Development of high speed waveform sampling ASICs

S. Ritt
Paul Scherrer Institute, CH-5232 Villigen, Switzerland
E-mail: stefan.ritt@psi.ch

Abstract

This paper describes recent developments using Switched Capacitor Arrays (SCA) for ultra-fast waveform sampling. This technology had a renaissance recently and many groups are developing chips now using modern chip technologies. Compared with traditional flash analog-to-digital converters, SCAs have the advantage that they can sample analog signals at several giga-samples per second with resolutions exceeding 12 bits at a power consumption of only a few ten milli-watts. Packing many channels onto a single chip, SCAs are currently replacing ADCs in several fields, such as gamma ray astronomy and rare decay particle physics experiments.

Introduction

Modern data acquisition in particle physics experiments is getting more and more challenging. On one hand accelerator based experiments tend to go to higher particle fluxes, causing significant pile-up in the detectors [1]. With a finer segmentation of the detectors the amount of data channels increases, causing severe problems of power consumption and cost issues. On the other hand many experiments require an excellent resolution, which can sometimes not be realized with flash analog-to-digital converters (FADC) in the GHz range, since they have usually only eight or ten bits. An alternative approach to tackle these requirements are the so-called Switched Capacitor Arrays (SCA). These integrated circuits use a series of small capacitors, connected via switches to an input bus to sample an analog signal as shown in Fig. 1.

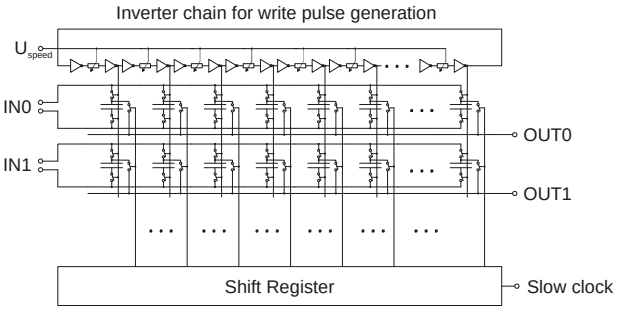

Fig. 1: Simplified schematic of a typical switched capacitor array chip

Instead of using a sampling clock in the giga-hertz range which would be hard to handle, the write switches at the sampling capacitors are operated via a chain of inverters. A short pulse travels down this inverter chain and operates one write switch after the other, where the sampling frequency is given by the signal propagation time through one inverter stage. To make the sampling frequency adjustable, the power supply of the inverters is reduced ("starved" operation) or a variable resistor is placed between two inverters, thus forming with the parasitic input capacitance of the following inverter a variable delay line. The write pulse travels through the chain of inverters in a circular fashion, so that the input is continuously sampled until the inverter chain is stopped. The maximum sampling depth is then given by the number of capacitors. Once a trigger occurred and the writing is stopped, the capacitors can be read out through a second analog switch, operated by an ordinary shift register, which can be clocked with a few megahertz. While some SCAs digitize the signals on-chip, others send the analog signals off-chip, where they can be digitized with commercial ADCs. The SCA then operates as a kind of "time-stretcher" which samples the input signal in the GHz range and outputs it in the MHz range, where one can use cheap and precise ADCs. Multiplexing several outputs into a single external ADC reduces the number of required ADCs for a given number of input channels.

The drawback of this architecture is that it can only operate in a triggered mode. The input signal cannot be digitized continuously like with a FADC, but the sampling has to be stopped for the readout process, thus introducing dead time. It can therefore only be used in applications which are triggered, such as particle physics experiments. This is similar to a digital oscilloscope, which needs also an external or internal trigger to acquire a waveform. One design goal of the SCA is therefore to minimize the readout dead time, in order to keep the data acquisition efficiency as high as possible. Another problem comes from the fact that the propagation delay through the inverter chain depends on the power supply and the temperature, so the sampling frequency is not constant over time. The design goals for a SCA chip are summarized as following:

- Fast sampling speed
- High analog input bandwidth
- Good signal-to-noise ratio
- Short readout time
- Temperature stability
- Low power consumption

The following paragraphs will describe different methods to meet these partially conflicting design goals, in order to make this architecture attractive for many applications.

