Digital Silicon Photomultipliers with OR/XOR Pulse Combining Techniques

Digital Silicon Photomultipliers with OR/XOR Pulse Combining Techniques

Salvatore Gnecchi*†, Neale A.W. Dutton*†, Luca Parmesan*†,
Bruce R. Rae*, Sara Pellegrini*, Stuart J. McLeod*, Lindsay A. Grant*, Robert K. Henderson†
*ST Microelectronics Imaging Division, Edinburgh, United Kingdom
†The University of Edinburgh, Edinburgh, United Kingdom

***Abstract*—A recently proposed XOR-based Digital Silicon Photomultiplier is compared against the OR-based counterpart. We show experimental data from a set of SPAD pixel arrays in 130nm CMOS process with selectable OR tree and XOR tree for direct comparison. We demonstrate how XOR-based dSiPMs solve the limitation caused by monostable circuits and reach higher maximum count rates compared to optimised OR-based dSiPMs. The increased throughput of the SPAD array allows higher sampling rates for the digitisation of the light signal enhancing dynamic range and linearity.**

***Index Terms*—Single Photon Avalanche Diodes, SPAD, digital SiPM, dSiPM, XOR tree, OR tree, pile-up, PET**

I. Introduction

DIGITAL Silicon Photomultipliers (dSiPMs), [1], have now gained popularity in many applications including Positron Emission Tomography (PET) [2], visible light communications [3] and time-of-flight LIDAR [4]. Compared to their analogue equivalent (aSiPMs, see Fig. 1(a)), dSiPMs have many additional advantages such as the inclusion of in-pixel CMOS circuitry and on-chip timing and counting abilities without the need of external converters, see Fig. 1(b).

The linearity and timing resolution of dSiPMs are limited by the maximum rate at which the sensor electronics can count and/or timestamp single photons (*photon throughput*). In PET, loss of photons from the short ($\sim$ 100ns) but intense bursts of light (few thousand photons) from gamma scintillation events occurs primarily at the leading edge of the optical waveform. At this peak in photon arrival rate, a reduction in the detection rate of incident photons can occur due to the limited bandwidth of the digital processing electronics or due to photon *pile-up* within the SPAD detector array, i.e. undetected photons due to non-zero dead time of the sensor [5]. This degrades both the energy resolution linearity and the time of flight or coincidence resolving time (CRT), [6].

S. Gnecchi, N. A. W. Dutton and L. Parmesan are with the Imaging Division, STMicroelectronics, Edinburgh EH12 7BF, U.K. and also with the CMOS Sensors and Systems Group, The University of Edinburgh, Edinburgh EH9 3JL, U.K. (e-mail: s.gnecchi@ed.ac.uk; n.dutton@ed.ac.uk; l.parmesan@ed.ac.uk).

R. K. Henderson is with The University of Edinburgh, Edinburgh EH9 3JL, U.K. (e-mail: robert.henderson@ed.ac.uk).

B. R. Rae, S. Pellegrini, S. J. McLeod and L. A. Grant are with the Imaging Division, STMicroelctronics, Edinburgh EH12 7BF, U.K. (e-mail: bruce.rae@st.com; sara.pellegrini@st.com; stuart.mcleod@st.com; lindsay.grant@st.com).

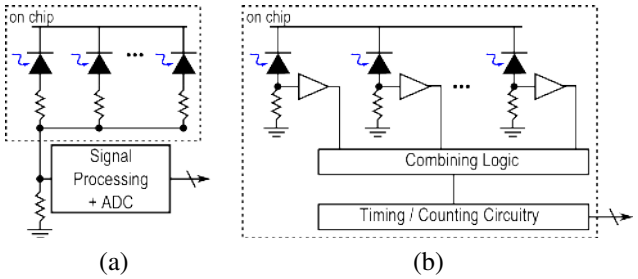

Figure 1. **Silicon Photomultipliers** - SPADs are aggregated into arrays. The outputs can be combined in an analogue way (a) or digitally (b).

Recent works have shown single photon avalanche diode (SPAD) array designs where pixels have achieved low dead times, [4], [7]. Typically, balanced "all-to-one" OR trees digitally combine SPAD pixels. Monostable pulse shortening circuits per input, [8], improve the throughput by reducing the dead time of the tree. Furthermore, an alternative combining approach has been proposed by the replacement of the OR tree with an XOR tree and the input monostable circuits replaced by toggle-type flip flops (TFFs) encoding SPAD events on both rising and falling edges, [9]. Fig. 2 summarises the variants of digital combination logic of SPAD pixels.

