Large-Format Geiger-Mode Avalanche Photodiode Arrays and Readout Circuits

Brian F. Aull, Senior Member, IEEE, Erik K. Duerr, Jonathan P. Frechette, K. Alexander McIntosh, Daniel R. Schuette, and Richard D. Younger

(Invited Paper)

Abstract—Over the past 20 years, we have developed arrays of custom-fabricated silicon and InP Geiger-mode avalanche photodiode arrays, CMOS readout circuits to digitally count or time stamp single-photon detection events, and techniques to integrate these two components to make back-illuminated solid-state image sensors for lidar, optical communications, and passive imaging. Starting with 4 × 4 arrays, we have recently demonstrated 256 × 256 arrays, and are working to scale to megapixel-class imagers. In this paper, we review this progress and discuss key technical challenges to scaling to large format.

Index Terms—Image sensors, Geiger-mode avalanche photodiodes, photon counting, lidar.

I. INTRODUCTION

SOLID-STATE image sensors have a multi-billion-dollar global market. The microelectronics revolution has enabled rapid progress in CMOS image sensors, which are now mass produced for cell phones, consumer electronics, and medical, automotive, and industrial applications. Most of these image sensors are based on the integration of photocurrent by a capacitor in the pixel circuit. The photocharge is sensed by an analog amplifier and digitized at the periphery of the array, a process that adds readout noise [1], [2].

If photons are scarce (or, equivalently, if integration time is short), it is desirable for the pixel to be able to digitally count or time stamp individual photon arrivals. This capability eliminates the need for analog sensing circuitry in the detection and readout path, and thus eliminates the readout noise.

Airborne flash lidar systems, for example, can achieve high area coverage rates by using large arrays of detectors, each of which precisely time stamps individual photons [3], [4]. However, specialized photodetectors with fast, single-photon response were until recently available only in the form of discrete devices with single detectors or small-scale arrays.

A Geiger-mode avalanche photodiode (GmAPD) can detect a single photon and produce an electrical pulse of sufficient amplitude to directly trigger a logic element. This “photon-to-digital conversion” makes for simple interfacing of the GmAPD to an all-digital CMOS pixel circuit [5]. There are mature fabrication techniques to make dense large-format arrays of GmAPDs. Arrays and image sensors based on this technology have many scientific applications such as biomedical imaging [6], spectroscopy [7], time-resolved fluorescence [8], and astronomy [9]. It is now even finding its way into commercial applications, such as lidar for self-driving cars [10].

There are, however, some challenges to scaling to large format and high pixel density. First, the avalanche discharge in a GmAPD produces a small amount of hot-carrier light emission that triggers spurious detection events in neighboring pixels [11]. Mitigation of this optical crosstalk is therefore necessary for successful scaling. Second, while the design of an all-digital CMOS pixel circuit may seem to be a straightforward engineering task, scaling to large arrays requires solutions to issues such as data readout bandwidth and power distribution. Third, hybrid integration of custom GmAPD arrays to CMOS readouts is often favored over monolithic approaches, either for performance optimization or for the use of non-silicon detector arrays. This presents non-trivial fabrication issues for large-format back-illuminated devices. The next three sections of this paper respectively discuss the detector physics and technology, CMOS readout architectures, and hybridization techniques. The final section discusses the technical progress and future research directions in the path to high-performance large-format Geiger-mode image sensors.

II. GEIGER-MODE APD TECHNOLOGY

A. Background and Pixel Architecture

An avalanche photodiode (APD) is a pn junction designed to support high electric fields that cause the primary photogenerated carrier to initiate an avalanche, that is, a chain of impact ionization events. The electron-hole pairs generated in this chain mediate internal gain. At biases below the breakdown voltage the chain is self-terminating; the average value and the variance of the gain are determined by the randomness of the avalanche process. This way of using the APD is referred to as linear-mode operation, because one obtains a continuous-time photocurrent...
proportional to the intensity of the optical signal. Gains in the range $10^{-3}$ are typically used.

It is also possible to operate the APD above the breakdown voltage and obtain discrete electrical pulses in response to individual photons. This is known as Geiger-mode operation. When an avalanche is first initiated, the average rate at which carriers are generated exceeds the rate at which they are collected at the device terminals. Barring an early random fluctuation that terminates the avalanche, this imbalance causes the current to grow until the internal fields are reduced by series resistance. At this point, the current reaches a self-sustaining steady state. The total number of electron-hole pairs generated can in principle be infinite. To be useful, however, the APD is connected to a circuit that detects the current flow and reduces the bias to below breakdown long enough to terminate the avalanche and shut the device off. This bias reduction is commonly referred to as quenching and the subsequent bias restoration as either arming or reset.

A number of circuit architectures have been demonstrated for arming, quenching, and event sensing [12], [13]. Geiger-mode operation, however, lends itself to a simple architectural concept: direct connection of the APD to a logic circuit. Fig. 1 illustrates the concept. The n side of the APD is initially armed to a voltage level $V_{DD}$ corresponding to a logic high. The p side is tied to a negative voltage supply whose amplitude $V_A$ is slightly less than the avalanche breakdown voltage. After arming, the reverse bias on the APD is $V_{DD} + V_A$, or roughly one logic swing above breakdown. When a detection occurs, the avalanche current discharges the parasitic capacitance at the n side. When the APD reverse bias falls below breakdown, the avalanche is no longer self-sustaining and the APD shuts off. At this point, the n side is at a voltage level close enough to zero to be sensed as a logic zero, and the APD is said to be disarmed.