Sampling Speed

While it is simple to get fast sampling speeds from the inverter chain, this technique has certain problems. Some chip designs use an odd number of inverters, which causes first a zero-to-one transition to travel through the inverter chain, followed by a one-to-zero transition (see Fig. 2 left). Some additional circuitry (not shown in the figure) converts both transitions into a short pulse, which opens the write switches of the sampling cell. The problem of this technique is that the two edges have different propagation speeds through the inverters, since the PMOS and the NMOS transistors in an inverter don't have the same driving strength. In order to determine the sampling speed for a given acquisition, one has to know in which of the two phases the sampling chip had been during acquisition and correct for it off-line.

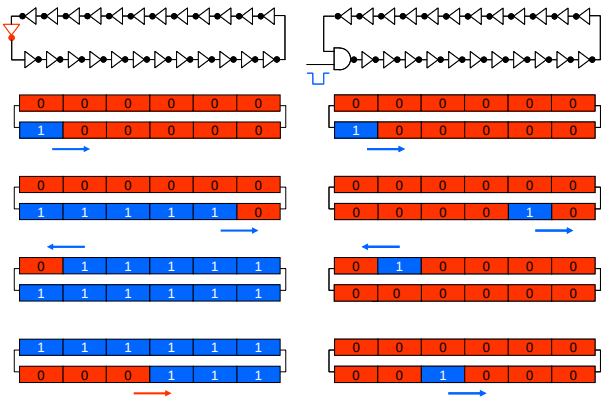

Fig. 2: Different methods of propagating a pulse though an inverter chain

Another approach is to produce a small pulse which travels down the inverter chain (Fig. 2 right). This pulse can be generated by using an AND gate and a short external "Start" pulse. Again we have the problem that both edges of the pulse travel at different speeds, so the pulse gets either wider or narrower over time, depending on the exact parameters of the inverters. One possible solution to this is to use a "tail-biting" circuitry (Fig. 3). The inverters are designed such that the pulse gets wider over time, but the "tail" of the pulse gets truncated at each cell, thus ensuring a constant width over time.

While this circuitry keeps the pulse width fixed over time, the actual sampling speed changes largely with temperatures and power supply voltage. Some SCA designs try to measure the sampling speed constantly during the data acquisition and correct for any change offline.

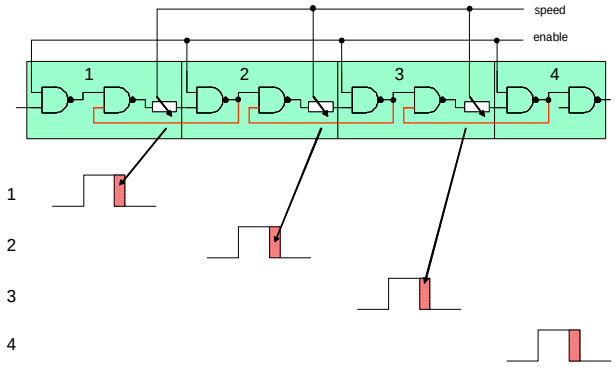

Fig. 3: "Tail-biting" circuitry which truncates a pulse to keep the width constant over time. The variable resistors between the gates allow controlling the pulse propagation speed over a wide range.

Other designs use an internal or external phase locked loop (PLL) to lock the sampling frequency to an external clock. The write pulse is tapped at one of the inverter stages and converted into a 50% duty cycle signal using a T flip-flop. The frequency of this signal is then equal to the sampling frequency divided by the number of inverter stages divided by two (for the T flip-flop) and lies typically in the MHz range. Therefore it can be compared easily to an external reference clock which is generated directly from a FPGA or a clock generator. The output of the phase generator is fed into a charge pump, which drives the analog speed control voltage for the inverter chain (Fig. 4).

