CMOS Integrated Circuits for the Quantum Information Sciences

JENS ANDERS1 (Senior Member, IEEE), MASOUD BABAIE2 (Member, IEEE), JOSEPH C. BARDIN3,5,4 (Member, IEEE), IMRAN BASHIR3 (Senior Member, IEEE), GÉRARD BILLIOT6, ELENA BLOKHINA5,7,8 (Senior Member, IEEE), SHAI BONEN9 (Student Member, IEEE), EDOARDO CHARBON10 (Fellow, IEEE), JOHN CHIAVERINI11,12, ISAAC L. CHUANG11,13,14, CARSTEN DEGENHARDT15, DIRK ENGLUND11, LOTTE GECK15,16, LOÏCK LE GUEVEL3,6, DON Hee HAM17 (Fellow, IEEE), RUONAN HAN14 (Senior Member, IEEE), MOHAMED I. IBRAHIM1,4,18 (Member, IEEE), DANIEL KRÜGER1,17, KA MENG LEI19 (Senior Member, IEEE), ADRIEN MOREL6, DENNIS NIELINGER15, GÄEL PILLONNET6, JEREMY M. SAGE20, FABIO SEBASTIANO21 (Senior Member, IEEE), ROBERT BOGDAN STASZEWSKI7,8 (Fellow, IEEE), JULES STUART11,12,13, ANDREI VLADIMIRESCU21 (Fellow, IEEE), PATRICK VLIEX15, AND SORIN P. VOINIGESCU9 (Fellow, IEEE)

1Institute of Smart Sensors, University of Stuttgart, 70049 Stuttgart, Germany
2Delft University of Technology, 2628 CD Delft, The Netherlands
3University of Massachusetts Amherst, Amherst, MA 01003 USA
4Google LLC, Goleta, CA 93117 USA
5Equal1 Labs, Fremont, CA 94536 USA
6Université Grenoble Alpes, CEA-Leti, F-38000 Grenoble, France
7School of Electrical and Electronic Engineering, University College Dublin, 4 Dublin, Ireland
8Equal1 Labs, 4 Dublin, Ireland
9Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada
10École Polytechnique Fédérale de Lausanne, 2002 Neuchâtel, Switzerland
11Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA
12Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA 02421 USA
13Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA
14Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA
15Electronic Systems (ZEA-2), Central Institute of Engineering, Electronics and Analytics, Forschungszentrum Jülich GmbH, 52428 Jülich, Germany
16Faculty of Electrical Engineering and Information Technology, RWTH Aachen University, 52062 Aachen, Germany
17John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 USA
18School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853 USA
19State Key Laboratory of Analog and Mixed-Signal VLSI, Faculty of Science and Technology, Department of Electrical and Computer Engineering, Institute of Microelectronics, University of Macau, Macau 999078, China
20IonQ, Inc., College Park, MD 20740 USA
21University of California, Berkeley, Berkeley, CA 92720 USA

Corresponding authors: Joseph C. Bardin; Shai Bonen; Edoardo Charbon; Carsten Degenhardt; Ruonan Han; Donhee Han; Gæl Pillonnet; Robert Bogdan Staszewski; Jules Stuart (e-mail: jbardin@engin.umass.edu; shai.bonen@mail.utoronto.ca; edoardo.charbon@epfl.ch; c.degenhardt@fz-juelich.de; ruonan@mit.edu; donhee@seas.harvard.edu; gael.pillonnet@cea.fr; robert.staszewski@ucd.ie; jules.stuart@nist.gov).

The work at Electronic Systems (ZEA-2), Forschungszentrum Jülich, was supported by the Helmholtz Association’s Initiative and Networking Fund “Scalable Solid State Quantum Computing.” The work at the University of Toronto was supported in part by the National Sciences and Engineering Research Council of Canada and in part by the EU Horizon 2020 FET OPEN project “IQubits” under Grant 829005. The work at Harvard University was supported in part by the Advanced Research Projects Agency—Energy under Grant DE-AR0001063 and in part by the Center for Integrated Quantum Materials, National Science Foundation (NSF) under Grant DMR-1231319. The work at the Lincoln Laboratory at the Massachusetts Institute of Technology was supported by the Assistant Secretary of Defense for Research and Engineering Air Force under Grant FAFST72-05-C-0002. The work at University College Dublin was supported in part by the Science Foundation Ireland under Grant 14/RP/12921 and in part by the Enterprise Ireland through Disruptive Technologies Innovation Fund under Grant DT 2019-0090C. The work at the Massachusetts Institute of Technology was supported in part by the NSF through the Interdisciplinary Science and Engineering Transformational Advances in Quantum Systems under Grant 1839159, in part by the Army Research Office Multidisciplinary University Research Initiative on “Imaging and Control of Biological Transduction Using NV-Diamond,” and in part by the Gordon & Betty Moore Foundation. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.
ABSTRACT Over the past decade, significant progress in quantum technologies has been made, and hence, engineering of these systems has become an important research area. Many researchers have become interested in studying ways in which classical integrated circuits can be used to complement quantum mechanical systems, enabling more compact, performant, and/or extensible systems than would be otherwise feasible. In this article—written by a consortium of early contributors to the field—we provide a review of some of the early integrated circuits for the quantum information sciences. Complementary metal–oxide semiconductor (CMOS) and bipolar CMOS (BiCMOS) integrated circuits for nuclear magnetic resonance, nitrogen-vacancy-based magnetometry, trapped-ion-based quantum computing, superconductor-based quantum computing, and quantum-dot-based quantum computing are described. In each case, the basic technological requirements are presented before describing proof-of-concept integrated circuits. We conclude by summarizing some of the many open research areas in the quantum information sciences for CMOS designers.

INDEX TERMS CMOS integrated circuits, quantum computing, quantum sensing.

