Q} = \frac{C_\mathrm{G}}{C_\mathrm{G} + 2C_\mathrm{X}} = \frac{1}{W} \,. \tag{3.10}$$

The remaining charge is located on the right (and left) C_X:

$$q_x = \frac{1}{2}\,(1 - q_0) = \frac{W-1}{2W}. \tag{3.11}$$

C_X can be replaced by the equivalent circuit of Fig. 3.13 for the calculation of the charge on C_G of pixel 1. This results in

$$q_1 = q_x \, \frac{C_\mathrm{G}}{C_\mathrm{G} + C_\mathrm{X}} = \frac{1}{W}\,\frac{W-1}{W+1}, \tag{3.12}$$

and, more generally,

$$q_i = \frac{1}{W}\left(\frac{W-1}{W+1}\right)^{|i|} . \tag{3.13}$$

The q_i sum up to 1 for $i = -\infty, \ldots, \infty$ as expected. The q_i basically represent the charge fractions seen by the preamplifiers because C_G is dominated by C_eff.

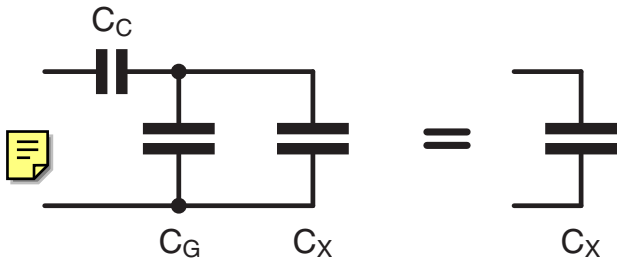


Fig. 3.13. Adding another C_C–C_G pair in front of C_X must again give C_X in an infinite arrangement

The cross talk can be quantified by the fraction of charge seen on the first neighbors, i.e. q_1. Using the approximation (3.8) leads to $W \approx 1 + 2C_\mathrm{C}/C_\mathrm{G}$ and so (3.12) becomes, to the first order,

$$q_1 \approx \frac{C_\mathrm{C}}{C_\mathrm{G}} \approx \frac{C_\mathrm{C}}{C_\mathrm{eff}}. \tag{3.14}$$

A realistic cross-talk value can be estimated as follows: The total capacitance of the rectangular pixels considered here is dominated by C_C, and so $C_\mathrm{C} \approx C_\mathrm{det}/2$. A typical preamplifier has an effective input capacitance which is significantly larger than the pixel capacitance, say, $C_\mathrm{eff} \approx 25 C_\mathrm{det}$. This leads to a cross talk q_1 of 2 %, a value which is acceptable in most applications.

In reality, the cross-talk issue is very complicated because the transient signals induced on neighboring pixels by the drifting charges must also be taken into account.

3.2.1.5 Speed

Depending on the application, the speed requirement for the preamplifier–shaper–discriminator chain can be different. At least two aspects must be considered.

The first important figure of merit is the *timing precision* with which the arrival time of a hit is determined. This is important for instance in particle physics experiments at the LHC where hits must be associated with one particular bunch crossing with a precision of better than 25 ns. Although constant delay contributions can, in principle, be calibrated out, the time resolution is improved by a short delay between the charge deposition and the reaction of the discriminator. A short preamplifier rise time, a high shaper bandwidth (or no shaper at all), and a fast discriminator are therefore desirable. This generally requires more current in the corresponding circuit blocks. The response time is a function of the input charge and so the "time walk" curve, as illustrated in Fig. 3.8, can be used to characterize the behavior. Charge depositions just above the threshold generate only a small discriminator overdrive so that the response becomes slower. If the amplitude of a hit is known, the measured time can, in principle, be corrected for on the basis of the known time walk characteristic. This task has already been performed on the front-end chip [211]. Note that a good timing precision can also be obtained with a "slow" preamplifier using special techniques like the deconvolution of the analog pulse shape [214] or the detection of the zero crossing of an appropriately shaped pulse.

A second important characteristic is the maximum possible *hit rate* which is limited by the time required to process one hit. This "dead time" ΔT can be dominated by the analog section but it may also be determined by the time required in the following readout circuitry to process the hit. A very complex behavior is introduced when the readout speed depends on buffers being filled, on the average occupancy, etc., and so Monte Carlo simulations are required to estimate the fraction of lost hits in a realistic environment.

In order to determine the fraction of hits lost due to a dead time ΔT with random events at an average rate of r hits per second, two different (idealized) cases can be distinguished:

When a hit occurring during the dead time interval *extends* the dead time by another ΔT, no hit pair with an arrival time difference smaller than ΔT can be detected. The loss fraction can therefore be found by integrating the normalized probability density of time intervals t between consecutive hits

$$p(t) = r\, e^{-r\, t} \tag{3.15}$$

from zero to up to ΔT. This leads to the loss fraction for *extended dead time*,

$$\text{loss}_{\,\text{ext}}(r,\,\Delta T) = \int_0^{\Delta T} p(t)dt = 1 - e^{-r\,\Delta T}, \tag{3.16}$$

which is $\approx r\Delta T$ for $\Delta T \ll 1/r$. Nearly all hits are lost when $\Delta T \gg 1/r$ because the dead time is permanently extended by new hits. An example for an extended dead time is the recovery time required in the preamplifier to discharge the feedback capacitor.

The situation is slightly different when hits occurring during the dead time do *not* extend the dead time. After a dead time of length ΔT, the detector finds a new hit after a time $1/r$, on average, and so the loss fraction for *nonextended dead time* is

$$\text{loss}_{\,\text{nonext}}(r,\,\Delta T) = \frac{\Delta T}{\Delta T + 1/r} = \frac{r\Delta T}{1 + r\Delta T}\ . \tag{3.17}$$

This is $\approx r\Delta T$ for $\Delta T \ll 1/r$, as before. The losses for the nonextended case are always smaller than for the extended case because the system is still able to detect a new hit after the fixed dead time even if the rate is high. An example for a nonextended dead time can be the digital hit processing.

A low dead time is, for instance, important in counting pixel chips where rates of up to 10 MHz per pixel can be required. The dead time often depends on the amplitude of a hit. Furthermore, the threshold for the following hit is clearly a function of the time delay as the shaper output must come back to its initial value. Higher order shapers (see Fig. 3.30) and a careful cancellation of the undershoot due to the feedback circuit (see Sect. 3.1.3) are important in such applications.

3.2.1.6 Power Consumption

The maximally allowed power dissipation in the pixel chip also depends on the application. The front-end chips can be cooled relatively easily in systems for X-ray detection because these consist of only one sensor layer in which the particles are absorbed. Components used for cooling underneath the front-end chips, hence, do not degrade the system performance. In detector assemblies in particles physics, on the contrary, several layers of sensors with front-end chips are usually used to measure several points of the track of penetrating particles (see Sect. 4.1). The flight path of such particles is influenced by multiple scattering in the material which therefore must be kept to a minimum. The cooling (see Sect. 4.5.2) is simplified if less power must be taken out of the system. As the dissipation of the front-end chips is usually the dominant heat source (other sources are the module control chip, the interface chips, the power supply cables, and the sensor itself), their power consumption must be reduced. A typical figure for the power used in particle

physics applications is ≈50 μW/PUC. The total current drawn by preamplifier, shaper, discriminator, readout, etc., should therefore not exceed 25 μA if a supply voltage of 2 V is used.

This limitation constitutes a challenge to achieve the analog performance goals for noise and in particular for speed. In order to obtain an impression of the significance of the problem, note for instance that a single D-flip-flop implemented in a 0.25 μm technology using enclosed NMOS devices for radiation hardness (see Sect. 3.2.2) and clocked at a frequency of 40 MHz at a supply of 2 V already draws an average current of ≈10 μA. A signal trace running across a pixel of 100-μm height has a capacitance of the order of 20 fF in a 0.25 μm technology and so a power of 3 μW is dissipated when the trace is clocked at 40 MHz at a supply of 2 V. Another source of power dissipation in the digital circuitry can be short circuit currents between power and ground during the switching of CMOS inverters or gates. The transconductance of the transistors offered by modern technologies is high, and so short circuit currents of several ~~10 μA~~ are common. The duration of the short circuit currents must be minimized by keeping the rise time of all digital signals very fast. This makes the design delicate and can cause increased cross coupling to the analog section. Possible remedies to this problem are discussed in Sect. 3.2.3.

3.2.1.7 Analog Charge Measurement

For monitoring purposes and in order to improve the spatial resolution, the measurement of the signal charge with a modest resolution is desirable. An amplitude information with a resolution of only a few bits can lead to a noticeable improvement in the spatial resolution in some applications. A better amplitude precision often does not significantly improve the spatial resolution [13, 215].

A straightforward approach is the sampling of the peak amplitude on a storage capacitor and the subsequent readout of the stored voltage (or charge) through a multiplexer. This approach has been implemented, for instance, in the CMS pixel readout [196] described in Sect. 3.4.2.

The BTeV readout chip uses a 2 bit-FADC (flash analog-to-digital converter) to get a very low resolution pulse height measurement [195].

The analog signal processing can be avoided by measuring the width of the discriminator output signal. This ToT is an almost linear function of the charge if a constant current feedback is used. The added complexity is small if the rising and the falling edge of the discriminator output are measured in units of the system (bunch crossing) clock. The quality of the analog information depends on the uniformity of the preamplifier/discriminator chain and on the maximum duration allowed for the hit signal (long signals introduce dead time). A resolution of ≈4 bits can be easily achieved. The ATLAS pixel chip described in Sect. 3.4.2 implements such a ToT readout.

3.2.1.8 Summary

Some typical specifications of readout chips for pixel detectors in particle physics are summarized in Table 3.2.

Table 3.2. Typical specifications for pixel chips used in particle physics

Quantity	Specification
Pixel area	2,500–40,000 μm^2
Noise (ENC)	<200 e^-
Threshold dispersion	<200 e^-
Power dissipation per pixel	<50 μW
Sensor leakage current tolerance	0–100 nA
Hit time measurement	<25 ns

3.2.2 Radiation-Tolerant Design

Pixel detectors are often used to detect ionizing radiation and highly energetic photons which traverse both the sensor and the readout chips. The MOS devices on the chips can therefore be exposed to high levels of radiation. When low-energetic radiation is detected (e.g. synchrotron radiation of some 10 keV), a significant fraction is absorbed in the sensor and so the readout chip is partially shielded. In particle physics experiments at the LHC, fluences of $10^{15}\,n_{eq}/\,cm^2$ are expected during 10 years of operation. This corresponds to a radiation dose of 500 kGy.

The dominant effect of ionizing radiation on CMOS electronics is the shift of device thresholds mostly due to the accumulation of positively charged holes in the gate and field oxide [216]. It has been known for a long time that the electrons can tunnel out of the oxide from a thin surface layer and consequently it was observed that the upcharging is significantly reduced when the oxide becomes as thin as ≈12 nm [32]. Modern "deep submicron" (DSM) CMOS technologies with gate lengths of 0.35 μm and below have even thinner gate oxides. The thin oxides and the high-quality processing steps employed lead to devices which show nearly no more threshold shifts for fluences of 300 kGy and above [217]. Parasitic NMOS transistors with thick field oxide can, however, develop thresholds low enough to open up parasitic leakage current paths. They must therefore be avoided. The current path can be interrupted by p^+ guard rings [218], so that a design with enclosed NMOS devices separated by guard rings can make the chip radiation-tolerant. This

has been successfully verified for several modern DSM technologies. No special measures need to be taken for PMOS transistors because the thresholds of parasitic devices in an n-type substrate or in an n-well increase rather than decrease with irradiation. An alternative geometry suited to implement NMOS devices with small W/L ratios has been suggested in [219]. Figure 3.14 shows schematically a radiation-tolerant layout for one or two series connected NMOS transistors.

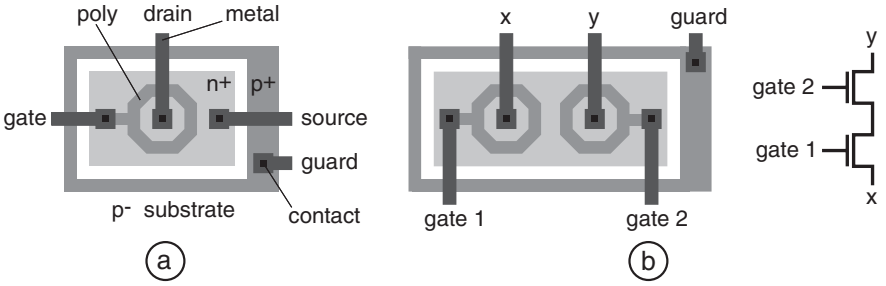


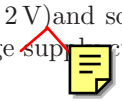
Fig. 3.14. Schematic layout of a (**a**) radiation-tolerant annular NMOS with guard ring and (**b**) series connection of two devices

In the particle physics community the very encouraging results of [217] have triggered a migration from the so far used specialized radiation-hard technologies to DSM technologies. The expected high yield for large area chips, the very high integration density, and the many levels of routing metal render them particularly well suited for pixel chips.

The design of mixed mode circuits using DSM technologies with radiation-tolerant design techniques, i.e. annular NMOS devices with guard rings, has several consequences some of which are mentioned in the following:

- The *device models* provided by the vendor are usually not suited for devices with enclosed geometry. Test structures must therefore be prototyped in order to extract parameters for reliable circuit models. Among the quantities to determine are the effective W/L ratio, output conductance, capacitances, noise, and matching, which has been measured to be worse for enclosed devices than for rectangular layouts [217].
- The *W/L ratio* of NMOS devices with enclosed layouts cannot be decreased below $\approx$2 [217] because the effective width of the device increases with increasing length. This large W/L ratio can be a serious limitation in analog design. High-quality NMOS current sources are difficult to implement and their noise contributions must be taken into account carefully. As a further consequence, NMOS devices are often operated in weak inversion at the currents used in pixel analog sections.

- The *capacitances and output conductances* of the inner and outer terminals of the enclosed devices as well as the gate overlap capacitances are different. This must be taken into account in the circuit models. The association of one or the other terminal to drain and source is an additional degree of freedom in the design. The inner terminal is normally used as the drain because of its significantly smaller capacitance. The asymmetry can cause problems, for instance, of charge injection in MOS switches.
- The *series connection of NMOS devices* must follow the rule that n^+ regions on different potentials must be separated by a gate or by a p^+ guard ring. Two devices connected in series are possible within one guard ring when the inner terminals are used as illustrated in Fig. 3.14b. The intermediate node has a very high parasitic capacitance in this case due to the large area of the n^+ region. More than two devices require additional guard rings. Serially connected NMOS devices should therefore be avoided if possible, for instance by replacing CMOS NAND gates by NOR-type logic with parallel NMOS devices.
- The *input capacitance* of CMOS logic gates is large, in particular if the W/L ratio of the PMOS devices is matched to the mostly unnecessarily large W/L value of the enclosed NMOS devices. This leads to a significantly increased dynamic power consumption as compared to "minimum-size" logic. Furthermore, the short circuit current in such gates during the input signal transition is large (up to 200 μA for a "matched" inverter operated at 2 V)and so *all* signal transitions must be very fast in order to avoid large supply current spikes.



3.2.2.1 Single Event Upset

Strongly ionizing particles (in particular slow heavy ions) can deposit very large charges in a very small volume of silicon. If the charge happens to be deposited on a storage node (DRAM or SRAM cell, latch, flip-flop), this node can flip such that the stored information gets corrupted. If such a "single event upset" (SEU) occurs in a state bit of a state machine, a wrong behavior with possibly serious consequences may follow. If the state machine enters a "forbidden" state, a permanent lockup may completely block the system. (State machines should therefore be designed such that they recover from all possible states.) The rate for SEU events depends very much on the circuit schematic and on the physical layout and on weather the bit flips from 0 to 1 or from 1 to 0. Shift registers can be more sensitive when being clocked as compared to the static state. The sensitivity of a circuit to SEU is often quoted as a cross section. Measured values for various flip-flop designs for instance range from 10^{-12} to 10^{-16} cm^2 [220–223]. The cross section decreases rapidly with smaller energy depositions and becomes negligible below a certain threshold energy.

Standard techniques to address SEU are the use of error correcting logic (e.g. Hamming codes), for instance, in RAMs or self-correcting triple redundancy cells for individual bits. These solutions require significant additional hardware so that they can be used only when really needed. In some applications it may be sufficient to detect an SEU by a checksum mechanism (e.g. a parity bit) and react externally by a reconfiguration of the circuit. A fairly simple and efficient approach to reduce SEU in a standard SRAM cell consisting of two cross-coupled inverters is the addition of a capacitor between the two storage nodes as illustrated in Fig. 3.15. The capacitor slows down the cell and so short transients on the storage nodes cannot easily flip the cell. A reduction of the SEU cross section by 2 orders of magnitude has been reported [222](see also Fig. 5.4).