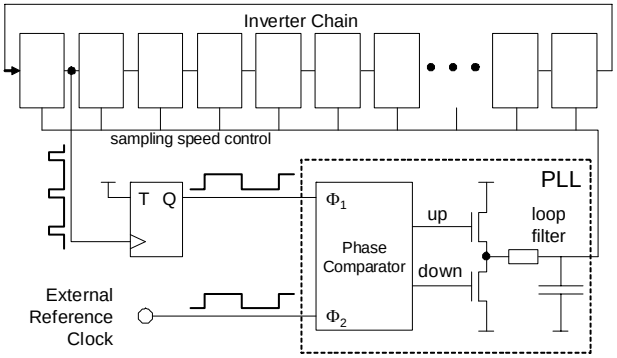

Fig. 4: Control of the sampling speed using a PLL and an external reference clock

This works very similar to a standard PLL, where the voltage controlled oscillator (VCO) is replaced by the inverter chain. By tuning the loop filter correctly, a very small residual phase jitter can be obtained. Several SCA designs report a value in the 20-30 ps range over a wide temperature range.

An associated problem comes from the fact that the propagation time of a pulse through an inverter stage differs on a chip. This comes from small geometrical and doping uncertainties of the used CMOS transistors inside the inverters. In modern deep sub-micron processes this effect can change the propagation time by 10-30% easily.

While there is a certain random pattern between neighboring inverters, there is usually also a gradient of the propagation time from one side of the chip to the other, coming from gradients of the process parameters in the wafer production. While the variation between two neighbors (differential temporal nonlinearity) is usually small, it can add up to much bigger values when summing up these values over many inverter stages (integral temporal nonlinearity). If the propagation time of inverter cell i and therefore the sampling bin width is Δt_i, then the differential temporal nonlinearity can be expressed with $\Delta T_i = \Delta t_i - \Delta t_{nominal}$, where $\Delta t_{nominal}$ is the nominal bin width and the inverse of the sampling speed. At 5 GSPS sampling speed for example, $\Delta t_{nominal}$ is 200 ps. The deviations from the nominal bin width ΔT_i contain usually two parts. The part which is fixed over time and originates from the geometric uncertainties in the inverters is called "fixed aperture jitter". The other part is called "random aperture jitter" and varies from measurement to measurement. Since the fixed aperture jitter is constant over time, it can be measured and corrected for. This can be done for example by sampling a sine wave with a well known frequency of maybe 500 MHz produced by a function generator. The effective bin widths Δt_i are then modified in an iterative process over many sampled sine waves, until the deviation between the sampled sine waves and the theoretical sine waves become minimal. The integral temporal nonlinearity ΔTI_i can then be expressed as

$$\Delta TI_i = \left(\sum_{j=0}^{i} \Delta t_j \right) - i \cdot \Delta t_{\text{nominal}}$$

An example of such an integral temporal nonlinearity is shown in Fig. 5. It shows the values ΔTI_i for a SCA with 1024 sampling cells running at 5 GSPS.

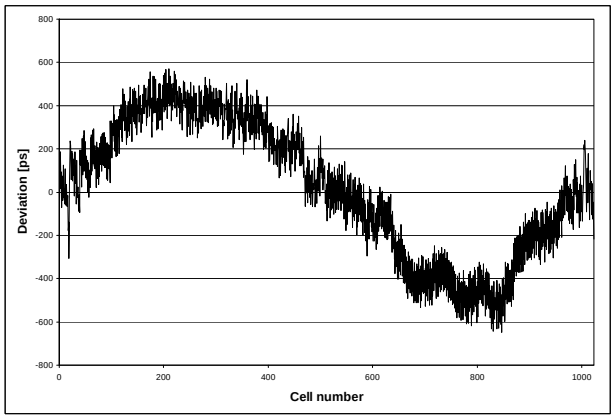

Fig. 5: Integral temporal non-linearity of a typical SCA chip with 1024 sampling cells

While the neighbor-to-neighbor variations are relatively small, adding them up leads to an integral nonlinearity of up to ±500 ps. The chip used in this case has actually a folded inverter chain, where cells 0 to 511 go in one direction from one side of the chip while cells 512 to 1023 go back in the opposite direction. The mentioned gradient of the process parameters in the wafer production can therefore clearly be seen as a slow variation of the integral nonlinearity over all cells.