I. INTRODUCTION

The ability to coherently control and faithfully measure the state of quantum mechanical systems is a central requirement in quantum sensing, quantum communication, and quantum computing. For many of today’s quantum platforms, these functions are at least partially, if not fully, carried out electrically, and a large component of the implementation of these quantum-based systems lies in the design of electrical control and measurement techniques. Early proof-of-principle systems have generally employed laboratory-grade hardware for control and measurement. For instance, today’s state-of-the-art quantum computers—which contain on the order of 100 qubits—employ rack-mount electronics implemented using commercial-off-the-shelf (COTS) components for generation and digitization of the control and readout signals, respectively [1], [2]. This approach has been justified as the field of quantum computing is still in the proof-of-concept phase, where architectures are rapidly evolving and the basic principles required for practical fault-tolerant quantum computing are still being verified. For instance, the first demonstration of postclassical computation was just reported [3] in 2019, and realizing postclassical performance is still an active research area being pursued by many groups [4], [5]. While several experiments with scalable quantum error correction (QEC) codes have been recently reported [6], [7], [8], [9], QEC codes that sufficiently suppress errors with respect to code distance, as predicted by theory [10], are necessary for fault-tolerant quantum computing; today, state-of-the-art QEC codes are just able to reach the break-even point where errors do not grow with increasing code distance [11], and the realization of fault-tolerant quantum computers will require either improved physical qubit error rates or new QEC codes that can tolerate higher error rates. Even so, scaling the technologies required for quantum control and measurement to the degree required for fault-tolerant quantum computing will necessitate the miniaturization and optimization of these electrical interfaces [12], [13].

When considering implementation of the electrical control and measurement systems required for quantum information systems, silicon complementary metal–oxide semiconductor (CMOS) and bipolar CMOS (BiCMOS) technology platforms are a natural candidate. These technologies benefit from decades of intense development driven by the economic demand to improve computer and communication systems, and today’s CMOS and BiCMOS technology platforms support both dense and reliable digital circuitry as well as transistors with speeds permitting analog operation well above 100 GHz [14], [15], [16], [17], [18]. As we will see, both of these features are critical when considering the signal generation and sensing requirements of quantum computers and, as such, there is a growing interest in studying the use of CMOS and BiCMOS technologies in quantum-related applications.

This article provides a review of the current state-of-the-art in silicon CMOS and BiCMOS integrated circuits for application in the quantum information sciences (QIS). It was written collaboratively between a consortium of authors with early contributions in this field. However, this is a dynamic field, and this article is only meant to provide a snapshot of this nascent field; even while preparing the manuscript, a new batch of results has appeared in the literature [19], [20], [21], [22], [23], [24], [25], [26], [27]. As such, this article is not intended to be a comprehensive summary, but rather a snapshot of this nascent field.

The article is broken up into two main sections encompassing the key themes currently predominant in quantum information systems. First, work related to quantum sensing is described, including material analysis and magnetometry systems based on nuclear magnetic resonance (NMR) and electron spin resonance (ESR). Next, research related to quantum computing is described. Circuits for use with trapped-ion qubit, superconducting qubit, and spin qubit systems are described. While detailed treatments of the individual quantum technologies are beyond the scope of this
**Table 1** List of Notation and Associated Definitions

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0, 1</td>
<td>Qubit quantum states</td>
</tr>
<tr>
<td>2</td>
<td>Qubit second excited quantum state</td>
</tr>
<tr>
<td>(\omega_0)</td>
<td>0-1 transition angular frequency</td>
</tr>
<tr>
<td>(\omega_1)</td>
<td>1-2 transition angular frequency</td>
</tr>
<tr>
<td>(\eta = \omega_1 - \omega_0)</td>
<td>Qubit anharmonicity</td>
</tr>
<tr>
<td>(T_1)</td>
<td>Relaxation time from 0 to 1</td>
</tr>
<tr>
<td>(T_2)</td>
<td>0-1 dephasing time</td>
</tr>
<tr>
<td>(R)</td>
<td>Rabi rate</td>
</tr>
<tr>
<td>(B_0)</td>
<td>DC Magnetic field</td>
</tr>
<tr>
<td>(\gamma)</td>
<td>Gyromagnetic ratio</td>
</tr>
<tr>
<td>(k)</td>
<td>Boltzmann constant</td>
</tr>
<tr>
<td>(\hbar)</td>
<td>Reduced Planck’s constant</td>
</tr>
<tr>
<td>(\chi)</td>
<td>State-dependent read-out resonator frequency shift</td>
</tr>
<tr>
<td>(e)</td>
<td>Elementary charge</td>
</tr>
</tbody>
</table>

The main focus of this article is to describe the integrated circuits that have been developed for each of these areas, as well as the related research that has been carried out to enable the implementation of these devices. To assist the reader throughout this article, important definitions and notation are given in Table 1.

**II. CMOS MICROSYSTEMS FOR COHERENT QUANTUM CONTROL IN NMR AND ESR**

Imagine a quantum system with energy eigenstates \(|0\rangle\) and \(|1\rangle\) with a splitting energy \(\hbar \omega_{01}\), subjected to a time-varying field—electric or magnetic, depending on the nature of the system—with a frequency close to \(\omega_{01}\). The state of the quantum system will evolve as \(c_0(t)|0\rangle + c_1(t)|1\rangle\) with complex coefficients \(c_0(t)\) and \(c_1(t)\) being determined by the Schrödinger equation. The time course of \(c_0(t)\) and \(c_1(t)\) is such that up and down transitions between \(|0\rangle\) (\(c_0 = 1, c_1 = 0\)) and \(|1\rangle\) (\(c_0 = 0, c_1 = 1\)) indefinitely repeat as far as the time-varying field is sustained. This is the Rabi oscillation, whose frequency \(\omega_R / 2\pi\) is proportional to the time-varying field’s amplitude. In reality, due to dephasing effects, the evolution will eventually deviate from the ideal prediction, by losing the phase information of \(c_0(t)\) and \(c_1(t)\), with the characteristic dephasing time denoted as \(T_2\). If \(\omega_R T_2 \gg 1\), many Rabi oscillations will manifest before dephasing. This is the coherent regime. If \(\omega_R T_2 \ll 1\), the state will lose the phase information of \(c_0(t)\) and \(c_1(t)\) even before completing a small fraction of one Rabi cycle. This is the noncoherent regime, where Rabi oscillation cannot be observed.

