System-on-Chip Considerations for Heterogeneous Integration of CMOS and Fluidic Bio-Interfaces

Datta-Chaudhuri, Timir

http://hdl.handle.net/10945/52401
System-on-Chip Considerations for Heterogeneous Integration of CMOS and Fluidic Bio-Interfaces

Timir Datta-Chaudhuri, Member, IEEE, Elisabeth Smela, Member, IEEE, and Pamela A. Abshire, Member, IEEE

Abstract—CMOS chips are increasingly used for direct sensing and interfacing with fluidic and biological systems. While many biosensing systems have successfully combined CMOS chips for readout and signal processing with passive sensing arrays, systems that co-locate sensing with active circuits on a single chip offer significant advantages in size and performance but increase the complexity of multi-domain design and heterogeneous integration. This emerging class of lab-on-CMOS systems also poses distinct and vexing technical challenges that arise from the disparate requirements of biosensors and integrated circuits (ICs). Modeling these systems must address not only circuit design, but also the behavior of biological components on the surface of the IC and any physical structures. Existing tools do not support the cross-domain simulation of heterogeneous lab-on-CMOS systems, so we recommend a two-step modeling approach: using circuit simulation to inform physics-based simulation, and vice versa. We review the primary lab-on-CMOS implementation challenges and discuss practical approaches to overcome them. Issues include new versions of classical challenges in system-on-chip integration, such as thermal effects, floor-planning, and signal coupling, as well as new challenges that are specifically attributable to biological and fluidic domains, such as electrochemical effects, non-standard packaging, surface treatments, sterilization, microfabrication of surface structures, and microfluidic integration. We describe these concerns as they arise in lab-on-CMOS systems and discuss solutions that have been experimentally demonstrated.

Index Terms—Biosensor, heterogeneous integration, lab-on-a-chip, lab-on-CMOS, microfluidics, system on a chip.

I. INTRODUCTION

INCORPORATING CMOS chips for direct sensing and interfacing in fluidic or biological systems introduces a number of constraints and associated technical challenges spanning the fields of integrated circuits, microelectromechanical systems (MEMS), and biosensors. While some, such as interference and coupling between different signals, power limitations, mixed-signal processing, communications interfaces, and floor-planning, are familiar in system-on-chip design, a number of

II. SYSTEM DESIGN AND MODELING

A. Local Versus Remote Computation

Lab-on-a-chip systems fall into two broad categories, as indicated schematically in Fig. 1. These categories are distinguished by whether the sensing and the signal processing functions are co-located or separate, a decision that occurs early in the design. In the first, an interface chip (not necessarily CMOS, often passive) accommodating the sensors is the only part that touches the fluid or biological system. It is connected to one or more ICs that perform signal processing and computation.

1) Remote Computation, a Multi-Chip Solution: Multi-chip solutions are generally easier to implement and less costly than single chip solutions (Fig. 2). This approach allows the use of widely available sensing and packaging technologies as well as commercially available components, resulting in designs that are more modular in nature.
There are many examples of successful 2-chip systems. For example neural amplifiers are typically implemented using a 2-chip configuration [1]-[4]. There are also examples of electrochemical sensors that are implemented using passive arrays in conjunction with external instrumentation [5], [6]. In a 2-chip system an IC designer can rely on well-known design and validation methods, and the biological interface chip can be realized independently, with an agreement between designers on the number of wires needed to connect the two chips. In addition, the electronics can be readily and rapidly tested with other biological interfaces, either custom-made or commercially available.

The size of the IC is defined by the necessary circuits, and there are no inherent limits on the size and shape of the passive chip. This allows the cost of the IC to be minimized. Using a second non-CMOS chip also permits passive electrical components to be placed off of the expensive CMOS chip, for example using an integrated passive device (IPD) process, allowing higher circuit density, but at the cost of adding input/output (I/O) pads. The passive chip can be as large as needed to accommodate fluidics (as discussed in Section II-C4). At the same time, the size, form factor, and material of the second chip can be optimized for the biological application. For example, neural probes must be long, thin, and flexible.

The most important advantage to using more than one chip is that sensing density is limited by the availability of input-output connections. Sensing sites are typically connected to off-chip active circuits using wires, leading to a 1:1 mapping between sensing sites and dedicated CMOS bond-pads in the case of completely passive sensors. Another disadvantage is the opportunity for noise to enter the measurements. Clearly, unshielded wires between the sensors and electronics can result in strong interference from ambient electromagnetic (EM) radiation (e.g., power supplies [1]). Even with shielded wires, the wire length should be minimized.

For a single-chip approach, noise can be reduced by actively buffering weak signals.

### 2) Heterogeneous Integration, a 1-Chip Approach:

There are also numerous examples of successful integrated systems designed for direct interaction with biology or fluids [7], such as neural amplifiers [8]-[10], magnetic sensors [11]-[13], optical sensors [14], [15] and electrochemical sensors [16]-[18]. The main reasons for combining CMOS with MEMS surface structures and fluidics are: weak signals, sensing that requires active circuits, extreme size constraints, applications for which wires are not possible, and desired additional functionality or increased density of sensing sites.

A significant motivation for creating an integrated system is the minimization of electrical noise and parasitics. Most single-chip solutions for neural amplifiers exhibit exceptional signal clarity with high SNR. Sensing modalities that utilize inherently passive devices, such as optical sensors or magnetic sensors, produce weak signals that suffer from interference or attenuation due to parasitic loading if connected by wire to a remote chip. These types of sensors benefit from close integration with signal conditioning and readout electronics, and perhaps computation.

Some sensing mechanisms, wherein the circuit operation plays an integral role in the sensing mechanism, inherently require close integration with active circuits. Examples include active pixel sensors [19] and charge-based capacitance measurement approaches [20]. In other cases the quality of sensing can be enhanced through close integration of sensors and circuits. One method for improving the detection of very weak light, such as bioluminescence [21], is to minimize optical loss and maximize geometrical collection efficiency by placing the sample volume and optical detectors as close as possible to each other [22].

Some applications, such as microrobots and implantable devices, are highly sensitive to size and power requirements associated with sensing. For systems that operate under extreme size constraints, it is necessary to integrate sensing together with circuitry in order to minimize the footprint associated with sensing, since two chips take more space than one. In a similar vein, we anticipate the development of future applications in which it is simply not possible to establish connections from the sensing region to the outside world using wires. In such applications, sensors can be integrated closely with circuitry in order to provide power transfer and communication.