After the fixed patter jitter is measured and correct for, the remaining residual random jitter is usually very small and determines the timing resolution which can be achieved with this technology. Various groups reported values for the residual jitter below 10 ps, making this technology very attractive for precise timing measurements. It has been shown that high speed waveform digitizing gives the best timing resolution compared with other technologies such as constant fraction discrimination or multi-level discrimination in combination with high resolution time-to-digital converters (TDC).

Sampling Depth and Analog Bandwidth

A critical issue in SCA technology is to obtain a high analog bandwidth. Even with a 10 GSPS sampling speed, the achievable timing resolution will be low if the analog bandwidth is low. This is because the analog bandwidth determines the maximum rise-time of a sampled signal. Timing jitter arises primarily from voltage jitter, mediated through a finite rise time. This is illustrated in Fig. 6.

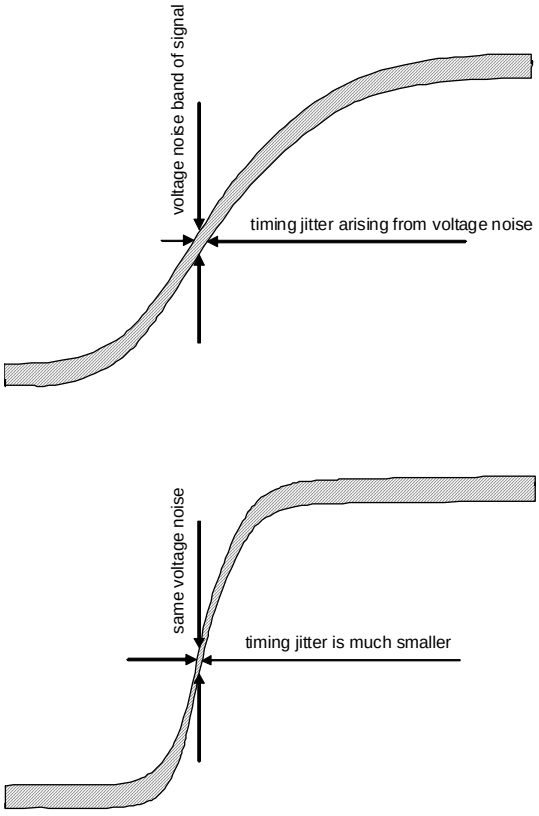

Fig. 6: Any voltage jitter of a signal leads to timing jitter. For a slow rising signal (top) the associated timing jitter is larger than for a fast rising signal (bottom).

As can be seen, the associated timing jitter for the same amount of voltage jitter is smaller for a faster rise-time of the signal. This is relevant both for discrimination (fixed threshold or constant fraction) as well as for direct waveform sampling.

The analog bandwidth of a SCA chip depends largely on the capacitance of the input bus (INx in Fig. 1). Each analog switch (which usually consisted of one NMOS and one PMOS transistor forming a transmission gate) has a typical parasitic capacitance in the order of 20 fF in a 250 nm CMOS technology. On a chip with 1000 sampling cells, this can easily add up together with the parasitic capacitance of the input bus to an input capacitance of 30-40 pF. The bond wire of the input line has a resistance of typically 2-3 Ohms, which forms with the input capacitance a low pass filter with a cut-off frequency of:

$$f_{3dB} = \frac{1}{2\pi RC} = 1.8\ \text{GHz}$$

If the signal source has a non-zero impedance, this value is reduced further. For a 20 Ω source impedance, the bandwidth is already reduced to 200 MHz.

The input capacitance can be reduced of course if one uses an active input buffer. It is easy to design an Operational Transconductance Amplifier (OTA) in a 350 nm or 130 nm CMOS process with a bandwidth above one Giga-hertz and an input capacitance in the order of a few pF, but the problem is the output. The OTA has to drive the above mentioned 30-40 pF, which reduces its bandwidth dramatically. It is of course possible to daisy-chain several OTAs in series, so that each of them has to drive only a fraction of the input bus. But in this case the transfer functions of each OTA have to be convoluted with each other, shifting the 3dB point to lower frequencies. If several OTAs are used in parallel as a fan-out of the input signal to several smaller bus segments, this is not the case, but the input capacitance is then multiplied by the number of OTAs, thus increasing the input capacitance and reducing again the bandwidth. Most existing SCA chips with internal OTAs have therefore a bandwidth not exceeding 300 MHz.

