Vertically integrated deep N-well CMOS MAPS with sparsification and time stamping capabilities for thin charged particle trackers

L. Ratti, L. Gaioni, M. Manghisoni, V. Re, G. Traversi

Abstract

A fine pitch, deep N-well CMOS monolithic active pixel sensor (DNW CMOS MAPS) with sparsified readout architecture and time stamping capabilities has been designed in a vertical integration (3D) technology. In this process, two 130 nm CMOS wafers are face-to-face bonded by means of thermocompression techniques ensuring both the mechanical stability of the structure and the electrical interconnection between circuits belonging to different layers. This 3D design represents the evolution of a DNW monolithic sensor already fabricated in a planar 130 nm CMOS technology in view of applications to the vertex detector of the International Linear Collider (ILC). The paper is devoted to discussing the main design features and expected performance of the 3D DNW MAPS. Besides describing the front-end circuits and the general architecture of the detector, the work also provides some results from calculations and Monte Carlo device simulations comparing the old 2D solution with the new 3D one and illustrating the attainable detection efficiency improvements.

1. Introduction

Deep N-well CMOS monolithic active pixel sensors (DNW CMOS MAPS) were proposed a few years ago as possible candidates for charged particle tracking applications. They were specifically developed with the aim of enabling fast readout of large detector matrices through sparsification techniques [1]. In DNW-MAPS, the collecting electrode consists of a deep N-well integrating also NMOS devices from the analog front-end (AFE) section in the internal P-well. The collected charge is read out by a classical optimum chain for capacitive detectors, including a fully CMOS charge preamplifier which makes the charge sensitivity independent of the detector capacitance. Also CMOS digital blocks for sparsified data readout and time stamping are laid out in the elementary pixel cell. Based on the proposed device, the first ever MAPS detectors with pixel level sparsification have been fabricated and successfully tested at the Proton Synchrotron facility at CERN [2–4]. New technology options have been considered recently, in particular with the aim of improving such DNW MAPS properties as spatial resolution and detection and charge collection efficiency. Among the investigated technologies, vertical integration (also known as 3D) processes [5] seem the most promising ones. By stacking two or more standard CMOS layers (or tiers) one on top of the other, 3D processes may be very effective in providing, at affordable costs, increased functional density, physical separation of the analog front-end from the digital blocks and reduction of the area covered by competitive N-wells in the sensor layer. This paper describes the main design features and discusses the expected performance of a vertically integrated DNW CMOS MAPS detector, the SDR1 (sparsified digital readout) chip, specifically aimed at vertexing applications to the International Linear Collider (ILC) facility. The chip was designed in the framework of the 3DIC Consortium, a collaboration among Fermilab and French and Italian institutions pursuing the development and fabrication of vertically integrated front-end circuits and monolithic detectors. The SDR1 MAPS sensor represents the evolution of another monolithic detector, the SDR0 chip, previously designed and fabricated in a planar, 130 nm CMOS technology [2], to which reference will be made throughout this work for the sake of comparison. The paper is structured as follows. After this introduction, Section 2 will be concerned with a short description of the two-tier vertical integration technology used for the design of the 3D DNW MAPS. Section 3 will describe the detector architecture, providing some details on the pixel level analog and logic processors and on the digital back-end. Circuit simulation results will be also discussed in the same section. Finally, Section 4 will present some data from calculations and Monte Carlo device simulations emphasizing the achievable detection efficiency improvements through a comparison between the SDR0 and the SDR1 chip performance.
2. Vertical integration technologies

The success of a vertical integration process relies upon three fundamental steps, namely:

- fabrication of electrically isolated connections through the silicon substrate (through silicon vias, TSV);
- substrate thinning (below 50 μm);
- layer-to-layer alignment and mechanical bonding.

Different approaches have been proposed and are available for their implementation [6]. The technology cross-section shown in Fig. 1, in particular, points to the main features of the extremely cost-effective process provided by Tezzaron Semiconductor [7] which was used for the design of the SDR1 chip. The Tezzaron process can be used to vertically integrate two (or more) layers, specifically fabricated and processed for this purpose by Chartered Semiconductor in a 130 nm CMOS technology. In the Tezzaron/Chartered process, wafers are face-to-face bonded by means of thermo-compression techniques. Bond pads on each wafer are laid out on the copper top metal layer and provide the electrical contacts between devices integrated in the two layers. The top tier is thinned down to about 12 μm to expose the through silicon vias (TSV), therefore making connection to the buried circuits possible. Among the options available in the Chartered technology, the low power (1.5 V supply voltage) transistor option was chosen. The technology also provides six metal layers (including two top, thick metals), dual gate option (3.3 V I/O transistors) and N- and P-channel devices with multiple threshold voltages [8].