CMOS circuits from the foundry can be modified using post-processing to add sensing domains. Intermediate structures, between the IC and the sample, can transduce physical phenomenon beyond those sensed by CMOS alone, such as pressure or mass. Micromachined channels and structures also offer additional capability. An impressive early example was an implantable active probe [23] that was developed into to a high density neural interface with microfluidic drug-delivery capability [24].

Sensing density in passive arrays is limited by wiring. Active circuitry can utilize addressing to support a high density of sites. In a binary switching scheme, the number of required wires scales logarithmically with the number of sensing sites, rather than linearly.

There are many obstacles to implementing a single-chip configuration. System design and modeling have completely new sets of constraints that may be unfamiliar to a circuit
designer (discussed in the remainder of this section). The tools for designing such systems have not reached a level of sophistication and reliability comparable with that of IC design software. Furthermore, heterogeneous integration of electronics and fluidics requires packaging, which cannot rely on industry-standard wire bonding or flip chip mounting (Section III). Finally, using a CMOS chip as a platform for successful cell culture and biological experiments requires modification of existing biological protocols (Section IV).

3) Comparison of Signal Quality With the Two Approaches: Notwithstanding the technical challenges, in many design scenarios the most important considerations are resolution and accuracy. Ideally, one would like to be able to quantitatively compare the signal quality of the one- and two-chip approaches to aid in the decision process. Unfortunately, differences between individual systems and a lack of testing data make this difficult. Typically, circuit designers report the electrical specifications of their circuits alone (no sensors), and in some cases the circuits are tested with sensors attached. It is uncommon to find a characterization of the system performance before and after the front-end sensors have been added.

A good example of a two-chip system is a neural amplifier chip connected to passive, remote electrodes. Significant design effort has been invested in the development of low power amplifiers (for example for implantation or portable use) that exhibit low input-referred noise. The input-referred noise amplitude must be lower than the neural signals, which are as low as 10 s of μV. Typically the neural amplifiers are characterized before they are connected to the electrodes, and the noise characteristics of the system are not usually specified.

The electrodes for neural amplifiers contribute both thermal noise and 1/f noise [23], as well as less-understood interfacial noise. The contributions of each of these is unknown. Two groups have characterized their systems with added electrodes, providing ballpark values that would be expected to be similar for other systems. Najafi et al. [23] found that the noise (at 1 kHz) in their system increased tenfold, from under 10 nV/√Hz for the preamplifier alone to 100 nV/√Hz, upon the addition of electrodes (of 5.8 MΩ). A band-limited integration under the noise curve provided in [23] shows that the rms noise value increases by a factor of 2 as the electrode impedance increased to 12.9 MΩ. Mohseni et al. [25] found that the input-referred noise increased, from 7.8 to 10 μV rms, when an electrode (177 μm²) grounded in saline solution was attached. (Surprisingly, the addition of an unusually large (3000 μm²) electrode decreased the input-referred noise below that of the amplifier alone, from 7.8 μV rms to 7 μV rms.)

System noise has been estimated by post-processing the amplified neural signal. The identified spikes were removed by filtering, and the background noise was estimated from the remainder. This method was utilized in [26], showing that the input-referred noise increased from 3.5 μV rms to 10 μV rms when considering the added effects of the electrodes and test setup. The electrodes again brought the noise floor of the system up to 10 μV rms.

There are no reports of noise in 1-chip systems. Since the electrodes are close to the active devices, presumably the only added noise would be front-end noise introduced by the sensing electrodes. For further information about noise contribution mechanisms from electrodes, we point the reader to the insightful discussion about electroencephalogram (EEG) electrodes in [27].

4) 1-Chip or 2-Chip Approach: Single chip systems are significantly more complicated to implement. In fact, the focus of many papers has been solely on the development of packaging and integration schemes for single chip systems because most packaging cannot support fluidic interfaces. The design process for lab-on-CMOS systems should consider from the outset not only the sensor operation but also all of the integration and application constraints.

The design flow begins with the decision to make either a 1-chip or 2-chip system (Fig. 2). Given the integration challenges of a single chip system, the decision to pursue this approach must be sufficiently justified. Since a 2-chip system includes one passive chip the decision process must revolve around the ability of the passive chip to perform the required sensing, and the feasibility of communicating high quality information from the passive chip to the active CMOS IC. Because the 2nd chip is passive it cannot contain any active sensing sites, switches, or signal buffers. Accordingly, the first criterion becomes whether the sensing can be performed with purely passive structures. For example, a simple photodiode could be placed on a passive chip, whereas an active pixel sensor could not. Similarly, if the signal source is weak and the signal path is subject to noise comparable with the signal power, then active amplification with a single chip solution would be required. Lastly, even if the sensing can be performed by passive devices, communicating the information between the passive chip and active chip can present issues with the number of electrical connections required, the possibility of creating the connections given the application, and the opportunity for noise to be coupled into the signal traces. Certain applications cannot accommodate a large number of connections. For example, in implantable systems the percutaneous wiring density is limited by the surgical site, and the expected lifetime is a function of the anticipated motion. Since each sensing site must be connected by a wire to the active IC, these wires will effectively act as antennas for noise from electromagnetic radiation. Therefore a single chip solution is necessitated when either the required number of wires cannot be implemented, when the wires will experience excessive mechanical forces, or when the signals are too weak and thus subject to excessive interference from noise.

B. Modeling of Heterogeneous Systems

The materials and components (including the biological ones) in a heterogeneous system often have wide ranges of parameters (such as dielectric constants and layer thicknesses) and batch-to-batch inconsistencies. Some CAD tools partially address the modeling of heterogeneous systems, e.g., finite element simulation of MEMS structures and SPICE-like simulation of biological components, but widespread adoption and standardization are not on par with IC CAD tools. As these tools develop to incorporate cross-domain simulation and modeling, they are expected to play an increasingly important role in the design and optimization of future heterogeneous systems.
1) MEMS Modeling and Simulation: There was a push in the early 1990s to develop standardized goal-oriented CAD tools for MEMS [28], with lumped-parameter models for various mechanical components that could be subsequently integrated into an electrical simulator to model their interaction with circuits [29]. This effort and others have led to the development of modern MEMS simulators that are capable of aiding the entire MEMS design process from mask design to fabrication. Software can model the results of sequential deposition and etching steps and provide a 3D model of the resulting physical structures. However, MEMS processing is susceptible to variations, especially within a university research environment. A few CAD platforms are capable of modeling process uncertainty and error propagation using Monte Carlo methods (e.g., MEMS Pro). However, even these advanced CAD tools are incapable of modeling the interactions of MEMS with CMOS and biology.