This scheme enables scaling to large-format small-pixel all-digital image sensors for several reasons. First, the interface between the APD and the pixel circuit is simple and has a small transistor count. Second, the APD itself functions as a CMOS-compatible logic element. While a voltage $V_A$ of several tens of volts is required for APD biasing, the voltage swing on the CMOS side needed for efficient photon detection is in the range of 3–5 V. Thus the APD is CMOS-logic-compatible in many older CMOS processes. In advanced process (180 nm and below) that have smaller voltage swings, circuit techniques are used to augment the voltage swing, such as cascading the transistors used for arming and disarming. Like a CMOS logic element, the APD draws negligible static bias current.

B. GmAPD Structure and Operation

A structure typically used in our GmAPDs is the separate-absorber-multiplier (SAM) structure [14]. Fig. 2 shows an idealized one-dimensional doping profile of our silicon GmAPDs and electric field distributions in the armed and disarmed states. In this device, a p-type separator layer partitions the diode into separate regions for photon absorption and avalanche multiplication. Specific layer thicknesses, doping polarities, and doping levels depend on the APD material system (Si or InGaAsP) and optimization criteria.

A photon enters the absorber layer and is absorbed, generating an electron-hole pair. In the silicon SAM structure shown in Fig. 2, the hole is collected at the p+ contact. When the APD is in the armed state, a modest electric field exists in the absorber that sweeps the electron into the multiplier region, where the electric field is high enough to induce an avalanche. After the discharge, the field in the multiplier is reduced to below the breakdown value. The doping polarities favor electron-initiated avalanches over hole-initiated avalanches. This choice maximizes avalanche probability in silicon, because electrons have a higher ionization coefficient than holes.

C. Figures of Merit

An important figure of merit for a single-photon-sensitive image sensor is photon detection efficiency (PDE). This is defined as the probability that a single incident photon is detected by the pixel circuit. PDE is the product of the external quantum
efficiency, which is the probability of primary photocarrier generation, and avalanche probability, which is the probability that the avalanche is initiated and is sustained long enough to trigger a logic event. The avalanche probability is a monotonically increasing function of the overbias, which is the amount by which the APD reverse bias exceeds the breakdown. In a silicon SAM with a 0.7-μm multiplier thickness, for example, the voltage drop across the multiplier must exceed the breakdown by about 3 V in order to get an avalanche probability of 0.8.

The dark count rate of a GmAPD is the frequency of spurious detection events caused by thermally generated carriers, tunneling currents, and other mechanisms. In a passive imaging application, the average dark count rate can be subtracted from the total event rate, but its variance is a source of shot noise that can limit sensitivity.

When the pixel is designed to time stamp the photon arrival, the timing jitter of the APD is an important performance metric. This is the variation of the time delay between the arrival of the photon and the triggering of the detection event. Sources of jitter include variation of the time required for the primary photocarrier to drift to the multiplier and for the avalanche current to build up.

In a detectable Geiger-mode avalanche discharge, the number of electron-hole pairs produced depends on the voltage swing and the capacitance, but can easily be in the range $10^6$–$10^8$. Some of these carriers can be trapped in the multiplier layer, and then released later. If the quenching time is shorter than the detrapping time constant, these carriers can trigger spurious events, known as afterpulses. Prevention of afterpulsing therefore determines the minimum quenching time. The GmAPD is unresponsive when disarmed, and in applications requiring continuous detection of a stream of photons, this signal blockage sets the photon flux level above which count rate saturation occurs.

The Geiger-mode avalanche process creates some near-infrared hot-carrier light emission, which can trigger spurious events in nearby pixels. As discussed in section V, this optical crosstalk is still a major issue in image sensors with pixel pitches below 50 μm. In devices requiring sub-μs quenching times, the photocarriers created by this light emission can even retrigger the emitting APD. Such optical “selftalk” is another potential source of afterpulsing.

### D. Silicon and III-V GmAPD Arrays: A Comparison

Lincoln Laboratory has developed custom GmAPD arrays hybridized to digital CMOS readout circuits using bump bonding or 3D integration techniques. Silicon is the material of choice for ultraviolet, visible, and near-infrared wavelengths up to approximately 900 nm. InGaAsP/InP APDs have been developed for short-wave infrared detection, with particular emphasis on 1064-nm and 1550-nm wavelengths. (See Table I.) The fabrication techniques and device structure details are very different in these two materials systems. Some of the design tradeoffs also differ [15].

Fig. 3 shows a two-dimensional cross sectional view of the silicon SAM APD used in lidar applications. The APD is a planar structure fabricated in a silicon layer, typically 15 μm thick, homoepitaxially grown on a silicon substrate. During processing of the front side of the wafer, the n+ region and the p separator are formed by photolithographically patterned ion implantation steps, and Al contact pads made. Ultimately, the silicon substrate is removed and a thin p+ contact layer, common to all the APDs, is formed on the back side, either by ion implantation and rastered laser annealing or by molecular beam epitaxy (MBE).