Noncoherent transitions are commonplace (as one of numerous examples, virtually, all laser transitions are in this regime). Obtaining coherent transitions with \(\omega_R T_2 \gg 1\) tends to demand more efforts to increase \(\omega_R\) or \(T_2\). Masers and atomic clocks in molecular beam arrangements are an example of coherent-regime machines, where \(T_2\) is elongated. For another example, an optical material can manifest Rabi dynamics, such as self-induced transparency, but only when driven by an ultraintense light to increase \(\omega_R\). In contrast, liquid-state NMR readily occurs in the coherent regime due to its naturally long \(T_2\). Being in the coherent regime, before dephasing, nuclear spins can be manipulated into any desired quantum state \((c_0, c_1)\) by applying a sequence of pulses of time-varying magnetic fields. Exploiting such coherent quantum-state control, NMR experiments can determine molecular structures at atomic resolution, which has revolutionized chemistry and biology. Finally, while ESR is harder to observe in the coherent regime, in particular, in solid states, ESR in diamond nitrogen vacancy (NV) centers has been amply demonstrated in the coherent regime in recent years, offering another quantum-mechanically coherent system.

Here, we will review recent advances in using CMOS radio frequency (RF) transceivers to coherently manipulate and read the quantum states of nuclear spins in NMR, and CMOS microwave transmitters for coherent ESR excitation in the NV center.

**A. CMOS-BASED NMR (HARVARD UNIVERSITY, UNIVERSITY OF STUTTGART, AND UNIVERSITY OF MACAU)**

In NMR, the Zeeman-split states of a nuclear spin 1/2 (e.g., from \(^1\)H, \(^{13}\)C, \(^{19}\)F, and \(^{31}\)P) created in a static magnetic field \(B_0\) serve as the two energy eigenstates, \(|0\rangle\) (spin up) and \(|1\rangle\) (spin down), with an energy splitting of \(\hbar \omega_{01}\), where \(\omega_{01} = \gamma B_0\) (\(\gamma\): gyromagnetic ratio of the nuclear spin used) falls into the RF region (for example, for the \(^1\)H proton spin, for \(B_0 = 1\) T, \(\omega_{01} / 2\pi \approx 42\) MHz). In this case, the time-varying field at a frequency \(\omega_0\) to cause state evolution is a magnetic one.

Liquid-state NMR occurs in the deep coherent regime due to its exceptionally long \(T_2\): e.g., for \(^1\)H proton spins in water at room temperature, \(T_2 > 1\) s. Thus, when an RF magnetic field is continuously applied, a large number of Rabi oscillation cycles manifest before dephasing. Or by applying RF magnetic fields in a particular sequence of pulses, the nuclear spin can be manipulated into any desired quantum state \((c_0, c_1)\). A broad palette of RF pulse sequences for such coherent quantum state control has been developed from the decades of NMR research. In fact, the pulse sequence technique is generally applicable to any coherent-regime dynamics and, thus, serves as an indispensable tool in experimental quantum information processing. In this sense, NMR has benefited the development of quantum computing. NMR itself can be used for quantum computing [28], but as it is typically performed at room temperature on a collection of nuclear spins and the two-level splitting is far smaller than the room temperature thermal energy, it involves mixed rather than pure quantum states. NMR quantum computing has, thus, proven increasingly difficult with a growing number of qubits.

In fact, the main usage of NMR is the study of structures of molecules containing NMR-active nuclei (e.g., organic molecules). The coherent quantum state control using various RF pulse sequences enables the determination of the molecular structures at atomic resolution. As such, NMR has transformed organic chemistry, medicinal chemistry, and...
structural biology. In addition, the measurement of the long dephasing time $T_2$ itself offers a window into gross material compositions (e.g., more fat or less fat) that affect the degree of dephasing. This line of measurements opened up new avenues for medical imaging (magnetic resonance imaging—MRI) and geological interrogation, in particular, the study of subsurface fluids for oil exploration.

An NMR instrument consists of a magnet to produce $B_0$ for the Zeeman splitting, a sample-surrounding coil to apply the RF magnetic field and to pick up the spin motions, and an RF transceiver to drive the sample coil for the RF magnetic field generation and to process the signal (spin motions) picked up by the coil. Traditional NMR instrumentation is large, heavy, and expensive due to the use of superconducting magnets and discrete RF electronics. The NMR work is, thus, largely confined within dedicated facilities. To make the benefit of NMR more broadly available, a wealth of efforts have recently been dedicated to miniaturization by replacing the superconducting magnet with a permanent magnet and by integrating the RF transceiver on a CMOS chip [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39]. While the permanent magnet has limited $B_0$ strength and homogeneity and, thus, cannot be used to analyze large molecules such as proteins, many NMR applications are aimed at the analysis and identification of small molecules, for which a permanent magnet can be sufficient. Also, since the permanent magnet is not a considerable size bottleneck, integrating the RF transceiver on a chip is justified. The resulting small system can make NMR more accessible and may enable new on-site applications for chemical reaction monitoring, molecular fingerprinting, diagnostic molecular sensing, quality control, and biomedical imaging of small organisms and tissues.

Fig. 1 shows two small NMR system examples, which were developed at Harvard University and are based on CMOS transceivers and permanent magnets. The system in Fig. 1 (top), by Sun et al. [29], combines a permanent magnet ($B_0 \sim 0.5$ T, < 0.1 kg) with a CMOS RF transceiver ($1.4 \times 1.4$ mm$^2$). Due to a significant $B_0$ inhomogeneity in the permanent magnet, this system cannot perform spectroscopy to resolve chemical shifts and $J$-coupling, but it can measure $T_2$, which was used for biomolecular sensing. The one in Fig. 1, bottom, by Ha et al. [30], uses a 7.3-kg Halbach magnet ($B_0 \sim 0.5$ T) and a CMOS RF transceiver ($2 \times 2$ mm$^2$). It is capable of not only $T_2$ measurements and MRI [31] but also 1-D and 2-D spectroscopy such as $J$-resolved spectroscopy, correlation spectroscopy, and heteronuclear single/multiple-quantum coherence spectroscopy via various RF pulse sequences for quantum control [30], while many orders of magnitude smaller than traditional NMR spectroscopy systems.