The main MEMS CAD software packages are listed in Table I. Some of these tools, such as MEMS Pro, are built as modular extensions to existing circuit design software. Other tools, such as those offered by ConventorWare and IntelliSense, are more complete packages that include multi-physics capabilities and MEMS process simulation, and they can be used to create parametric behavioral models for circuit simulators.

2) Two-Step System-Level Modeling: CAD tools for each domain can be used to simulate particular aspects of an integrated system. Unfortunately, no tool adequately addresses all of these issues in one platform. The fundamental obstacle to modeling an entire system with heterogeneous integration of CMOS, MEMS, and biology is that each one of the domains is sufficiently complex that modeling interactions between them presents a significant challenge. A preliminary approach to integrated modeling may be to use domain-specific modeling systems to develop macro-level behavioral models that can be subsequently imported into a simulator to provide a limited level of cross-domain simulation capability. Even then, model validation for lab-on-CMOS components will be non-trivial: MEMS devices can vary significantly, and biological processes are notoriously stochastic.

Once the decision to pursue a 1-chip system has been made, a coupled iterative simulation process is required, such as the one introduced in [30] for electro-thermal circuit simulation, or the method demonstrated in [29] that utilizes the results from physics simulations and empirical validation to create behavioral SPICE models of RF-MEMS devices. Fig. 3 illustrates the two-step design and simulation methodology for the multiple domains in which a lab-on-CMOS device must operate. The core idea is that the results from simulation in one domain are used to inform the others.

The process begins with the biological question to be answered. This sets the requirements for the transduction as well as the packaging and fluidic integration. The MEMS and circuit schematics are simulated in SPICE, while the multi-physics models are built around the chip’s environment. The models for the MEMS structures are developed on their own and then used as parametric inputs to both the circuit and multi-physics models. Then, the circuit simulations and multi-physics simulations are performed iteratively.

C. CMOS IC Design

The rest of this paper pertains to lab-on-CMOS systems: fully integrated systems with CMOS, post-fabricated MEMS structures, and cells cultured on the surface of the chip. IC design for such an integrated system needs to take into account thermal effects, changes in electrical characteristics, electrochemistry, and additional floor planning constraints.

1) Thermal Effects:

Circuit design and simulation: In many cases, lab-on-CMOS systems incorporating living biological cells will operate in a 37 °C environment. Circuit simulators can account for this, i.e., via the “.TEMP” statement in SPICE. The temperature dependence of MOS transistors has been well characterized: the threshold voltage varies with temperature [31], as does the carrier mobility [32]. Some advanced simulation packages can even compensate for Joule heating and heat transfer (e.g., the Heatwave module for Cadence), but if they do not, the onus is on the designer to take these electro-thermal effects into account [33] using a secondary simulation platform. The option also exists to build temperature compensation into the circuits [34] using feedback from on-chip temperature sensors [35]-[38].

Modeling temperature variations: Ceramic dual inline packages (CerDIPs) are designed to act as heat sinks [36],...
but lab-on-CMOS packages present new thermal boundary conditions. Typical thermal analysis of IC operation assumes a boundary condition of room temperature. A secondary or modular plug-in simulation platform can be used to model Joule heating effects as well as heat conduction by the package. This two-step iterative modeling approach requires an initial SPICE simulation to find the power dissipation of the circuit blocks. Power dissipated over the circuit layout areas is used to model thermal heating effects using a finite element simulator [36] such as COMSOL, adding information on materials and the environment. Then, the results of the thermal simulation can be fed back into the SPICE simulation to specify the operating temperature of the circuits for determination of changes in operation and power consumption.

Biological consideration of thermal effects: Cell culture requires strict temperature control: for mammalian cells, the surface should be at 37 ± 2 °C [39], a condition that is typically maintained by an incubator. However, when the IC is operated, power dissipation raises the chip surface temperature. This causes hot spots that can exceed the acceptable range for cell culture [36].

Fig. 4 shows the surface temperature predicted by an electro-thermal simulation of a mixed signal IC within an epoxy package at a boundary condition of 37 °C. The IC was designed as an active microelectrode array for cultured neural cells. It comprised an array of pixels to amplify bio-potential signals, perform nonlinear signal processing to enhance spikes, and output the spatial location of spiking events as they are identified in real time using an asynchronous readout scheme. An earlier version of this chip was reported in [40]. The analog array included amplifiers with in-pixel spike detection circuits, and the digital portion of the IC included address event representation circuitry and high powered buffers to enable the chip to directly drive external hardware. Fig. 4(a) shows the predicted temperature of the chip when water covers it entirely, with the surface temperature indicated by the color. It is evident that high power dissipation in the digital circuits results in temperatures that are too high over the entire surface, despite the high thermal conductivity of silicon and conduction of heat by direct contact with water. This issue could be partially resolved by changing the thermal boundary condition: specifically, by lowering the incubator temperature by an amount corresponding to the local heat generated by the IC. Alternatively, the results of an electro-thermal simulation can be used to inform the design of microfluidics to actively cool the IC. Fig. 4(b) shows the resulting temperature profile when a fluidic channel is used to transport heat away from the IC. Although the temperature of the fluid at the inlet is 37 °C, the active flow reduces the surface temperature of the IC to a level that is within the tolerance for cell culture.

For use of the lab-on-CMOS system outside an incubator, on-chip heaters can be employed advantageously to warm the cell medium [36], at the expense of significant power draw. However, many biological experiments require fluid flow over the cells. In such cases, it is necessary to preheat the fluid because on-chip heaters cannot warm the fluid sufficiently without operating at much higher temperatures than cells can tolerate. Without flow, and if the system is well insulated, on-chip heaters may be sufficient, but they should be uniformly spatially distributed throughout the layout to minimize thermal gradients. It has been observed that relatively small variations in temperature play a large role in cell viability across the chip [36]. Power dissipation causes a similar concern for \textit{in-vivo} implantation, for which there are varying reports of the maximum acceptable power density. The reported values range from 40 mW/cm² [41] to 80 mW/cm² [42] based on the type of tissue and heat dissipation from blood flow around the implantation site.