Fig. 4 shows a cross section of a SAM APD designed for operation at 1064-nm wavelength. The layers are grown by organometallic vapor phase epitaxy (OMVPE) or MBE and then the individual diodes defined and isolated by etching mesas and passivating the sidewalls with polyimide.
There a number of key differences between the silicon and InP/InGaAsP APD structures. First, the InP doping polarity favors hole-initiated avalanches because holes have higher ionization coefficient than electrons.

Second, alloy composition, and therefore bandgap, can be varied in the compound semiconductor system. This allows for the InGaAsP absorber bandgap to be tuned for the wavelength of operation. Most of the rest of the structure is InP, which has a higher bandgap and is therefore relatively transparent at the wavelength of operation. The device can be illuminated through a substrate that has been partially thinned to allow efficient optical coupling with microlens arrays. In contrast, the substrate must be completely removed in the case of the homoepitaxial silicon structure.

Third, OMVPE or MBE allows for precise control over layer thicknesses with abrupt interfaces between regions of different doping or alloy composition. The InP separator layer, for example, is typically \( \sim 70 \) nm in thickness and this thickness can be finely tuned to optimize the design for a given temperature and bias. The doping profile of the separator in the silicon device, on the other hand, is determined mostly by the dose and energy of an ion implant. Implant range straggles produces effective separator thicknesses in the 300–400 nm range. The depth and thickness of the separator are not independently controllable. In operation, additional voltage drop is required to support the thick transition from the high field needed in the multiplier to the lower field needed in the absorber.

Fourth, the InP materials have a direct bandgap whereas silicon has an indirect gap. Consequently, InGaAsP has a much higher absorption coefficient and the absorber can be much thinner than in silicon. A typical InGaAsP absorber thickness is 1.5 \( \mu \)m. The silicon absorber thickness depends on the intended wavelengths of operation, but is typically in the range 5–10 \( \mu \)m.

III. CMOS READOUT CIRCUITS

Image sensors based on the photon-to-digital conversion concept can be customized for various applications, including lidar, laser communications, wavefront sensing, and passive imaging. The digital CMOS readout circuit architecture is usually tailored to one of these applications.

A. Lidar

A flash lidar system acquires three-dimensional images by illuminating a scene with a short laser pulse and imaging it onto a focal plane in which each pixel precisely time stamps the return pulse. The GmAPD array is an enabling technology because it can time stamp individual photons, leading to efficient use of transmitter power and rapid acquisition of imagery [16]. The pixel circuit implements a stopwatch function; a fast digital counter is used to measure the time interval between the transmission of the laser pulse and the detection event. The timing data is read out in the relatively long time interval (10’s of \( \mu \)s) between laser pulses, so the GmAPDs are operated with low duty cycle.

We first demonstrated a proof of this concept in 1997 by wirebonding a front-illuminated 4 \( \times \) 4 silicon GmAPD array to a CMOS chip fabricated in a 500-nm CMOS technology. The chip comprised 16 pseudorandom counters driven by a common 350 MHz clock. A detection event in an APD generates a stop signal that freezes the corresponding counter and measures two vernier bits by capturing the states of the clock and a delayed copy of the clock. The same architecture, illustrated in Fig. 5, was used in a series of 32 \( \times \) 32 arrays (designated by year as CMOS200x) with 100-\( \mu \)m pitch fabricated in a 350-nm technology. These CMOS chips were hybridized to APD arrays to make back-illuminated focal planes for the Jigsaw system, the first of a series of airborne lidar systems that could perform foliage penetration by fusing multiple 3D images taken from different look angles [17].

Over the past decade we have developed larger-format lidar readout circuits while also scaling down the pixel pitch and reducing the per-pixel power dissipation. The most recent framed chip has a format of 256 \( \times \) 128 with 50-\( \mu \)m pixel pitch, fabricated in a 180-nm CMOS process. A number of architecture improvements enable greater power efficiency. The pixels in early CMOS designs effectively contained a “stop watch” circuit operating at fast clock frequencies required to generate precision photon timing information. Each pixel generated a timestamp per frame, which represents either the time an avalanche was detected, or an “end of gate value” latched in when the exposure window of the time gated sensor was disabled. In most lidar applications, the individual frames have spatially sparse detection events and the image is built up by aggregation of many frames. The latest CMOS designs exploit this with architectures that leverage the sparse detections to make the power and data volume proportional to the activity on the array. Timing circuitry with extensive clock gating is utilized, and the designs also use data thinning in the readout stream; only pixels that have had a detection event report out a timing value. Previous designs were limited to 20 kframes/s frame rates, while the frame rate of the new design can be increased as the frames become sparser, approaching 150 kframes/s for the sparsest returns, while consuming only one third the power. (See Table II.) Power dissipation of the latest framed readout chips depends on frame rate and activity level, but 0.5 W is a typical value for the R7 chip operating at 20 kframes/s with 1000 events per frame.

B. Laser Communications

A focal plane array based on GmAPDs can be used as a receiver in a free-space optical communications link. By using
a pulse-position-modulation (PPM) format, the time stamping of a single photon can convey multiple bits of information. The use of an array enables a receiver system to acquire and track a moving transmitter platform with less reliance on gimbal or other moving parts. The required quenching time of a GmAPD, however, is too long to support Gbit/s data rates, as there would be severe signal blockage while the detector is unresponsive. This problem is solved by spreading the optical signal over a collection of multiple pixels, often referred to as a macropixel. If one signal photon triggers a pixel, the next photon most likely triggers a different pixel. The effective dead time is therefore shortened by a factor equal to the number of pixels in the macropixel. This obviously trades off the spatial resolution available for acquisition and tracking. Another design tradeoff is reset time and false detection probability, because the aggregate dark count rate increases with the size of the macropixel without a concomitant increase in signal.