Due to the generally low $B_0$ values of permanent magnets (<1.5 T), the NMR signal is very weak in these small systems. This limitation has been overcome by making the CMOS RF receivers sensitive enough. But, there is another avenue to further improve the sensitivity, which can also be implemented using CMOS chips. Concretely, the NMR signal can be enhanced by transferring the much larger magnetization of the electron spins from free radicals dissolved in the sample to the nuclear spins. In this dynamic nuclear polarization based on the Overhauser effect [40], one can fully saturate the electron spin states, in which case the NMR signal is boosted by two orders of magnitude. But this full saturation scheme comes at the cost of a high-power (~kW or more) driving at the Zeeman-split frequency of the electron spins in the microwave regime. Alternatively, one can use a power small enough to be handled by a CMOS microwave transmitter to drive the electron spins without aiming at full saturation: this still boosts the NMR signal usefully, as recently shown using a CMOS chip that integrates the microwave electron spin driver and NMR RF transceiver [38].

While we have so far discussed the recent and future CMOS-based NMR systems from the viewpoint of instrument miniaturization, we may also think of the CMOS NMR development from the following viewpoint: while CMOS technology holds its most significant success in computing, a wealth of efforts has recently been dedicated to interfacing CMOS chips with biological and chemical systems for applications in fundamental and applied biology, e.g., CMOS ion-sensitive field-effect transistor arrays for DNA sequencing and detection [41], [42], CMOS electrochemical cell arrays for pH control and DNA synthesis [43], and CMOS electrode arrays for recording neurons and cardiomyocytes [44], [45], [46], [47], [48]. CMOS-based NMR represents an addition to this line of effort that interfaces CMOS electronics with material systems for life and material science applications.

### B. HYBRID CMOS-DIAMOND SPIN CONTROL AND MAGNETOMETRY

Solid-state color centers are attracting broad interest for quantum sensing, due to their long spin coherence time and efficient optical interface for spin polarization and readout. For instance, NV center in diamond has emerged as a leading room temperature platform for sensing and imaging of temperature [49], electric fields [50], and magnetic fields [51], [52], [53]. For magnetometry, picotesla magnetic...
field sensitivity under ambient conditions has been demonstrated [54], [55], [56]. However, conventional approaches for NV sensing involve bulky and discrete off-the-shelf instruments. These instruments are required for the manipulation and readout of the spin states of the NV centers. The NV-based sensors consist of a number of components [52]: 1) microwave signal generator and delivery structure to control the NV spin state; 2) optical filter to reject the green laser pump and a photodetection subsystem for NV spin-dependent red fluorescence measurement; and 3) green pump laser. The current discrete realization of the above-mentioned system limits practical applications and is difficult to scale up. However, the room temperature operation and solid-state nature of NV centers lower the barriers for miniaturization through on-chip hybrid integration.

1) RESEARCH AT MIT
To address this challenge, a team at the Massachusetts Institute of Technology (MIT) for the first time demonstrated custom chip-scale CMOS platforms that tightly integrate electronics and NV centers in diamond for magnetometry at room temperature [57], [58], [59], [60]. In [57] and [58], the basic concept of chip-scale miniaturization of NV-center quantum sensors is described. Fig. 2 shows a custom CMOS architecture consisting of the required components to perform optically detected magnetic resonance (ODMR) in a hybrid chip-NV integration. Within a 200×200 μm² footprint, a stack of a microwave inductor, a photonic filter, a photodiode is realized. The chip is fabricated using TSMC 65-nm CMOS technology. An on-chip signal generator, a current driver, and a loop inductor deliver a vertical ac magnetic field to excite the |m_f = ±1⟩ ground levels of the NV centers. The diamond layer, attached to the top of the chip, is illuminated with a green light, and the red fluorescence is detected by a p+/n-well/p-sub photodiode placed under the loop inductor.

It is noteworthy that the majority of the green light is not absorbed by the diamond, but transmits into the chip, and potentially generates a strong background photocurrent on top of the red-light-induced signal if not filtered away. This would increase the shot noise of the photodiode, limiting the magnetic sensitivity of the sensor. To reduce such noise, a plasmonic nanophotonic filter, using a grating of the CMOS interconnect metal (M8), is implemented above the photodiode (see Fig. 2). This filter is based on the concept of wavelength-dependent plasmonic loss [61]. The measured suppression of green light is ∼10 dB. Using this hybrid CMOS-NV-center integration platform, the ODMR spectrum of a film of bulk diamond (with uniform and well-defined lattice structure) was attached to the same CMOS chip and demonstrated ODMR with vector-sensing capabilities, with 32 μT/Hz^{1/2} sensitivity, as presented in [58].

In [59] and [60], the above hybrid CMOS-NV architecture is extended to a scalable design with enhanced sensitivity. The design shown in Fig. 3 can interact with NVs across a large diamond area, hence achieving a higher SNR in ODMR. A chip with an integrated 2.87-GHz phase-locked loop (PLL) was implemented using TSMC 65-nm process. To generate a homogeneous microwave field, an array of current-driven linear wires are implemented using M8 of the chip. By controlling the current flowing in each conductor using active current sources, we can generate >95% field homogeneity over ∼50% of the area of the array compared to 25% in [57]. The current-driven wire array implemented on M8 with extra two layers of metal gratings (M3 and M6) forms an enhanced photonic filter. This filter utilizes wavelength-dependent plasmonic loss and diffraction-based Talbot effects. The extra two grating layers (M3 and M6) are placed close to the maxima (minima) of the green (red) light diffraction pattern of the grating on M8. This results in green rejection of 25 dB. The demonstrated vector magnetic field sensitivity of 245 nT/Hz^{1/2} is 130× better than the previous prototype in [57] and [58], which is mainly due to the scalable microwave launcher and the multilayer nanophotonic filter. These structures are co-designed with the on-chip electronics, thanks to the multifunctional CMOS technology. The scalable architecture demonstrated in [59] and [60].
provides a clear pathway to further push the sensitivity of the proposed sensor to sub-nT/Hz$^{1/2}$ (see Fig. 4). Homogeneous microwave field generation over large space also promises a significant sensitivity enhancement via spin coherent control [62] (see Fig. 4). Ibrahim et al. [60] demonstrated Rabi oscillations on the whole NV ensemble using the magnetic field generated on-chip and off-chip optical detection. The results shown in Fig. 5 confirm that all the NV centers in the ensemble exhibit the same Rabi frequency ($\sim$1.2 MHz) determined by the local microwave field and that the NVs are flipped at the same rate and their spin-dependent fluorescence signals are added coherently. The ability to manipulate ensembles of NV centers with a scalable and compact chip-scale platform opens the door for on-chip realizations capable of performing gradient magnetometry and multiplexed analytical NMR spectroscopy, among others.