3. The SDR1 chip

The features of the Tezzaron/Chartered process have been exploited in the design of the SDR1 chip, which is based on the same readout architecture as the SDR0 monolithic sensor [2]. SDR1 is a two-tier, vertically integrated 240 × 256 MAPS matrix with a 20 μm pixel pitch, token passing binary readout architecture and the capability for storing two hits and the relevant 5-bit time stamps. The step from the DNW MAPS in a planar CMOS technology to its vertically integrated version is illustrated in Fig. 2, showing a cross-sectional view of a 2D MAPS and of its 3D translation. In the SDR1, the deep N-well sensor and the analog front-end are integrated on a different layer from the digital front-end (DFE). The main benefits deriving from such an approach can be summarized as follows:

- all the PMOS devices used in digital blocks are integrated in a different substrate from the sensor, therefore significantly reducing the amount of N-well area (and its parasitic charge collection effects) in the surroundings of the collecting electrode and improving the detector charge collection efficiency (CCE);
- although in the SDR1 chip the pitch is 20% smaller than in the SDR0 monolithic sensor (in which the pitch was 25 μm), the effective area available for device integration is larger (625 μm² in the SDR0 MAPS, 800 × 400 μm² for each tier—in the SDR1 sensor); this has been exploited to increase the functional density of the sensor (in particular, to implement the capability for double-hit storing) and to reduce threshold dispersion by designing large transistors where required (see Section 3.1);
- from the previous point it should be apparent that a better trade-off between integrated functionality and detector pitch can be achieved, yielding a smaller point resolution;
- separating the digital front-end from the analog processor and the sensor can effectively prevent digital blocks from interfering with the analog section and from injecting charge into the sensor through parasitic capacitive coupling; nevertheless, it is worth emphasizing here that the SDR0 test structures were not affected to a significant extent by this sort of problems.

The following sections provide a detailed description of the circuits integrated in the pixel cell, made up of 276 transistors, and of the general architecture of the detector.
3.1. Analog front-end

The bottom (analog) layer of the SDR1 elementary cell includes the deep N-well sensor (its layout features are outlined in Section 4.1), whose signal is processed by a charge sensitive amplifier (CSA). The CSA is followed by a threshold discriminator, which has been only partially integrated in the bottom tier. Fig. 3 shows the block diagram, with some transistor-level details, of the analog front-end electronics (the deep N-well detector is represented by means of its technology cross-section). The CSA is based on a folded cascode scheme and represents a shaperless version of an optimum channel for capacitive sensors [9], hence the name shaperless front-end (SFE). This reduction in the analog front-end complexity makes it possible to comply with the point resolution constraints set for the ILC vertex detector [10]. The SFE input NMOS device (whose dimensions were chosen based on criteria for optimum detection efficiency [10]) features a W/L = 20/0.18, and a drain current of 1.4 μA. Charge restoration in the SFE feedback network is obtained through a PMOS current mirror stage, providing a linear discharge of the metal-oxide-metal capacitor C_F (about 1 fF). The bandwidth of the SFE has been purposely limited by loading its high impedance node with an NMOS capacitor (about 50 fF) in order to reduce high frequency noise effects. The power dissipation of about 5 μW, together with a 1% duty-cycle operation (controlled by the PowerDown command in Fig. 3) makes the analog front-end compatible with the ILC power specifications.

Fig. 4 shows the response, obtained from circuit simulations, of the SFE to an 800 electron charge pulse for different values of the current I_F biasing the feedback current mirror. As it can be detected in Fig. 5, an integral non-linearity of about 2% has been obtained from circuit simulations over an input dynamic range of 2000 electrons. The expected charge sensitivity is close to 800 mV/fC. A value of about 35 electrons has been obtained for the equivalent noise charge (ENC) at a detector capacitance of 200 fF and at a 1.2 nA current biasing the feedback network. Parallel noise contribution from the feedback current mirror is actually negligible at the amplifier peaking time, approximately 1 μs. Simulations show an increase of about 2 electrons for I_F going from 1 to 3 nA. The threshold dispersion, about 35 electrons as for the noise, is mainly contributed by the SFE input device (28 electrons) and by the threshold discriminator (22 electrons). Taking into account a most probable value of 800 electrons, obtained from device simulations, for the charge collected by the pixel, the expected signal-to-noise ratio is about 23. The threshold discriminator consists of a differential NMOS pair, integrated on the analog layer, with a mirrored PMOS load followed by a second gain stage both integrated on the top (digital) tier together with the logic blocks. Electrical connection between the discriminator devices is guaranteed by inter-tier bond pads. Note that, since the PMOS devices in the discriminator stage need to be designed with quite a large area for matching purposes, moving them to the digital layer is expected to improve the charge collection efficiency by further reducing the area of charge-stealing N-wells on the sensor tier.
3.2. Digital section

