A 32×32 Single Photon Avalanche Diode Imager with Delay-Insensitive Address-Event Readout

Joseph Lin and Andreas G. Andreou
Department of Electrical and Computer Engineering
Johns Hopkins University, Baltimore, MD 21218
Email: {linjh,andreou}@jhu.edu

Abstract—We report the design and test of a 32×32 array of single photon avalanche diodes. The imager uses a delay-insensitive address-event link for the readout. The chip is fabricated in 0.18\(\mu\)m CMOS in an area of 1.886×1.866 of which 4% is occupied by the readout circuits.

I. SINGLE PHOTON IMAGING
Imaging light intensity down to single-photon levels has many applications in science and the military. In fluorescence microscopy [1], applications such as Förster Resonance energy transfer and fluorescence lifetime microscopy can benefit if the lifetime of the photon emitted by a molecule is measured. Military applications include night vision, range-finding and 3D-imaging [2].

Modern photon counters are generally expensive, delicate equipment found only in scientific laboratories. They are made from photomultiplier tubes (PMTs) or microchannel plates (MCPs), whose cost and size forbid their use in large integrated arrays. Photon counting can be accomplished with EM-CCD imagers, which have high resolution and low noise, however these imagers still remain very expensive and suffer from low frame-rates (typically 10 frames per second). Researchers are increasingly using avalanche photodiodes (APD), which are the solid state equivalent of PMTs. Recently it has been shown that they can be fabricated in a complementary metal-oxide-semiconductor (CMOS) process—the same processes used to fabricate commercial chips [3]. This opens the door for large scale integration and focal plane processing with CMOS circuits.

In this paper we present an image sensor composed of a 32×32 array of single photon avalanche diodes (SPAD). It is unique in that it uses a delay-insensitive asynchronous protocol for the readout, where the readout circuits occupy just 4% of the chip area. This is a significant reduction when compared to [4], where the readout circuits comprise over 44% of the chip area (accounting for technology scaling). The event-driven readout takes advantage of the asynchronous nature of the SPAD. Furthermore, imagers with event-driven readout have been shown to have high dynamic range [5].

The next section discusses SPAD properties and Section III describes the structure of the SPAD used in our sensor array. Section IV describes the architecture of the asynchronous readout circuits. In Section V we report test results and Section VI concludes the paper.
imagers are the N-well to P-substrate junction or the P+ implant to N-well junction. These structures work well below breakdown, however, they are not suited for Geiger-mode operation since they suffer from premature breakdown around the edges. The consequence of this edge breakdown is an insufficient area for avalanche multiplication. With the introduction of deep N-well modules in standard CMOS processes, it has become possible to create a guard ring around the junction which prevents premature breakdown [7], [8]. Fig. 1 shows this structure that is used in the pixel.

IV. ADDRESS-EVENT READOUT

Address-event representation (AER) is the standard protocol used for communicating spikes between two-dimensional arrays of silicon neurons. When a neuron generates an event, a transmitter encodes this with the address of that neuron. That address is sent over a physical link to another chip where a receiver decodes it then sends an event to a neuron with the same address. In this manner, multiple virtual point-to-point connections can be made between neurons in an array using the same physical link.

For this imager, we co-opt this architecture such that SPADs replace neurons as the event generators. The receiver is used to send events to the SPADs which toggle them on or off. This gives the array the ability to focus on a region-of-interest. Also SPADs with high output rate can be silenced to give a chance for SPADs with a lower rate to use the bus.

We use a modification of the original AER protocol called delay-insensitive (DI) word-serial AER [9], which has superior throughput. In word-serial AER, the addresses are sent in bursts: row address followed by multiple column addresses. Delay-insensitivity is added by using a one-hot code; these codes use $2^b$ lines to encode $b$ bits. The $i$th line is raised to output the $i$th bit of the address [10]. Fig. 2 shows a diagram of the timing protocol. In our design $b = 2$ for a 1-of-4 DI encoding.