The obvious thing to do of course is to reduce the number of sampling cells. Some extreme designs with only 32 sampling cells show a bandwidth above 5 GHz, but of course such designs are only usable for a limited number of applications. The general trend is therefore the design of so-called “Cascaded Switched Capacitor Arrays” (CSCA). They consist of a short fast sampling stage with typically 32 sampling cells being constantly updated during the sampling time $t_{sampling}=1/f_{sampling}$. This fast sampling stage consists of the normal inverter chain and sampling capacitors used in a traditional SCA chip. Between two updates (the “hold” phase), the contents of a sampling cell is transferred into a secondary sampling stage. Since this transfer process can be done in $31*t_{sampling}$, the involved buffers can be much slower than the input bandwidth. The write switches for the secondary sampling stage can be operated by a simple shift register, which is clocked by the tapped write pulse of the fast sampling stage (Fig. 7).

Several groups are currently designing this new class of SCAs, some of them using a 130 nm CMOS process. Together with the speed gain of this process, SCAs with a sampling speed of 10 GSPS and an analog bandwidth of 5 GHz should be possible.

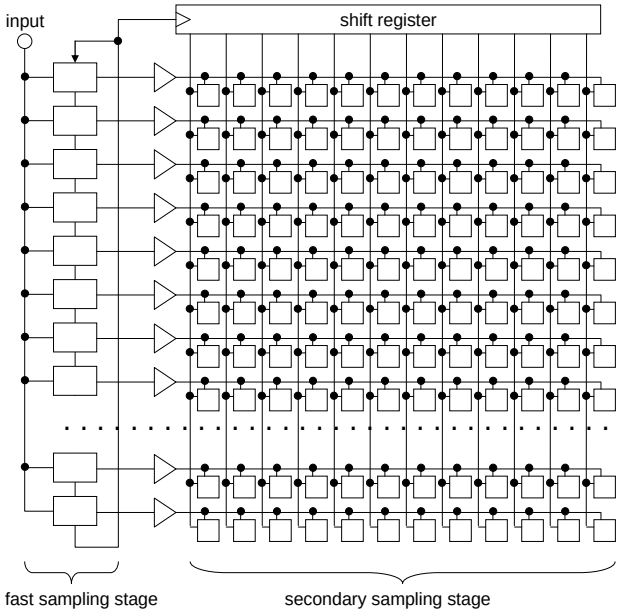

Fig. 7: Simplified schematic of a Cascaded Switched Capacitor Array

Readout Techniques

Once the analog signal is stored in the SCA chip, the next challenge is to read it out of the chip. In order to reduce the dead time of the system, this should be done as fast as possible without compromising the achievable signal-to-noise ratio.

There are two possibilities, either analog or digital readout. In analog readout, the cell contents is sent off-chip and digitized with an external ADC, typically with resolutions of 10-12 bits. The readout speed of SCA chips is in the range of a few MHz up to about 30 MHz. This allows the usage of cheap commercial ADCs. Several SCA channels can be multiplexed into a single ADC, reducing the system costs significantly. If a SCA with 1000 sampling cells for example is completely read out with 30 MHz, the readout dead time is 33 μs. In order to reduce the readout dead time, segments of the sampling cells can be read out and digitized in parallel, which then of course requires more external ADCs.

One problem of the analog readout is the fact that the storage capacitors are very small (usually between ten and a few hundred fF), thus the stored charge is too small to be sent directly off chip. Simple designs use therefore source followers in each sampling cell, which convert the stored voltage into a current, which can be sent off the chip or converted at the output pin back into a volt-

age using a shunt resistor. The problem here is that a simple source follower has a significant temperature dependence, which requires elaborate calibration procedures. Better designs use OTAs for each cell or for small groups of cells. In order to keep the power consumption of the chip low, these OTAs must be powered on only during the readout of the connected sampling cell. Some chips use more than 10'000 OTAs per chip, which makes it necessary that a minimal OTA design is chosen. This causes some non-negligible offset variation between the OTAs, but this is normally constant over time and temperature and can there fore be corrected for easily.