2) Fluidic Effects on Communication and Computation: Communication traces are needed for coupling signals and controls in and out of the system and for internal signal routing for processing or computation. By changing the medium that surrounds the chip from air to water, the effective dielectric constant between areas of the chip is increased (since the dielectric constant of water is approximately 80 times that of air), making the IC more susceptible to coupling of interfering signals. This can be illustrated by examining pad-to-pad capacitance. In a typical CMOS IC, the pad-to-pad impedance is large and well characterized, but in the presence of an aqueous solution such as cell medium, the capacitance between pads increases. Using
a finite element model we show (Fig. 5) that this increase can be over an order of magnitude. (The relative permittivity of saline solution varies with salt concentration, but for the levels of solutes used in biological media the relative permittivity remains close to that of water [43]). The larger capacitance will increase pad-to-pad signal coupling, which is a problem when one pad has a digital signal and the other an analog signal from a high impedance source.

Due to the increased likelihood of signal coupling, it may be necessary to use standard analog circuit techniques such as using driven guards for immunity to interference [44] and differential processing for immunity to correlated noise sources, or to introduce additional microfabrication to mitigate the noise, such as the introduction of an electrical shielding layer between the surface and the underlying circuits [44]. We have previously observed this coupling empirically in two different lab-on-CMOS systems. In one example, the lab-on-CMOS system was designed to measure the capacitive coupling between anchorage-dependent cells and their substrate, as a means to assess cell health [45]. Early versions of the sensor utilized a single-ended measurement technique, and significant coupling of signals from adjacent traces was observed. A subsequent version of this chip used differential measurements and driven guards to reduce this coupling [44]. In another example, the lab-on-CMOS system was designed to measure the weak bio-luminescence signal from genetically engineered cells in response to binding by a target analyte [46]. The cells were cultured on the surface of the chip, but the high dielectric constant of the culture medium introduced unacceptable signal coupling, which degraded the measurement quality. In this case signal quality was restored by introducing an optically transparent indium tin-oxide shielding layer between the culture volume and the underlying optical sensors.

The results of FEM simulation can again be used in a two-step modeling process: the effect of coupling through the fluid can be taken into account during circuit simulation by altering model parameters. Unfortunately, the parameters will differ among CMOS fabrication processes, which utilize different materials and layer thicknesses. It would likely require collaboration with the CMOS foundry for accurate modeling.

Increasing inter-pad spacing at the cost of expensive silicon (Si) real estate is not an optimal way to alleviate signal coupling. FEM simulation can be used to determine the packaging materials and appropriate layer thicknesses to mitigate the increased coupling. For the results shown in Fig. 5, a three dimensional model of an IC was used for electrostatic simulation. An initial simulation was performed to determine the difference in capacitance when air was replaced with fluid. Further simulations were then performed to determine the necessary thickness of passivation material (parylene) to minimize the effects of signal coupling through the fluid.

3) Electrochemical Effects: Microfluidic and cell culture environments are inherently electrochemical, and CMOS chips are inherently electrical. This difference imposes new challenges, such as unknown electrode potentials that vary substantially over time due to electrochemical effects and unintentional activation of electrostatic discharge protection structures due to DC differences between the electrochemical “ground” and on-chip electrical potentials.

Electrochemical potentials may generate large DC offsets, which saturate the inputs of electrical circuits. One possible way to reject such offsets would be to use an auto-zeroing amplifier, such as the one proposed by Hasler et al. [47], which is capable of rejecting DC offsets in the input signal by using an adaptive floating-gate circuit. For neural amplifiers, another approach to this issue has been the use of differential amplifiers with large coupling capacitors at the inputs [1]. These amplifiers work well but use a large amount of space, decreasing the density of active microelectrode arrays. This is a case where a 2-chip solution may be advantageous (Section II-A1).

Another problem occurs when electrical circuits are used to drive electrochemical reactions. Electrochemical reference electrodes provide a well-defined potential in electrochemical space, but thin-film versions are short-lived. Without the use of a true reference electrode, the potential of the CMOS circuits relative to the electrochemical potentials in the fluid is not known exactly. Thus, redox potentials may appear to be scaled and shifted. This places dynamic range constraints on the circuits, in some cases requiring them to provide driving signals that are below the supply ground. In such cases, the circuits must be designed to have positive and negative supplies with a reference ground (connected to the fluid) at the middle of the supply rails [48].

The design of on-chip potentiostats for electrochemistry requires consideration of not only the materials and environment, but also informed design of the CMOS circuits that drive the reactions. We point the reader to [49] for a thorough discussion of the various circuit configurations and readout techniques, including temperature compensation for precise current measurement.

4) Floor Planning and Area Resources: Limited Si area always poses a challenge to designing CMOS systems. The integrated nature of lab-on-CMOS design complicates floor planning further because portions of the chip need to be utilized for reference electrodes, fluidic integration, cell culture, and MEMS structures.

![Fig. 5. Results of an FEM simulation showing the change in pad-to-pad capacitance when water ($\varepsilon_r \approx 80$) replaces air ($\varepsilon_r \approx 1$) on the chip surface. Also shown is the effect of varying the thickness of an intermediate passivation layer ($\varepsilon_r = 3$). Values have been normalized to the capacitance in the presence of air.](image)
Surface topography: Topographical variations in the CMOS surface can be problematic for post-processing. The thickness of structures fabricated on top of the chip may be less than the height variations of unplanarized layers, distorting the structures as well as creating difficulties in employing sacrificial layer release. Patterning fine features may also be a problem due to inconsistent focus during photolithography. The occurrence of topographical artifacts varies between CMOS processes; in some technologies, the internal layers are planarized, but the surface is not. If the topmost metal layer introduces unacceptable surface topography, it cannot be utilized, reducing design flexibility and possibly system capability.

Surface structures: The physical layout of overlying electrical or capacitive structures will affect the placement of underlying electronics. At least the top layer of CMOS metal may be required for interfacing, and will thus not be available for signal routing.