III. QUANTUM COMPUTING

Despite the exciting opportunities related to hybrid CMOS/quantum sensing systems, the vast majority of research efforts spanning the quantum and CMOS disciplines that have been reported to date focus on quantum computing.

There are many different qubit modalities that are available, each with their own advantages and disadvantages. Examples include trapped ions, cold atoms, photons, superconducting circuits, and semiconductor quantum dots (QDs). Here, we focus on the technologies for which CMOS circuits have already been reported—trapped ions, superconducting circuits, and semiconductor QDs—first describing the high-level requirements and then reviewing the circuit results. While a detailed comparison of the pros and cons of these technologies from a control-related perspective is beyond the scope of this article, we refer the interested reader to [63] for such a discussion.

A. ION-TRAP-BASED QUANTUM COMPUTING

Trapped atomic ions offer a promising platform for quantum computation, due to their natural reproducibility and to the maturity of the experimental tools used for controlling
their quantum states. In this review, we will focus on voltage generation requirements for ion experiments; for a detailed primer on trapped ions for quantum computation and additional context for the other experimental requirements, we refer the reader to [64].

Surface electrode ion traps [65], which are typically constructed using standard lithographic fabrication techniques [66], [67], [68], can enable the monolithic integration of control technology into the trap substrate. Commercial CMOS foundries have been used to fabricate surface electrode ion trap chips [69], which demonstrates the creation of ion traps in standard multiproject runs without modification. This CMOS compatibility allows trap designers to leverage decades of classical CMOS electronics development and integrate compact, fast, and robust components. Monolithic integration can potentially benefit array-based ion trap architectures that have been proposed to increase the number and connectivity of ion qubits on a chip [70], [71], [72]. To date, the integration of detectors [73], laser focusing and routing [74], and voltage generation [75] have all been separately demonstrated.

A variety of electrical controls are commonly employed in trapped ion quantum computing. In a surface electrode Paul trap, confinement of ions requires a combination of RF and quasi-static control voltages applied to trap electrodes patterned on the trap surface. For a linear trap, the RF potential provides confinement in two directions and causes ions to line up in chains along a central axis. The other control voltages are used to provide confinement along the final (axial) direction and can also be used to transport ions along the trap axis, or to separate, merge, or rotate chains of ions [76], [77], [78], [79]. Typically, the required RF amplitude (∼100 V) prohibits the direct synthesis of voltage, but the other control electrodes may be driven directly by digital-to-analog converters (DACs), which can be placed inside the vacuum apparatus [80] or even be integrated into the substrate of an ion trap [75]. Most ion species and trap geometries require control voltages in the range of ±10 V or higher, which limits the available processes for DAC fabrication. Since the speed at which ions can be moved around the trap depends on the voltage update rate, a high-bandwidth DAC may be desired in some applications. When many ion transport operations are required, dedicated control voltage systems, consisting of separate elements for voltage generation, amplification, and filtering, are typically employed; see [81] and [82] for details on the design of such systems.

Voltage noise on the DACs will translate into electric field noise sensed by the ion [83], and this can lead to heating of the ion’s motional state, which negatively affects the fidelity of quantum gates that use the shared motion of ions to transmit information between qubits [84], [85]. As a rule of thumb, a voltage noise of 1 nV/√Hz or lower is desired at the frequencies of the motional modes used for quantum gates, which are typically ∼1 MHz [86]. In the case of a device with integrated voltage sources, electric field noise from other parts of the CMOS circuitry must also be shielded from the ion, and likewise, the high-voltage RF trap potential should be decoupled from the DAC circuit. Both of these challenges can be addressed by the addition of a ground plane, between the DAC circuit and the trap electrodes, which can even be implemented as a mesh, since the RF signal has a relatively low frequency (tens of megahertz for typical ion species and geometries) and correspondingly long wavelength. Aside from the fast voltage noise that directly drives the ion’s motion, slow voltage variation can also introduce errors in the form of unwanted ion displacement and drift in the ion’s motional frequencies; for typical trap geometries and ion species, a long-term voltage stability <1 mV is preferred. In addition to electric field noise, magnetic field noise, originating from spurious current as digital electronics are switching states, could potentially decohere quantum information stored in magnetic-field-sensitive states. A ground plane fabricated out of a superconducting material will provide shielding against dc and ac magnetic fields, though this will require low-temperature operation; a ground plane made of a high-conductivity normal metal can also shield ac magnetic field fluctuations.

Trapped-ion experiments are often operated in cryogenic vacuum chambers, in which the trap temperature drops to near 4 K. At low temperatures, electric field noise originating from the trap material is reduced, and collisions with background gas molecules are strongly suppressed [87], [88], [89]. Integrated DACs should ideally be compatible with this cryogenic operation. Traditionally, commercial CMOS foundries may not supply process characterizations at cryogenic temperatures, so some calibration will be required to ensure that devices function reliably at low temperature. Some ion operations and experiments can accept small variations in voltage, as long as these can be compensated or trimmed using data collected with the ion. Applications requiring deterministic control of voltages may benefit from the addition of an in situ method for measuring DAC voltages to ensure linear behavior and calibrate any voltage offsets. Finally, when using a cryogenic system, the power dissipation of the DACs must not exceed the cooling power available from the cryostat, which is typically around 1 W at the coldest stage. Power dissipation may also be an important constraint in room temperature systems where heat sinking is difficult or in applications where total power consumption is critical.

1) RESEARCH AT MIT AND MIT LINCOLN LABORATORY

In the trap shown in Fig. 6, monolithically integrated DACs fabricated by a team at MIT and MIT Lincoln Laboratory provide the voltages on the control electrodes [75]. This trap, produced as part of a multiproject wafer through the MOSIS service in the GlobalFoundries CMHV7SF180-nm node and operated in a cryogenic (~4 K) ion trap system, enabled the precise control of the axial confinement frequency and ion position via serial control lines. The chosen process allowed the inclusion of high-voltage transistors, with which 16 integrated 12-bit DACs with a range of ±8 V were created;
FIGURE 6. Ion trap chip with integrated DACs. The small squares around the edge of the chip are wire bond pads for connecting power (four redundant groups of 4) and digital control (one group of 5). Ions are confined approximately 50 μm above the surface of the metal electrodes (appearing gold in color here, but made from aluminum). The two largest parallel electrodes in the center of the trap are the RF electrodes, which provide ponderomotive radial confinement and are controlled by an external source. The other 14 electrodes at the top and bottom of the chip plus two additional electrodes in the middle of the trap are the dc electrodes, which provide axial confinement and are controlled by the integrated voltage sources. The cutaway reveals the layout of the circuit in the internal layers. Each DAC channel provides 12 bits of resolution and takes up 130 μm × 270 μm in the CMOS layers. The different colors in the drawing represent different signal planes.