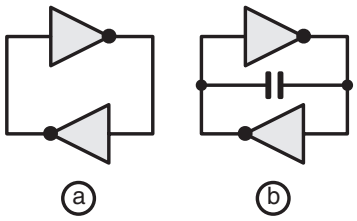


Fig. 3.15. Standard (**a**) and SEU-protected SRAM cell (**b**)

The ingenious DICE cell [224] drawn schematically in Fig. 3.16 uses a different concept to address the problem. The information (and its inverse) is stored on 2+2 independent storage nodes which are cross-coupled in such a way that a temporary flip of one node due to an SEU does not permanently flip the cell. The circuit is very simple and compact. The layout must assure that corresponding storage node is not flipped simultaneously by the same charge deposition.

3.2.3 Cross Talk

Pixel electronics is particularly affected by cross talk from the digital to the analog section because these are densely packed in the active area. The intrinsic noise and the achievable charge thresholds of the analog section are low and so already very small disturbances degrade the overall performance. A voltage step of 1 V, for instance, injected through an extremely small parasitic capacitance of only 1 fF generates a cross-talk charge of 1 fC or 6,250 electrons, which is much more than the noise and typical thresholds. Cross-talk signals can be injected capacitively (across the chip, via the substrate or via the sensor) or through spikes in supply or bias signals. Several precautions can be taken to address the problem.

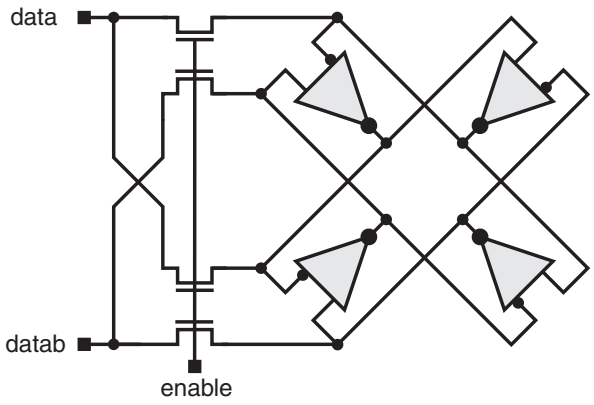


Fig. 3.16. SEU-tolerant DICE storage cell, redrawn from [224] to stress the symmetry. The *triangular symbol* is an inverter consisting of two transistors with separate inputs for the gate of the NMOS and the PMOS transistor

3.2.3.1 Shielding

A metal *shield* should cover the chip so that signal swings do not couple capacitively to the sensor. A capacitance of $1.8\,\mathrm{aF}/\,\mu\mathrm{m}^2$ between chip and sensor has been measured [195]. This value depends only on the distance between sensor and chip, i.e. on the height of the final bump connections. A hermetic shield often requires more than one metal layer due to technology limitations in the individual layers (maximum metal density, slots in metal). The shield must be connected to a "clean" net because voltage spikes on the large area would be detrimental. The connection of the shield must have low resistance and low inductance because its large area is itself prone to pickup from the signals below.

The *input pad* is the most sensitive node due to its relatively large area. It must be well shielded. The integration of the feedback and test charge injection capacitors in the metal stack underneath the pad can be a solution (see Sect. 3.3.2).

Guard rings are often used around the analog parts. They may be less effective in DSM technologies, however, because the substrate resistance can be very low, allowing an efficient signal path underneath the guard implantations.

3.2.3.2 Power Supplies

Separate power supplies for analog and digital parts and possibly also for large buffers and CMOS output pads are commonly used. Low inductance supply traces and multiple bond pads are standard techniques.

On-chip voltage regulators can be used to provide clean voltages and to compensate for voltage drops on the supply cables. The CMS pixel chip, for instance, uses four regulators for two analog and two digital supplies [204].

The *injection of switching noise* from the digital supplies into the substrate can be reduced if a separate net is used for the bulk and possibly the well connections [225].

Spikes on the digital supply are generated by the charging and discharging of capacitive loads during signal transitions, but also by the short circuit current during transitions, which can be quite important when enclosed devices with large W/L are used to achieve radiation hardness. While slow rise times would decrease the magnitude of capacitive spikes, the DC-currents would become unacceptable if ratioed CMOS logic (i.e. with equally strong PMOS and NMOS devices) were used. Other logic families should therefore be considered:

- Small PMOS devices (i.e. not matched to the large W/L of the NMOS) reduce the input capacitance and the short circuit current which lasts, however, longer due to the slow signal rise times. This approach requires very careful design because rise and fall times and propagation delays are very asymmetric. The noise immunity is reduced because the switching thresholds are low.
- The addition of a current limitation (current source) into the CMOS logic reduces the short circuit current and the rise times [226]. The limited drive strength must be individually adapted to the capacitive load and so a very careful extraction of all parasitics is necessary. The underestimation of a load can limit the operation speed of the whole chip. Programmable drive strength in crucial places are conceivable.
- Differential full swing logic with adjacent signal pairs produces balanced emissions (toward shield and substrate), but the supply spikes are even increased, because more lines must be driven. The differential voltage swing is twice the supply voltage and so a large charge is stored in the capacitance between the two traces.
- Differential logic with a reduced signal swing supplied with a constant current is a very attractive candidate because spikes are small and emissions are low. The limited drive strength, however, requires a careful design, as mentioned above. The total power dissipation of such a logic is not necessarily higher than for CMOS, because the dynamic contribution is significantly reduced by the low signal swing.

It may be sufficient to use low noise logic at particularly critical places, like in the PUCs.

A *reduction of the switching noise* is achieved by avoiding the simultaneous flipping of signals with large loads (busses), for instance by using Gray-encoded counters. Less (more optimized) logic produces less spikes.

Periodic signals may be less dangerous than intermittent activity because their effect on the analog part is constant.

3.2.3.3 Further Remarks

A *high power supply rejection* of the analog section reduces the sensitivity to spikes on the supply. This can be achieved, for instance, with a differential preamplifier design [226] at the expense of an increased analog power dissipation. Note that the power supply rejection of the circuit must still be significant in the signal frequency band. The second input of the differential amplifier is an extremely sensitive node and must be treated with care. *Decoupling capacitors* can be added locally on the chip. Although the achievable total capacitance is not very large (tens of nanoFarads), they can be very efficient because of the small inductances of the local connections. A possible yield problem introduced by the large area of the thin gate oxide can be overcome by using thin metal traces for the connection acting as fuses and by a special circuitry to switch off defective devices [227]. These techniques are commonly used in industrial designs. They lead to an area overhead of roughly a factor of 2, depending on how sophisticated the test circuitry is. A large chip capacitance in conjunction with the wire bond inductance can lead to oscillations. These are very difficult to predict because the damping elements are difficult to model.

The *input/output signals* going to and coming from the chip during data taking should use low swing signals, slew rate limited signals, current outputs, or differential signaling (LVDS).

3.2.4 Testability and Ease of Operation

3.2.4.1 Testability

The large number of channels on pixel chips makes it very time-consuming to perform precise analog measurements for every pixel. These are required, however, to characterize the design and to sort out known good dies before mounting them onto the sensor. It is therefore crucial to foresee internal circuitry to simplify testing. These features can also be used to monitor correct operation of mounted chips.

- The *injection of known test charges* into the preamplifiers must be possible. The simultaneous stimulation of several pixels, preferably in a freely programmable spatial pattern, should be possible so that a realistic signal activity can be generated on the chip. This feature is useful, for instance, to measure weather a significant activity in the digital readout degrades the analog performance by cross talk. Simultaneous injections can also be used to characterize several pixels simultaneously.

  The injection circuitry must be designed very carefully in order not to increase the noise due to the extra components connected to the amplifier input. The test charge is often generated by applying a voltage step to small injection capacitors located in the pixels. This solution guarantees

fairly identical charges in all pixels because the matching of on-chip capacitors is usually very good. The value of the injection capacitors must be determined with high precision in order to calculate the absolute amount of injected charge. An elegant method is an on-chip measurement for the very simple charge pump circuit presented in [228]. The calculated charge can be cross calibrated with the charges deposited by monoenergetic γ-rays once a sensor has been connected to the chip.

Another method to inject charges are current pulses of well-defined amplitude and duration. They can be generated easily by steering a known current to the preamplifier input with a differential pair. This concept allows the generation of multiple successive hits in short time intervals. This is not easily possible with capacitive injection (it requires staircase-shaped signals), but can be a valuable feature in applications where a high count rate is required. A good matching of the charges injected into different channels is more difficult to guarantee, however, because active MOS devices are involved.

- A *leakage current injection* into every pixel is useful to verify the correct operation of the leakage compensation circuitry.
- The injection of *digital test patterns* can be valuable for quick tests of the digital section. This feature can also be used, for instance, to quickly fill data buffers with known contents and to study overflow conditions etc.
- *Monitoring* of internal DC-levels, of bias currents, supply voltage drops, etc., can be useful. The interesting signals can be connected to few test pads with analog multiplexers so that the external test overhead is small.
- The observation of *analog waveforms* can be accomplished by multiplexers in the pixels which connect selectable buffered signals to a readout bus.
- It should be possible to *verify* the contents of all internal registers.

3.2.4.2 Ease of Operation

The operation and test of chips before their assembly and inside a sensor system can be simplified and accelerated by additional features. For instance, defective pixels which show hits with no signal applied can mask themselves off automatically. A very time-consuming operation is the local fine-tuning of the pixel thresholds. This is usually done by an iteration of threshold scans followed by a change of the trim DAC settings. Some intelligence inside the pixel can be used to search for the DAC setting which corresponds to a 50% hit fraction. The continuous injection of the desired threshold charge is then sufficient to correctly set the thresholds in all participating pixels. Such an "auto-tune" feature has been implemented, for instance, in the ATLAS FEI3 chip [211]. Further self-test features could be implemented to verify buffers, check overflow conditions, and others.

3.3 Analog Signal Processing

This section presents some commonly used circuits for the most important building blocks of the analog part of the PUC (see Fig. 3.4). A low-noise, low-power amplifier fed back by a small capacitance is used to convert the input charge to a voltage. The feedback capacitor must be discharged after a hit by an appropriate circuit. The leakage current of the sensor flows into the pixel input in DC-coupled arrangements. It must therefore be absorbed by a suited compensation circuit. A discriminator is used to convert the analog pulse to a digital hit signal. Some chips use a bandwidth-limiting shaping amplifier between preamplifier and discriminator to reduce the noise and to increase the double hit capability. Preamplifier, feedback, and discriminator with threshold trim are briefly discussed in Sects. 3.3.1 to 3.3.3. A rough calculation of the noise expected at the output of a CMOS charge amplifier with no shaping is presented in Sect. 3.3.5. The effect of a simple shaper is analyzed in some detail in Sect. 3.3.6.

3.3.1 Charge Amplification

A simple single-ended cascoded amplifier is very often used in pixel and strip readout chips [229] owing to its simplicity and current efficiency. In the "direct" or "straight" cascode configuration shown in Fig. 3.17a the input device $M_{\rm in}$ is mainly biased by the current source $I_{\rm B1}$, which is often operating from an additional, lower supply voltage to save power. The cascode device $M_{\rm casc}$ keeps the drain of $M_{\rm in}$ at a constant potential so that the signal current flows to node v_1 where it generates a voltage signal. The smaller current $I_{\rm B2}$ in this branch makes it easier to achieve a high output impedance of the current source and thus a high DC-gain. A source follower is often added to reduce the capacitive loading of the dominant node v_1 in order to increase the bandwidth. The very popular [194, 230, 231] "folded" arrangement shown

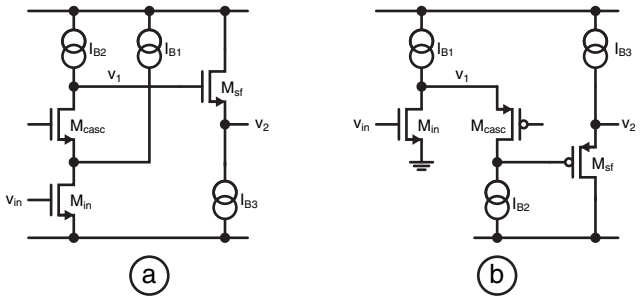


Fig. 3.17. The "direct" cascode (**a**) and the "folded" cascode (**b**) are commonly used single-ended amplifier configurations. They are often decoupled from the following circuits with voltage followers which are also indicated in the circuit diagrams

in Fig. 3.17b is slightly less current efficient when the same supply voltage is used (bias current I_{B2} does not flow through the input device) but it is better suited for DC-feedback and low supply operation owing to its higher useful signal range. Variations of this basic cell using regulated cascode structures have been used as well.

Both circuits can of course also be implemented with PMOS input devices. The choice to be made is influenced by many factors, some of which are briefly listed here.

- The transconductance of the input device should be high in order to increase the bandwidth or to reduce the channel noise (see Sect. 3.3.5). This favors NMOS devices.
- The $1/f$-noise, however, is usually better for PMOS transistors [232].
- The coupling of substrate noise into the sensitive input node through the input device can be reduced if it is located in a separate well. This can be connected to a "clean" potential.
- When designing for radiation hardness, NMOS devices require special attention in order to avoid leakage after irradiation (see Sect. 3.2.2). One approach is the use of an enclosed geometry for this transistor type. This approach leads to NMOS devices with fairly large transconductances and so current sources with a large output resistance are difficult to implement. The crucial current sources I_{B2} in Fig. 3.17 are therefore better implemented with PMOS devices if a high DC-gain is desired. This fixes the input device type for a given topology. (When using devices with large transconductance for current sources, their noise contribution may be significant.)
- The source followers in Fig. 3.17 have asymmetric (large signal) rise and fall times. The polarity of the input charge signal determines the direction of the leading edge and therefore constraints the design.

The power supply rejection of the amplifier is improved if a differential pair is used [202]. The current required for the same noise or speed is, however, increased. One input of the differential amplifier is connected to the sensor. The second "reference" input is an extremely sensitive node and must be connected to a very clean potential. In the LHC1 chip, this potential is generated internally.

3.3.2 Feedback and Leakage Compensation

The value of the feedback capacitor C_f is chosen as a compromise, among others, between a high gain (small C_f) and a large effective input capacitance (larger C_f) as discussed in Sect. 3.1.3. Typical values range from a few to tens of femtoFarads. All parasitic contributions from other devices, in particular in the feedback circuit, must be carefully considered. Such small capacitors can easily be implemented as metal–metal structures with an area of ~~some~~ $<100\,\mu m^2$. Their geometry must be designed carefully in order to reduce

additional parasitic capacitances to the input node. One possibility is to use the bump pad as one capacitor plate as illustrated in Fig. 3.18. This also provides a shield for the bump pad. If the value of C_{f1} is not sufficient, another capacitor C_{f2} can be added, its bottom plate being shielded by the injection capacitor in this example.

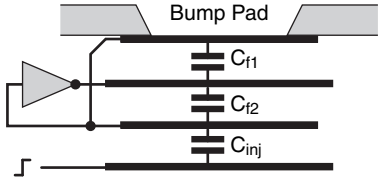


Fig. 3.18. A sandwich structure of metal plates can be used to implement the feedback capacitor while simultaneously shielding the input pad and the preamplifier input from noise sources

The charge deposited on the input node must eventually be removed, i.e. the feedback capacitor C_{f} must be discharged. Some general requirements for the reset circuit are a reasonably fast discharge, linearity, a small noise contribution, and the response to leakage current at the input. Some commonly used approaches are briefly discussed here (see also [233]).

3.3.2.1 Resistive Feedback (MOSFET in Linear Region)

If a resistor in parallel to the feedback capacitor is used to discharge for instance $C_{\mathrm{f}} = 5$ fF in $\tau = 100$ ns, a value of $R_{\mathrm{f}} = 20$ MΩ is required. This is too high to be implemented as a passive device, and the associated capacitive parasitics would be prohibitive. A MOSFET operated in the linear region (see Fig. 3.19a) has a channel resistance of $R = [K(W/L)(V_{\mathrm{GS}} - V_{\mathrm{T}})]^{-1}$. For a reasonably long device with $W/L = 0.1$ and a transconductance parameter of

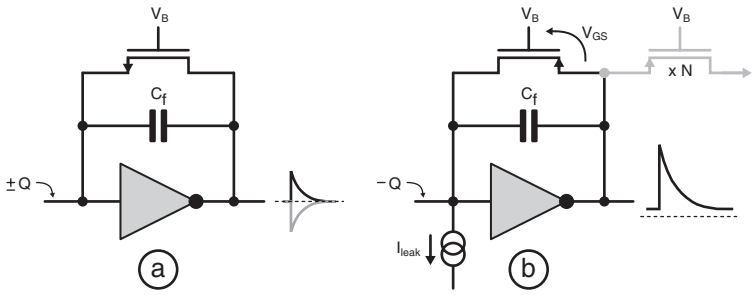


Fig. 3.19. Feedback (**a**) with a long NMOS device operated in the linear region and (**b**) with a PMOS operated in saturation

$K = 50\ \mu\mathrm{A}/V^2$, the required gate overdrive would be only $V_{\mathrm{GS}} - V_{\mathrm{T}} = 10\,\mathrm{mV}$. This small value leads to a very low saturation voltage and so only very small output voltages are possible if a true RC-discharge is desired. Larger output signals show very different behavior depending on the polarity. Sensor leakage current flowing through the feedback changes its behavior [230]. Transistor mismatch can lead to significant channel to channel variations. The simple solution of a MOSFET operated in the linear region works fairly well for AC-coupled strip readout chips where feedback capacitors are larger (required by the larger sensor capacitance) and output swings are smaller. It is difficult to bias in pixel chips, however, and saturates early unless very long devices are used.