Metal layers commonly used in CMOS are not chemically inert, and therefore cannot be used to connect to the fluid directly (Section IV-C). Ideally, the top electrode should be covered with an inert metal or conducting polymer. Furthermore, aluminum (Al) (which is commonly used for on-chip routing) is susceptible to etching by both wet and dry processes used during post-fabrication. To mitigate accidental etching of the CMOS top metal layer, it is advisable to use a metal stack with arrayed vias (tungsten (W), a common material for vias, is impervious to many etchants that attack Al), to ensure connectivity with underlying metal layers.

Fluidic integration: One of the major hurdles in integrating microfluidics arises from the mismatch in sizes between microfluidic systems and CMOS ICs [16]. Although microfluidics take advantage of fluid dynamics at the micro-scale, the fluid path and the inlets and outlets are relatively large. Typical ICs are millimeter scale, while microfluidic systems are centimeter scale. Additionally, providing access to larger fluid volumes makes it easier to maintain the cell culture environment [36].

If wire-bonding is used for packaging, the fluid path must avoid the bondpads entirely. One approach has been to remove bondpads from two sides of the IC to make room for a fluid path down the center of the chip [50]. Other packages use patterned thin-film metal traces to provide I/O [16], [51]. Systems that employ flat interconnects significantly ease adding complex fluidics. If the surface is flat enough to employ microfabrication techniques to create I/O paths, then other post-processing, such as the addition of MEMS structures and microfluidics, is also possible.

Placement considerations for biology: It is known that electric fields influence cell morphology, particularly nerve cells [52]. For cell culture or device implantation, it may be necessary to use one metal layer to shield the biology from the underlying electronics, again using one of the metal layers in the CMOS process.

Although some of the effects of EM radiation from the IC itself can be modeled, the interaction of the IC with the environment cannot. Basic EM radiation modeling can be performed in COMSOL with parameters such as locations of potential sources and RF signal generation sites. However, the effect that the EM radiation will have on the cells and the medium cannot be modeled fully because in many cases these effects remain unknown. Certain well-understood effects could be predicted, for example dielectrophoresis [53], but modeling other effects, such as electrochemical interface effects and ionic gradients, is likely to be too complicated for a general FEM solver.

III. PACKAGING AND INTEGRATION CONSTRAINTS

CMOS chips exposed to fluidic environments have packaging requirements not satisfied by standard methods developed for ICs. Packaging that integrates electronics with fluidics remains an ongoing area of research with different approaches demonstrated in the literature [9], [11], [12], [15], [16], [54]–[56].

Unlike traditional electronic packaging using CerDIP or similar methods, fluidic packaging has not been standardized by industry. Few commercial products exist that incorporate CMOS with fluidics. One notable exception is the BioCAM system offered by 3Brain, which allows neural cell cultures to be plated on the surface of a CMOS sensing array. Although the packaging for the commercial product is proprietary, some related early work by Berdondini et al. [57] indicates that the CMOS chips were wire-bonded to a PCB and then passivated using epoxy. Another example of an industrial solution that was successfully adapted for fluidic integration was based on the commonly used low temperature co-fired ceramic (LTCC) package. CMOS chips were bump-bonded to ceramic carriers, and passivation material was applied to the bonds to isolate them from fluid, allowing cell culture on the surface of the IC [10]. LTCC packages themselves can be designed to incorporate microfluidic channels (in addition to the electrical traces) within the stacked layers that form the ceramic substrate [58].

In the absence of standardized fluidic packaging, there are two options to accommodate lab-on-CMOS requirements: adapt a pre-existing standard package or develop a custom package. The primary requirement for any fluidic integration is that the IC surface must be selectively passivated, allowing isolated electrical connections from the I/O region to the outside world while exposing the “active” areas of the chip to the fluid to enable sensing. Inadequate passivation results in system failure and undesired electrochemical effects, as shown in Fig. 6.
traces provide electrical I/O, and a passivation material isolates them from the embedded in a handle wafer and leveled using a filler material. Patterned metal Fig. 8. Schematic of an example custom chip-in-hole package. The IC is embedded in a handle wafer and leveled using a filler material. Patterned metal traces provide electrical I/O, and a passivation material isolates them from the sensing region.

The issue of protecting the I/O from the fluid is easily solved for large CMOS chips, on which spatial separation of the sensing region and the I/O allows the use of physical structures such as glued-on wells. If the chips are only mm scale, other techniques must be used for integrating the fluidics. A review of some common methods can be found in [51].

If wirebonds are used to provide I/O, isolation is challenging because of their topography. Wirebond passivation can be done with a conformal material such as parylene (which can be selectively patterned) or epoxy (which can be selectively applied or UV exposed). In a standard chip carrier (Fig. 7), the entire cavity can serve as a fluid well. Alternatively, if the passivation material can be patterned during deposition (for example by using a UV-curable polymer or sacrificial material (see Section III-F), it can be used to define the fluidic path. Similarly, if the IC is wire-bonded to a PCB, a patternable passivation material can be used to protect the wirebonds, after which additional fluidic structures can be added around the chip. However, patterning the passivation material photolithographically is difficult because of the large step height between the chip and the carrier.

The second category of packages, the custom “chip-in-hole” type, allow significantly more freedom for integrating not only fluidics but also additional structures. The defining characteristic of such packages is that the chip surface is level with the carrier. Post-processing is straightforward when the IC is embedded in a “handle” substrate (Fig. 8) [16], [51], [55]. This allows microfabrication techniques to be used not only for passivation, but also for constructing the I/O and the fluidic structures.

MEMS-based post-processing can also be used to increase system functionality in embedded chip packages by adding structures such as optical filters or actuators for sample manipulation. Handle wafers further allow placement of off-the-shelf components away from the CMOS real estate. If the handle substrate is a semiconductor, the possibility is opened for integrating large-area components (resistors, capacitors, and inductors), and even diodes and transistors.

One issue with creating chip-in-hole carriers is that the size of the chip from the foundry is inconsistent. There can be a substantial gap between the edge of the IC and the handle wafer, which inhibits conformal application of photoresist and results in discontinuity of the metal traces used for I/O connections. An elegant solution is to create a custom hole for the IC using the chip itself to define the lateral dimensions of the cavity (self-aligned masking) [59]. Other approaches to the gap issue have been to create holes that are larger than the chip and use material such as spin-on-glass (SOG) or polymers to bridge the gap and bring the chip flush and level with the surface [16]. The issues caused by non-uniformly sized chips are mitigated if, instead of a hole being created within the carrier, the carrier is formed around the chip, e.g., by using a polymeric carrier [51]. Epoxy yields a rigid substrate that can be treated like a wafer. An interesting alternative is the use of an elastomer such as PDMS [12], which allows for flexible substrates. Forming the carrier around the chip has the additional benefit of enabling integration of multiple chips and electrical components into a single carrier wafer.