The sparse readout technique implemented in the SDR1 pixel sensor is based on a token passing scheme, which is similar to the one developed for the BTeV pixel and silicon strip readout chips [11,12]. The operation of the detector has been tailored on the beam structure of the International Linear Collider, which will feature 2820 bunches per train, each train lasting slightly less than 1 ns, with an inter-bunch period of 330 ns and a repetition rate of 5 Hz (corresponding to a duty cycle of 0.5%) [13]. Each pixel can retain data for the entire bunch train period (detection phase), when a pixel is hit for the first time, the set-reset flip–flop (SR FF, FFSRK) is set (the NLatchEnable signal being low and the relevant time stamp could be stored). During the readout phase (or inter-train period), data (X and Y coordinates and time stamps) are sent off to the detector periphery and then output after serialization. A detailed discussion of the detector operation is carried out in the following, where the digital front-end and back-end and the general readout architecture of the sensor are thoroughly described.

3.2.1. Digital front-end

Besides the PMOS load and the second gain stage of the threshold discriminator, the top tier of the SDR1 elementary cell also includes a number of digital blocks taking care of double-hit detection, time stamping, data sparsification and pixel masking. A block diagram of the digital front-end is shown in Fig. 6. During the bunch train period (detection phase), when a pixel is hit for the first time, the set-reset flip–flop (SR FF, FFSRK) is set (the NLatchEnable signal being low and the NMasterReset signal being high) and the relevant time stamp register (TSR-1) gets frozen. The Q output of FFSRK is also used as the D input of a delay FF (D FF, FFDR), which is bound to get set at the next low–high transition of the discriminator. Therefore, upon a second hit, FFDR is set and the relevant 5-bit time stamp (TSR-2) gets frozen. At the end of the detection phase, when the readout phase begins, the NLatchEnable signal is switched on (therefore preventing the SR and/or D flip–flops from getting set in those pixels which were not hit during the bunch train period) and the token is launched through the MAPS matrix (see also Section 3.2.2). Each hit pixel, after receiving the token (TokenIn signal), gets hold of the column and row buses (getX and getY signals are pulled down) at the next cell clock (CellClk signal) rising edge, and releases position and time stamp information (by acting on the readout enable input signal of the time stamp register, NRO_EN) within a cell clock period. As soon as the cell starts sending off data, the SR FF is reset (this is possible if the NMasterReset signal is high, which is the common operating condition in both detection and readout phases) and the token is released and sent to the other token passing core in the same cell. If the cell has registered a second hit, the same readout procedure is followed as for the first hit. Otherwise, the token very rapidly passes through the token passing core and is sent (TokenOut signal) to the next hit pixel or to the matrix output (LastTokenOut signal, see Section 3.2.2 and Fig. 7). Before starting chip operation, a kill mask can be loaded to the detector to disable noisy pixels.

3.2.2. Digital back-end and readout architecture

The digital readout operation of the SDR1 detector can be discussed at the matrix level on account of the block diagram shown in Fig. 7, referring to a 240 \times 256 DNW MAPS matrix. As already mentioned in the previous section, at the beginning of the readout phase, a token is launched through the matrix starting from the cell in the first row and first column. This is done by setting the FirstTokenIn signal. The token scans the matrix in a row by row fashion and stops in the first hit cell it finds along its path. Then, at the next rising edge of the cell clock (CellClk signal), the pixel gets hold of the column and row buses, pointing to the x and y registers at the periphery of the sensor matrix, and sends off the time stamp register content. Almost immediately, the token is released and searches for a second hit in the same cell. If a second hit is found, then data are read out as in the case of the first hit. After reading out the second cell hit, or if no second