3.3.2.2 Feedback with MOS in Saturation

The feedback device of Fig. 3.19b can be kept in saturation if a small current (from sensor leakage or sometimes as input bias of the amplifier) always flows. The device type must be chosen such that the source is at the output. This requires the use of a PMOS for negative input charges (and therefore also negative leakage current). The stationary output voltage increases with higher input leakage currents. A negative charge deposition at the input leads to a positive edge at the output and so $|V_{\mathrm{GS}}|$ of the PMOS rises. The increasing drain current discharges C_{f}. Due to the nonlinear relationship between the output voltage swing and the drain current, the shape of the discharge curve depends on the signal amplitude. The discharge time constant is also determined by the characteristics of the feedback transistor, by the value of the feedback capacitor, and by the leakage current. It cannot be adjusted externally. The time integral over the discharge current in the feedback device equals the input charge and so a second, N times wider transistor (shown in gray) can be used as a "current mirror" to provide a precisely N times larger replica of the input charge. Its gate is connected to V_{B}, the source to the output of the amplifier of Fig. 3.19b and the drain to a voltage equal to the input voltage [234].

3.3.2.3 Constant Current Feedback

The circuit shown in Fig. 3.20a discharges C_{f} with a nearly constant current and so the triangular pulse shapes of Fig. 3.21 are observed. The simple circuit [195,210] is, at the same time, able to sink leakage currents I_{leak} much larger than the bias current I_{b}.

The function of the circuit is derived here for identical devices M_1 and M_2 operating in strong inversion with the usual square law current relationship $I_{\mathrm{D}} = K'(V_{\mathrm{GS}} - V_{\mathrm{T}})^2$ in saturation and $I_{\mathrm{D}} = 2K'V_{\mathrm{DS}}(V_{\mathrm{GS}} - V_{\mathrm{T}} - \frac{1}{2}V_{\mathrm{DS}})$ in the linear region [235]. The abbreviation $K' = \frac{K}{2}\frac{W}{L}$ is used, with K being the transconductance parameter and W and L the width and the length of the transistors, respectively. The substrate effect is neglected.

The bias current is denoted by both I_{B} and I_{b}. Kindly check this and do the needful.

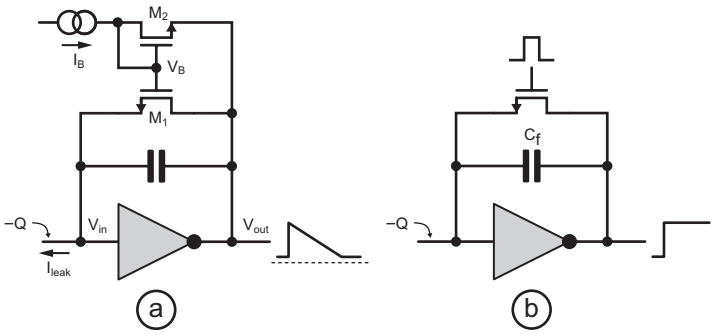


Fig. 3.20. Feedback (**a**) with a constant current source and (**b**) with a MOS switchd

The configuration shown in Fig. 3.20a is suited for negative input charges which produce positive output signals. A leakage current I_{leak} leads to a slightly positive potential at the output so that the source of M_1 is actually at the input. V_{GS} of M_1 can be expressed as

$$V_{\mathrm{GS},1} = V_{\mathrm{GS},2} + V_{\mathrm{DS},1} = V_{\mathrm{T}} + \sqrt{\frac{I_{\mathrm{b}}}{K'}} + V_{\mathrm{DS},1} \tag{3.18}$$

with I_{b} flowing through M_2 in saturation. (3.18) shows that M_1 operates in the linear region ($V_{\mathrm{DS},1} < V_{\mathrm{GS},1} - V_{\mathrm{T}}$) when no output signal is present and so

$$I_{\mathrm{leak}} = 2K'V_{\mathrm{DS},1}\left(V_{\mathrm{GS},1} - V_{\mathrm{T}} - \frac{1}{2}V_{\mathrm{DS},1}\right). \tag{3.19}$$

Injecting (3.18) into (3.19) and solving for $V_{\mathrm{DS},1}$ leads to

$$V_{\mathrm{DS},1} = \sqrt{\frac{I_{\mathrm{leak}} + I_{\mathrm{b}}}{K'}} - \sqrt{\frac{I_{\mathrm{b}}}{K'}}\,, \tag{3.20}$$

and, again using (3.18),

$$V_{\mathrm{GS},1} = V_{\mathrm{T}} + \sqrt{\frac{I_{\mathrm{leak}} + I_{\mathrm{b}}}{K'}}\,. \tag{3.21}$$

When a negative charge is deposited at the input, the output becomes positive and M_1 goes into saturation. The potential at the input is held nearly constant by the feedback action of the amplifier. $V_{\mathrm{GS},1}$ remains unchanged as long as the voltage at node V_{b} does not vary (this can be guaranteed by appropriate decoupling). The current through M_1 becomes

$$I_{1,\mathrm{active}} = K'(V_{\mathrm{GS},1} - V_{\mathrm{T}})^2 = I_{\mathrm{leak}} + I_{\mathrm{b}}\,.$$

The feedback in this situation provides the leakage current and exactly I_{b} to discharge C_{f}. A very similar result is obtained, surprisingly, if both devices

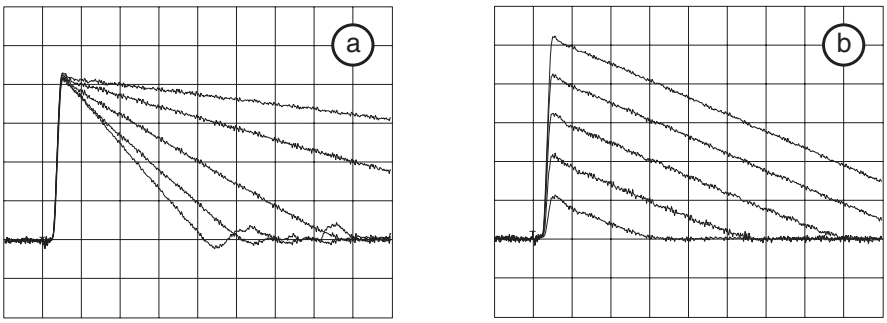


Fig. 3.21. Measured output signals of a charge preamplifier with constant current feedback (200 ns resp. 200 mV per division) [236]. The feedback current is varied in (**a**) from small values for the topmost traces to large values for the fast discharges. The injected charge is varied in (**b**)

operate in weak inversion. The increase of the feedback current by I_b is independent of the output amplitude and so the feedback capacitor C_f is discharged with a constant slope. This result is confirmed by the measurement in Fig. 3.21a where the feedback current I_b is varied. The output pulse comes back to the baseline after a time

$$T = Q/I_\mathrm{b} \; . \tag{3.23}$$

This expression remains valid even if the output saturates. Figure 3.21b shows the response of the circuit to increasing input charges for fixed I_b. The width of the pulse increases linearly with the injected charge and so a measurement of the width of the discriminated output signal can be used to determine the analog charge deposition. This method is used in the ATLAS FEI pixel chips [203]. Note that the leakage current I_leak can be significantly larger than the feedback current I_b. The only consequence in this simplified analysis is a DC-shift of the output potential according to (3.20).

3.3.2.4 Reset Switch

A simple MOS switch as shown in Fig. 3.20b can be used in applications where a continuous reset is not needed. This method has been successfully used in one of the first chips for microstrip readout [229, 237, 238]. Input and output of the amplifier are shorted and C_f is discharged before a signal arrives. The circuit becomes an ideal integrator once the switch is opened. Regular resets are required in order to avoid saturation of the amplifier output. Design issues are the charge injection from the gate to the input, when the switch is opened, and also the switching noise ("kTC-noise").

3.3.2.5 The Krummenacher Scheme

The circuit in Fig. 3.22 proposed already in 1991 [209] compares the DC-value of the preamplifier output to a reference voltage (which can be the voltage at the input of the amplifier) with the differential pair M_{1a}/M_{1b}. The current in the leakage compensation device M_2 is regulated such that it equals $I_{leak} + I_b$ in the equilibrium state. When a positive charge is deposited on the input node, the output goes negative as indicated. The complete bias current $2I_b$ is steered through M_{1b} while the current through M_{1a} is turned off. The input node is therefore discharged with a net current of I_b, independent of I_{leak}. The current through M_2 is not changed significantly because the large capacitor C keeps the gate voltage nearly constant. The noise contribution of M_2 is reduced by using a long device with a small transconductance. This scheme or one with minor modifications has become very popular and is used in several chips for strip and pixel readout [193, 206, 239].

Au: Should it be I_b as given in the artwork?

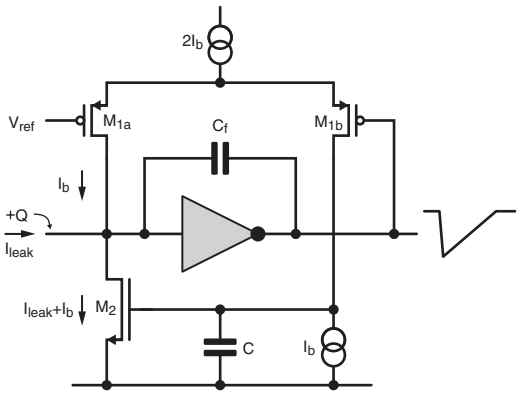


Fig. 3.22. Circuit to adjust the leakage current compensation in M_2 so that the output reaches a given reference level [209]. When a signal arrives, the current through M_{1a} is cut off and the input node is discharged with I_B

3.3.3 Hit Discrimination

A discriminator is used to detect a preamplifier (or shaper) output signal above a given threshold. It generates a digital hit signal which is fed to the readout (see Sect. 3.4) and which can also be used, for instance, to sample the analog signal amplitude for a later readout [196]. Important design aspects are power dissipation and speed, layout area, the achievable threshold range, and the homogeneity of thresholds in different pixels. Some proposed implementations are briefly presented here.

3.3.3.1 Differential Pair

A very popular circuit for the discriminator is the differential amplifier because of its simplicity and constant current operation. The threshold can be set, for instance, by generating a voltage offset between the two inputs. The circuit shown in Fig. 3.23a achieves this by pulling a current I_{thr} through a resistor R, leading to an offset RI_{thr}. The actual implementation [207] uses a PMOS device operated in the linear region for R and a weak PMOS current source for I_{thr}. The offset is a linear function of the threshold current and can be set all the way down to zero.

The threshold can be set directly if the dc output level of the driving signal can be controlled, for instance with a feedback circuit of the type shown in Fig. 3.22 where the stationary output level equals V_{ref}.

Another possibility to influence the switching point of the differential amplifier is the injection of an additional current in one of the branches as illustrated very schematically in Fig. 3.23b. This concept can be improved by using a transconductance amplifier with a wider linear range and a current comparator at the output [193].

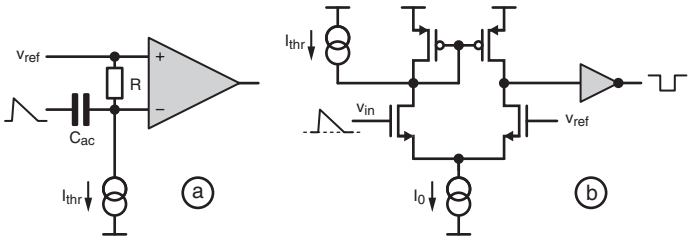


Fig. 3.23. Discriminators using differential amplifiers. The threshold can be set by introducing a DC-offset between the two inputs with the help of a resistor supplied by a current (**a**) or by unbalancing the currents in the two branches (**b**)

3.3.3.2 Cascade of Low Gain Limiter Stages

The task of the discriminator is to amplify a small ("analog") voltage difference to a digital full swing signal. It can therefore be considered as a high gain amplifier. It is well known that fast high gain amplifiers can be obtained by cascading several low gain stages. The switching is further improved by limiting the output swing of the individual stages. This concept has been proposed for pixel detectors in [209].

3.3.3.3 Diode Biassed Inverter

The very simple circuit illustrated in Fig. 3.24 has been used in the PILATUS chip [198]. An offset voltage is introduced between the input and

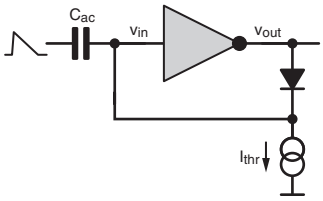


Fig. 3.24. Discriminator of the PILATUS chip. The threshold is set by the voltage across a diode supplied with current I_{thr}. The input signal is AC-coupled through a sufficiently large C_{ac}

the output of a simple CMOS inverter by pulling a current I_{thr} through a forward biassed pn-diode connecting output and input. The input signal is AC-coupled through a sufficiently large capacitor. The generated threshold is very constant because the properties of the pn-junction are much less subject to process variations than, for instance, MOS threshold voltages. The dependency of the threshold on mismatch in the inverter and in the current source is very small, and so a dispersion of only 120 e^- has been achieved without threshold trim. A drawback of this circuit is the limited range of possible thresholds.

3.3.4 Threshold Trim

A narrow threshold distribution without large tails is required in order to guarantee a constant hit efficiency and a low noise hit rate (see also Sect. 3.2.1). A careful design of the discriminator and all circuit components affecting the threshold is therefore necessary. Transistor matching and, possibly, voltage drops in supply and bias lines must be taken into account. If the required maximum threshold dispersion for a given application cannot be guaranteed by design, circuitry to fine-tune the threshold(s) in every single pixel must be included. Two basic approaches for the threshold "trim" have been pursued.

3.3.4.1 Digital Trim

Digital correction values of 3 bit [193, 196, 198, 206] to 7 bit [211, 240] resolution can be stored in every pixel. Voltage- or current-mode DACs are used to convert the binary information to control voltages or currents which influence the threshold of the discriminator(s). The "coarse" threshold is often set by a global control signal generated in the periphery of the chip with a high-resolution DAC (see Fig. 3.4). The fine adjustment in every pixel is a small correction to the global setting. The threshold change per trim step should be adjustable in order to find a good compromise between a wide trim range required to correct largely offset thresholds and a fine step size to achieve a low

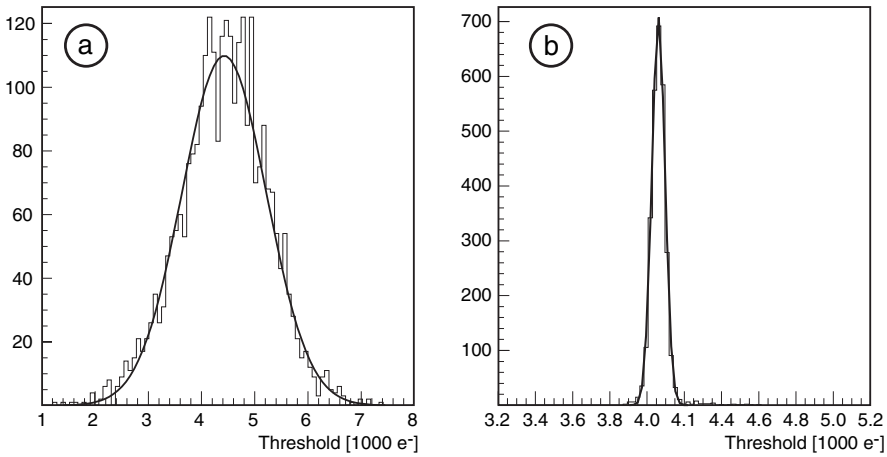


Fig. 3.25. Threshold distribution of the 2,880 pixels of a FEI chip before (**a**) and after (**b**) adjustment. The initial dispersion of $\sigma = 796\,\mathrm{e}^-$ drops to $\sigma = 35.5\ \mathrm{e}^-$

net dispersion after the trimming procedure. Figure 3.25 shows the threshold dispersion of a FEI2 chip before (a) and after (b) trim. The digital approach provides stable trimming once the correction values have been determined and downloaded to the pixel registers. The area required for the storage cells and the DAC is significant, however, so that the resolution is limited.