As in the lidar case, the CMOS readout circuit for this application time stamps arriving photons. However, the optical communication signal photons arrive at time intervals that are typically three orders of magnitude shorter than the time between lidar pulses. Each detector requires a quench time that is longer than the duration of a PPM frame. In addition to needing a macropixel to avoid signal blockage during quench times, the readout of timing values must not introduce blocking losses. To address this need an $8 \times 8$ array of photon timing circuits with an event-driven readout was demonstrated in 2004. During each PPM frame, the pixel time stamps an arriving photon using the same approach as the lidar circuit. When each row of pixels is queried for readout, pixels with events to report are read out through a serial readout path that bypasses pixels that have no events to report. The pixels with no events to report remain active as the next PPM frame begins, while the pixels that report events enter a wait mode for a prescribed hold-off time in order to quench. The chip was fabricated in a 350-nm CMOS technology and bump bonded to an $8 \times 8$ array of InP-based GmAPDs. A benchtop experiment for the NASA Mars Laser Communications Demonstration system achieved a data rate of 2 bits per photon in the presence of a very high background light flux [18].

As in the case of the lidar application, it is desired to scale to larger formats while also keeping power dissipation low. This requirement motivated the development of a series of $32 \times 32$ arrays that achieve power efficiency by utilizing an event-driven readout architecture with region-of-interest processing [19]. A block diagram of the architecture is shown in Fig. 6. Output formatting circuitry on each row collects timestamp data and formats it in to two data streams. The first data stream is a fire map showing the location of all of the events on the array, and the second is time stamp data packets with embedded position information. The ROIC also features a pixel disable option to prevent individual pixels from arming and a pixel mask option to designate pixels that should only output fire map data and not timing information. These features can be used to reduce power and output data when the full array is not required. Over several revisions of the chip, the design migrated from 350-nm to 180-nm CMOS technology and implemented architecture improvements to substantially lower power dissipation. (See Table III.)

### C. Wavefront Sensing and Passive Imaging

Another major class of applications requires counting rather than time stamping of photon arrivals. These applications do not require fine timing resolution, but rather high PDE and low dark count rate.

The first application pursued was low-latency wavefront sensing for adaptive optics. Shack-Hartmann wavefront sensors [20] were developed using silicon GmAPDs laid out as an array of quad cells ($2 \times 2$ subarrays) [21]. Each pixel in the CMOS readout circuit has a 10-bit counter that is incremented when a signal over a collection of multiple pixels, often referred to as a macropixel. If one signal photon triggers a pixel, the next photon most likely triggers a different pixel. The effective dead time is therefore shortened by a factor equal to the number of pixels in the macropixel. This obviously trades off the spatial resolution available for acquisition and tracking. Another design tradeoff is reset time and false detection probability, because the aggregate dark count rate increases with the size of the macropixel without a concomitant increase in signal.

As in the lidar case, the CMOS readout circuit for this application time stamps arriving photons. However, the optical communication signal photons arrive at time intervals that are typically three orders of magnitude shorter than the time between lidar pulses. Each detector requires a quench time that is longer than the duration of a PPM frame. In addition to needing a macropixel to avoid signal blockage during quench times, the readout of timing values must not introduce blocking losses. To address this need an $8 \times 8$ array of photon timing circuits with an event-driven readout was demonstrated in 2004. During each PPM frame, the pixel time stamps an arriving photon using the same approach as the lidar circuit. When each row of pixels is queried for readout, pixels with events to report are read out through a serial readout path that bypasses pixels that have no events to report. The pixels with no events to report remain active as the next PPM frame begins, while the pixels that report events enter a wait mode for a prescribed hold-off time in order to quench. The chip was fabricated in a 350-nm CMOS technology and bump bonded to an $8 \times 8$ array of InP-based GmAPDs. A benchtop experiment for the NASA Mars Laser Communications Demonstration system achieved a data rate of 2 bits per photon in the presence of a very high background light flux [18].

As in the case of the lidar application, it is desired to scale to larger formats while also keeping power dissipation low. This requirement motivated the development of a series of $32 \times 32$ arrays that achieve power efficiency by utilizing an event-driven readout architecture with region-of-interest processing [19]. A block diagram of the architecture is shown in Fig. 6. Output formatting circuitry on each row collects timestamp data and formats it in to two data streams. The first data stream is a fire map showing the location of all of the events on the array, and the second is time stamp data packets with embedded position information. The ROIC also features a pixel disable option to prevent individual pixels from arming and a pixel mask option to designate pixels that should only output fire map data and not timing information. These features can be used to reduce power and output data when the full array is not required. Over several revisions of the chip, the design migrated from 350-nm to 180-nm CMOS technology and implemented architecture improvements to substantially lower power dissipation. (See Table III.)