This work presents the first direct comparison between these techniques in dSiPM design. We provide experimental data from a test chip manufactured in 130nm CMOS process with selectable on-chip OR tree and XOR tree. We demonstrate the efficiency of XOR-based dSiPMs compared to OR-based topologies showing higher throughput, enhanced dynamic range and linearity. This would benefit PET applications, by improving gamma time of flight estimates by allowing more photons from the leading edge of the scintillation to be resolved, [6], [9]. In optical communications this allows increased data rate through complex modulation schemes and greater tolerance of background light. In LIDAR distance estimate distortion due to pile-up in the combining electronics will be reduced. A theoretical model for SPAD arrays is derived in Section II. The test chip is described in Section III with results following in Section IV. The final Section V provides conclusions and outlook of this work.

II. SPAD Array Detector Model

A SPAD array detector is made of N identical pixels described by the same dead time τ_d. In dSiPMs they form a single detector since the pixel outputs are combined

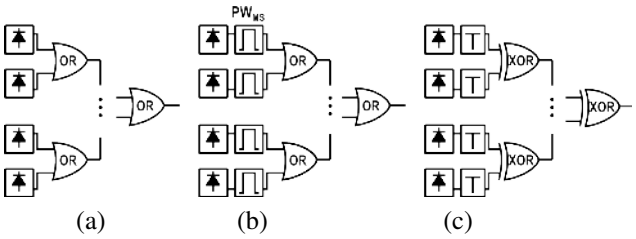

Figure 2. **Digital combination logic networks** - The single outputs are combined into: (a) single channel through an OR tree, (b) a monostable pulse shaper $\mathrm{PW_{MS}}$ + OR tree, (c) a toggle + XOR tree.

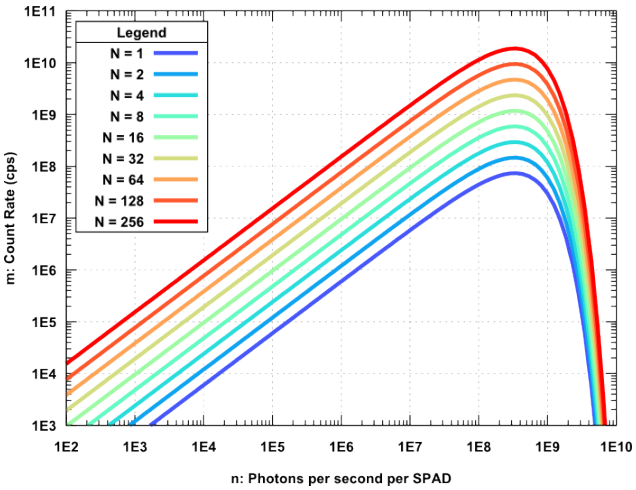

Figure 3. **Multi channel dSiPM** - The count rate is only limited by the paralysis of each pixel. The availability of N counters does not limit the maximum count rate. The parameters used are $\tau_d = 5\mathrm{ns}$ and PDE = 25%.

together. In the case of multi-channel dSiPMs, where each pixel has its own dedicated counter/converter [10], [11], the average count rate of the array is N times the count rate of each pixel which, in the case of paralysable SPAD pixels [12], can be written as:

$$m_0(N) = N \cdot m_{\mathrm{pixels}} = N \cdot n \cdot e^{-n \cdot \tau_d} \tag{1}$$

with a maximum equal to:

$$\mathrm{max_{M-C}} = \frac{N}{e \cdot \tau_d} \tag{2}$$

However, the count rate registered by an individual channel is limited either by the channel itself or by the counting circuitry. In fact, each channel needs to be fast enough to switch its electrical status (high to low and vice-versa) at every photon event. Moreover, the counting circuitry is required to have a bandwidth which allows all the detected events to be recorded.

In the case of a multi channel dSiPM, each counter has to handle a maximum count rate given by $\frac{1}{e \cdot \tau_d} \sim 200\mathrm{Mcps}$ [7]. Counters are typically able to process such frequencies, therefore the total count rate of an N-SPAD multi channel dSiPM is then limited only by the number of SPADs and the dead time τ_d as expressed by (1) and shown in Fig. 3.

When N pixels are aggregated together, limitations arise. The N-to-1 combining network might not propagate events to the single channel if their inter-arrival time is too low. The loss of counts due to the combining network is a process similar to the reduction of photon detection in the pixel itself. It is often referred to as *channel pile-up* [5]. This mechanism must be studied according to the network. The following sections provide analysis on tested OR-tree and XOR-tree.