3.3.4.2 Analog Trim

Another approach is the dynamic analog storage of a correction voltage on a capacitor. The precision of the adjustment depends only on the precision of the supplied voltage, so that very high tuning ranges with extremely high resolution can be achieved. The drawback of this approach is the unavoidable droop of the stored voltage due to leakage. This drift of the thresholds can be acceptable in applications where the chip is operated for short time intervals only, such that the values can be refreshed (like for instance in medical X-ray applications). The concept is less suited for continuous operation. The drift of the thresholds can be kept at a very low level if the leakage from the storage node is reduced. The simple circuit in Fig. 3.26 decreases leakage through parasitic diodes and the turned-off channel of the switch transistor M_2 by keeping the potentials of its drain and bulk terminals equal to the stored voltage. This is achieved by a unity gain buffer with a very small output current and so writing of an external voltage is possible through the open switches M_1 and M_2. Subthreshold leakage can be further reduced by overdriving the switch gates when they are off. A droop of the threshold of $\approx 0.1\ \mathrm{e}^-/\mathrm{s}$ has been measured at room temperature [207]. Figure 3.27 demonstrates the threshold precision which can be obtained with the analog approach. An initial rms

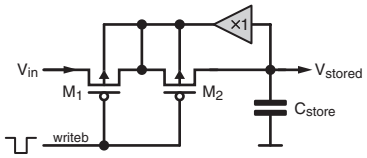


Fig. 3.26. Circuit to reduce leakage from an analog storage node. The bulk node connection and the intermediate node of two serial switches are kept at the same potential as the stored voltage by a very weak unity gain buffer so that currents flowing through leakage elements become small

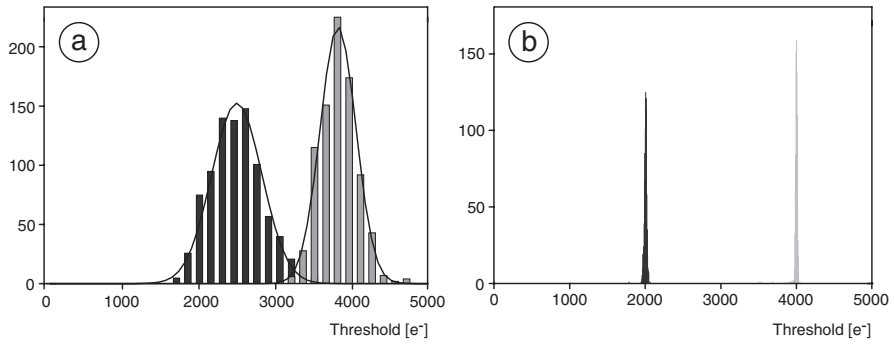


Fig. 3.27. Threshold distribution for the two discriminators in every pixel of the MPEC2.1 chip without sensor before (**a**) and after (**b**) analog trim. The rms of the distributions drop from 327 (233) electrons for the low and high threshold to 13 (11) electrons, respectively

dispersion of 327 (233)electrons, (for the two discriminators in every pixel) is reduced to 13 (11) electrons, which is basically the error in the determination of the thresholds.

3.3.5 Noise in a Simple FET Amplifier

This section gives a brief overview of the methods used to determine the noise at the output of the amplifier. A simplified circuit is treated here for illustration to give an idea of the noise values that can be achieved. It is assumed here for simplicity that the input capacitance is dominated by the sensor so that stray capacitances and in particular the input capacitance of the amplifier itself can be neglected. Only the dominant noise contributions from the sensor leakage current, from thermal and $1/f$-noise in the channel of the input transistor, and from the thermal noise of the feedback resistor are considered for simplicity. A more detailed treatment of a system with preamplifier and shaper is found in Sect. 3.3.6.

3.3.5.1 Transfer Function

Figure 3.28 shows the circuit used for the calculation in this section. The inverting preamplifier Fig. 3.28a consists of a transistor with transconductance g_m driving into a resistor R_o. The capacitance at the preamplifier output node is C_o. This circuit has the frequency-dependent voltage gain

$$v(s) = \frac{v_\mathrm{out}(s)}{v_\mathrm{in}(s)} = -\frac{v_0}{1 + s/\omega_0} \tag{3.24}$$

with the complex frequency variable $s = \mathrm{i}\omega$ and with the DC-gain and bandwidth constants

$$v_0 := g_\mathrm{m} R_\mathrm{o} \quad \text{and} \quad \omega_0 := \frac{1}{R_\mathrm{o} C_\mathrm{o}} \,. \tag{3.25}$$

The voltage gain of this amplifier is v_0 up to the bandwidth ω_0 from whereon it drops inversely proportional to the frequency until unity gain is reached at the unity gain bandwidth $\omega_\mathrm{GBW} = g_\mathrm{m}/C_\mathrm{o}$.

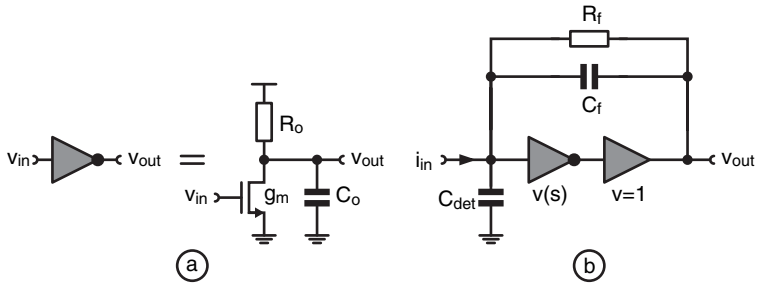


Fig. 3.28. Circuit used for the noise calculation. The preamplifier (**a**) is a simple gain stage with a capacitive load C_o. A unity gain buffer is added in the charge-sensitive amplifier configuration (**b**). The feedback capacitor C_f is discharged with a resistor R_f

The preamplifier is followed by an ideal unity gain buffer (source follower) with a small output impedance in the charge-sensitive configuration shown in Fig. 3.28b. A feedback capacitor C_f is discharged by a large resistor R_f in this example. The buffer is often added to minimize the bandwidth-limiting load at the output of the gain stage. It also simplifies the gain transfer function because a direct signal path through the feedback network is disabled. The feedback capacitor C_f is sometimes connected directly behind the preamplifier to avoid extra phase shifts introduced by the buffer. Summing currents at the input leads to

$$i_\mathrm{in} = sC_\mathrm{det} v_\mathrm{in} + (v_\mathrm{in} - v_\mathrm{out}) \left(\frac{1}{R_\mathrm{f}} + sC_\mathrm{f} \right) .$$

The elimination of v_{in} by (3.24) results in

$$-\frac{v_{\mathrm{out}}}{i_{\mathrm{in}}} = \frac{v_0}{\frac{1+v_0}{R_{\mathrm{f}}} + s\left[\frac{1}{\omega_0 R_{\mathrm{f}}} + C_{\mathrm{det}} + (1+v_0)C_{\mathrm{f}}\right] + s^2\frac{C_{\mathrm{det}}+C_{\mathrm{f}}}{\omega_0}}. \tag{3.26}$$

As explained in Sect. 3.1.3.2, the effective input capacitance of the preamplifier should be significantly larger than the sensor capacitance. Furthermore, a realistic amplifier has a gain much larger than unity. C_{f} should be usually much smaller than C_{det} so as to achieve a high charge gain. With the conditions

$$v_0\, C_{\mathrm{f}} \gg C_{\mathrm{det}} \gg C_{\mathrm{f}} \quad \text{and} \quad v_0 \gg 1$$

and using (3.25), expression (3.26) simplifies to

$$H(s) = -\frac{v_{\mathrm{out}}}{i_{\mathrm{in}}} \approx \frac{R_{\mathrm{f}}}{1 + s\left(\frac{C_{\mathrm{o}}}{g_{\mathrm{m}}} + R_{\mathrm{f}}C_{\mathrm{f}}\right) + s^2\frac{R_{\mathrm{f}}C_{\mathrm{det}}C_{\mathrm{o}}}{g_{\mathrm{m}}}}. \tag{3.27}$$

This transfer function describes how a current signal of frequency $s = \mathrm{i}\omega$ at the input is converted to a voltage signal at the output of the preamplifier. For DC-signals ($s = 0$), the gain is simply R_{f}.

The term $R_{\mathrm{f}}C_{\mathrm{f}}$ is the discharge time constant τ_{f} of the feedback capacitor. It should be larger than the rise time of the output signal τ_{r} in most applications as explained in Sect. 3.1.3. τ_{r} is the reciprocal of the closed loop bandwidth which is given by the unity gain bandwidth of the amplifier $g_{\mathrm{m}}/C_{\mathrm{o}}$ multiplied by the feedback factor $C_{\mathrm{f}}/C_{\mathrm{det}}$. It is therefore reasonable to use

$$R_{\mathrm{f}}C_{\mathrm{f}} \gtrsim \tau_{\mathrm{r}} = \frac{C_{\mathrm{o}}}{g_{\mathrm{m}}}\frac{C_{\mathrm{det}}}{C_{\mathrm{f}}} \gg \frac{C_{\mathrm{o}}}{g_{\mathrm{m}}}$$

so that (3.27) simplifies to

$$H(s) = -\frac{v_{\mathrm{out}}}{i_{\mathrm{in}}} \approx \frac{R_{\mathrm{f}}}{1 + as + bs^2} = \frac{R_{\mathrm{f}}}{(1 + s/\omega_{\mathrm{l}})\,(1 + s/\omega_{\mathrm{h}})}\ , \tag{3.28}$$

with

$$a = R_{\mathrm{f}}C_{\mathrm{f}} \quad \text{and} \quad b = \frac{R_{\mathrm{f}}C_{\mathrm{det}}C_{\mathrm{o}}}{g_{\mathrm{m}}}. \tag{3.29}$$

The two poles of the transfer function at ω_{l} and ω_{h} have been written down explicitly in (3.28). They can be determined from a and b. Simple expressions are obtained if the low-frequency pole ω_{l} and the high-frequency pole ω_{h} are widely separated, i.e. $\omega_{\mathrm{h}} \gg \omega_{\mathrm{l}}$, so that

$$\left(1 + \frac{s}{\omega_{\mathrm{l}}}\right)\left(1 + \frac{s}{\omega_{\mathrm{h}}}\right) = 1 + s\left(\frac{1}{\omega_{\mathrm{l}}} + \frac{1}{\omega_{\mathrm{h}}}\right) + \frac{s^2}{\omega_{\mathrm{l}}\omega_{\mathrm{h}}} \approx 1 + \frac{s}{\omega_{\mathrm{l}}} + \frac{s^2}{\omega_{\mathrm{l}}\omega_{\mathrm{h}}}$$

and therefore

$$\omega_{\mathrm{l}} \approx \frac{1}{a} = \frac{1}{R_{\mathrm{f}}C_{\mathrm{f}}} = \frac{1}{\tau_{\mathrm{f}}} \quad \text{and} \quad \omega_{\mathrm{h}} \approx \frac{a}{b} = \frac{g_{\mathrm{m}}}{C_{\mathrm{o}}}\frac{C_{\mathrm{f}}}{C_{\mathrm{det}}} = \frac{1}{\tau_{\mathrm{r}}}. \tag{3.30}$$

3.3.5.2 Noise from Sensor Leakage

The sensor leakage current I_{leak} has a white noise spectrum with a spectral density (with units $\text{A}^2/\,\text{Hz}$) of [235]

$$\frac{\mathrm{d}\langle i^2_{\text{leak}}\rangle}{\mathrm{d}f} = 2qI_{\text{leak}}. \tag{3.31}$$

This input noise spectrum is filtered by the transfer function (3.28) of the amplifier and so the squared rms noise at the output is obtained by integration of the filtered noise over all frequencies:

$$\begin{aligned}\langle v^2_{\text{out}}\rangle_{\text{leak}} &= \int_0^\infty |H(s)|^2 \,\frac{\mathrm{d}\langle i^2_{\text{leak}}\rangle}{\mathrm{d}f}\,\mathrm{d}f = 2qI_{\text{leak}}R_{\text{f}}^2 \int_0^\infty \left|\frac{1}{1+as+bs^2}\right|^2 \mathrm{d}f \\ &= 2qI_{\text{leak}}R_{\text{f}}^2\,\frac{1}{4a} = \frac{qI_{\text{leak}}}{2}\,\frac{R_{\text{f}}}{C_{\text{f}}}.\end{aligned} \tag{3.32}$$

The integral can be solved in the complex plane. This noise voltage is usually referred back to the input by division by the output signal q/C_{f} of a single electron (the shaping loss is small in this example). The resulting ENC characterizes the system by the fluctuation of the input charge (in electrons) required to cause the observed voltage noise at the output. It can be compared directly with the signal charge to determine the signal-to-noise ratio. The ENC caused by a leakage current I_{leak} in this simple system is

$$\text{ENC}_{\text{leak}} = \frac{C_{\text{f}}}{q}\sqrt{\langle v^2_{\text{out}}\rangle_{\text{leak}}} = \sqrt{\frac{I_{\text{leak}}}{2q}\tau_{\text{f}}} = 56\ \text{e}^- \times \sqrt{\frac{I_{\text{leak}}}{\text{nA}}\,\frac{\tau_{\text{f}}}{\mu\text{s}}}. \tag{3.33}$$

Example: A leakage current of $I_{\text{l}} = 5\,\text{nA}$ at a fall time of $\tau_{\text{f}} = 500$ ns leads to an ENC of 88 electrons.

3.3.5.3 Transistor Channel Noise

The white thermal noise in the channel of the input transistor leads to an equivalent white noise voltage source at its gate with a white spectral density of [235]

$$\frac{\mathrm{d}\langle v^2_{\text{therm}}\rangle}{\mathrm{d}f} = \frac{8}{3}\,\frac{kT}{g_{\text{m}}}. \tag{3.34}$$

A small signal serial voltage v at the input leads to an input current v/Z_{in} where the lumped input impedance Z_{in} contains the parallel connection of the detector, feedback, stray, and gate capacitance of the input device. This can be verified by explicitly writing down the transfer function for a serial voltage source (see also Sect. 3.3.6.2). Z_{in} is dominated by the sensor capacitance in this example and so the equivalent input noise current is $\langle i^2_{\text{therm}}\rangle \approx \langle v^2_{\text{therm}}\rangle|(sC_{\text{det}})^2|$. The output noise becomes

$$\langle v_{\mathrm{out}}^2\rangle_{\mathrm{therm}} = \int_0^\infty |H(s)|^2 \frac{\mathrm{d}\langle i_{\mathrm{therm}}^2\rangle}{\mathrm{d}f}\,\mathrm{d}f = \frac{8}{3}\frac{kT}{g_{\mathrm{m}}} C_{\mathrm{det}}^2 R_{\mathrm{f}}^2 \int_0^\infty \left|\frac{s}{1+as+bs^2}\right|^2 \mathrm{d}f$$
$$= \frac{8}{3}\frac{kT}{g_{\mathrm{m}}} C_{\mathrm{det}}^2 R_{\mathrm{f}}^2 \frac{1}{4ab} = \frac{2}{3} kT \frac{C_{\mathrm{det}}}{C_{\mathrm{f}} C_{\mathrm{o}}}$$

and so

$$\mathrm{ENC}_{\mathrm{therm}} = \frac{C_{\mathrm{f}}}{q}\sqrt{\langle v_{\mathrm{out}}^2\rangle_{\mathrm{therm}}} = \sqrt{\frac{kT}{q}\frac{2C_{\mathrm{det}}}{3q}\frac{C_{\mathrm{f}}}{C_{\mathrm{o}}}}$$
$$= 104\ \mathrm{e}^- \times \sqrt{\frac{C_{\mathrm{det}}}{100\ \mathrm{fF}}\frac{C_{\mathrm{f}}}{C_{\mathrm{o}}}}. \tag{3.35}$$

This expression does not contain the transconductance of the input transistor because the decrease in noise for higher g_{m} is cancelled by the increase of the bandwidth. The situation is different when a shaper limits the system bandwidth independently of g_{m}. An increase of the transistor current in that case would lead to a lower noise (see Sect. 3.3.6).

Example: For a detector capacitance of $C_{\mathrm{det}} = 500$ fF, a feedback capacitor $C_{\mathrm{f}} = 5$ fF, and a load capacitor $C_{\mathrm{o}} = 50$ fF at the bandwidth-limiting output node of the amplifier, the ENC due to the transistor channel noise is 74 electrons. With an input device with a transconductance of $g_{\mathrm{m}} = 500\ \mu$S, the rise time would be $\tau_{\mathrm{r}} = (C_{\mathrm{o}}/g_{\mathrm{m}} \cdot C_{\mathrm{det}}/C_{\mathrm{f}}) = 10$ ns.