### TABLE II

<table>
<thead>
<tr>
<th>Designation</th>
<th>Format</th>
<th>Timing Pixel Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMOS2002</td>
<td>$32 \times 32$</td>
<td>0.5 100 350</td>
</tr>
<tr>
<td>CMOS2005</td>
<td>$32 \times 32$</td>
<td>0.3 100 350</td>
</tr>
<tr>
<td>R1 through R3</td>
<td>$128 \times 32$</td>
<td>0.5 50 180</td>
</tr>
<tr>
<td>R1 through R7</td>
<td>$256 \times 64$</td>
<td>0.5 50 180</td>
</tr>
<tr>
<td>R1 through R3</td>
<td>$256 \times 128$</td>
<td>0.5 50 180</td>
</tr>
</tbody>
</table>

### TABLE III

<table>
<thead>
<tr>
<th>Designation</th>
<th>Format</th>
<th>Timing Pixel Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLCD_R1</td>
<td>$8 \times 8$</td>
<td>3 100 350</td>
</tr>
<tr>
<td>MLCD_R2</td>
<td>$8 \times 8$</td>
<td>1.5 100 350</td>
</tr>
<tr>
<td>ASYNC_R1</td>
<td>$32 \times 32$</td>
<td>1 100 350</td>
</tr>
<tr>
<td>ASYNC_R2-R5</td>
<td>$32 \times 32$</td>
<td>1 50 180</td>
</tr>
<tr>
<td>ASYNC_R6</td>
<td>$32 \times 32$</td>
<td>1 50 180</td>
</tr>
</tbody>
</table>

Fig. 6. Block diagram of the $32 \times 32$ asynchronous ROIC architecture.
the PCROIC, was fabricated in a 180-nm CMOS technology. The readout circuitry operates independently of the circuitry that queries and rearms the APDs, so that image data can be read out without blinding the array. Each pixel has a detection event flip-flop that is set when a detection event occurs. In one mode of readout, these detection bits are read out directly as binary frames. In addition, the pixel has a 7-bit counter that counts events and supports a readout bandwidth compression scheme. There is an overflow bit that is set when the 7-bit counter overflows. During an extended integration time, the rows are successively addressed and the overflow bits read out and reset. At the end of the integration time, the residual contents of the 7-bit counters can be either read out or discarded if their contribution to the integrated signal is less than the shot noise.

The PCROIC overflow readout scheme effectively extends the 7-bit counter, adding more significant bits that reside in an external frame storage device. The next step is to put a frame store on the imager chip and add processing circuitry that allows for preprocessing and readout bandwidth reduction. This was implemented on the Digital Vision Sensor (DVS), fabricated in a 180-nm CMOS process.

Fig. 7 shows a block diagram of the DVS architecture. Raw image data from a 256 × 256-pixel array is transferred to an SRAM memory through a row-wise bank of programmable processing elements. These processors can be used for temporal or spatial filtering or more advanced functions such as optical flow extraction. Raw image data can be generated at 1 kframe/s, for example, while the processed images can be read out at 30 frames/s.

D. Advanced Readout Architectures

Currently we are developing focal planes with 256 × 256-pixel format using a 65-nm CMOS process. The migration to the 65-nm node allows for much higher transistor density and therefore more sophisticated functionality in the pixel. One chip under development is capable of performing passive imaging over the whole array while time stamping photons for lidar or communication over a dynamically programmable region of interest. Others scale the lidar and async architectures to the 256 × 256 format while implementing further improvements to decrease power dissipation. Additionally we are exploring 3D-integrated circuit architectures that split the APD front-end interface circuit and digital pixel logic onto two separate circuit tiers. This enables using a deeply scaled submicron technology for the digital logic circuit, since the higher voltage biases would be contained on the front-end tier.

IV. INTEGRATION OF APD ARRAYS TO CMOS READOUTS

Many groups have developed monolithic CMOS photon-sensitive imagers based on incorporation of a silicon GmAPD into the pixel circuit [22]. The GmAPD is formed using doped structures already present in the standard CMOS foundry process, such as n-wells and source-drain implants. These APDs are most commonly termed single-photon avalanche diodes (SPADs). This approach offers rapid, low-cost prototyping and avoids the electrical parasitics associated with bump bonds. The SPAD is a thin, shallow-junction device. This is advantageous because the photocarrier is generated close to the avalanche layer, giving very low timing jitter.

Monolithic CMOS SPAD arrays have a number of disadvantages. The thin detector structure has poor detection efficiency at red and near-infrared wavelengths. The area of the pixel circuit is partitioned between the SPAD and the readout circuit, and this limits detector fill factor. Microlenses can partly compensate for this, but their use places constraints on the f-number of the imaging optics. The use of a standard foundry process does not allow for customization of the APD doping and layer thicknesses. Such customization was needed for both the quad-cell arrays and the passive imagers described above. The APDs for these devices were fabricated with a custom stepped separator implant that creates band bending that guides the photoelectrons to the multiplier, giving high-fill-factor response [23]. Finally, monolithic SPADs work only for silicon APDs, and our goal is to develop a common focal plane technology for a variety of detector materials systems and wavelength ranges. Hence, we are pursuing hybrid integration.

Bump bonding has been the most common approach to integrating detector arrays to CMOS readout circuits and many of our image sensors have been bump bonded. It is desired, however, to scale to large formats and small pixel pitches (<20 μm), and to reduce the parasitic capacitance of the APD. This has motivated the pursuit of advanced 3D integration techniques.