An alternative approach is to digitize the sampling cell contents directly on the chip. One interesting design is to use one ADC per sampling cell, which can be easily achieved by using a Wilkonson type ADC. This ADC consists of a simple comparator and a voltage ramp. The voltage ramp is produced by a counter connected to a DAC. When the signal from the voltage ramp exceeds the voltage stored in the sampling cell, a digital strobe signal is generated which is used to latch the contents of the counter generating the ramp (Fig. 8).

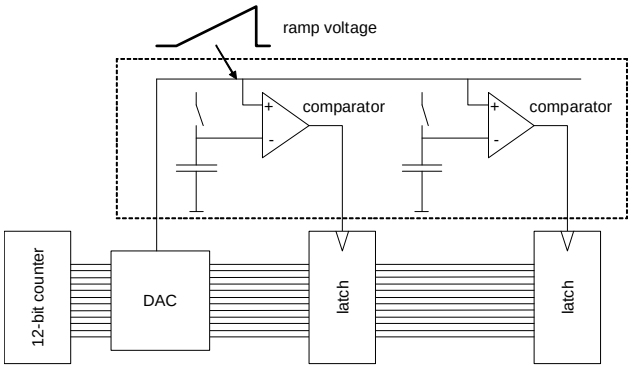

Fig. 8: Schematic of a SCA using Wilkinson-type ADCs to digitize the sampling cell contents

The nice feature of this technique is that all comparators can work in parallel, and in principle there is only a single ramp needed to digitize all cells. Some designs have the latches directly on the chip and read them out via large shift registers, while other designs implement the latches in the associated FPGA. This makes the SCA design very simple, since all it needs is one comparator per sampling cell (dashed line in Fig. 8). If the counter is clocked with 100 MHz for example, all 4096 codes can be ramped through in 40 μs. The problem is that each sampling cell then needs a separate data line from the chip to the FPGA. Given the limited number of pins possible on a chip, one cannot read out all comparators in parallel, but has to multiplex banks of comparators to the output pins, which increases the read-out time accordingly.

In order to reduce the readout time for both the analog and the digital case, it is possible to read only those sampling cells which contain a nonzero value. This weill be described in more detail under "Advanced Techniques".

SCA Examples

The first SCA chips were developed in the 1970's for time projection chambers. Since then, the CMOS processes have been improved constantly and SCA chips with increased sampling speed and larger sampling depth have been designed. All of the designs had to make certain choices in the readout method (analog vs. digital), the inputs (active or passive, single ended or differential), and stabilization (PLL). The designs are normally limited by the chip area needed by the sampling capacitors. Given a maximum chips size of typically 20-40 mm^2, different compromises of sampling depths vs. number of channels per chip were made. Following table gives a very limited overview of currently existing SCA chips in the GHz range and their basic characteristics.

Chip family	SAM [2]	LAB [3]	DRS [4]
Max. sampling speed	2.5 GSPS	3.7 GSPS	6 GSPS
Analog Bandwidth	300 MHz	900 MHz	950 MHz
Number of channels	2	1-16	9
SNR	13.4 bits	10 bits	11.4 bits
Sampling depth	144-2520	256-64k	1024-8k
Readout time	650 μs	150 μs - 10 ms	30 ns * $n_{samples}$
Input buffers	Yes	Yes	No
Stabilization	Int. PLL	Ext. DAC	Int. PLL
ADC	External	Internal Wilkinson	External
Supply voltage	3.3V	2.5V	2.5V
Power/channel	150-500 mW	15-50 mW	14-45 mW
Process	AMS 0.35	TSMC 0.25	UMC 0.25

Table 1: Basic characteristics of some SCA chip families

The achievable resolution with these chips depends on the signal-to-noise ratio, which is given by their intrinsic noise and the usable input range.

All groups involved in the development of the chip families listed in Table 1 are currently developing new versions, so new exciting SCA chips are supposed to show up in the next years.