3.3.5.4 Transistor $1/f$-Noise

Various expressions are used to describe the $1/f$-noise in MOS devices. For instance, the $1/f$-noise voltage (in V^2/Hz) at the gate of a FET operated in strong inversion can be modeled by the expression [232]

$$\frac{\mathrm{d}\langle v_{1/f}^2\rangle}{\mathrm{d}f} = \frac{K_{\mathrm{f}}}{C_{\mathrm{ox}} W L}\frac{1}{f}, \tag{3.36}$$

with K_{f} being the device- and technology-dependent constant. C_{ox}, W, and L are the gate oxide capacitance per unit area and the (effective) transistor width and length, respectively. This expression is also used in some SPICE implementations for circuit simulation (in PSPICE for instance with switch NLEV = 2 and AF = 1).[4] The noise density depends on the frequency f and is largest at low frequencies. As before, the output noise is

[4] Another common parametrization of $1/f$-noise uses C_{ox}^2 in the denominator and a coefficient usually named K_{a}

$$\begin{aligned}\langle v_{\mathrm{out}}^2\rangle_{1/f} &= \frac{K_{\mathrm{f}}\, C_{\mathrm{det}}^2\, R_{\mathrm{f}}^2}{C_{\mathrm{ox}} W L} \int\limits_0^\infty \left|\frac{s}{1+as+bs^2}\right|^2 \frac{\mathrm{d}f}{f} \\ &\approx \frac{K_{\mathrm{f}}\, C_{\mathrm{det}}^2\, R_{\mathrm{f}}^2}{C_{\mathrm{ox}} W L} \frac{a^2}{a^4-b^2}\ \ln\left(\frac{a^2}{b}\right) \\ &\approx \frac{K_{\mathrm{f}}}{C_{\mathrm{ox}} W L} \frac{C_{\mathrm{det}}^2}{C_{\mathrm{f}}^2} \ln\left(\tau_{\mathrm{f}} \frac{g_{\mathrm{m}}}{C_{\mathrm{o}}} \frac{C_{\mathrm{f}}}{C_{\mathrm{det}}}\right).\end{aligned}$$

The approximation of widely separated poles has again been used in this calculation and some approximations have been made. The ENC becomes

$$\mathrm{ENC}_{1/f} \approx \frac{C_{\mathrm{det}}}{q} \sqrt{\frac{K_{\mathrm{f}}}{C_{\mathrm{ox}} W L}} \sqrt{\ln\left(\tau_{\mathrm{f}} \frac{g_{\mathrm{m}}}{C_{\mathrm{o}}} \frac{C_{\mathrm{f}}}{C_{\mathrm{det}}}\right)}. \tag{3.37}$$

Example: The expression under the logarithm is 50 with the values used in the previous section, and so the last term becomes ≈2. In a 0.25 μm technology with $C_{\mathrm{ox}} = 6.4\ \mathrm{fF}/\mu\mathrm{m}^2$ and $K_{\mathrm{f}} = 33 \times 10^{-25}\,\mathrm{J}$ for an NMOS device [232] and a transistor with $W = 20\ \mu$m and $L = 0.5\ \mu$m, the ENC is

$$\mathrm{ENC}_{1/f} \approx 9\ \mathrm{e}^- \times \frac{C_{\mathrm{det}}}{100\,\mathrm{fF}} \quad \text{(NMOS input device)}. \tag{3.38}$$

The $1/f$-noise contribution for the presented example with $C_{\mathrm{det}} = 500$ fF is therefore ≈45 electrons. This value can be decreased in particular by using a PMOS input device with a 27 times smaller K_{f} [232] or, for instance, by increasing the input transistor size.

3.3.5.5 Thermal Noise of the Feedback Resistor

The white thermal noise current

$$\frac{\mathrm{d}\langle i_{R\mathrm{f}}^2\rangle}{\mathrm{d}f} = \frac{4kT}{R_{\mathrm{f}}} \tag{3.39}$$

of the feedback resistor has the same effect on $\langle v_{\mathrm{out}}^2\rangle$ as noise from the sensor leakage current, and so the term qI_{leak} in (3.32) can be simply replaced by $2kT/R_{\mathrm{f}}$, yielding

$$\langle v_{\mathrm{out}}^2\rangle_{R\mathrm{f}} = \frac{kT}{C_{\mathrm{f}}}.$$

The ENC of

$$\mathrm{ENC}_{R\mathrm{f}} = \frac{C_{\mathrm{f}}}{q} \sqrt{\frac{kT}{C_{\mathrm{f}}}} \approx 13\ \mathrm{e}^- \times \sqrt{\frac{C_{\mathrm{f}}}{\mathrm{fF}}} \tag{3.40}$$

is relatively small (29 e^- for the values above). It does not depend on the value of the feedback resistor in this particular case, because the increase of the thermal noise for smaller resistor values is cancelled by the simultaneous reduction of the integration time.

3.3.5.6 Summary

The approximative calculations of this section have illustrated some general facts:

- The noise is determined by the frequency-dependent transfer function of the system and by the magnitude of individual primary noise contributions.
- Noise can be reduced by manipulating the transfer function, for instance, by a shaping amplifier as it will be described in the following section.
- The fundamental noise due to sensor leakage current increases in slower systems.
- The contributions from transistor thermal channel noise and from $1/f$-noise become larger with increasing detector capacitance.
- $1/f$-noise is often small compared to other sources. It becomes important, however, if very small output noise must be achieved.
- For pixel detectors, ENC values in the order of 100 electrons can be achieved without too much effort for sensors with several 100 fF capacitance and input devices with moderate g_{m}.

The effect of a shaper is discussed in some detail in the next section.

3.3.6 Noise in Charge-Sensitive Amplifier/Shaper Combination

The noise at the output of the amplifying chain can be decreased by the reduction of the bandwidth with appropriate filters, commonly referred to as "shapers." The system depicted in Fig. 3.29 consisting of a preamplifier and a "semi-Gaussian" shaper modeled by a combination of N high-pass and M low-pass stages is treated here as a simple example [241]. The effect of a feedback circuit is neglected. An ideal, buffered preamplifier is assumed, a realistic assumption as the shaper is usually the bandwidth-limiting element. The case of a simple *CR-RC*-shaper is studied in more detail. A MOS input device and a bipolar input transistor are considered for comparison.

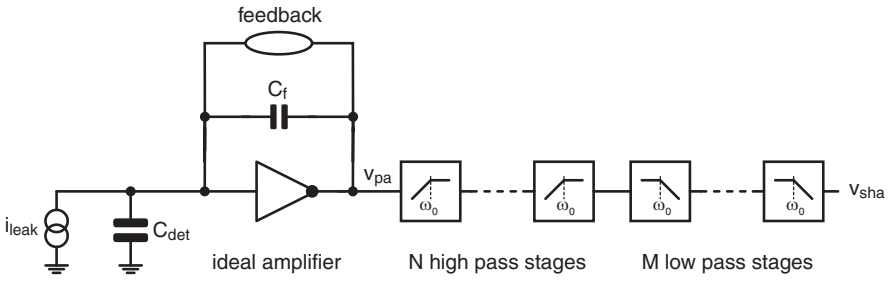


Fig. 3.29. Ideal charge preamplifier followed by N high-pass and M low-pass stages for signal filtering

3.3.6.1 Pulse Shapes after the Shaper

The output signal of the preamplifier can be approximated by a step function with amplitude $U = Q_{\mathrm{in}}/C_{\mathrm{f}}$ if the rise time of the amplifier is small and the discharge time constant $C_{\mathrm{f}}R_{\mathrm{f}}$ is large compared to the filter time constant of the shaper. The shaper is described by a Laplace transform $\mathcal{L}_{\mathrm{HP}} = s\tau(1 + s\tau)^{-1}$ for each of the N high-pass stages and by $\mathcal{L}_{\mathrm{LP}} = (1 + s\tau)^{-1}$ for each of the of M low-pass stages, where $s = \mathrm{i}\omega$ is the complex frequency variable. The filter time constant $\tau = 1/\omega_0$ is the reciprocal of the corner frequency ω_0. It is assumed identical for all high- and low-pass sections for simplicity. The complete Laplace transform of the output signal of the (N, M)-shaper with a unity step as an input signal (with Laplace transform $\mathcal{L}_{\mathrm{Step}} = 1/s$) therefore is

$$\mathcal{L}^{(N,M)}(s) = \frac{1}{s}\left(\frac{s\tau}{1+s\tau}\right)^N \left(\frac{1}{1+s\tau}\right)^M = \frac{\tau^N s^{N-1}}{(1+s\tau)^{N+M}}. \tag{3.41}$$

The output signal in the time domain is calculated as the inverse Laplace transform which can be carried out by determining the residuum of the $(N + M)$-fold pole at $s = -1/\tau$.

$$\begin{aligned} f^{(N,M)}(t) &= \mathrm{Res}\left.\frac{\tau^N s^{N-1}\,\mathrm{e}^{st}}{(1+s\tau)^{N+M}}\right|_{\mathrm{s}=-1/\tau} \\ &= \frac{\tau^N}{(N+M-1)!}\lim_{s\to -1/\tau}\frac{\mathrm{d}^{N+M-1}}{\mathrm{d}s^{N+M-1}}\left[\frac{s^{N-1}\,\mathrm{e}^{st}}{(1+s\tau)^{N+M}}\left(s+\frac{1}{\tau}\right)^{N+M}\right] \\ &= \frac{1}{(N+M-1)!}\left(\frac{t}{\tau}\right)^M \sum_{i=0}^{\infty}\frac{(-t/\tau)^i}{i!}\frac{(M+i+N-1)!}{(M+i)!}\,. \end{aligned} \tag{3.42}$$

For a shaper with only one high-pass section ($N = 1$), this simplifies to

$$f^{(1,M)}(t) = \frac{1}{M!}\left(\frac{t}{\tau}\right)^M \mathrm{e}^{-t/\tau} \tag{3.43}$$

with peaking time and maximum amplitude of

$$t_{\mathrm{peak}}^{(1,M)} = M\tau = \frac{M}{\omega_0} \quad \text{and} \quad f_{\mathrm{max}}^{(1,M)} = \frac{1}{M!}\left(\frac{M}{e}\right)^M. \tag{3.44}$$

For the simple CR–RC-filter with only one low pass ($M = 1$) this becomes

$$f^{(1,1)}(t) = \left(\frac{t}{\tau}\right)\mathrm{e}^{-t/\tau} \tag{3.45}$$

with peaking time and maximum amplitude of

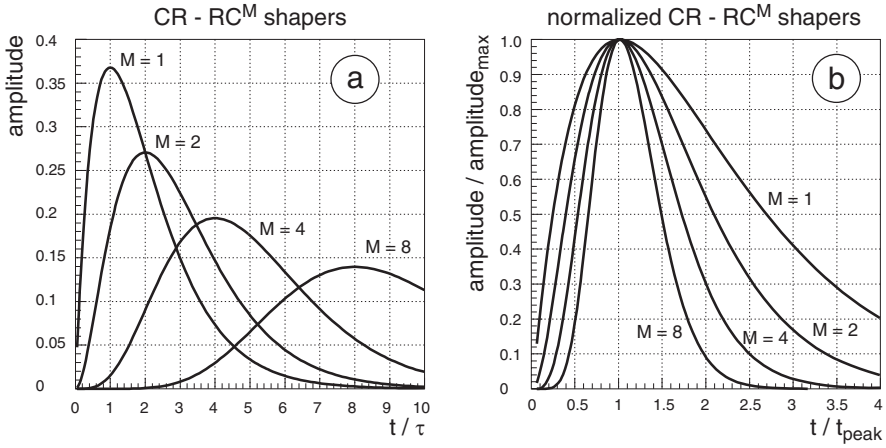


Fig. 3.30. Step responses of CR–RC^M-shapers for $M = 1, 2, 4, 8$ (**a**) and pulse shapes normalized to same peak amplitudes and times (**b**)

$$t_{\text{peak}}^{(1,1)} = \tau = \frac{1}{\omega_0} \quad \text{and} \quad f_{\max}^{(1,1)} = \frac{1}{e}. \tag{3.46}$$

Figure 3.30a shows $f^{(1,M)}(t)$ according to (3.43). The shape of the pulses is compared in Fig. 3.30b where all curves are normalized to the same peak amplitude and peaking time. This requires the corner frequencies to be increased by a factor of M. The width of the normalized pulses decreases with increasing M and so higher order shapers are better suited if good double pulse resolution is required. An increase of M above 4 does not shrink the pulse width very much further. For a shaper with two high-pass sections ($N = 2$), expression (3.42) becomes

$$f^{(2,M)}(t) = f^{(1,M)} \left(1 - \frac{t/\tau}{M+1}\right) . \tag{3.47}$$

This function crosses zero at $t = (M + 1)\tau$ and has a negative undershoot (Fig. 3.31). This behavior is often unwanted and so additional (parasitic) high-pass stages must be avoided.

3.3.6.2 Total Noise

The noise of the amplifier is modeled by a serial noise voltage source and a parallel noise current source at its input as shown in Fig. 3.32. The noise sources are characterized by the frequency spectrum of the (squared) rms voltage $\langle v^2(f)\rangle$ or current $\langle i^2(f)\rangle$. White noise sources have frequency-independent noise while the spectrum of $1/f$-noise sources increases at low frequencies. In the following calculation, the parameterized spectral noise densities

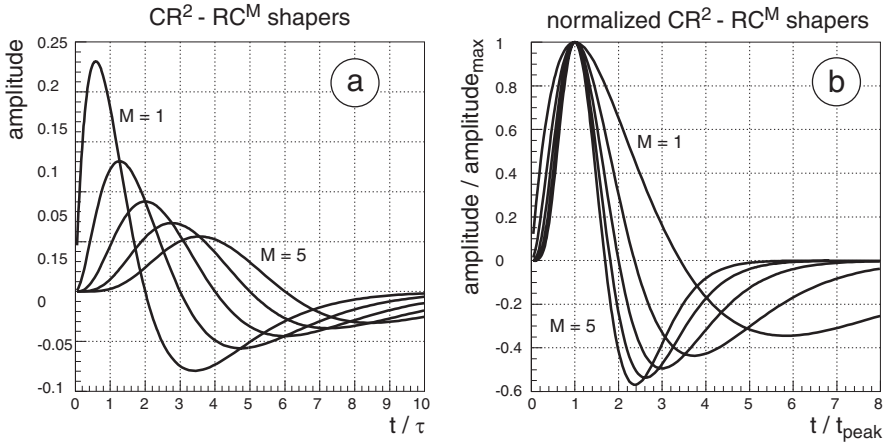


Fig. 3.31. Step responses of CR^2–RC^M-shapers for $M = 1$–5 according to (3.47) (**a**) and normalized to the same peaking amplitude and time (**b**)

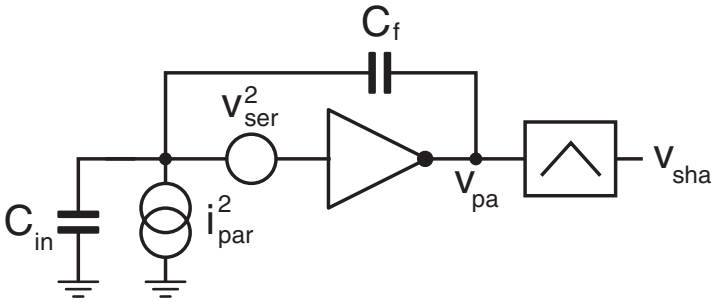


Fig. 3.32. Charge-sensitive preamplifier and shaper with input noise sources and capacitive input load

$$\text{Serial noise voltage:} \quad \frac{\mathrm{d}\langle v^2(f)\rangle}{\mathrm{d}f} = V_0 + V_{-1}f^{-1} \tag{3.48a}$$

$$\text{Parallel noise current:} \quad \frac{\mathrm{d}\langle i^2(f)\rangle}{\mathrm{d}f} = I_0 \tag{3.48b}$$

are used. V_0 and I_0 are the coefficients for the white noise contributions and V_{-1} is for the $1/f$-noise. Specific values for these parameters are given later in (3.57a) and (3.59a).