For silicon detector arrays, there is yet another motivation for 3D integration. Back illumination presents a greater challenge for bump bonded silicon detectors than for InP. This is illustrated in Fig. 8. The opaque silicon substrate must be entirely removed, entailing fabrication and handling difficulties with a thin (<10 μm) detector layer.

Our early silicon lidar focal planes were hybridized by a die-to-wafer process termed bridge bonding [5]. Individual CMOS chips are epoxied face to face with detector arrays on an APD wafer. The CMOS chip and epoxy layer then serves as a mechanical support for APD substrate removal. The electrical connections are made at the end of the process by etching vias through the APD and epoxy layers in between pixels, and patterning metal “bridges” that connect each APD to its pixel circuit. The
process worked for small-format low-fill-factor detector arrays, but is not scalable.

Another approach, used for our first 256 x 256 passive image sensor, is to transfer the silicon detector layer to a fused silica substrate. This involves bonding the front side of the APD wafer to a temporary support wafer, removing the substrate, passivating the back surface, bonding the back surface to the fused silica wafer, and then removing the temporary handle from the front side. This process produces a detector wafer with a transparent substrate; the arrays can then be bump bonded to CMOS readouts by the same process used for InP-based detectors [24].

The next advance is to use 3D integration techniques that replace the indium bump with a much smaller via connection. It is also desirable to avoid the use of adhesives and materials that create mismatches in thermal expansion coefficients. Lincoln Laboratory developed a 3D integration process based on its in-house fully depleted silicon-on-insulator (SOI) CMOS process. The SOI wafer is bonded to the detector wafer using oxide-to-oxide wafer bonding. The SOI handle wafer is then removed and electrical connections made by etching vias and filling them with metal in a process similar to that used to connect successive layers of metal in a CMOS process. Once the bonded wafer pair completes these steps, another SOI wafer can be bonded to it and the process repeated to add another “tier” of CMOS circuitry. In 2004 this process was used to fabricate a 64 x 64 focal plane for lidar with one GmAPD tier and two CMOS tiers. This was to our knowledge the first three-tier 3D integrated imager with per-pixel 3D via connections [25].

The Lincoln 3D integration process is categorized as a “via last” process because the connection vias between the tiers are formed after the wafers are bonded together. Ziptronix pioneered a “via first” process, known as Direct Bond InterconnectTM (DBI) [26], illustrated in Fig. 9. In the DBI process, nickel or copper posts are formed on each wafer before the wafers are bonded. The posts are capped with a bonding oxide and polished to planarity. Once the wafers are bonded, an annealing step strengthens the oxide-oxide bond and simultaneously causes the posts on each wafer to expand and fuse with the corresponding posts on the other wafer. Then the CMOS wafer serves as a permanent mechanical support both during back illumination processing and subsequent device operation. While this process is not readily applicable to multi-tier imagers, it has a number of compelling advantages. First, it is much simpler than the detector transfer or via last processes described above. No temporary support wafers are needed. Mechanical bonding and electrical interconnection are achieved in the same step. Second, it uses the same kinds of high-temperature-tolerant materials that are used in an integrated circuit fabrication process flow. This allows for much more flexibility in the final stages of processing, allowing, for example, an end-of-process sinter to reduce detector dark current. Third, since the detector wafer does not need a transparent substrate for mechanical support, thermal expansion mismatch is avoided. Fourth, if microlenses are used, they can be put directly on the back surface of the detector layer and they can be fast. Fifth, DBI posts can be scaled to µm sizes, facilitating dense pixel-level interconnects and low parasitic capacitance.

V. THE PATH TO MEGAPIXEL GEIGER-MODE IMAGE SENSORS

Much progress has been made over the past decade in understanding the characteristics of large dense arrays of GmAPDs, demonstrating a variety of CMOS readout architectures, and developing 3D integration techniques. We now discuss achievements and challenges in these three areas and relate them to scaling to larger format.

A. APD Arrays

Direct detection lidar has been the “low hanging fruit” for the GmAPD technology, largely because it is not demanding on many of the detector performance parameters. The dark count rate need be only low enough to give a small false alarm probability during a sub-µs gate time. Time stamps are collected over multiple frames so that the PDE need not be very high. The duty cycle imposed by the lidar range gate gives ample quench time, so afterpulsing is not an issue. In many systems the use of microlenses is a viable option. Because the detectors operate at a given transmitter wavelength, the III-V detector structure can have built-in spectral filters to reduce crosstalk.

We have successfully fielded a number of lidar systems based on GmAPD arrays. The latest airborne system, MACHETE, uses dual 256 x 64 InGaAsP/InP arrays operating at 1064-nm wavelength. The GmAPD has a PDE of 0.45 with microlenses and a timing jitter of 400 ps. The ability to digitally time stamp
single photons over a large field of regard enables outstanding area coverage rates. The MACHETE system has a collection rate of over 1300 km²/hr, almost two orders of magnitude higher than the best commercial systems.

Broadband passive imaging presents greater challenges. PDE and DCR directly determine signal-to-noise ratios under low-light conditions. A high-fill-factor design is needed to avoid the drawbacks of microlenses. Rapid reset is needed to avoid signal blockage and the detector needs to operate with high duty cycle. It is desirable to scale to large format and small pixels.