A. Architecture

In the transmitter (Fig. 3a), event generators (E) make requests to the row arbiter through an interface (I). The row arbiter selects one row and directs an encoder to output that row’s address, which is buffered (A). Selection by the arbiter also enables the row’s event generators to raise their column lines, which are latched before being relayed to the column arbiter. As was done for the row, this arbiter directs an encoder to output these columns’ addresses to a buffer (A) one by one. After prompting the first buffer to output the row address, the sequencer (SEQ) repeatedly prompts the second one to output the column addresses—until the latch signals that it is empty. At which point the sequencer causes the tailword (TB) to be output and prompts the latch to load the next row, which was read out when the burst was being sent.

In the receiver (Fig. 3b), the sequencer directs buffers (A) to load incoming row and column addresses. They are decoded, stored (L and Latch), and sent to the event-recipients (E) when the tailword shows up. Additional buffering (B) allows a burst to be decoded while the previous one is being written.

The architectures for the transmitter and receiver were generated through process decomposition, the first step of Martin’s three-step synthesis procedure for asynchronous circuits [11]. The next step is HSE, where each communication is expanded into a pair of four-phase request-acknowledge sequences:

```
# Four-phase #
*[ao+;{ai};ao-;{'ai}]
*[{p};p0+;{'p};p0-]
```

The active port (A) asserts the request (ao+) and waits for the acknowledge ({ai}), then deasserts the request (ao-) and waits for the acknowledge to clear ({ai})—a lazy-active port checks immediately before raising the request to start the handshake. The passive port (P) waits for the request ({p}) before asserting the acknowledge (po+), then waits for the request to clear ({"p"}) before deasserting the acknowledge (po-). Note that the first two-phase communication synchronizes A and P; the second is superfluous—it just returns the signals to their initial state. For more on HSE notation, see Table I. Due to space constraints, we refer to [9] for both the transmitter and receiver’s HSE and circuits.

B. SPAD Pixel

Each pixel contains a SPAD (circular structure) with 10 µm diameter photosensitive area. The total layout measures
SIC makes a request to SIC passive, XIC the request when and there is an incoming pulse from the SPAD (passive, and T XIC quenched. The first and second two-phase communications signal that which brings the bias voltage below breakdown. As the V node lowers, the current through the HVPMOS serves to signal (XIC XIC communicates with the transmitter bus can be controlled through RIC and MEM. RIC interfaces with the receiver circuits and MEM toggles between enabling or disabling the SPAD interface. With R passive, C passive, and P active, RIC’s HSE is:

When the row and column (ri and ci) are selected, RIC communicates with MEM and relays its acknowledge to the receiver (po;[pi];ro+). With M passive and S active, MEM’s HSE is:

MEM lowers or raises s to respectively disable or enable SIC. m is used as a state bit.

Once we have HSE sequences, we proceed to the final step, compiling them into production-rule sets (PRS), which are straightforward to implement with CMOS transistors [11]. Due to space constraints, only the synthesized CMOS circuits are shown (Fig. 5).

V. TEST RESULTS

We present test results for a 32×32 array fabricated in 0.18µm CMOS technology (Fig. 6). The chip occupies an area of 3.52mm² of which 0.144mm² and 0.150mm² is used by the transmitter and receiver, respectively.

An important measure of a SPAD detector is its photon detection efficiency (PDE). This is simply the ratio of output pulses to incident photons. Fig. 7 shows the PDE of a representative SPAD for various excess bias voltages and wavelengths. The excess bias voltage (Vp) is defined as the voltage over the breakdown voltage. The breakdown voltage of this SPAD was measured at 10.3 V. The SPAD was illuminated with monochromatic light in 10nm increments from a Newport 74100 monochromator. The pulses were counted with an Opal