The effect of the two noise sources on the preamplifier output voltage can be calculated separately if they are uncorrelated: The parallel noise current must flow through the feedback capacitor assuming a perfect virtual ground at the input of the amplifier so that

$$\text{Parallel noise}: \quad \frac{\mathrm{d}\langle v_{\mathrm{pa}}^2(\omega)\rangle}{\mathrm{d}\omega} = \frac{\mathrm{d}\langle i_{\mathrm{par}}^2(\omega)\rangle}{\mathrm{d}\omega}\frac{1}{(\omega C_{\mathrm{f}})^2} = \frac{I_0}{2\pi}\frac{1}{(\omega C_{\mathrm{f}})^2}, \quad (3.49)$$

where the spectrum has been expressed as a function of the angular frequency $\omega = 2\pi f$. The serial noise voltage is related to the preamplifier output voltage through the capacitive divider made up of C_{f} and $C_{\mathrm{in}} = C_{\mathrm{det}} + C_{\mathrm{parasitic}} + C_{\mathrm{preamp}}$ by

$$v_{\mathrm{pa}}^2 = v_{\mathrm{ser}}^2 \left(\frac{C_{\mathrm{in}} + C_{\mathrm{f}}}{C_{\mathrm{f}}}\right)^2 .$$

Under the realistic assumption that $C_{\mathrm{f}} \ll C_{\mathrm{in}}$, the output noise spectrum from this noise contribution becomes

$$\text{Serial noise}: \quad \frac{\mathrm{d}\langle v_{\mathrm{pa}}^2(\omega)\rangle}{\mathrm{d}\omega} \approx \left(V_{-1}\omega^{-1} + \frac{V_0}{2\pi}\right)\left(\frac{C_{\mathrm{in}}}{C_{\mathrm{f}}}\right)^2 . \quad (3.50)$$

Equations (3.49) and (3.50) can be combined to

$$\frac{\mathrm{d}\langle v_{\mathrm{pa}}^2(\omega)\rangle}{\mathrm{d}\omega} = \sum_{k=-2}^{0} c_k \omega^k, \quad (3.51)$$

with

$$c_{-2} = \frac{I_0}{2\pi C_{\mathrm{f}}^2}, \quad c_{-1} = V_{-1}\frac{C_{\mathrm{in}}^2}{C_{\mathrm{f}}^2}, \quad \text{and} \quad c_0 = V_0\frac{C_{\mathrm{in}}^2}{2\pi C_{\mathrm{f}}^2}. \quad (3.52)$$

The preamplifier output $v_{\mathrm{pa}}^2(\omega)$ is filtered by a shaper with N differentiating high-pass stages and M integrating low-pass stages with identical corner frequencies ω_0. The squared transfer function of such a shaper with gain A is given by

$$H_{N,M}^2(\omega) = A^2 \frac{(\omega/\omega_0)^{2N}}{[1 + (\omega/\omega_0)^2]^{N+M}}. \quad (3.53)$$

The total squared shaper output voltage becomes

$$\begin{aligned}\langle v_{\mathrm{sha}}^2\rangle &= \int_0^\infty H_{N,M}^2(\omega)\,\mathrm{d}\langle v_{\mathrm{pa}}^2(\omega)\rangle = \sum_{k=-2}^{0}\int_0^\infty c_k\,\omega^k H_{N,M}^2(\omega)\,\mathrm{d}\omega \\ &= \frac{A^2}{2}\frac{1}{\Gamma(N+M)}\sum_{k=-2}^{0} c_k\,\omega_0^{k+1}\Gamma\left(N+\frac{k+1}{2}\right)\Gamma\left(M-\frac{k+1}{2}\right), \quad (3.54)\end{aligned}$$

where Γ is the gamma function with $\Gamma(x+1) = x\Gamma(x)$, $\Gamma(1) = 1$, and $\Gamma(1/2) = \sqrt{\pi}$.

3.3.6.3 Equivalent Noise Charge with a Simple *CR–RC*-Shaper

For the simple CR–RC-shaper with $N = M = 1$, (3.54) simplifies to

$$\langle v_{\mathrm{sha}}^2 \rangle = A^2 \frac{\pi}{4} \left(\frac{c_{-2}}{\omega_0} + \frac{2}{\pi} c_{-1} + \omega_0 c_0 \right) . \tag{3.55}$$

This noise voltage at the output must be normalized to a typical signal. If the peak output signal amplitude for an input charge of a single electron $V_{\mathrm{max}} = \frac{1}{e} A \frac{q}{C_{\mathrm{f}}}$ from (3.46) is used, the ENC referred back to the input (the "ENC") becomes

$$\mathrm{ENC}_{\mathrm{CR-RC}}^2 = \frac{\langle v_{\mathrm{sha}}^2 \rangle}{V_{\mathrm{max}}^2} = \frac{e^2}{4q^2} \left(\frac{\tau}{2} I_0 + \frac{1}{2\tau} V_0 C_{\mathrm{in}}^2 + 2\, V_{-1} C_{\mathrm{in}}^2 \right) , \tag{3.56}$$

where the definitions from (3.52) have been used and the corner frequency of the bandpass filter ω_0 has been replaced by the reciprocal of the peaking time τ according to (3.46). (Note that this relation is valid only for $N = M = 1$.)

3.3.7 FET Preamplifier

For an amplifier with a FET input stage, the three contributions in (3.56) are

$$\text{From leakage current } I_{\mathrm{leak}}: \qquad I_0 = 2qI_{\mathrm{leak}} , \tag{3.57a}$$

$$\text{From transistor channel noise:} \qquad V_0 = \frac{8}{3} \frac{kT}{g_{\mathrm{m}}} , \tag{3.57b}$$

$$\text{From } 1/f\text{-noise:} \qquad V_{-1} = \frac{K_{\mathrm{f}}}{C_{\mathrm{ox}} WL} . \tag{3.57c}$$

For the 0.25 μm technology used in Sect. 3.3.5 with $C_{\mathrm{ox}} = 6.4$ fF/μm^2, $K_{\mathrm{f}} = 33 \times 10^{-25}$ J, and the same NMOS input device with $L = 0.5$ μm and $W = 20$ μm, the ENC becomes

$$\left(\frac{\mathrm{ENC}}{\mathrm{e}^-} \right)^2 = 115 \, \frac{\tau}{10\,\mathrm{ns}} \, \frac{I_{\mathrm{leak}}}{1\,\mathrm{nA}} \tag{3.58a}$$

$$+ 388 \, \frac{10\,\mathrm{ns}}{\tau} \, \frac{\mathrm{mS}}{g_{\mathrm{m}}} \left(\frac{C_{\mathrm{in}}}{100\,\mathrm{fF}} \right)^2 \tag{3.58b}$$

$$+ 74 \left(\frac{C_{\mathrm{in}}}{100\,\mathrm{fF}} \right)^2 \tag{3.58c}$$

at room temperature. For a sensor with $C_{\mathrm{in}} = 200$ fF, a leakage current of $I_{\mathrm{leak}} = 1$ nA, a shaper peaking time of $\tau = 50$ ns, and a transconductance of 0.5 mS in the NMOS input transistor, the three contributions to ENC^2 are

575, 621, and 296, respectively, and so the total theoretical ENC becomes $\mathrm{ENC} = 40\,\mathrm{e}^-$.

The sensor must have a small leakage current and a small capacitance in order to reduce the noise. Note that other noise contributions like the intrinsic input capacitance of the amplifier due to its gate capacitance, stray capacitances, etc., have been neglected in this treatment and so the above result presents only a lower limit. The dominant contribution in this example is from white channel noise which can be reduced with more current in the device. A reduction of L also leads to an increase of g_m but short channel effects may then worsen the $1/f$-contribution. An increase in the input transistor width W also helps g_m, but the effective input capacitance is also increased. This capacitance adds to C_det so that an optimum W can be found for a given detector capacitance [241]. Using a PMOS input device with a smaller K_f decreases the $1/f$-noise, but the transconductance drops and so this would not be a good choice in this example. The contribution of the leakage current is independent of the amplifier details and depends only on the shaping time. A shorter shaping accumulates less leakage current noise, but unfortunately, the channel noise contribution increases in that case.

3.3.8 Bipolar Amplifier

A bipolar transistor can be characterized by its current gain $\beta = I_\mathrm{c}/I_\mathrm{b}$, where I_c and I_b are collector and base current, respectively. The transconductance is simply given by $g_\mathrm{m} = qI_\mathrm{c}/kT$ [235]. It is assumed here for simplicity that the detector leakage current is small compared to the base current, that the contribution of a feedback network can be neglected, and that thermal noise from the base-spreading resistance r_b can be ignored. (This is the case when $r_\mathrm{b} \ll 1/2g_\mathrm{m}$, which is usually the case for small collector currents.) $1/f$-noise is small in bipolar devices and is therefore neglected here ($V_{-1} = 0$). A more detailed treatment of noise in a bipolar charge amplifier can be found for instance in [242, 243]. The most important noise contributions in (3.56) are

$$\text{From base shot noise:} \qquad I_0 = 2qI_\mathrm{b} = \frac{2qI_\mathrm{c}}{\beta} \tag{3.59a}$$

$$\text{From transistor channel noise:} \qquad V_0 = \frac{2qI_\mathrm{c}}{g_\mathrm{m}^2} = \frac{2q}{I_\mathrm{c}}\left(\frac{kT}{q}\right)^2 \tag{3.59b}$$

and so

$$\mathrm{ENC}^2(I_\mathrm{c}) = \frac{e^2}{4q}\left[\frac{I_\mathrm{c}\tau}{\beta} + \frac{C_\mathrm{in}^2}{I_\mathrm{c}\tau}\left(\frac{kT}{q}\right)^2\right]. \tag{3.60}$$

The ENC becomes a linear function of the detector capacitance for large C_in as before. For a given C_in, the noise increases for small collector currents because the current noise referred back to the input becomes dominant. The base current shot noise, on the other hand, contributes significantly at large

collector currents. This behavior is particular for the bipolar input device where an optimum collector current

$$I_{\mathrm{c,opt}}(\tau) = \frac{kT}{q} \times \sqrt{\beta} \times \frac{1}{\tau} \times C_{\mathrm{in}} \approx 0.25\ \mu\mathrm{A} \times \sqrt{\beta} \times \frac{10\ \mathrm{ns}}{\tau} \times \frac{C_{\mathrm{in}}}{100\ \mathrm{fF}} \quad (3.61)$$

exists. More current is needed for faster shaping and larger input capacitance. Both noise sources give the same contribution for this optimal collector current. The resulting optimal total noise is independent of τ:

$$\mathrm{ENC}^2_{\mathrm{opt}} = \mathrm{ENC}^2(I_{\mathrm{c,opt}}) = \frac{\mathrm{e}^2}{2}\frac{kT}{q}\frac{C_{\mathrm{in}}}{q\sqrt{\beta}} \quad \text{for} \quad N = M = 1 \quad (3.62)$$

or

$$\mathrm{ENC}_{\mathrm{opt}} \approx \frac{115\ \mathrm{e}^-}{\sqrt[4]{\beta}}\sqrt{\frac{C_{\mathrm{in}}}{100\ \mathrm{fF}}}\ . \quad (3.63)$$

For example, for $C_{\mathrm{in}} = 400$ fF, a peaking time of 100 ns, and $\beta = 100$, the best possible noise of $\approx$75 e^- requires a collector current of $\approx$1 μA. Smaller noise values are not possible with this bipolar transistor and with CR–RC shaping.

3.3.8.1 Influence of the Collector Current

The noise varies only slowly with I_{c} and so a smaller collector current is often chosen [242] in order to reduce the power dissipation of the preamplifier. The resulting noise increase is shown in Fig. 3.33a where the excess noise factor $E = \mathrm{ENC}/\mathrm{ENC}_{\mathrm{opt}}$ is plotted as a function of $x = I_{\mathrm{c}}/I_{\mathrm{c,opt}}$. Equation (3.60) becomes

$$E = \frac{\mathrm{ENC}}{\mathrm{ENC}_{\mathrm{opt}}} = \sqrt{\frac{x + 1/x}{2}} \quad \text{with} \quad x = \frac{I_{\mathrm{c}}}{I_{\mathrm{c,opt}}}\ . \quad (3.64)$$

The current saving can be calculated by solving the above expression for x for a given excess noise factor $E > 1$. The resulting expression $x = E^2 - \sqrt{E^4 - 1}$ is plotted in Fig. 3.33b. A reduction of the power consumption by 50% leads to a noise increase of only 11%.

3.3.8.2 Effect of Higher Order Shapers for Bipolar Input

The output noise can be reduced by removing more high-frequency components with additional low-pass filters. Evaluation of (3.54) for $N = 1$ and general M with the bipolar noise coefficients from (3.52) and (3.59a),(3.59b) yields

$$\langle v^2_{\mathrm{sha}} \rangle = \frac{A^2\sqrt{\pi}}{2\pi C_{\mathrm{f}}^2}\frac{\Gamma\left(M - \frac{1}{2}\right)}{\Gamma\left(M + 1\right)}\left[\frac{qI_{\mathrm{c}}}{\beta\omega_0}\left(M - \frac{1}{2}\right) + \frac{1}{2}\frac{(kT)^2}{qI_{\mathrm{c}}}C_{\mathrm{in}}^2\omega_0\right]. \quad (3.65)$$

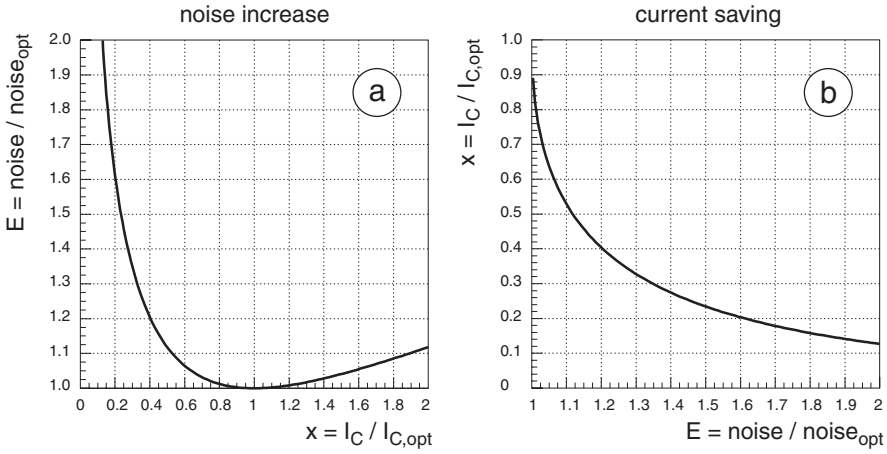


Fig. 3.33. Noise increase as a function of the normalized collector current in the bipolar input transistor (**a**) as given by (3.64), and current saving to achieve a given excess noise (**b**), calculated using the same expression

After normalization to the peak signal amplitude (3.44) for an input charge of ~~1~~ Λ electron

$$V_{\max}^{(1,M)} = A\frac{q}{C_\mathrm{f}}\frac{1}{M!}\left(\frac{M}{e}\right)^M$$

the ENC can be compared to the optimum ENC value of the simple *CR–RC*-shaper (3.62):

$$E = \frac{\mathrm{ENC}^{(1,M)}}{\mathrm{ENC}_{\mathrm{opt}}^{(1,1)}} = \frac{e^{M-1}\sqrt{(2M)!}}{(2M)^M}\sqrt{x + \frac{1}{x(2M-1)}}\,, \tag{3.66}$$

where $x = I_\mathrm{c}^{(1,M)}/I_{\mathrm{c,opt}}^{(1,1)}$ is the ratio between the collector current and the optimal current for the *CR–RC*-shaper (3.61). The noise is minimal for

$$x_{\min} = \frac{1}{\sqrt{2M-1}}, \quad \text{where} \quad E_{\min} = \frac{e^{M-1}}{(2M)^M}\sqrt{\frac{2(2M)!}{\sqrt{2M-1}}}\,. \tag{3.67}$$

$E_{\min}$ converges to $\frac{\sqrt[4]{8\pi}}{e} \approx 0.825$ for large M. This value is the ultimate noise improvement achievable with high order $CR–RC^M$-shapers. Figure 3.34a shows the excess noise factor E from (3.66) as a function of the normalized collector current for $M = 1$–4 for constant shaper corner frequency ω_0. It seems that low noise values can be achieved with very low currents if the order of the shaper is increased. This apparent improvement is, unfortunately, caused by the linear increase of the shaper peaking time with increasing M. For a *constant* peaking time the center frequency of the shaper must be increased correspondingly. The excess noise factor for the case of a constant

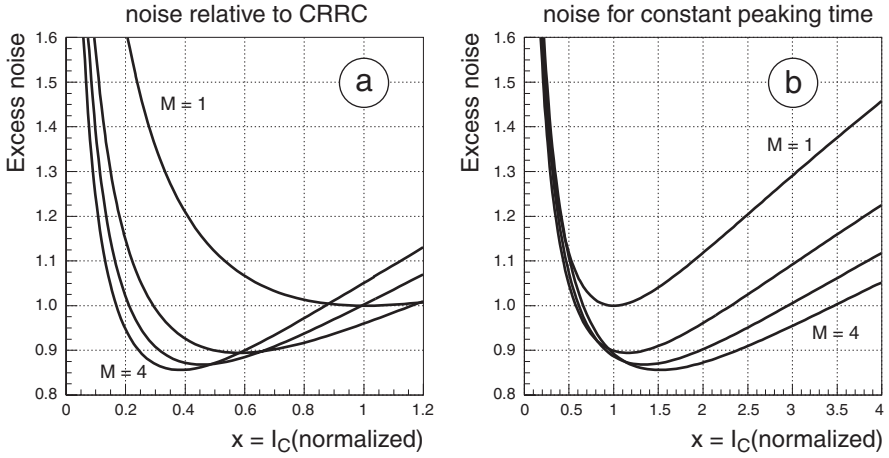


Fig. 3.34. Noise of CR–RC^M-shapers (M = 1–4) normalized to the noise of a CR–RC-shaper. The excess noise factor $E = \mathrm{ENC}^{(1,M)}/\mathrm{ENC}_{\mathrm{opt}}^{(1,1)}$ is plotted as a function of the normalized collector current $x = I_{\mathrm{c}}^{(1,M)}/I_{\mathrm{c,opt}}^{(1,1)}$ for constant corner frequency ω_0 (**a**) and for constant peaking time (**b**)

peaking time is plotted in Fig. 3.34b. It becomes apparent that the collector current must be increased for all higher order shapers in order to improve on the noise.