For silicon GmAPD arrays, optical crosstalk reduction will be essential for scaling. At low levels, optical crosstalk creates image blur. It has been found in the bump bonded 256 × 256 arrays that at overbiases sufficient to get PDE above the 10–20% range, crosstalk is severe enough to trigger chains of spurious events that dominate the dark count rate [11]. Crosstalk reduction is being pursued by several methods, including structures to block or absorb the emitted light and scaling down parasitic capacitance to reduce the amount of light emitted.

For InP GmAPDs, crosstalk is also an issue if broadband response is desired and when scaling to smaller pixel pitches. Because of the presence of direct interband optical transitions, these APDs emit light more efficiently than silicon APDs. Measurements of the emitted spectrum show a peak near the InP bandgap energy accompanied by a broad thermal emission component extending to longer wavelengths. The transparency of the InP below the band edge also allows lower-loss coupling to neighboring pixels through the substrate as compared to Si APD arrays. Analysis of the crosstalk in InP-based APD arrays has shown that the primary coupling modes are through the substrate [27]. Recent development efforts have focused on eliminating these substrate-coupled crosstalk contributions and have demonstrated significant additional reductions in crosstalk compared to the previous designs [28], as shown in Fig. 10. These advanced crosstalk-suppression techniques should enable the further reduction in pixel pitch necessary for scaling to megapixel-class imager arrays.

The InP APDs present other issues for scaling. At present they are mesa devices that rely on microlenses for light concentration. No high-fill-factor design currently exists. Other manufacturability issues will be the focus of ongoing research. Pixel yield is an issue because a failed pixel can expose the CMOS to destructive voltages. To reduce this risk, APD arrays fabricated in the past would be prescreened prior to bump bonding and leaky detectors would have the bump bond removed. More recently, the development of in-pixel fuses [29] has eliminated the need for this time-intensive process. Uniformity of the breakdown voltage can also be a limiting issue for scaling to very large arrays. Most work in the past decade has focused on fabrication of arrays on two-inch diameter InP wafers, where the variation in breakdown voltage would typically be less than 2 V across the wafer. This would lead, for example, to peak-to-peak variation of breakdown voltage of about 0.5 V for 256 × 64 arrays with 50-μm pixel pitch. Fabrication on larger-diameter wafers has demonstrated increased uniformity, with a similar 2 V total variation found across a six-inch diameter APD wafer. This enhanced breakdown voltage uniformity will become more important as the dimensions of future APD arrays increase. Finally, the migration from bump bonding to a more advanced 3D integration requires heterogeneous integration technique. We have demonstrated wafer bonding of InP detector arrays to a silicon readout wafer.

B. CMOS Readout Circuits

Scaling to megapixel format will drive further improvements and innovations in CMOS readout architectures and packaging. While the semiconductor industry has demonstrated yielding large-format die with billions of devices, digital imaging ROICs present several unique challenges.

Limitations on both the size of the system optics and the CMOS foundry reticle have constrained pixel sizes, thereby limiting the amount of circuitry and wiring that can be placed behind each detector element. When combined with reduced pixel sizes that limit real estate for power distribution in the ROIC, the increased power requirement makes voltage drop due to resistive loss (IR drop) a concern for large-format arrays. Average power is a concern on the core voltage rail that supplies power to the digital circuits behind each pixel. Additionally, any current spikes that occur due to simultaneous switching of transistors can lead to dynamic IR drop that can cause issues ranging from missed clock edges to data loss and corruption. Dynamic IR drop also becomes a concern for the higher-voltage rails that are used to provide an arming pulse to the GmAPD array to bring it into Geiger mode.

Packaging and power delivery are also a concern. While large-format CMOS chips such as processors typically use area array electrical connections (such as C4) spread throughout the chip, our digital ROICs still rely on peripheral wire bonds for...
electrical connectivity, due to the need to flip-chip the GmAPD array onto the device.

Scaling to megapixel format arrays will require advances in packaging and low-power circuit architectures. For packaging, leveraging advances in 2.5 D and 3 D integration and Si interposer technology will be key in developing new power delivery solutions that move away from wire bonds for electrical connectivity. For low power circuits, as mentioned above, we have been pursuing architectures that are event driven, have region-of-interest processing, and even multi-modes of operation. These techniques reduce the amount of power consumed across the array and break the paradigm where total power consumption grows linearly with array size. The other benefit is that these techniques also reduce the amount of output bandwidth required for the ROIC.

Total output bandwidth is the other major challenge for moving to large-format ROICs. High-speed serializer and deserializer cores utilized for commercial chips high power, expensive, and complex to implement. In addition to limiting the amount of data generated on the array through region-of-interest processing, adding more computation on to the array is another key technical thrust, to preprocess data prior to outputting off chip and limit the total bandwidth required.

C. 3D Integration

The evolution of advanced hybridization techniques has lowered one of the major obstacles to scaling silicon GmAPD arrays to megapixel format and dense pixel pitch. The DBI process in particular has been recently used to fabricate arrays on the PCROIIC readout. The process is producing arrays with outstanding pixel yield and image cosmetics, even in images with no correction, as shown in Fig. 11.

Because this was our first use of the DBI process, it was carried out with conservative design rules; the posts have a 6-μm diameter and they connect 12-μm landing pads. Therefore, we have not yet achieved the capacitance reduction needed to reduce crosstalk. The imagers also showed elevated dark count rates believed to be due to gold contamination at the foundry doing the DBI processing. These imagers are undergoing extensive characterization aimed at refining the design and scaling down the 3D via connections. Future DBI integration will be done in a contamination-free facility.