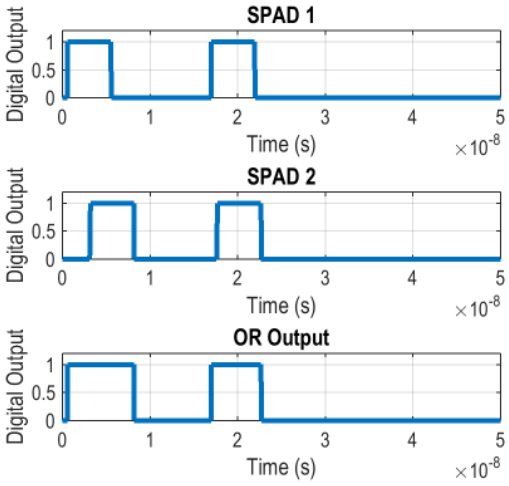

Figure 4. **OR Gate - Example waveforms**. Every photon event is encoded on each SPAD output as a rising edge. Through a 2-input OR gate, the second event within the pixel dead time is lost.

A. OR-based dSiPM

A digital SPAD pixel output is represented by a square pulse having a width equal to the dead time of the pixel, τ_d. When two different pixels detect a photon, the OR-gates need to propagate the two rising edges through all the tree to avoid count loss. However, an OR gate propagates only the first in time photon detected by any of its input pixels within a dead time window. As shown in Fig. 4, a 2-input OR gate merges two pulses together if they happen within the same τ_d extending the output pulse. Therefore, an M-input OR gate will propagate only 1 photon per τ_d missing any later $(M-1)$ photons in τ_d. This behaviour appears to be similar to the pile-up in passive recharge pixels. In a similar way of a paralysable detector [12], the propagated count rate of the OR tree can be written as:

$$m_{\mathrm{OR}}(n) = m_0(n) \cdot \exp(-m_0(n) \cdot \tau_d) \tag{3}$$

The maximum detection rate of the OR tree $\mathrm{max_{OR}}$ therefore equals the maximum count rate of a single SPAD, $\mathrm{max_{SPAD}}$:

$$\mathrm{max_{OR}} = \mathrm{max_{SPAD}} = \frac{1}{e \cdot \tau_d} \tag{4}$$

Such solution is commonly used in low light applications where the main goal is to cover a large active area with a high number of pixels [13], [14].

If the number of photons detected by the array exceeds $1/\tau_d$, simple OR gates are not sufficient. To overcome the dead time bottleneck, additional monostable circuits have been adopted in recent works [8], [15]. The example waveforms in Fig. 5 show how temporally compressing the SPAD pulses can increase the detection rate. In the provided example, the 2-input OR gate is now able to

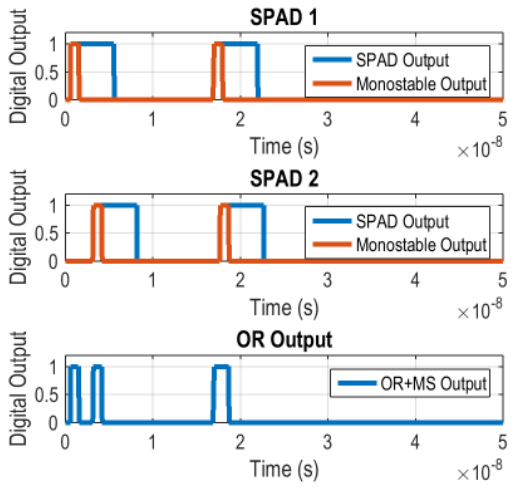

Figure 5. **Monostables + OR Gate** - The addition of monostable circuits cancels the limitation of the pixel dead time reaching higher detection rates.

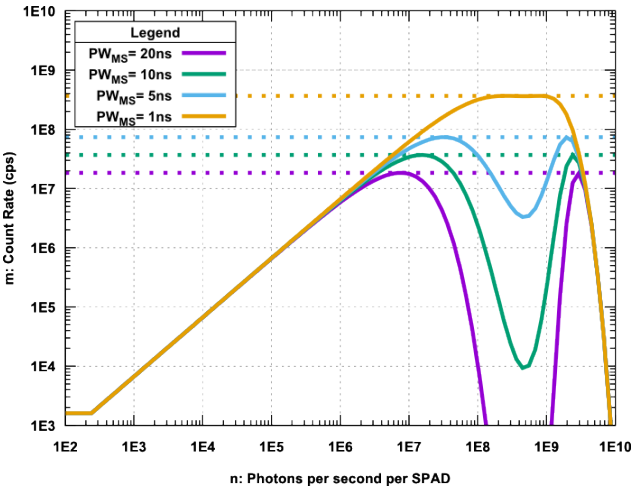

Figure 6. **Modelled count rate for monostables + OR tree** - Different pulse widths for the monostable circuits are simulated. The plot shows data from (5) with N = 16, PDE = 25% and $\tau_d = 5\text{ns}$. Dashed lines show calculated maxima.

propagate the first two detected photons. However, due to the pulse width of the monostable output, the last photon (second rising edge on SPAD 2) fails to propagate due to the pulses being merged together. The monostable output represents the main limitation of this architecture.