FIGURE 7. Measured noise power density (square of voltage noise spectral density) on one of the trap electrodes. The voltage noise spectrum is measured, while the trap is immersed in liquid nitrogen (77 K). The noise measurement with the EIS open is limited by the noise floor of the spectrum analyzer dashed blue (lower) line. The instrument’s input capacitance causes the measured noise to roll off when the EIS is closed, due to the pole formed at 120 kHz with the finite “on” resistance of the FETs in the EIS (R_{EIS} = 3.3 kΩ). The unfiltered noise is extrapolated from data taken at frequencies below this pole, assuming a 1/f frequency scaling dashed green (upper) line.

critically, the DACs made in this process kept their function at low temperature and retained a sufficient voltage range to permit operation with typical trap frequencies. While this demonstration focused on the quasi-static control voltages, similar trap-chip-integrated electronics may also find application in control of integrated optical modulators and biasing and readout of on-chip photodetectors, further enabling practical trapped-ion quantum information processing.

The measurement and mitigation of voltage noise was a major focus of the work presented in [75]. By adding an integrated analog switch to the DAC output, directly before the connection to the trap electrodes, a compromise between voltage noise and speed can be achieved. Here, this device is referred to as an electrode isolation switch (EIS). The variable resistance of the EIS makes a low-pass filter in conjunction with a parallel capacitor, which allows the amount of filtering to be modified in situ. Due to the relatively high voltage present at trap electrodes, the EIS must be composed of a complementary pair of high-voltage field-effect transistors. In Fig. 7, the voltage noise of an example DAC with an EIS both open and closed is reported, showing a noise power attenuation of up to six orders of magnitude after opening the EIS. Since trap electrodes behave as capacitors with very low leakage, trap voltages can be quickly manipulated with the EIS closed and then effectively disconnected from the amplifier noise by opening the EIS, while the trap voltages are maintained by the charge stored on the electrodes. In more complex designs, the EIS could also be used for switching between multiple integrated voltage sources, which has been shown to be useful in experiments with fast ion transport [90], [91]. Regardless of the operation of the EIS, traps with integrated voltage sources also have the natural benefit of reduced environmental noise picked up on long wires into the chamber.

Operational amplifiers in the design of the DAC circuit may draw current in steady-state operation, which leads to joule heating. In a room temperature experiment, substantial heat can be dissipated using active cooling or heat sinking. However, in a cryogenic application, where only limited cooling power is available, this heat may lead to increased trap temperature. A power down circuit, which quenches current sources in the amplifier, can reduce power consumption and lower the trap temperature in situ. By combining this power down with the EIS architecture, trap voltages may be maintained, even as the rest of the DAC is shut down.

Integration of voltage sources offers the potential to increase bandwidth, due to decreased parallel capacitance from wiring and chamber feedthroughs. Digital communication rates in the gigahertz range may be achieved using controlled-impedance lines and differential signaling. Ultimately, the total voltage update rate may be limited by the number of bits in each DAC and the number of DACs placed in sequential serial buses. In practice, voltage updates on trap electrodes may be limited by the low-pass filter consisting of the output impedance of the DAC and the parallel capacitance on the trap electrode. To shunt voltage pickup from the RF electrode that would otherwise couple onto the dc electrodes, a parallel capacitance of tens of picofarads on each trap electrode is desired, depending on the trap geometry [92]; thus, an output impedance in the kilohm...
range or below will be necessary for applications requiring high bandwidth. The use of an EIS or partial power down circuit also affects the timing between voltage updates, so the tradeoff between noise and speed must be considered in the DAC design.

Combining integrated voltage sources with other integrated technology can offer further benefits. Adding in classical control, like a microprocessor or RAM, can bring voltage update rates into the range of hundreds of megahertz or more, without requiring management of vacuum feedthrough impedance. A local processor could also run a feedback loop, using, for instance, the fluorescence signal from a detector to control the position of an ion, without requiring external calibration by the user. If all required control technology is integrated into the substrate beneath the trap, it may then be possible to tile many of these cells into a larger architecture of connected traps.

B. SUPERCONDUCTING-CIRCUIT-BASED QUANTUM COMPUTING

In superconducting quantum computing [93], qubits are engineered as nonlinear microwave resonators, which are implemented using one or more Josephson junctions (JJs) embedded in a superconducting passive circuit. When cooled to temperatures in the range of 10 mK, these devices display quantum mechanical behavior and can be used as qubits. A quantum processor (QP) can then be created by introducing electromagnetic coupling between these circuit-defined qubits, similar to how coupling between resonators is controlled in microwave filter circuits. Since superconducting QPs can be realized monolithically and engineered at the circuit level, they are currently one of the most popular approaches to quantum computing.