VI. Conclusion

Starting from a wire-bonded 4 × 4 GmAPD array to prove the “photon-to-digital conversion” concept, two decades of progress have brought us to the threshold of fabricating back-illuminated high-fill-factor arrays with Mpixel format. Further discussion of ongoing research on the silicon GmAPD array technology can be found in a previous review article [30]. Our long-term vision is to develop adaptive multi-function imaging systems. Rather than merely acquiring raw image data, such systems will extract useful information from the incident light field. The imager could have multiple tiers of processing that perform functions such as temporal change detection, spatial filtering, or histogram accumulation on a pre-readout basis. While this review focused on custom hybridized image sensors, many of these capabilities will be supported by monolithic CMOS SPAD arrays. Currently the latter is much cheaper to develop and the user must do careful cost/benefit analysis in choosing between the two approaches.

Acknowledgment

Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force.

REFERENCES

Brian F. Aull (M’75–SM’06) received the B.S. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, and the Master’s and Ph.D. degrees in electrical engineering from Massachusetts Institute of Technology, Cambridge, MA, USA. He is currently a Technical Staff Member in the Advanced Imaging Technology Group, MIT Lincoln Laboratory, Lexington, MA, USA. He led the development of Geiger-mode avalanche photodiode (GmAPD) arrays hybridized to all-digital CMOS circuitry. These imagers were used by the Laboratory to demonstrate lidar imaging systems based on measurement of photon flight time. He led the successful demonstration of a photon time stamping imager based on the Lincoln Laboratory’s three dimensional (3-D) integrated circuit fabrication process; this was the world’s first three-tier 3-D integrated circuit. Later, he extended the GmAPD technology to adaptive optics applications, where he led the development of Shack–Hartmann wavefront sensing arrays based on GmAPD quad cells. More recently, he has developed large-format GmAPD arrays for passive imaging.

Erik K. Duer received the B.S. degree from the University of Virginia, Charlottesville, VA, USA, in 1995, and the S.M. and Ph.D. degrees from Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 1998 and 2002, respectively, in electrical engineering.

In 2003, he joined MIT Lincoln Laboratory, Lexington, MA, USA, where he has been developing avalanche photodiodes for single-photon detection using a variety of III–V material systems. He is currently the Assistant Leader of the Advanced Imager Technology Group.

Jonathan P. Frechette received the B.S. and M.S. degrees in electrical engineering from Northeastern University, Boston, MA, USA, in 2005 and 2009, respectively. He is currently a Technical Staff Member in the Advanced Imager Technology Group, MIT Lincoln Laboratory, Lexington, MA, USA. For the past 12 years his primary focus has been the development of readout integrated circuits that hybridize to Geiger-mode avalanche photodiodes to enable imagers that are single photon with sub-nS timing resolution. Over this time he has led the design of ROICs for applications including lidar and optical communications. His most recent research has focused on power efficient timing architectures, event driven readouts, region of interest processing, and multimode imaging.

K. Alexander McIntosh received the B.S. degree in physics from the University of California, Davis, CA, USA, in 1986.

He is a Technical Staff Member of the Advanced Imager Technology Group, Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA, USA, where he has investigated high-speed infrared detectors, terahertz photonixers, and photonic crystal structures for infrared and microwave applications. His current research interests include the development of arrays of Geiger-mode avalanche photodiodes.

Daniel R. Schuette received the B.A. degree in physics and the B.Sc. degree in mathematics from the University of Chicago, Chicago, IL, USA, and the M.Sc. and Ph.D. degrees in physics with a minor in electrical engineering and computer science from Cornell University, Ithaca, NY, USA.

He is currently a Technical Staff Member of the Advanced Imaging Technology Group, MIT Lincoln Laboratory, Lexington, MA, USA. His primary technical interests are in optical systems, electrical device and circuit design, and embedded computer systems. His work currently covers a portfolio of image sensor programs that span from extension of established technologies (e.g., charge-coupled devices) to the development of emerging technologies (e.g., photon-to-digital Geiger-mode avalanche photodiode image sensors). This work is held together by two common themes—extension of scientific image sensors into new sensitivity and performance regimes, and greater integration of control, data processing, and analysis into scientific quality sensors. His career has focused on developing novel imaging detector technologies; from early career work on cameras for ground-based very-high-energy gamma-ray astronomy, continuing during his Ph.D., into development of next-generation camera systems for the study of biophysical systems at synchrotron X-ray facilities, to his present work at MIT Lincoln Laboratory. Over the course, he has been a member of research groups in the Department of Physics, University of Chicago, the Department of Astronomy, University of California, Los Angeles, CA, USA, and the Department of Physics, Cornell University, Ithaca, NY, USA.

Richard D. Younger received the B.S. degree in engineering physics from the University of Colorado, Boulder, CO, USA, in 2000, and the M.A. degree in physics from Boston University, Boston, MA, USA, in 2006. In 2000, he joined the Lincoln Laboratory, where he helped to develop RF photonic sampling techniques and laser coherent combination methods. He is currently in the Advanced Imaging Technology Group developing large arrays of avalanche photodiodes for a variety of applications.