To describe such process, it is sufficient to replace τ_d with the shortened pulse width $\mathrm{PW_{MS}}$ in (3) which then becomes:

$$m_{\mathrm{OR+MS}}(n) = m_0(n) \cdot \exp(-m_0(n) \cdot \mathrm{PW_{MS}}) \quad (5)$$

The maximum count rate this time is dependent on the pulse width $\mathrm{PM_{MS}}$ and the number of SPADs: if the pulses are sufficiently compressed by the monostable circuits, the ideal maximum is given by (2), otherwise it is limited by the pulse width. In the latter case the maximum count rate is:

$$\mathrm{max_{OR+MS}} = \frac{1}{e \cdot \mathrm{PW_{MS}}} \quad (6)$$

Predictions of this model for 16 SPADs with 5ns dead time and 25 % PDE are presented in Fig. 6 (maxima are highlighted with dashed lines). Pulse widths larger than the SPAD dead time are drawn for the purpose of showing the effect of monostable cells and /or emulating longer dead times. The plotted lines show an increase of the counts at high photon rates due to the paralysis of the single pixels being attenuated by the non-paralysable network.

B. XOR-based dSiPM

Monostable circuits have been designed to reach pulse widths as low as few hundreds of picoseconds, [2] although routing a bias voltage down to each monostable cell is always necessary.

Recent works have proposed to replace the monostable circuits with toggle cells followed by an XOR-tree replacing the OR gates [3], [9]. The pulse train of each SPAD pixel is toggled to generate a signal where each transition contains time information of the photon events.

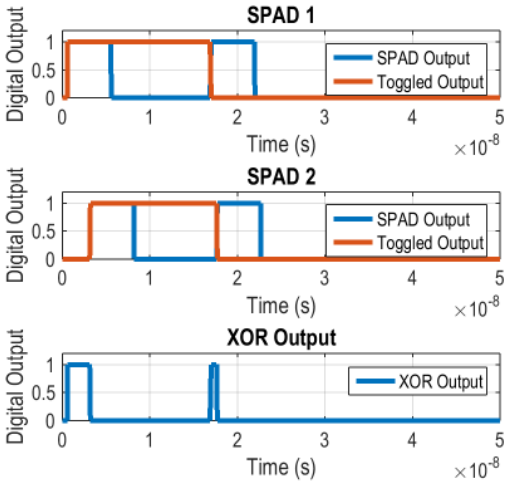

Figure 7. **XOR Gate - Example waveforms** - The toggle cells generate both rising and falling edges at each photon event. The edges are propagated through the XOR gates.

Both signal edges need to be successfully propagated to the single channel by XOR gates, refer to Fig. 7 as an example. This eliminates the need of shrinking the SPAD pulses since the combination is now done through XOR gates.

The maximum detection rate for a toggle + XOR-tree network, $\mathrm{max_{XOR}}$, is limited by the ability of the electrical signal (the single output channel) to create a certain minimum pulse width $\mathrm{PW_{MIN}}$ to be then processed by the counter/converter. Since both edges are representative of photon events, the maximum detection rate can be written as:

$$\mathrm{max_{XOR}} = \frac{1}{\mathrm{PW_{MIN}}} \quad (7)$$

This limitation to the count rate can be modelled in a similar way of a non-paralysable detector model[12]:

$$m_{\mathrm{XOR}}(n) = \frac{m_0(n)}{1 + m_0(n) \cdot \mathrm{PW_{MIN}}} \quad (8)$$

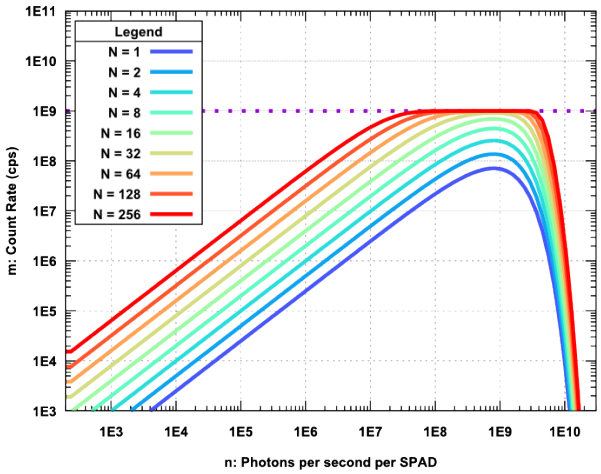

Figure 8. **Modelled count rate for XOR tree** - The count rate tends asymptotically to the maximum allowed by the technology. The modelled behaviour resembles non-paralysable detector models. The maximum count rate is set to $^1/_{\mathrm{PW_{MIN}}} = 10^9$.