![Fig. 6. Block diagram of the digital front-end electronics integrated in the elementary cell of the SDR1 MAPS detector.](image-url)
hit is found, the token scans ahead, searching for the next hit pixel, which will be read out during the next cell clock period. When the token gets out of the last cell, the \texttt{LastTokenOut} signal goes high. \texttt{X}, \texttt{Y} and time stamp data are serialized and sent off the chip by means of a 24 input (8 bits for the \texttt{X} and the \texttt{Y} coordinates, 5 bits for the time stamp and three sync bits) multiplexer within a cell clock period. The readout clock (\texttt{ReadOutClk}) frequency is an integer multiple of the cell clock (\texttt{CellClk}) and is synchronous with it. Pull down transistors and tri-state buffers in the time stamp registers have been designed to enable readout clock operation at 100 MHz (about 4.2 MHz for the cell clock). During the detection phase, a 5-bit Gray counter is used to broadcast the time stamp to all of the cells in the matrix. The 5-bit time stamp allows the bunch train interval to be split into 32 time slices, providing useful complementary information to other detectors for track reconstruction. Given the ILC bunch train period, about 930 ms, each time slot amounts to about 29 ms, corresponding to a time-stamp clock frequency of about 34.4 kHz.

4. Detection efficiency improvement through 3D processes

Use of vertical integration technologies is expected to provide significant benefits to the DNW MAPS sensor performance in terms of detection efficiency. The overall detection efficiency \( \epsilon_T \), of the DNW monolithic sensor integrated in the chips of the SDR series, under the hypothesis of factorizability, can be written as

\[ \epsilon_T = \epsilon_{DNW} \cdot \epsilon_{AFE} \cdot \epsilon_{DFE} \cdot \epsilon_{ROA}. \]

In particular:

\( \epsilon_{DNW} \) represents the contribution to the overall efficiency provided by the sensor and is given by the ratio between the number of particles (assumed as minimum ionizing ones) producing a signal which exceeds the discriminator threshold and the number of particles hitting the detector; it depends on the layout and area of the DNW sensor and of the N-wells, on the threshold level and on technology parameters, such as the substrate resistivity;

\( \epsilon_{AFE} \) represents the contribution to \( \epsilon_T \) provided by the SFE and the threshold discriminator; it is mainly related to the AFE dead time (the time during which the SFE output is over threshold and the cell is blind) and depends on the threshold level and on technology parameters, such as the substrate resistivity;

\( \epsilon_{DFE} \) represents the contribution to the overall efficiency provided by the DFE and is given by the ratio between the number of hit stored in the pixel and the number of signal induced low–high transitions in the threshold discriminator in each single bunch train; it depends on the storing capability of the cell, the hit occupancy (i.e., the number of particles hitting the detector per
bunch crossover (BCO) per mm²), the cluster multiplicity and the detector pitch;

\( e_{ROA} \) represents the contribution to \( e_T \) provided by the readout architecture and is given by the ratio between the number of readout hits and the number of stored hits; it depends on the readout clock frequency, the number of pixels per chip, the degree of readout parallelism (parallel readout of the matrix is an option, although it has not been implemented in the SDR1 prototype) and the hit occupancy.

Use of vertical integration technologies may have a direct, significant impact especially on \( e_{DNW} \) and \( e_{DFE} \). In the following sections, such an impact will be evaluated, through Monte Carlo simulations and analytical calculations, by comparing the detection efficiency performance in the SDR1 with that in the SDR0 MAPS sensor.

4.1. Sensor detection efficiency

As already mentioned in Section 3, in the design of DNW MAPS, 3D technologies can be exploited to move a large part of the PMOS devices and their N-wells to a different layer from the collecting electrode. Fig. 8 shows the layout of the DNW sensor and of the N-wells for PMOS transistors in the SDR0 and SDR1 (bottom tier) MAPS detectors. In the SDR0 pixel, the DNW sensor covers just 35% of the cell area, whereas it covers more than 50% in the case of the vertically integrated version. Moreover, a nonnegligible fraction of the cell area (more than 30% of the total area covered by DNW and N-well diffusions) is taken, in the SDR0 pixel, by the N-wells, whereas they take a small portion (amounting to about one tenth of the total area covered by DNW and N-well diffusions) in the case of the SDR1 detector. In order to fully appreciate the beneficial effects of 3D design on the sensor collection efficiency, two series of Monte Carlo simulations have been performed based on a random walk algorithm modeling the carrier motion in the undepleted substrate of monolithic pixel detectors [14]. Each series consists of a set of 10,000 particles impinging on the central element of a 3 × 3 matrix in a random position. The simulated detector matrices were implemented using the sensor layouts of Fig. 8. The resulting detection efficiency \( e_{DNW} \) as contributed by the sensor at varying discriminator threshold levels is displayed in Fig. 9 in the case of the SDR0 and of the SDR1 chip. Sensor detection efficiency is still well over 99% at a discriminator threshold of 300 electrons in the SDR1 MAPS. Fig. 10 shows the average cluster multiplicity as a function of the discriminator threshold again for the SDR0 and the SDR1 MAPS. The SDR0 cluster size is always smaller than in the SDR1 case. This may result as a consequence of the larger area taken by N-wells in the SDR0 chip, which reduces the effective amount of charge available for the DNW collecting electrodes in the matrix, therefore reducing, at each threshold level, the average number of pixels over threshold.