While there are a number of emerging superconducting qubit topologies, today’s state-of-the-art superconducting QPs [3], [94], [95] employ transmon qubits [96], which consist of capacitively shunted JJs [see Fig. 8(a)]. In this context, the JJ serves the function of a lossless nonlinear inductance, and the transmon can be understood as an LC resonator that is nonlinear at the single photon level, resulting in the anharmonic energy diagram of Fig. 8(b). Compared to competing qubit technologies, the transmon is only weakly anharmonic, with typical values of $\eta/2\pi = \omega_{12}/2\pi - \omega_{01}/2\pi$ and $\omega_{01}/2\pi$ in the range of –200 to –350 MHz and 4–8 GHz, respectively. The $\omega_{01}$ transition can be excited resonantly through an interface that can be engineered as a weak inductive or capacitive coupling. Thus, provided that the bandwidth of the electrical drive signal is limited to avoid exciting higher level transitions (e.g., $\omega_{12}$), the transmon can be used to approximate an ideal two-level qubit. While this places a constraint on the maximum Rabi frequency that can be achieved (or equivalently, the minimum duration that can be used for a $\pi$ pulse), with proper shaping of the complex baseband envelope, it is still feasible to carry out microwave (XY) gate operations on timescales as short as 10 ns [97], [98]. The loaded quality factor of the nonlinear resonator is directly related to the qubit’s relaxation time constant $T_1$, and significant work has gone into the optimization of superconducting qubits over the past decade, with isolated devices achieving $T_1 > 300 \mu s$ [99] and those within moderate-size QPs achieving $T_1 > 100 \mu s$ [100]. Care must be taken in coupling a microwave excitation to a transmon to prevent loading associated with the real impedance of the generator from limiting the transmon quality factor (and hence $T_1$), and coupling quality factors on the order of $50 \times 10^6$ are typical. For this level of coupling, a Rabi frequency of 50 MHz is achieved with an available signal power of about $-70 \text{ dBm}$, referenced to the qubit drive port. Compared to other qubit technologies, this represents very tight coupling [63], and the tradeoff for such low drive power is a sensitivity to noise on the drive line. To prevent the drive port from contributing noticeably to the qubit’s decoherence (e.g., during idling), the noise level on the drive line must be kept at or below the single photon level ($T_{\text{eff}} \leq \hbar \omega_{01}/k$). Thus, a signal to noise of about $130 \text{ dB/Hz}$ is required on the microwave control signal line. Additional specifications include an integrated amplitude accuracy of 0.25% and a phase accuracy of $0.2^\circ$ [101]. These specifications each correspond to contributions to the average error rates of $10^{-5}$.

In a practical QP, the JJ in the transmon is often replaced by a JJ loop, also known as a superconducting quantum interference device, which serves as a flux-tunable nonlinear inductance. In this configuration, the flux-bias line—also referred to as the $Z$ control line—is used to tune the qubit frequency, whereas the RF drive line—also referred to as the $XY$ control line—is used to drive the qubit resonantly. While incorporating frequency control into the transmon circuit does introduce a mechanism for degradation of both the relaxation
time $T_1$ and the dephasing time $T_2$,\footnote{Typical values of $T_1$ for a frequency-tunable transmon in a large processor are $\approx 20 \, \mu$s [3]. The lower numbers in comparison to fixed frequency transmons are partially related to the extra $Z$ tuning port.} it also provides several benefits. First, individual control of each qubit frequency through a static $Z$-bias current allows for qubit idling frequencies to be optimized for maximum $T_1$, which is an important feature that allows for one to avoid the detrimental effects of parasitic two-level systems, which can introduce loss at unpredictable (and time-varying) frequencies [102]. Second, dynamic control of the qubit frequency can be used for a variety of purposes such as phase gate operations, two-qubit interactions, and reset operations. While the absolute value of current depends upon the mutual coupling of the $Z$ drive line to the qubit, currents of 500 $\mu$A or less are typically required. Since the flux sensitivity of a transmon increases as the bias moves away from the flux-insensitive (e.g., zero-bias) points, the resolution required for the $Z$ currents depends upon the desired range of operating points. However, to enable operation over a wide range of biases, a DAC with 14 or more bits of resolution is typically required [2]. Noise on this line is also critical, and it is common practice to apply 20 dB or more of cryogenic attenuation to the output of a room temperature DAC used for driving the $Z$ control ports.

Fig. 9 shows the block diagram of Google’s control and measurement system for a 54-qubit QP (see [2] for more details). Room temperature baseband and RF arbitrary waveform generators are used to generate each of the $XY$ and $Z$ control signals, respectively. These signals are heavily attenuated and filtered along the thermal gradient to reduce noise levels before being multiplexed to a single $XYZ$ line per qubit at the 10-mK stage of the dilution refrigerator. Note that the filters marked “IR” are of critical importance, as they block propagation high-frequency thermal photons down the interconnects via higher order modes.

While there are many approaches to interacting transmon qubits, one technique that has been gaining in popularity in recent years is to employ tunable couplers between the qubits [103], [104], which permits realization of a QP with qubits configured in a 2-D grid, as shown in Fig. 8(c) [3]. This configuration has the advantage of being forward-compatible with the surface code. An example circuit design of a tunable coupler appears in Fig. 8(d). Here, an additional transmon circuit is used as a frequency-tunable resonance that is used to control coupling through a capacitive network. With such a device, it is possible to deterministically enable ZZ or iSWAP-like interactions through dynamic control of the qubit and coupler frequencies [105]. Example electronics used to drive the coupler lines appear in Fig. 9.

The state of a superconducting qubit is typically measured using dispersive readout [106], [107]. In this approach, each qubit is coupled to a linear resonator that is significantly detuned from the qubit frequency. The effective resonance frequency seen looking into the linear resonator then depends upon the state of the qubit, with a shift up or down in frequency of $\pm \chi$ being observable when the qubit is in the $|0\rangle$ and $|1\rangle$ state, respectively. The frequency shift $\chi$ depends upon the detuning and resonator–qubit coupling strength, but is typically in the range of a few megahertz [13]. Thus, the state of the qubit can be measured via a reflection coefficient measurement of the linear resonator. A Purcell filter is typically incorporated into the readout circuit both to further isolate the qubit from the impedance of the 50-Ω transmission line employed for readout and also to permit multiple qubits to be readout using frequency-division multiplexing. The excitation tones used for measuring the qubit states must be extremely weak to ensure a quantum nondemolition measurement; the readout resonator is typically populated with ten or fewer photons. To enable quick (100s of nanoseconds) and high-fidelity readout, a quantum-limited amplification chain is required, with a first-stage parametric amplifier at the base temperature providing about 20 dB of quantum-limited gain and a cryogenic semiconductor amplifier thermalized around 4 K providing additional gain while adding minimum noise. As shown in Fig. 9, typical readout configurations require significant attenuation as well as numerous circulators and/or isolators, which separate forward and reflected waves and prevent amplifier noise from being back-injected into the readout resonators.