3.3.9 Summary

The noise of an idealized preamplifier followed by a semi-Gaussian CR^N–RC^M-shaper has been studied. The total rms shaper noise voltage is given by (3.54) as a function of the input noise sources and of the shaper characteristic. Equation (3.56) gives the ENC for the special case of a CR–RC-shaper with $N = M = 1$. The consequences for a FET input device and for a bipolar input transistor have been discussed. The latter case allows a simple optimization of noise for a given detector capacitance by an appropriate choice of the collector current. The two dominating noise contributions in this case are the parallel noise current source due to the base current and a serial noise voltage source from the collector current referred back to the input. The former increases with increasing collector current while the latter decreases so that an optimum collector current $I_{\mathrm{c,opt}}$ can be found, the exact value depending on the type of shaper. The simplest case of a CR–RC-shaper gives an ENC of 75 e^- for a detector capacitance of 400 fF and a bipolar current gain of $\beta = 100$. This value is achieved with a collector current of $I_{\mathrm{c}} \approx 1\ \mu$A for a peaking time of 100 ns. More current is needed for faster shapers with shorter peaking times but the minimal noise value remains the same. Higher

order CR–RC^M-shapers can reduce the noise only by at most ≈15% with slightly increased collector current for a given peaking time.

3.4 Readout Architectures

The digital hit signals of the discriminators must be further processed by circuitry in the pixel and at the chip periphery. The architecture of this readout depends very much on the target application. The counting of the number of hits during a given time interval in every pixel can be sufficient in medical applications. This requires simple counters in the pixels and a mechanism to transfer the counter values to the periphery. More detailed information is required in applications in particle physics. The positions, often also the times and possibly the corresponding pulse amplitudes, of all hits belonging to an interaction must be provided. This requires a timing precision of 25 ns (the bunch crossing interval) for the detectors at LHC. Some experiments (like BTeV at FERMILAB) start the readout of every single event immediately after the interaction. Very often, however, a trigger system selects only a fraction of the events for readout in order to reduce the data volume sent to the DAQ. All hits must be identified and buffered for some time, in this case, because the trigger signal arrives with a significant delay. At the LHC experiments, the trigger latency is in the order of 2–3 μs which corresponds to ≈100 interactions. Almost all architectures perform an immediate zero suppression (i.e. process only pixels with amplitudes above a threshold) to reduce the size of the required buffers. The limited buffer space available can lead to a loss of hits.

The choice of a suited architecture mainly depends on the available chip technology and on the acceptable hit losses which can have very different characteristics for different readout concepts. Detailed simulations of the hit losses are therefore required before a choice can be made. An important decision to be made is whether the analog pulse height information of every hit is required. Some architectures are not suited for an analog readout.

3.4.1 Chips Without Data Buffering

The first pixel readout chips were designed for the relatively low interaction rates of the LEP accelerator so that every event could be read out. Figure 3.35 shows schematically the x–y-scanning scheme [244] which has been implemented in the DELPHI experiment to find the hit pixels in the matrix. The discriminator signals set the hit flip-flop in the pixels. Horizontal "stop"-lines are pulled high if a hit is present in the row. An asynchronous vertical scan is initiated by injecting a "start row scan" token into a scan chain. The token propagates through the scan units until a stop signal is encountered in the first row with hits. The "stopped" output signal from the scan unit is used to generate the row address and to select the active row

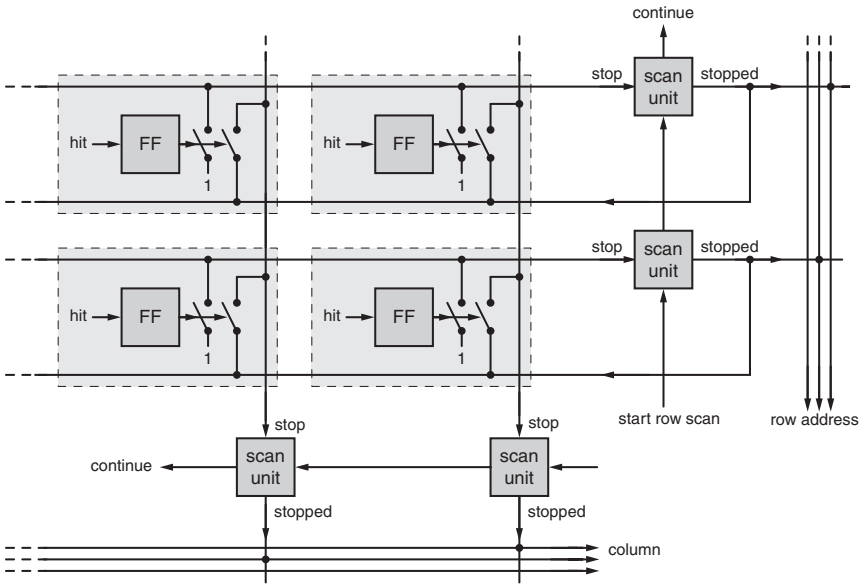


Fig. 3.35. The readout architecture of the DELPHI pixel chip

for the secondary column scan. This scan stops at the first hit column so that the x–y-coordinate of the first hit is obtained. The hit flipflops are reset successively so that the scan skips from hit pixel to hit pixel.

3.4.2 Chips with Zero Suppression and Data Buffering

A large family of pixel readout chips is used for tracking in high energy physics where accelerated particles collide with a fixed target or with other particles from a beam running in the opposite direction. The collisions occur at regular, well-known intervals of 25 ns in the case of the LHC collider at CERN, and so a bunch crossing clock (of 40 MHz at LHC) can be used to synchronize the data taking.

Most collision events do not contain relevant information, however, because the interesting physics processes are very rare. A selection of potentially interesting events is therefore made by other detectors of the experiment. This *trigger* decision, which is based on the analysis of many thousands of detector channels, requires several microseconds of processing time and so the trigger signal, which selects the interesting events for readout, arrives at the pixel chips with a delay of typically more than 100 bunch crossing clock cycles, i.e. some microseconds. The hit information of many events must therefore be stored *on the pixel chip* during this fixed *latency* interval. The event buffering is possible because only a small number of pixels ($\approx 10^{-4}$) are hit in one event. The storage of the hits only requires their real time detection, or in

other words, a "zero suppression." Only triggered hit information is sent to the DAQ of the experiment; all other hit information is discarded after the latency. Although the trigger rate is low (<1% of the events), several nearly consecutive readout requests may occur. The chips must therefore be capable of accepting new triggers before the data of a previous request has been completely sent out.

Several different readout concept that have been implemented to solve the problems of hit buffering during the trigger latency are presented in the next sections. A brief comparison can also be found in [245].

3.4.2.1 Timer in the Pixel

A very simple and elegant approach is to store the hit information in the pixel by starting a timer which elapses after the trigger latency. A hit belongs to the bunch crossing of interest when the falling edge of the timer coincides with the trigger signal which is sent to all pixels. The readout of the valid hit pattern after a trigger is achieved with a shift register in the OMEGA [194] chip family. The timers in this chip are implemented as chains of current-deprived inverters (Fig. 3.36b) so that several pulses can travel along the delay line simultaneously, thus reducing the dead time. The timer delay must be precise enough to elapse with an error of at most one clock cycle, however. This requires a precision of better than 1% which is difficult to achieve in

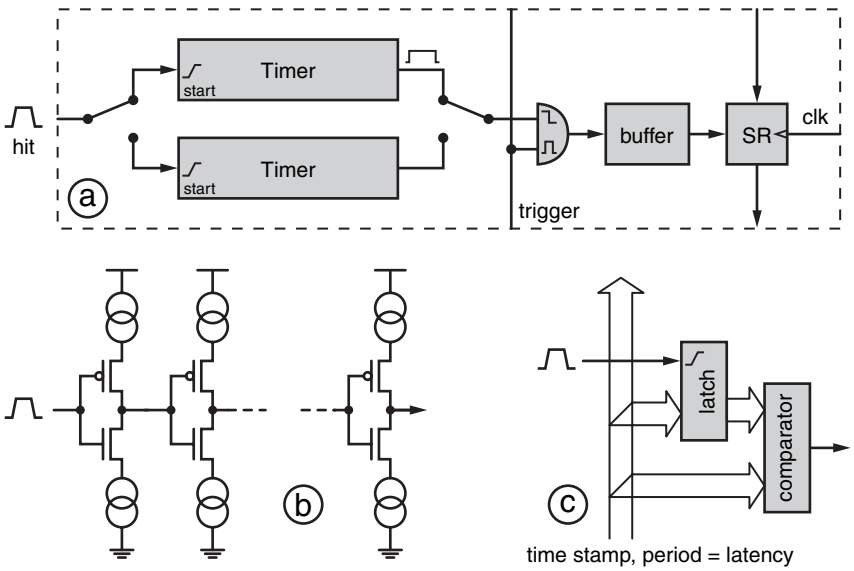


Fig. 3.36. Readout using timers in the pixel (**a**). The timers can be implemented as chains of current-deprived inverters (**b**) or by using an external time stamp in a purely digital solution (**c**)

all pixels due to component mismatch. The practical implementations of the timer therefore offers a trimming possibility in every pixel to fine-tune the delays. A peak-to-peak variation of 30 ns has been achieved with 3 trim bits in the OMEGA3/LHC1 chip [202]. More recent implementations of the timer concept (for the ALICE and LHCB experiments [226]) use two timers per pixel activated alternately to further decrease the dead time in the pixel (Fig. 3.36a). Multiple hit flags can be buffered in the pixel to reduce dead time during the readout and to allow several closely spaced triggers. In order to solve the problem of delay time variations, a digital circuit (Fig. 3.36c) can be used for the timers at the expense of increased power consumption. It uses a Gray-coded time stamp with a period of the latency distributed to all timers. The time stamp is latched when a hit occurs. The timer has elapsed when the distributed time stamp coincides again with the stored value.

3.4.2.2 Conveyor Belt Architecture

Another simple readout concept uses a vertically running shift register in every column to clock the row number of a pixel down the column (therefore the name "conveyor belt" proposed in [245]) as soon as a rising edge of the pixel discriminator occurs [246]. The shift register must be 6 bit wide, for instance, if the column is 63 pixels high. A shift register value ID arriving at the bottom signifies that the pixel in row ID has been hit. Furthermore, it is known that the hit has occurred ID clock cycles in the past (provided that the row numbering starts at the bottom), i.e. that the "age" of the hit is ID. The hit position is stored in a latch in one of several buffers at the bottom of the column as illustrated in Fig. 3.37. The age of the hit is reduced by the trigger latency and the result is written to a counter in the same buffer unit. The counter is incremented with the system clock. The total time since the hit has occurred equals the latency when the counter value reaches zero. The trigger coincidence is therefore made at that moment. The hits are flagged as valid for later serial readout. Several modifications (not shown in Fig. 3.37) have been made to this scheme in the ATLAS FEA chip [246] in order to improve the performance, to squeeze the logic in the available space, and to reduce power consumption. The presence of a valid ID in the shift register is flagged with a separate "full" bit which is also used to enable the clock of the shift register. The cells are therefore inactive when no hits are present, thus reducing the dynamic power consumption. A hit would be lost in this architecture in the event that the shift register cell is full at the moment where a new ID is to be written. Several tries to write the ID are therefore allowed, hoping that an empty shift register cell will pass by. The number of tries required is sent down the column in additional "late" bits for a correction of the hit time. The implementation in FEA shares the shift register between a pair of columns and between two consecutive rows in the column, i.e. one shift register cell serves four pixels with additional "left/right" and "up/down" flags to identify the hit cells.

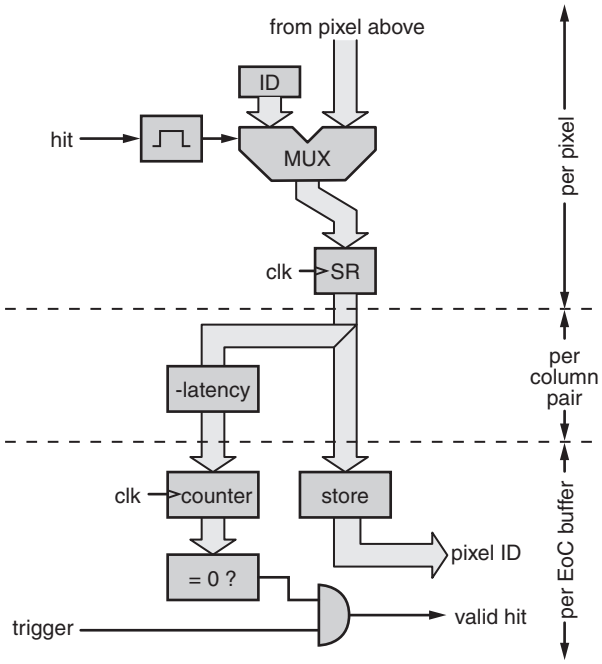


Fig. 3.37. The "conveyor belt" architecture uses a digital shift register to transport the ID of a hit pixel to the bottom of the column where it arrives exactly after ID clocks pulses. The trigger coincidence is performed after (latency-ID) further clock pulses in buffers at the bottom of the chip

The falling edges of the pixel discriminators can be used to determine the width of the discriminator output pulse (ToT) as a measure of the primary charge. The readout of the falling edges can be implemented as before, with an additional flag to distinguish the type of information. The buffers become more complicated in this case because the correct buffer unit for storing a falling edge information must be found.