4.2. DFE detection efficiency

The capability for double-hit counting implemented in the DFE of the SDR1 chip is supposed to improve the detection efficiency of the overall system. The advantage provided by 3D over planar technologies may be evaluated by directly calculating the detection efficiency contributed by the digital front-end. For this purpose, let \( h_o \) be the hit occupancy (i.e., the average number of
On cell being hit exactly during a bunch train period. The probability distribution for the number of times an elementary cell is hit If a pixel cell has the capability for storing the relevant DFE detection efficiency corresponds to the probability of an elementary cell being hit exactly \( n \) times in a bunch train is then given by

\[
P(n) = \frac{O_c}{n!} \exp(-O_c).
\]

If a pixel cell has the capability for storing \( m \) hits in a bunch train, the relevant DFE detection efficiency corresponds to the probability of the pixel being hit no more than \( m \) times,

\[
\varepsilon_{\text{DFE}} = P(n \leq m) = \sum_{i=0}^{m} P(i).
\]

Therefore the DFE detection efficiency for the SDR0 MAPS, \( \varepsilon_{\text{DFE,SDR0}} \), corresponding to the probability of an elementary cell being hit no more than once, is given by

\[
\varepsilon_{\text{DFE,SDR0}} = \exp(-O_{c0}) + O_{c0} \cdot \exp(-O_{c0})
\]

where \( O_{c0} \) is the occupancy for the SDR0 pixel. On the other hand, the DFE detection efficiency for the SDR1 MAPS, \( \varepsilon_{\text{DFE,SDR1}} \), corresponding to the probability of an elementary cell being hit no more than twice, is given by

\[
\varepsilon_{\text{DFE,SDR1}} = \exp(-O_{c1}) + O_{c1} \cdot \exp(-O_{c1}) + \frac{O_{c1}^2}{2} \cdot \exp(-O_{c1})
\]

where \( O_{c1} \) is the occupancy for the SDR1 pixel. Fig. 11 shows the detection efficiency contributed by the digital front-end as a function of the hit occupancy in the case of the SDR0 (single-hit detection) and of the SDR1 (double-hit detection) chip at a discriminator threshold \( Q_t \) of 300 electrons. The curves take into account the cluster size as obtained from the data of Fig. 10. Note that the detection efficiency in the SDR1 DFE is larger than 99% at \( h_{ch} = 0.15 \) particles/BCO/mm\(^2\), which is five times the hit occupancy value foreseen for the innermost layer of the vertex detector when the ILC is operated at 500 GeV [15].

5. Conclusion

This paper has been devoted to discussing the main design features and expected performance of a deep N-well CMOS monolithic sensor designed in a 130 nm vertical integration, or 3D, technology. The use of a double-layer process in the development of DNW MAPS has the potential to address, at an affordable cost, the main issues of a planar 130 nm CMOS technology, in particular low charge collection efficiency due to the nonnegligible area covered by charge-stealing N-wells, low detection efficiency due to limitations in hit-storing capability, moderate-to-low point resolution due to the limited scale of integration and cross-talk phenomena between digital blocks and the sensor and/or the analog section, which are typical of mixed-signal circuits. Significant improvements in terms of detection efficiency have been confirmed by Monte Carlo simulations and analytical calculations. Several test structures have been designed to evaluate the suitability of the Tezzaron/Chartered technology for the fabrication of DNW MAPS. Full characterization is supposed to start by Fall 2009, when the 3D chips will be delivered. A 240 × 256 MAPS matrix with sparse readout and time stamping capabilities, included in the integrated 3D structures, should be ready together with the needed setup for a test beam in 2010.

Acknowledgments

The authors are indebted to Ray Yarema and the members of the 3DIC Consortium for the help provided during the design activity. They also acknowledge the contribution from Fabio Bellotti and Alessia Manazza to the design of the SDR1 chip.
Finally, the authors wish to thank Valeria Speziali for her support throughout this work.

References