The modelled equation is graphed for a different number of SPADs in Fig. 8. The results show a significant difference compared to the OR-tree: the elimination of the monostable takes away the limitation on the maximum count rate and it further changes the profile of the saturation region. As seen in the graphs, a saturated XOR tree shows a flat region when many SPADs are combined together. The reduction of count rates at high photon rates is due to the paralysis of the individual SPADs: the reduced $m_0(n)$ reflects on (8).

C. Low photon inter-arrival time

We here discuss the loss of photons at low inter-arrival time

$$\Delta t_p \ll \mathrm{PW_{MIN}} \tag{9}$$

to show the effect of combining two or more event with such inter-arrival time through an OR tree and an XOR tree. First, Fig. 9 shows a zoomed-in view of two very close-in-time photons incident on two separate SPADs. Somewhere in the N-to-1 network, these two events will be combined through an OR gate or an XOR. In the first case, assuming that a monostable cell is available to create a pulse width equal to the minimum pulse width of the single channel, then, as expected, the two events will be merged together into a single pulse and the timing information of the latter event is lost. In the same situation, the XOR gate shows a much more critical performance. In fact, the gate should ideally create two consecutive edges within a very short time, hence a pulse width $\mathrm{PW} = \Delta t_p$ but since the rise and fall times are not fast enough, no edges will be created at the output of a non-ideal gate. With no edges, the XOR output shows no trace of the either photon events.

We can conclude that OR-trees are able to preserve the information of the first detected photon while each XOR gate in the tree cancels each pair of photons but one photon has a chance to survive if the number of incident photons in a particular time window is odd. Contrary to the OR-tree example, where the propagated photon event is the

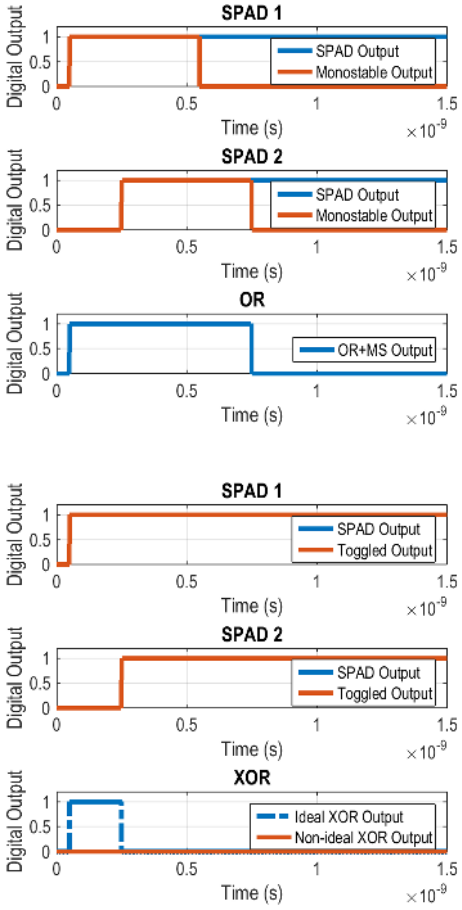

Figure 9. **Low photon inter-arrival time** - The OR gate manages to detect the first-in-time photon whereas the XOR gate misses both photons due to the inability to create a very narrow pulse width.

first in time detected photon, nothing can be said about the eventual propagated photon even in the XOR tree since the cancellation of the pairs is not predictable.

Both approaches share the common limitation that in applications where the system is required to detect short bursts of simultaneously emitted photons, the detection is not going to be successful. For such applications the most efficient architecture is represented by the multi-channel approach.