3.4.2.3 Time Stamp Readout

The basic idea of this approach is to record the time at which a hit has occurred in digital form, the so-called *time stamp* [247]. When the trigger signal selects a certain bunch crossing for readout, the time information of all accumulated hits is compared to the interesting time (or time interval) referred to by the trigger signal. Hits with the correct time stamp are read out, all older hits are rejected. The time of the trailing edge of the discriminator output can be memorized as well so that the pulse width (the ToT) can be determined digitally by calculating the difference of the two values. The time stamp for a hit could be stored, in principle, in the pixel and the trigger

coincidence made there. This would block a hit pixel, however, during the full latency so that a significant inefficiency would be introduced. The time stamp values are therefore transferred to buffers at the bottom of the pixel columns as fast as possible in the architecture of the chip used in the ATLAS experiment [203]. The most important elements of the digital readout are sketched in Fig. 3.38, where four elementary tasks are running in parallel:

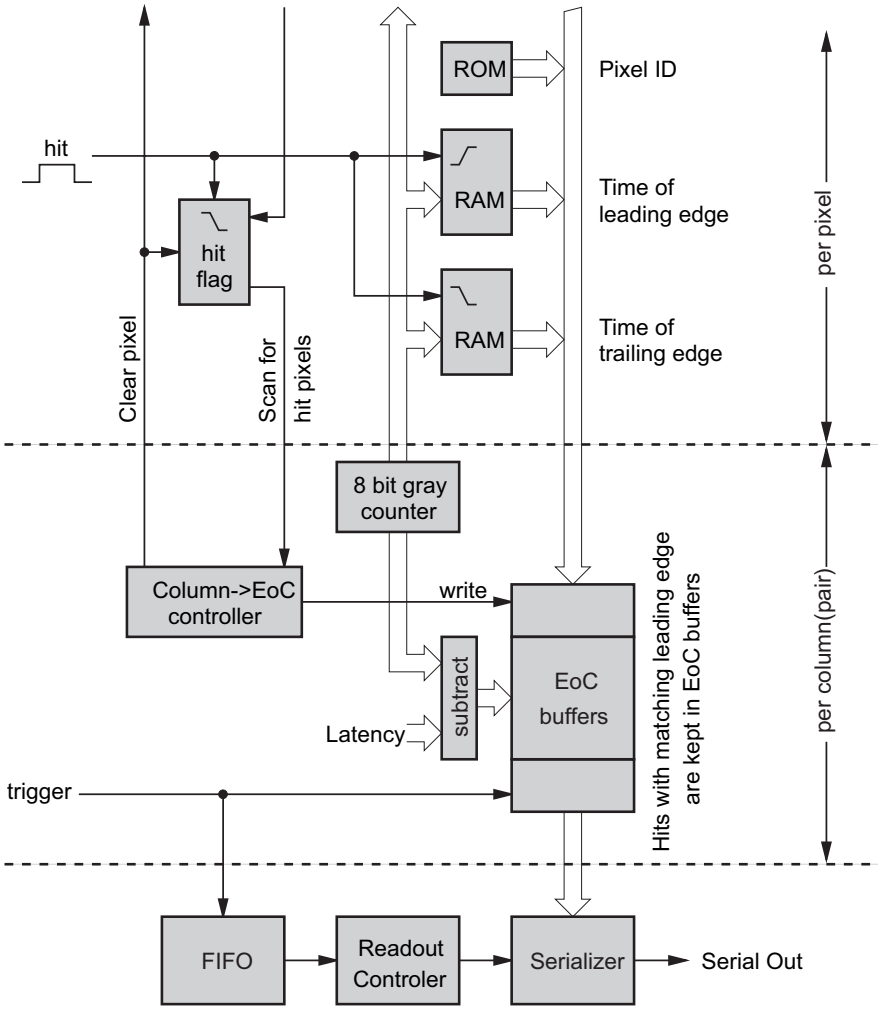


Fig. 3.38. Simplified diagram of the time stamp readout used in the FEI chip family for the ATLAS experiment

1. An 8-bit-wide time stamp counter operated at the 40 MHz bunch crossing rate generates the time reference which is distributed to all pixels in the chip. The counter uses Gray encoding to avoid out-of-sequence values during transitions and to minimize the number of bit transitions and thus the power consumption. When a pixel is hit, the time stamps of the rising and of the falling edge of the discriminator output signal are stored in memory cells in the pixel. The pixel data are ready to be processed when the falling edge has occurred and a "hit" flag in the pixel is set.
2. All hit flags in a column are connected by a fast asynchronous priority scan which indicates to a control logic at the bottom of the column that at least one hit is ready for readout. The control logic requests the uppermost hit pixel in the column to send its rising and falling edge time stamp and its row ID (hard coded in a ROM in the pixel) down the column on a 24-bit-wide bus. This information is stored in a free location of the end-of-column (EoC) buffer pool. The processed pixel is cleared and the scan continues the search for other hit pixels. Hits are thus transferred from the pixels to the EoC buffers at a programmable rate of 5–20 MHz. The left and the right halves of a column pair share the time stamp and readout busses and the EoC buffer pool consisting of 64 locations in the ATLAS FEI chip [211] to save layout area and power. The two sides are served alternately by the column control logic.
3. The hits must stay in the EoC buffers until the trigger latency, i.e. the time difference between the hit occurring in the sensor and the arrival of the trigger signal, has elapsed. The leading edge time stamp value in every EoC buffer is therefore permanently compared to the value of the time stamp counter minus the (programmable) latency value. When the values coincide and a trigger signal is present, the hit is flagged as "valid for readout." It is discarded (the EoC buffer location is freed) otherwise. A list of pending triggers is kept in a FIFO where they are distinguished by a 4-bit trigger number. The trigger number of a particular trigger is stored together with the "valid for readout" flag in the EoC buffer so that the hits can be associated with selected triggers later.
4. A readout controller initiates the serial readout of the hit data as long as pending triggers are present in the trigger FIFO. The EoC buffers are searched for valid data with the corresponding trigger number. The column and row address and the ToT of these hits are serialized and sent to the MCC. The readout controller adds a "start of event" and an "end of event" word (with error and status bits) to the data stream.

3.4.2.4 Column Drain Architecture

The "column drain architecture" [196] developed for the CMS experiment transfers all hits occurring within one clock cycle to a buffer pool in the periphery. The concept has been first implemented in a 0.8 μm technology with two metal routing layers only so that the number of devices and bus

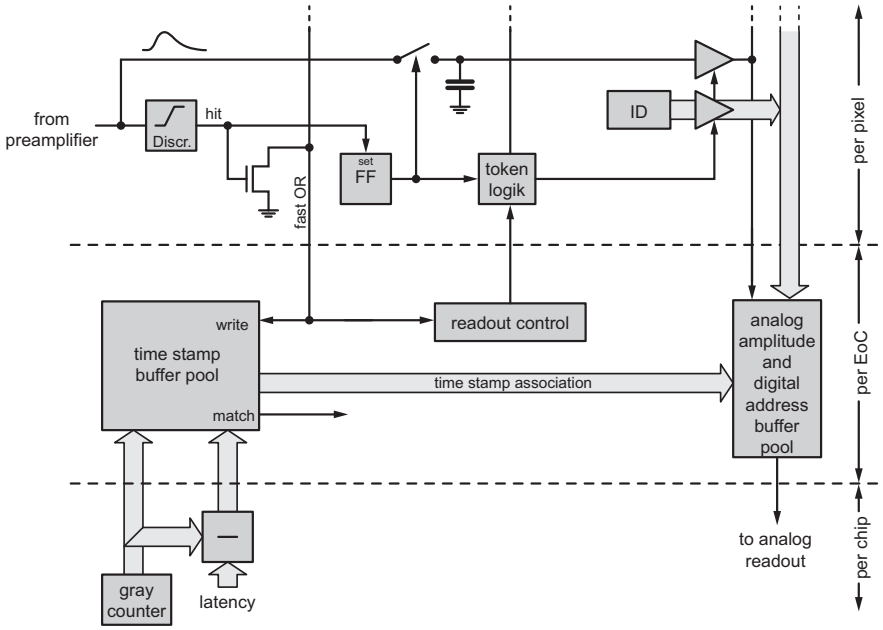


Fig. 3.39. The column drain readout transfers the amplitude and address of hits buffers in the EoC which are associated with a single digital buffer holding the time stamp of the event

signals could be minimized. The final readout chip was realized in a 0.25 μm technology [205] which allowed the implementation of additional features to ensure efficient operation of the innermost pixel layer at full LHC luminosity.

Au: Is this edit ok?

The main parts of the readout are illustrated in Fig. 3.39. One or several hits produced by the pixel discriminators in a column pair are flagged to the EoC by a fast OR. A single time stamp for *all* these hits is stored in 1 of 12 available digital buffer locations. The hit information of this group of pixels is now transferred immediately to a buffer pool, and hence the name "column drain." The hit pixels in the column pair are found by a fast scan mechanism which basically passes a token from cell to cell through closed switches until a hit cell is found. This mechanism has been designed to operate at above 1.5 GHz in order to reduce the dead time introduced by the scan. Pixels without a hit are still sensitive during the scan. The CMS readout provides the *analog* information stored by a sampling circuit on capacitors in the pixels. This analog value is output by the pixel which has been identified through the token scan. All hits are transferred sequentially to a pool of 32 buffers per column pair each containing an analog cell for the amplitude information and 9 digital cells for the pixel address. They are associated with the corresponding time stamp in the digital buffer by a pointer mechanism.

Many buffers are therefore available for a single event so that high local multiplicities (as they can occur in jets of particles) can be coped with.

The time stamps in the digital buffers are permanently compared with a second time stamp value offset by the latency. When a match occurs and no trigger signal is present, the event is discarded and all buffers are freed. A serial analog readout is started otherwise. For the readout the pixel address is translated into five analog signals (two for the double column and three for the pixel) with six possible levels. The data for one hit can be transferred within six clock cycles. The readout of several chips is controlled by a token which is passed from chip to chip. It is generated by a separate "token manager" chip.

3.4.2.5 BTeV Readout

The BTeV readout implemented in the FPIX chip family [195] is shown in Fig. 3.40. Four identical EoC readout controllers at the bottom of the columns communicate with the pixels in the column through 4×2 command lines. The controllers can issue one of the four states "look for data," "idle," "output," and "reset." These states are sent to a command interpreter in every pixel. The readout can be divided into several elementary steps:

- One of the EoC controllers is requested by a priority encoder to send the "look for data" command to the pixels on its pair of command lines. This controller will be responsible for processing the next event in the column.
- When a pixel is hit, the command interpreter in that pixel is linked to the command set which carries the "look for data" pattern. The hit fast-OR line is activated. This informs the EoC that one or more hits have occurred in a column. The active EoC controller stores the time stamp (bunch crossing number) in a register and switches from the "look for data" state to the "idle" state. Another free EoC controller is selected immediately by the priority logic to issue "look for data."
- All "idle" EoC controllers and the associated pixels wait until their readout is requested. In the triggered operation mode, a desired event time stamp is presented to a comparator in the EoC controller (not shown in Fig. 3.40) which compares it to the stored time stamp value. Readout is started if the values are equal. If no match is found after a programmable timeout, the event is discarded. An operation mode without any trigger has also been implemented. All events are read out in this case.
- When a readout is requested, at most one controller issues an "output" command to its associated pixel set. All hit pixels in the set pull the read fast-OR line low. The EoC bus controller starts a token scan to find the first pixel in the selected set. The first pixel found outputs its address and ADC data onto the column bus, resets itself, and withdraws its assertion of the read fast-OR.

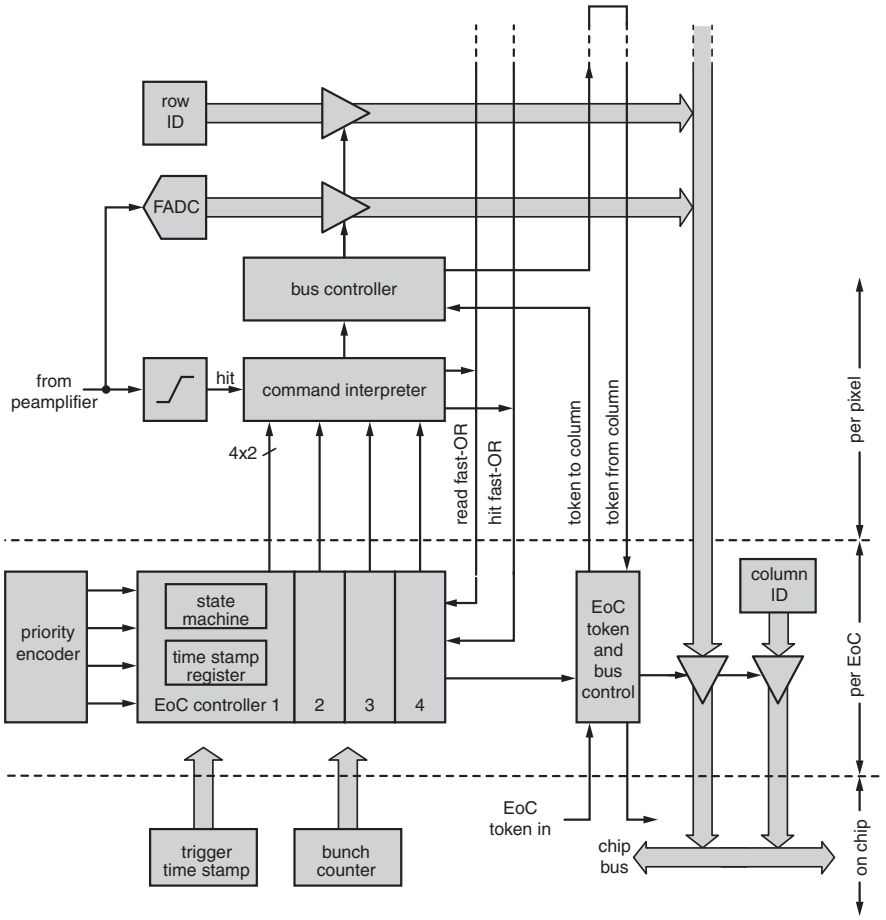


Fig. 3.40. The FPIX chip family uses several simple readout controllers in the end of the column part. All hits within a column occurring at the same bunch crossing are associated with one of the four available controllers where the time stamp is stored

- All pixels have been found when the read fast-OR goes high. The EoC controller informs the priority logic that it is ready again to issue a "look for data."
- The EoC controllers can also sent a "reset" signal to reset all associated pixels if an event has been discarded, for instance, due to a timeout.

The FPIX readout uses a low-resolution flash ADC (3 bit) to digitize the hit amplitude immediately in the pixel. This information is sent to the readout bus together with the pixel address. The architecture is implemented in such a way that all hits can be read out immediately as long as the rate is not too high. This allows the pixel information to be used for early trigger

decisions. A throttling mechanism is available to discard hits for some time when the hit rate exceeds the readout bandwidth.

3.4.3 Counting Chips

For applications in biomedical (X-ray) imaging, synchrotron radiation experiments, autoradiography, and others, the number of particles absorbed in every pixel during a given time interval must be determined. The hit signals are therefore counted in every pixel and read out after the measurement interval as indicated in Fig. 3.41a. The practical implementation of this simple goal requires very compact counters of $\geq$16 bit to cope with the hit rates in brightly illuminated pixels. Classical binary counters are space-consuming and require a dedicated readout. Any state machine which generates as many states as possible out of N available bits can be used, in principle, to implement the counter. A particular simple design is a linear shift register fed back with an exclusive OR gate from two or more taps [248]. If the taps are chosen correctly, this leads to $2^N - 1$ states ("maximum length shift register counter"), and so the bits are used very efficiently. The bit patterns generated in this arrangement are pseudorandom numbers, unfortunately, which makes it difficult to determine the number of clock cycles from a given pattern. Lookup tables are therefore used best to determine the corresponding number of counts. Different taps can lead to slightly shorter sequences which can be decoded more easily [249]. This might therefore be a better choice for very long counters. The shift counter has the additional advantage of a simple

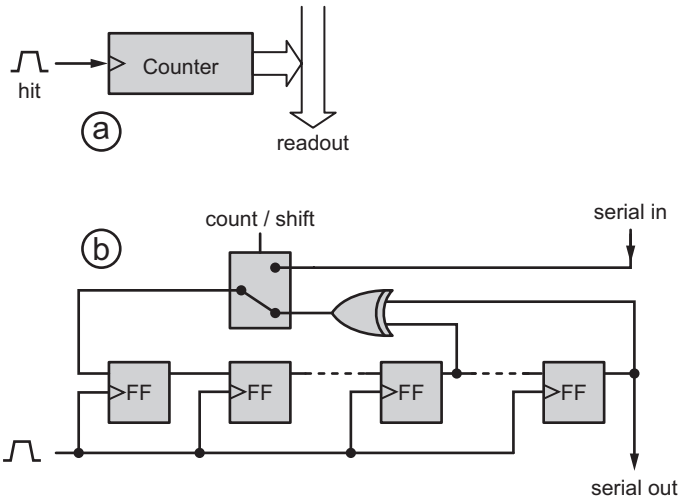


Fig. 3.41. (**a**) Pixel with a counter. (**b**) The counter can be implemented as a simple linear feedback shift register . This solution has the additional benefit of a simple serial readout

serial readout by means of a multiplexer as shown in Fig. 3.41b. The flipflops used can be very simple dynamic cells [207] if all counters are clocked at a minimum rate with a regular refresh clock during the acquisition interval. The number of refresh clock pulses is subtracted from the count value after the serial readout of the counters. Most counting pixel chips are using this technique [163, 193, 197, 198].

The limitation of the dynamic range given by the maximum number of counts can be overcome by letting the counter "wrap around." An overflow bit is set by the 16-bit wide counters in the XPAD chip [208, 250] when this happens. External circuitry regularly scans the overflow bits of the chips and increments external counters. A virtually infinite dynamic range can be achieved with this method if it is guaranteed that every overflow bit can be recorded and cleared before the counter wraps a second time. At the end of the measurement, all counters are read out (this requires 1 ms for the XPAD chips). The total pixel count is calculated from the number of overflows and the remaining counter value.

Some counting chips use two discriminators per pixel with different thresholds so that incoming hits can be sorted into low and high energies. The analysis of the two images obtained should lead to a significant contrast enhancement for monoenergetic illumination (synchrotron) and is also expected to improve images taken with normal X-ray sources with a broader energy spectrum [199]. The latest Medipix chip [193] has two discriminators but only one counter while the MPEC2 [199] has two truly independent channels albeit with larger pixels. Systems using counting pixel chips are presented in Sect. 5.4.1.