Given that lab-on-CMOS design should encompass not only lab-on-a-chip systems but also medical devices and environmental sensors, it would be beneficial if there were greater collaboration between academia and industry in developing standardized packages for fluidic interfaces.

A. Surface Materials

Information about the materials on the surface of the integrated circuit is needed for subsequent processing. For example, whether the uppermost layer is nitride, oxide, or polymer determines its resistance to etchants. Commercial CMOS chips come with a fixed, proprietary set of materials, and different foundries are known to utilize different materials for the metal stacks and final passivation layers. Since the materials and thicknesses are unknown to the end-user, it may be necessary to identify materials using elemental analysis. Biocompatibility poses additional constraints (see Section IV-C). Depending on the material placement, it may be necessary to have compatibility with cells growing directly on the material surface or just with cells growing nearby. (Some materials contain toxic additives that can leach into the sample volume.) Lab-on-a-chip diagnostic systems require surface functionalization with biological molecules, such as nucleic acids, proteins, or antibodies. In this case the chip surface must be amenable to the attachment chemistry, e.g., silanization, thiol bond formation, micro-printing, or electrodeposition.

B. Resist Edge Beads

Microfabrication requires that photoresist can be applied uniformly and defined photolithographically. Resist is applied by spin-coating to yield a particular thickness. Surface tension results in an unwanted thicker “edge bead” around the perimeter.
of the substrate [60]. For wafer-level processing, the edge bead is removed so that the photo-mask can make contact with the resist surface, since a gap degrades resolution. For chip-level processing, edge beads can pose a significant problem. In our own work we have measured (Tencor P-20 profilometer) edge beads as thick as 25 μm for a resist with a nominal thickness of 1.8 μm (Shipley 1813, 2000 rpm) on a 3 × 6 mm die. On mm-scale ICs, the bead can occupy most of the die area. Lin et al. measured resist uniformity under similar conditions (3000 rpm, 3 × 3 mm die), and it was limited to 50% of the chip surface [61]. Embedded chip packaging techniques circumvent this issue by shifting the edge bead to the perimeter of the handle wafer.

Another approach to mitigating the edge bead is to increase the centrifugal force on the resist by temporarily adhering the chip to the outer edge of a large wafer, thereby increasing the spin radius and corresponding centrifugal force. This technique has been applied at typical resist spin speeds [61], and dies exhibited less edge beading and greater uniformity. In our own work at higher spin speeds (8000 rpm) the edge bead was reduced in width and height, but the resist thickness in the center decreased.

C. Low Temperature Processing

Many standard processes used during MEMS fabrication, such as annealing polysilicon structures to relieve stress, entail temperatures that would damage underlying circuits [62]. The thermal budget for an integrated system with heterogeneous materials is limited by the material with the lowest thermal tolerance. Si-based CMOS chips with Al interconnects can handle temperatures as high as 450 °C [63]. Polymers have lower limits: parylene melts at 290 °C [64], polydimethylsiloxane (PDMS) deteriorates at 340 °C [65], and the limit for SU-8 (an epoxy-based negative resist typically used for permanent structures) is 200 °C [66]. The limits for individual materials do not account for interfacial stress arising from thermal expansion mismatch between materials. For example, Si expands 3 ppm/°C, while SU-8 2000 expands 52 ppm/°C. This difference of over an order of magnitude can result in the destruction limit for SU-8 structures from the Si substrate at temperatures below the destruction limit for SU-8.

D. Damage to Circuits During Post-Processing

IC design rules include considerations to mitigate damage to gate oxides by antenna effects (charging during plasma-based processing) [67], [68]. Post-processing steps that employ plasma can cause similar damage, which can be avoided by depositing a metal layer beforehand to electrically short all the exposed CMOS connections. This metal layer can later be removed, after the post-processing has been completed.

E. Interfacial Impedance

An electrode-electrolyte interface presents a large interfacial impedance due to the transition between electronic and ionic conduction. One traditional method to reduce the impedance is to use electrodes having a large area; this, however limits sensing array density. Another well-known method is to use platinum black or iridium for the electrodes [69]. Another class of materials recently employed for recording [70] and electrical stimulation [71] are conjugated polymers, applied as a coating over the metal electrode. They decrease interfacial impedance and increase the charge capacity of the electrode. One example is poly(3, 4-ethylenedioxythiophene) polystyrene sulfonate (PEDOT:PSS), which was found to increase the SNR for in-vivo implanted electrodes [72]. We found that PEDOT:PSS deposited on Au electrodes using the methods outlined in [71] resulted in a decrease in impedance of nearly two orders of magnitude [51].

F. Fluidics Integration

Cell culture requires a medium that maintains isotonicity and pH, and that provides nutrients and growth factors. Microfluidics can be used not only to provide the required environment for cell culture, but it can also be used to provide chemical stimulation to the cells. For fluidics that are open to the atmosphere, evaporation changes the osmolality. Evaporation is exacerbated when working with smaller volumes of fluids [73]. Sealing the system is likely to be required unless the fluid is continually replenished.

Microfluidic channels can be fabricated separately and then bonded to the chip surface (e.g., PDMS can be bonded to silicon dioxide (SiO2) using oxygen plasma), or materials like SU-8 can be patterned directly on the surface of the IC using photolithography [13]. Hybrid systems have also been demonstrated in which the microfluidic channels were patterned in a material such as acrylic or glass and the fluid manifold sealed to the surface of the IC using an intermediate gasket material such as PDMS [59]. Alternatively, microchannels can be fabricated within the encapsulation material itself by using a soluble patternable material around which the encapsulation material is allowed to cure [74].

Surface hydrophilicity is essential for both adherent cell culture and proper wetting of the surface by the fluid medium. CMOS ICs are often passivated at the foundry with a natively hydrophobic material like silicon nitride (Si3N4). Fortunately nitride is readily modified by using oxygen plasma to create an intermediate layer of oxide that is hydrophilic [75].