Advanced Techniques

As mentioned already, one design goal for SCA chips is to minimize the readout time. One possibility is to read out only a subset of the sampling cells. Particle detectors produce normally only short pulses, so most of the sampling cells usually contain zero or "baseline" values. Some designs use on-chip comparators to determine

which cells have a nonzero value, while other designs use external circuitry to determine which cells are of interest and should be read out. One interesting method is used in the DRS4 chip, called “Region-of-Interest” (ROI) readout mode. Let’s assume a short pulse from a particle detector in a wide sampling window (Fig. 9). The aim is now to read out only the hashed area containing the signal. Since the write pulse from the inverter chain runs constantly in a circular fashion and the trigger happens at random times, this region can be anywhere in the sampling window. In principle one could determine the position of the region with an external timing unit, but then the problem would be to program the chip correctly to address only the interesting area. Given the limited number of pins, the address of that window would have to be clocked in sequentially for each read-out, which would increase the readout time again.

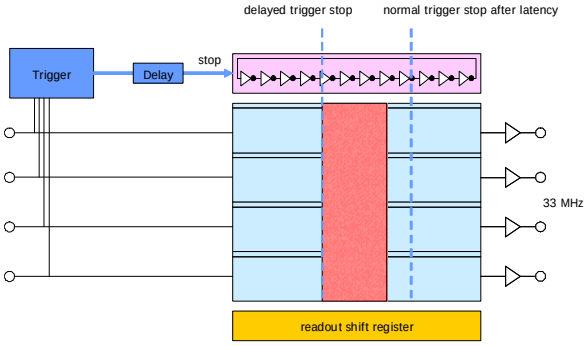

Fig. 9: Scheme of the Region-of-Interest readout mode in the DRS4 chip

The DRS4 chip solves this problem in an elegant way. If we consider some trigger latency, the inverter chain might be stopped at the right vertical dashed bar labeled “normal trigger stop”. With a simple external trigger delay, this stop position can be shifted the left vertical bar labeled “delayed trigger stop”, since the write pulse wraps around in the inverter chain. The external delay is designed such that this delayed stop position is exactly in front of the ROI part of the waveform. The DRS4 chip has now the possibility to transfer this stop position in a single clock cycle into the readout shift register, so that the readout is started exactly from this position. After the desired number of readout clock cycles (hashed area), the read-out is stopped and the chip is restarted. If the ROI consists for example of only 50 samples out of the 1024 samples of the full sampling window, the read-out time can be reduced from 30 µs to 1.5 µs when the external digitization is done with 33 MHz, which would allow several 100’000 acquisitions per second.

Most existing SCA chips arrange their sampling cells in rows and columns, in order to fill the rectangular chip area optimally. This gives the designer the choice to trade sampling depth versus the number of channels. Some applications require a very deep sampling depth, which can be achieved if all sampling cells are connected in a single sequence. One extreme case is the BLAB1 chip [5], which connects 128 rows of 512 samples to form a single array with 65536 sampling cells. Rather than fixing a design on a certain number of channels, it is however possible to assign channels dynamically. This can be achieved for example by using a shift register, which enables only certain rows of the sampling array as can be seen in Fig. 10.

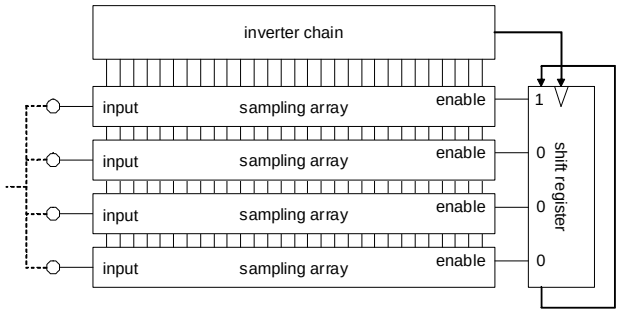

Fig. 10: Channel cascading by using a separate shift register

If the shift register in this example contains a “1111” pattern, all four rows are always enabled and the chip works with four channels. If it is loaded however with a “1000” pattern, and this pattern is shifted after each rotation of the write pulse in the inverter chain, then one has effectively a chip with a single channel with four times the sampling depth. Of course it is then necessary to connect the same analog source to all four inputs externally (dashed lines). The nice thing in this case however is the fact that the analog bandwidth is not compromised, because one has four times the input capacitance, but also four bond wires in parallel, so the analog performance is exactly the same as in the four channel configuration, given that the signal source impedance is negligibly small. This principle can be extended to span several chips, if the shift register input and output is connected to two pins of the chip and daisy-chained through all chips. This way one can obtain very deep sampling depths. Since the signal source impedance will then start playing a role, the signal must be split actively, so that each chip input is driven by a separate buffer.