<table>
<thead>
<tr>
<th>Operation</th>
<th>Notation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal</td>
<td>v</td>
<td>Voltage on a node</td>
</tr>
<tr>
<td>Complement</td>
<td>~v</td>
<td>Inversion of v</td>
</tr>
<tr>
<td>And</td>
<td>v &amp; w</td>
<td>High if both are high</td>
</tr>
<tr>
<td>Or</td>
<td>v</td>
<td>Low if both are low</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Operation</th>
<th>Notation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential</td>
<td>[u];v+</td>
<td>Drive v high after u is high</td>
</tr>
<tr>
<td>Concurrent</td>
<td>v, w+</td>
<td>Drive v and w high</td>
</tr>
<tr>
<td>Repetition</td>
<td>*(v;v−)</td>
<td>Repeats forever</td>
</tr>
</tbody>
</table>

36.225 × 36.225µm². Fig. 4 shows a block diagram of the pixel. A high-voltage p-type MOSFET (HVPMOS) is used for passive quenching [12]. The anode (Va) of the SPAD is set to a negative voltage such that the difference between VP and Vr is greater than the breakdown voltage. During an avalanche event, the reverse current through the SPAD discharges VP, which brings the bias voltage below breakdown. As the Vp node lowers, the current through the HVPMOS serves to passively recharge the node above breakdown.

The output pulse is read out by SIC. With S active and P passive, SIC’s HSE is:

\[
\text{# SIC} \#
\begin{align*}
+ &[[\text{si} & \text{vi} & \text{s}] ; \text{so} + ; [\text{si} & \text{vi} & \text{s}] ; \text{so} - ]
\end{align*}
\]

SIC makes a request to XIC only when it is enabled (s high) and there is an incoming pulse from the SPAD (vi). It releases the request when XIC acknowledges and the SPAD has been quenched. XIC interfaces to the transmitter circuits. With X passive, and T active, XIC’s HSE is:

\[
\text{# XIC} \#
\begin{align*}
+ &[[\text{x} i] ; \text{yo} + ; [\text{yi}] ; \text{to} + , \text{x} o + ;
\text{["x" i] } \text{yo} - ; [\text{"y" i} ] ; \text{to} - , \text{x} o - ]
\end{align*}
\]

The first and second two-phase communications signal that XIC’s row has been selected (yo;[yi]) and its column signal (to+) has been read (yo;[yi]), respectively. XIC communicates with SIC (xo+;["x" i]) after the first and clears its column signal after the second.

The pixel’s access to the transmitter bus can be controlled

![Fig. 4. Pixel Block Diagram. The receiver and transmitter communicated with the SPAD through RIC and XIC, respectively. MEM is a state bit that enables or disables the SPAD interface SIC.](image1)

![Fig. 5. Pixel Synthesized Circuits. The tilde denotes a signal’s complement. RIC’s pi is wired to ro.](image2)
Kelly XEM3010, which has a Xilinx Spartan-3 FPGA. Fig. 8 shows the dark count rate across the chip for an excess bias voltage of 1.5V. The dark count rate was measured by using the receiver to disable all but one pixel at a time.

VI. DISCUSSION

The address-event architecture provides several advantages for readout of SPAD arrays. The asynchronous on-demand nature of the circuits yields low power consumption at low light levels. The architecture is scalable as it uses very little chip area with regard to the area used for imaging. The receiver allows for the ability to control individual pixel access to the transmitter bus. Not only can this be used for region-of-interest, but also pixels with high event rates can be silenced. This mitigates the limitation of the fixed bandwidth of the address-event link. For this imager, dark count rates of over 1 million events/sec were measured. This can quickly use up the available capacity of the link. Further testing is required to find the correct operating regime through tuning of the pixel voltage biases. Otherwise, future versions of this chip will have to lower the dark count rate. This could be achieved by using a different technology process.

ACKNOWLEDGMENT

The authors would like to thank Dr. Philippe O. Pouliquen for helpful comments and suggestions, and Dr. Miriam A. Marwick for designing the HVPMOS and SPAD device structures.

REFERENCES