III. Test Chip

The test chip has been manufactured in STMicroelectronics 130nm CMOS process with five SPAD arrays, as pictured in Fig. 10. Moreover, the individual SPAD outputs of each array can be combined onto single channel through either monostable circuits with OR tree or toggles with XOR tree, as in Fig. 2(b) and (c). Both networks are placed on-chip, beside the pixel arrays to maximise the fill factor. The monostable cells are voltage-controlled by an off-chip DAC. The selected combination logic is attached to a 16bit ripple counter or, alternatively, off-chip via a buffered pad for characterisation with an oscilloscope. An FPGA-controlled enable signal for the counter adjusts the exposure time of the sensor. The chip schematic is provided in Fig. 11. Each array is composed by 4×4

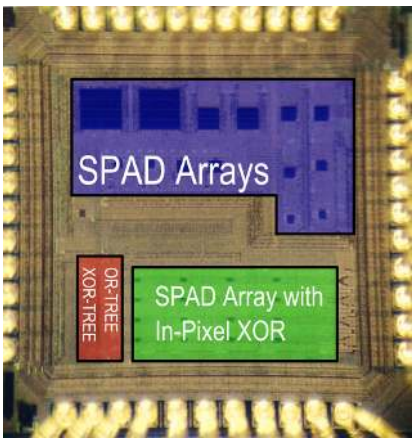

Figure 10. **Test chip** - A selection of 4×4 SPAD arrays were manufactured together with selectable combination logic. An optimised 16×16 array with XOR tree is also available.

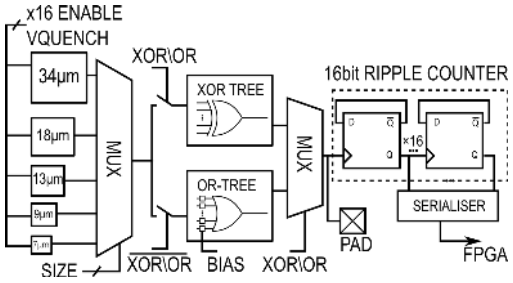

Figure 11. **Test chip schematic** - Selection of five pixel pitch variants for SPAD arrays plus selectable combination logic. On-chip counters stream out the total count for the chosen dSiPM configuration.

SPAD pixels. Enabling transistors allow the control of the number of activated SPADs, a quenching transistor acting as a voltage controlled resistor allows tunable SPAD dead time. The 16 outputs of each array are connected on a common bus through tri-state buffers for the selection of individual arrays. Five pixel-pitch variants were designed for the 4×4 arrays from 7µm pitch with 2µm SPAD diameter to 34µm pitch with 32µm SPAD diameter, see Table I.

An additional 16×16 array of 7µm pixel pitch with dedicated XOR tree is available for further investigations.

IV. Experimental Results

The SPAD array and combining logic tree configurations have been tested in a range of light levels set by a current controlled LED. For each light level, the average

Table I
SPAD ARRAY SET - SPAD ARRAY PARAMETERS

Pitch (μm)	**Fill Factor** (%)	**DCR*** (cps)	**Dead Time**
7	6.4	23.6	
9	18.7	278	
13	37.4	1.33×10^3	$\tau_d \simeq 5$ns
18.6	73.6	3.04×10^3	
34.6	85.4	3.22×10^3	

* *per SPAD*

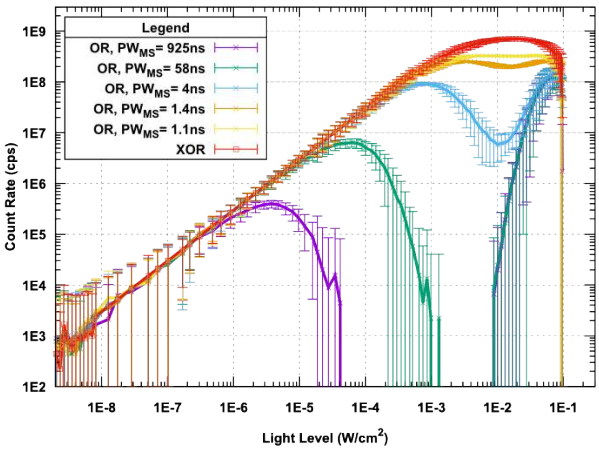

Figure 12. **Combining technique analysis** - The plot shows the count rates of the XOR tree (red line) and the OR tree with different pulse widths $\mathrm{PW_{MS}}$ set by the monostable circuits. The array consists of 16 of the 7 µm pitch SPADs. A reduced chi-squared of $\chi^2/DoF \simeq 1.49$ (worst case) indicates a good match of the experimental data with the predicted equation.

count rate is estimated by dividing the on-chip ripple counter output by the chosen exposure time. To improve the statistics, the measurements were iterated allowing a mean value and a standard deviation to be computed from the ensemble of registered values of count rate. The latter is used as indication of the uncertainty in the error bar plots throughout this section.