A common problem related to the readout of SCA ships is how to form a trigger. The SCA chip needs to be stopped for readout when a physical event occurred and the inputs carry some signal. The traditional way is to split the input signal, where one branch goes to the SCA chip and the second branch goes to a dedicated trigger electronics, which uses discriminators and logic elements to form a trigger. An alternative approach is to use some additional circuitry inside the SCA chip to form an on-chip trigger. One problem here is that the chip then becomes specific for the application it has been designed for. A more general solution can be achieved by using the external ADC together with the FPGA reading out that ADC (Fig. 11).

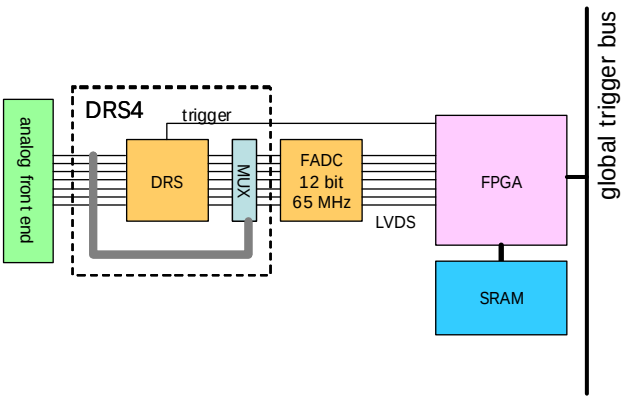

Fig. 11: Integrated readout and triggering with the DRS4 chip

The DRS4 chip has eight channels, which can be read out in parallel with modern 8-channel ADCs, such as the AD9222. The ADC is read out by a FPGA, which is also responsible of controlling the SCA chip. Using the same hardware, a trigger can be formed by utilizing the multiplexer built into the DRS4 chip. During the sampling phase, the input signals are connected to the switched capacitor array but a copy is also routed to the output, where it is digitized continuously by the ADC. The FPGA obtains therefore the signal of the inputs with a sampling rate of 65 MHz in this case. Based on this information, it can derive a trigger using some digital comparators and logic. Local trigger information can be sent to a global trigger bus, where it can be combined into a global trigger, which then stops the SCA chip. The multiplexer at the SCA chip is then flipped over and the capacitor array is read out though the same ADC and FPGA. This method allows for some sophisticated trigger algorithms without additional hardware, given that the timing resolution of 65 MHz is enough.

Conclusions

The development of SCA chips in the multi-GHz range has observed a big progress in the last couple of years, since it was realized that this technique can replace traditional ADCs and TDCs at a much lower cost and power consumption. The high channel densities of some of these chips make it possible to integrate many thousands of channels in a single rack, or even mount the chips directly on the associated detectors. It is considered to be the enabling technology for large gamma ray telescope arrays such as CTA and AGIS.

Future development goes into the direction of higher analog bandwidth and sampling speed, which are required to instrument upcoming particle detectors working in the few-ps timing resolution range. More designs will surface using on-chip preamplifiers, trigger circuitry and having integrated fast ADCs. As the community using these chips is growing, the costs per channel are dropping to a few dollars, making the cost of front-end electronics dominated by external components.

References

1. S. Ritt for the MEG collaboration, Nucl. Phys. Proc. Suppl. **162** (2006) 279
2. E. Delagnes *et al.*, NIM **A567** (2006) 21
3. G. Varner *et al.*, NIM **A583** (2007) 447
4. S. Ritt, NIM **A518** (2004) 470 and http://drw.web.psi.ch
5. L. Ruckman, G. Varner and A. Wong, NIM **A591** (2008), 534