G. Package Sterility

The risk of exposing cells to microbial pathogens or mold increases every time they are removed from the incubator. One approach to maintaining healthy cultures is to keep them in sealed chambers enclosing the sensing sites [76]. Microfluidics structures can also be employed to isolate the cells from the environment and have successfully been used for cell culture outside of an incubator [36]. Fluid flow within microfluidic channels can, however, apply large shear forces to cells. Stress due to shear has been shown to lead to abnormal morphology, induce changes in biomolecular processes [77], and cause apoptosis [78].

The lab-on-CMOS device should be sterilized initially and between experiments. The standard method of autoclaving
(120–190 °C) may be incompatible with heterogeneously integrated systems. Rinsing with alcohol or exposure to ultraviolet light (UV) are other common sterilization methods, but material compatibility should be known beforehand. Parylene and Si₃N₄, commonly used for surface passivation, are compatible with alcohol and bleach, allowing use of liquid disinfectants followed by UV light exposure.

It is also important to consider bio-fouling for chips that are to be reused. Cell culture on the surface of the chip will inevitably result in the formation of a film of cells and proteinaceous debris on the surface of the chip. This film degrades signal strength and can prevent proper cell attachment when the chip is reused. We have observed that these films must be removed as quickly as possible and not allowed to dry out, since this makes removal difficult. Rinsing with alcohol is not an effective method for cleaning chip surfaces because it can result in “fixation” of cellular debris on the chip surface [79]. Use of a cleaning agent that combines detergents with protease enzymes (such as Alconox Tergazyme) is recommended by manufacturers of multi-electrode arrays. We have achieved effective cleaning of active microelectrode array chips by soaking in a 1% solution of Tergazyme and applying gentle agitation over the course of two hours, followed by rinsing with water and sterilization using alcohol and UV light exposure.

H. Robustness

Materials in the system should ideally remain robust and stable over time, but when they are exposed to an aqueous saline environment, many chemically react (Section IV-C). Another pernicious way in which systems fail is by material absorption of water. If a polymer is used to encapsulate wire-bonds, swelling leads to electrochemistry (Fig. 6) and bond wire detachment. In our experience there are many underlying mechanisms that can lead to material changes and associated system failure, including: exposure to solvents, which can cause swelling and cracking; dissolution of metals over time, resulting in connection failures; and aging effects such as precipitation of functional species embedded in polymeric surface coatings [22].

IV. BIOLOGY ON CMOS

Living biological components, such as sensory cells, must be provided an environment that closely mimics physiological conditions.

A. Cell Plating

The culture of adherent cells can be facilitated by adhesion promoters. While some cells, e.g., muscle cells, do not require coatings for surface attachment, others, such as neurons, do. Natural adhesion promoters such as collagen, laminin, and fibronectin simulate the extracellular matrix found in most types of tissue. Cationic polymers such as polylysines and polyethylenimine create a positively charged surface, which facilitates attachment of negatively charged cell membranes [80]. Unbound synthetic polymer molecules can be toxic, so surface treatment with these requires thorough rinsing after coating [81].

Promoting intimate contact between cells and the underlying material has the added benefit of increasing the SNR for electrical sensing [82]. However, most adhesion promoters are insulating, so coatings must be kept thin to minimize interfacial impedance. Biology protocols do not consider the minimum thickness of adhesion promoter required for cell attachment, and the tradeoff between adhesion promoter concentration and signal strength has not yet been characterized. In our experience it has been necessary to experimentally determine the type and amount of adhesion promoter required to maintain cell adhesion in the presence of modest shear forces arising from fluid perfusion.

B. Cell Culture Outside of an Incubator

A bicarbonate buffering system [using carbon dioxide (CO₂)] is used for acid-base homeostasis by many animals, and it has been adopted for mammalian cell culture [83]. Bicarbonate buffered media require non-atmospheric levels of CO₂ to maintain physiological pH, provided by incubators. Lab-on-CMOS systems will likely need to work outside of an incubator. The Hibernate medium (Life Sciences) [86] can sustain mammalian neural cell culture at 2–8 °C and atmospheric CO₂ levels for up to 30 days, allowing experiments, and even culture, outside an incubator. We have successfully used the HEPES buffering system for culture of sensory neurons outside of an incubator. HEPES can be toxic to cells when exposed to light [85], but this was not a limiting factor in our work. However, although CO₂-independent media do exist, bicarbonate remains essential for certain types of cell culture [84].

Amphibian cells have less strict requirements (nutrients, temperature, pH) than mammalian cells and can survive hypoxic conditions. In our experience, primary amphibian neurons remained viable (determined by cilia movement and Trypan blue exclusion) for at least 7 days at 2–8 °C under atmospheric CO₂ when stored in a standard amphibian Ringer’s solution containing glucose, salts, and HEPES.

Incubators also maintain temperature for cell culture. When traditional CMOS packaging is used (e.g., CerDIP), a potential approach to culture outside an incubator may be to place a heating element on the package itself and use an on-chip temperature sensor to implement feedback control. Ceramic packages have high thermal conductivity as well as substantial thermal mass, allowing them to act as stable heat reservoirs.

C. Biocompatibility of Intrinsic and Added Materials

Materials such as Au, Si, SiO₂, Si₃N₄, and SU-8 have been shown to not inhibit cell culture and proliferation for durations of at least a week [87]–[89]. However, many of the materials commonly used in microfabrication have not been tested for compatibility with cell culture.

Al is reactive and subject to electrochemistry when in contact with cell media. Unfortunately, Al is commonly used for signal routing layers in CMOS, and active electrode arrays generally use the top metal from the CMOS process as the recording
focus is better and more uniform across the image. 1 mm of fluid. (b) The same region taken with a water immersion lens. The standard objective. The focus is the best possible given the optical path through electrode. Therefore, Al electrodes must be covered with an electrochemically inert electrode material. Au and platinum (Pt) have been used, but they both react with chloride (Cl\(^{-}\)). Both require an adhesion layer, such as chromium (Cr) or titanium (Ti). Ti is well established as a biocompatible material, with a long history of use in implants. However, both Cr and Ti react electrochemically with oxygen. Thus, it may be necessary to isolate the sidewalls of patterned Au/Cr or Pt/Ti layers from cell medium, e.g., with a dielectric film [90]. Patterning these covering metals also requires consideration. For example, wet etchants for patterning Ti generally contain hydrofluoric acid, which would also attack any SiO\(_2\) that may be on the IC surface. It may therefore be necessary to use lift-off for patterning instead of etching.