Fig. 12 shows results obtained by the 7µm pitch SPAD pixel array with all sixteen pixels enabled. Different monostable pulse widths, $\mathrm{PW_{MS}}$, have been used in OR dSiPM analysis (shown with coloured crosses with error bars). The red squares and error bars show the average count rate recorded by the XOR tree. The data shows the impact of the monostable circuit in the combining process as the maximum count rate is limited by its pulse width. The results resemble the model shown in Fig. 6. Furthermore, a higher count rate is registered by the XOR dSiPM which is mainly limited by the number of SPADs (no flat region).

To investigate the dSiPM which does not show flat saturation with 16 SPADs, the same experiment was repeated enabling the 16×16 array with XOR tree described in the previous section. Results of the measurements are shown in Fig. 13 where the number of activated SPADs has been swept from one single SPAD to the full array. The flat saturation starts to become significant for a number of activated SPADs $N > 32$. After that point, the dSiPM is not able to process the count rate of each SPAD giving as results a limited maximum count rate. These results match the predictions of the proposed model shown by (8) and Fig. 8.

As a final test, both architectures were tested with a common limitation. A dedicated output pad gives to the chip the ability to connect the final XOR/OR signal to external counters. FPGA ripple counters can be then used instead of the dedicated on-chip ones. This feature allowed a further limitation to be introduced on both combining networks to better understand eventual common bottle-

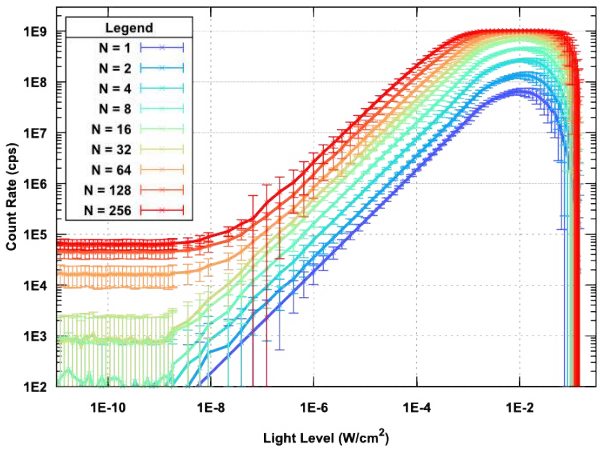

Figure 13. **XOR tree** - The data show the saturation of an XOR based dSiPM with tunable number of activated SPADs N. The modelled equation fits the data with a reduced chi-squared of $\chi^2/DoF \simeq 1.54$ worst case.

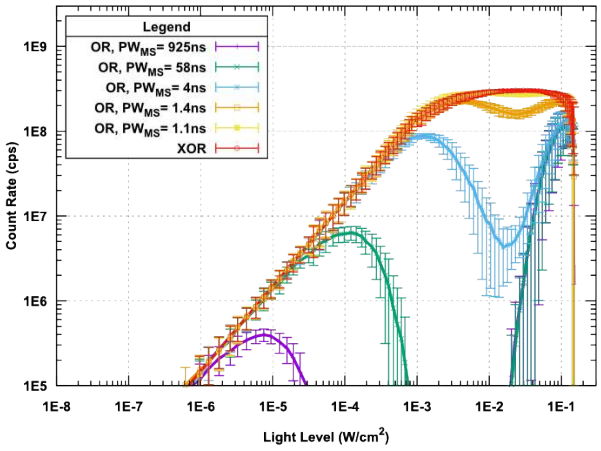

Figure 14. **External FPGA Counters** - The plot shows the average counts recorded by external counters. The array configuration is the same as the case shown in Fig. 12. For this experiment, the output has been brought through a dedicated pad to the FPGA where it has been connected with off-chip counters. The calculated reduced chi-squared is $\chi^2/DoF \simeq 1.49$ (worst case)

necks in dSiPM design. In Fig. 14 the recorded counts are shown. Two things are evident from this test. First of all the maximum registered count rate is much lower compared to the previous experimental set-up: $\sim$ 300Mcps against $\sim$ 320Mcps (OR tree) and $\sim$ 700Mcps (XOR tree). This important result shows how the output pad affected the signal of both network limiting the count rate to a common maximum registered by the FPGA counter proving the advantage of realising the counting system on-chip.

V. Conclusions and Outlook

We have demonstrated the efficiency of XOR-based dSiPMs reaching higher detection rates compared to optimised OR-based counterpart in the same CMOS process. We show the saturation region at high count rates and provide a model well verified by the experimental data. We furthermore prove the benefits of the elimination of monostable cells typical of OR-based dSiPMs. A full summary of dynamic range, linearity and throughput of the repsented dSiPMs is presented in Table II highlighting the performance of the proposed XOR-based dSiPM. This work looks towards a full characterisation of single-channel dSiPMs based on OR and XOR tree. The outlook of this work will be demonstrating the benefits of high detection rate SPAD arrays with high sampling rate timing circuits.