D. Imaging Cells on Opaque Substrates

It is often desirable to optically image cells plated on top of a CMOS substrate, for example, to correlate signals from an electrode array with cell locations or to perform calcium imaging. Imaging techniques that rely on an optical path through the substrate, such as differential interference contrast (DIC) or phase contrast microscopy, are not usable with opaque samples. Reflected light imaging techniques, such as confocal microscopy or reflected DIC, are necessitated.

An optical path through an air-liquid interface distorts the image and interferes with focus (Fig. 9) due to surface tension induced curvature of the liquid surface and uncorrected optical dispersion. Likewise, it is difficult to acquire high quality images through PDMS layers. Capping the microfluidics with a glass coverslip confines the fluid under a material with better optical properties than PDMS and also eliminates the meniscus, reducing distortion but not eliminating it. The best images are obtained with an immersion lens. The experimental configuration needs to accommodate the bulk of the lens, for instance with large-diameter shallow wells or removable wells.

Alternatively, determining cell location and detecting bioluminescence or fluorescence can be achieved with lensless imaging, using a CMOS optical detector as a contact imager [91]–[93]. Contact imagers forgo the large and often expensive optics required for traditional imaging and can be used in compact lab-on-a-chip systems.

V. CONCLUSION

We have reviewed many of the challenges associated with designing, packaging, and using a heterogeneously integrated system that interfaces CMOS with biology and fluidics. Splitting the sensing and signal processing functions onto separate chips alleviates many of the issues. There are, however, advantages to combining these functions in a single chip that make facing the integration issues worthwhile, including improved signal quality, increased sensing density, and added functionality.

During chip design, heat must be critically examined: power dissipation by circuits can easily be high enough to compromise cell viability, and operation in a 37 °C environment will need to be considered during circuit simulation. Exposure to fluids raises two other issues. The high dielectric constant of water directly over the circuits and routing traces increases signal coupling. Electrochemical potentials on the electrodes may saturate the inputs to amplifiers and impose dynamic range constraints. Unfortunately, system design and modeling remains a significant challenge because of the lack of tools that can address multi-domain integration, necessitating an iterative approach using both FEM and SPICE simulations. Additionally, floor planning must account for surface topography, connections to sensors and actuators produced by post-processing, and any necessary electrical shielding.

Packaging CMOS chips for fluidic interfacing must keep the I/O regions dry, but allow cell culture in the center of the chip. Fortunately this is an area that has seen much recent progress, especially from the lab-on-a-chip community, for whom integrating sensing, computation, and microfluidics enables not only miniaturization for portability but also massively parallel experiments and increased capability. The most versatile packages are those that embed the chip in a handle wafer, flush with the surface. The flat surface readily permits microfabrication procedures, such as the deposition of coatings to reduce interfacial impedance or the addition of microfluidics; it also opens the possibility to move components off the chip and onto the handle wafer.

Fundamentally, cell culture on CMOS has the same requirements as traditional cell culture, which means that the packaged chip may need to be coated with biomolecules and sterilized, and materials in the presence of cells must be nontoxic and nonreactive. Optical imaging of cells on the chip surface requires some consideration because the chip is opaque and is covered with a layer of fluid. One of the most important motivations for lab-on-a-chip systems is freedom from bench-top equipment, including incubators. Advancements in culture media have made it possible to keep some cell types alive without a CO\(_2\) environment, meaning that culture is possible if sterility and a temperature of 37 °C can be maintained. The former can be achieved with sealed wells or channels, and the
latter with heaters for the chip and fluid streams flowing over the cells.

We have outlined approaches that have allowed for successful implementation of lab-on-CMOS systems, both in our labs and many others. Lab-on-CMOS systems hold promise for future bio-electronic interfaces, both in the miniaturization of existing systems and in completely novel application domains.

References

Timir Datta-Chaudhuri (S’09–M’16) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, MD, USA, in 2007, 2013, and 2015, respectively. Currently, he is a member of the Neural Technologies Group within the Center for Bioengineering at Lawrence Livermore National Laboratory, Livermore, CA, USA. His research focuses on the development of implantable neural interfaces for restoring memory and the treatment of neuropsychiatric disorders. His research interests include brain-machine interfaces, lab-on-a-chip, cell-based sensing, biosensors, microsystem design and packaging, mixed-signal integrated circuit design, adaptive analog circuits, and neuromorphic engineering.

Elisabeth Smela (M’87) received the B.S. degree in physics from the Massachusetts Institute of Technology, Cambridge, MA, USA, and the Ph.D. degree in electrical engineering from the University of Pennsylvania, Philadelphia, PA, USA.

She worked at Linköping University, Linköping, Sweden, and at Risø National Lab, Roskilde, Denmark, developing microfabricated conjugated polymer devices, and then served as Vice President of Research and Development at the start-up company Santa Fe Science and Technology. She is a Professor in the Department of Mechanical Engineering and the Institute for Systems Research at the University of Maryland, College Park, MD, USA. Her recent research has focused on polymeric and biological microsystems, including cell-based sensing and mapping the positions of nucleic acids in tissue sections, as well as tactile skins for robots.

Pamela A. Abshire (S’98–M’02) received the B.S. degree (with honors) in physics from the California Institute of Technology, Pasadena, CA, USA, in 1992, and the M.S. and Ph.D. degrees in electrical and computer engineering from The Johns Hopkins University, Baltimore, MD, USA, in 1997 and 2002, respectively.

From 1992 to 1995, she worked as a Research Engineer in the Bradycardia Research Department of Medtronic, Inc. Since 2001, she has been with the Department of Electrical and Computer Engineering and the Institute for Systems Research at the University of Maryland, College Park, MD, USA, where she is currently an Associate Professor. Her research focuses on low-power mixed-signal integrated circuit design, adaptive integrated circuits, integrated circuits for biosensing, hybrid devices incorporating chips and biological components, and understanding the tradeoffs between performance and energy in natural and engineered systems.

Dr. Abshire serves on the Emerging Technologies and Research Advisory Committee for the U.S. Department of Commerce and on the Board of Governors of the IEEE Circuits and Systems Society.