Table II
SPAD ARRAY PERFORMANCE - SPAD PITCH = 7µm,

dSiPM Type	N_{SPAD}	PW_{MS}	Dynamic Range (dB)	Linearity (R^2)	Throughput
OR	**16**	**1ns**	**107**	**0.991**	**327 Mcps**
		1.4ns	107	0.997	252 Mcps
		4ns	96	0.997	91.7 Mcps
		58ns	74	0.989	6.37 Mcps
		925ns	51	0.979	398 kcps
XOR	1	-	140	1	64.9 Mcps
	2		130	1	136 Mcps
	16		**117**	**0.994**	**683 Mcps**
	128		87	0.987	992 Mcps
	256		83	0.987	998 Mcps

References

[1] T. Frach *et al.*, "The Digital Silicon Photomultiplier - Principle of Operation and Intrinsic Detector Performance," *2009 IEEE Nucl. Sci. Symp. Conf. Rec.*, pp. 1959–1965, oct 2009.

[2] L. H. C. Braga *et al.*, "A Fully Digital 8x16 SiPM Array for PET Applications With Per-Pixel TDCs and Real-Time Energy Output," *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 301–314, 2014.

[3] O. Almer *et al.*, "A SPAD-based Visible Light Communications Receiver Employing Higher Order Modulation Schemes," *IEEE Globecom*, 2015.

[4] I. Vornicu *et al.*, "A CMOS 8x8 SPAD Array for Time-of- Flight Measurement and Light-Spot Statistics," in *Proc. - IEEE Int. Symp. Circuits Syst.*, 2013, pp. 2626–2629.

[5] J. Arlt *et al.*, "A study of pile-up in integrated time-correlated single photon counting systems," *Rev. Sci. Instrum.*, vol. 84, no. 10, pp. 103 105–103 105–10, oct 2013.

[6] S. Seifert *et al.*, "The lower bound on the timing resolution of scintillation detectors," *Phys. Med. Biol.*, vol. 57, no. 7, pp. 1797–1814, 2012.

[7] A. Eisele *et al.*, "185 MHz Count Rate, 139 dB Dynamic Range Single-Photon Avalanche Diode with Active Quenching Circuit in 130nm CMOS Technology," in *2011 Int. Images Sens. Work.*, 2011, pp. 6–8.

[8] C. Niclass *et al.*, "A 100-m Range 10-Frame/s 340 × 96-Pixel Time-of-Flight Depth Sensor in $0.18\mu m$ CMOS," *IEEE J. Solid-State Circuits*, vol. 48, no. 2, pp. 559–572, 2013.

[9] N. A. W. Dutton *et al.*, "A Time-Correlated Single-Photon-Counting Sensor with 14GS/s Histogramming Time-to-Digital Converter," in *Solid- State Circuits Conf. - (ISSCC), 2015 IEEE Int.*, 2015, pp. 204–206.

[10] J. Richardson *et al.*, "A 32x32 50ps resolution 10 bit time to digital converter array in 130nm CMOS for time correlated imaging," in *Proc. Cust. Integr. Circuits Conf.*, no. 029217, 2009, pp. 77–80.

[11] D. Bronzi *et al.*, "100 000 Frames/s 64×32 Single-Photon Detector Array for 2-D Imaging and 3-D Ranging," *IEEE J. Selected Topics Quantum Electron.*, vol. 20, no. 6, pp. 354–363, 2014.

[12] D. F. Yu *et al.*, "Mean and variance of single photon counting with deadtime." *Phys. Med. Biol.*, vol. 45, no. 7, pp. 2043–56, 2000.

[13] T. Frach, "Optimization of the digital Silicon Photomultiplier for Cherenkov light detection," *J. Inst.*, vol. 7, no. 01, pp. C01 112–C01 112, 2012.

[14] M.-A. Tetrault *et al.*, "Real-Time Discrete SPAD Array Readout Architecture for Time of Flight PET," *IEEE Trans. Nucl. Sci.*, vol. 62, no. 3, pp. 1077–1082, 2015.

[15] L. H. C. Braga *et al.*, "A CMOS Mini-SiPM Detector with In-Pixel Data Compression for PET Applications," *IEEE Nucl. Sci. Symp. Conf. Rec.*, pp. 548–552, 2012.