Today’s state-of-the-art superconducting QPs have greater than 50 qubits and an example architecture\footnote{This is just one example architecture. At the other extreme from this configuration is a processor with fixed qubit frequencies (no $Z$ control) and fixed coupling between qubits. In such an architecture, all the gates are carried out via microwave control.} is that of Fig. 8(c). In this configuration, each qubit has one $XY$ drive line and three $Z$ control lines (one for the qubit and two for adjacent couplers). In addition, there is a readout line for every $M$ qubits (today, $M$ is limited to about six, but it may be possible to gain a higher multiplexing factor through system design). As shown in Fig. 9, the quantum controller.
for the 54-qubit Sycamore quantum computer requires a total of 54 XY control channels, 54 Z control channels, 88 coupler control channels, and nine readout channels. In general, the electrical controller for a processor of this architecture must have $N$ microwave (XY) control channels, $\approx 3N$ baseband control channels, and $N/M$ readout channels. Research into how to leverage semiconductor technology to build these control systems in a scalable approach is currently underway at several institutions. Here, we describe early results.

1) RESEARCH AT GOOGLE
Google has announced a roadmap to build an error-corrected quantum computer based on superconducting qubit technology [2] and is working on CMOS ICs in support of this goal. To date, work has focused on the exploration of cryogenic control electronics, optimized for operation while thermalized to the 3 K stage of a dilution refrigerator. Some potential advantages to cryogenic operation of the quantum controller include reduced waveform distortion since all interconnects between the controller and the QP can be superconducting and improved stability due to the tight temperature control. However, a downside of the cooling the quantum controller is that its power budget is much tighter, with compact coolers such as those used in today’s dilution refrigerators offering up to 2 W [108] of heat lift and larger systems offering 1 kW or more [109]. Thus, any cryogenic controller thermalized at $\approx 4$ K can dissipate at most of order 1 mW per qubit, so research is required to determine if high-performance quantum control and measurement systems can be implemented on this tight power budget. Preliminary work was reported in [101] and [110], where the design and characterization of a prototype cryo-CMOS XY quantum control IC was described.

A conceptual block diagram of the quantum control IC is shown in Fig. 10. Complex envelope generation is carried out using two arrays of 11 single-ended current-mode sub-DACs combined with the appropriate timing circuitry to generate symmetric pulses of current. These currents are then low-pass filtered and upconverted using a direct-conversion IQ mixer, driven from an off-chip local oscillator (LO). An additional polarity switch is employed in each channel to enable a full $360^\circ$ of phase coverage. The IQ mixer outputs are combined and bandpass filtered using a reconfigurable transformer network. By properly configuring the envelope currents, polarity settings, transformer tuning, and LO frequency, the chip can realize a wide range of symmetrically shaped pulses over the 4–8 GHz frequency range.

Each of the sub-DACs can be programmed to an accuracy of 8 bits, and an 8-bit master reference current provides control of sub-DAC full-scale range. Assuming that the sub-DAC responses are ideal, this level of resolution is beyond what is necessary to realize envelopes with sufficient accuracy to achieve a fidelity of $10^{-5}$. However, given that this chip was designed without cryogenic device models, the circuit was designed redundantly.

A block diagram of the implemented IC appears in Fig. 10(b). In addition to the DACs and direct-conversion system, the IC includes a small waveform memory capable of storing 16 different complex envelope waveforms, which are set prior to an experiment via a serial peripheral interface (SPI). This small instruction set is sufficient to program a universal single-qubit gate set, provided that virtual $Z$ gates are not required. During an experimental run, a waveform is selected via a 4-bit instruction interface and triggered via a trigger line. The circuit also features receivers on both the clock and LO lines, which were designed to limit the RF input power required on these lines to well below 100 $\mu$W.

The IC was fabricated in a 28-nm bulk CMOS process. A die micrograph appears in Fig. 10(c). As shown in Fig. 10(d), the chip was packaged within a module to facilitate testing. The module was then mounted on the 3 K stage of a dilution refrigerator and interfaced to a transmon qubit for testing. While full details of the test setup were provided in [101] and [110], it is important to note that a coupler was introduced at that output of the chip to allow for an amplitude-
Quantum control experiments carried out using Google’s cryogenic IC. (a) Experimental protocol for a Rabi experiment. (b) Measured Rabi oscillations as a function of pulse amplitude. (c) Experimental protocol for three-pulse experiment. (d) Measured ground-state probability as a function of pulse parameters.

A second experiment, whose protocol is described in Fig. 11(c), was used to demonstrate coherent control of the qubit using the cryo-CMOS quantum controller. For this experiment, the qubit was first reset to the ground state, being excited by a series of three pulses and being measured. The first and last pulses had fixed phase but swept amplitude, whereas the middle pulse had a nominal amplitude corresponding to a π rotation and the phase was swept. For each configuration, 5000 statistics were gathered. As described in [101] and [111], the measured ground state probabilities as a function of the swept variables appearing in Fig. 11 demonstrate the expected behavior. Several additional experiments were carried out, as described in [101]. Of note, it was shown that the quantum controller did not degrade the qubit relaxation time (indicating that the noise floor was acceptable), and it did not drive significant $|2\rangle$ state population so long as the pulse duration was greater than 15 ns. The worst-case power consumption of the IC was measured and found to be below 2 mW, including the LO, clock, and dc power (both analog and digital) delivered to the chip. While the IC described above served an important initial proof of concept, demonstrating that simple low-power direct-conversion quantum controllers are a viable solution for $XY$ control of transmon qubits, it is far from a complete solution enabling full control of large-scale QPs. For instance, the circuit only provides a single $XY$ channel and, therefore, cannot perform control of the $Z$ lines and coupler biases. In addition, more advanced pulse shaping (e.g., DRAG [97], [98]) is required to fully optimize the $XY$ gate performance, and features such as LO leakage nulling will be required if these systems are to be deployed at large scale. Developing integrated circuits that are able to overcome these and other limitations are topics that are currently under investigation at Google.

C. SEMICONDUCTOR-QD-BASED QUANTUM COMPUTING

At cryogenic temperatures, single electrons or holes can be trapped at the semiconductor–oxide interface of MOS structures biased at the onset of inversion. The quantum degrees of freedom of these charges can form a basis for quantum computation. Here, we focus on CMOS-compatible approaches, as they provide a plausible means to realize a closely packed quantum computer and, hence, are the motivation behind many CMOS circuit research efforts, some of which are described in this article. To motivate the circuit architectures being explored, we first briefly discuss the requirements for control and measurement of QD-based qubits. Other platforms sharing many similarities with CMOS semiconductor qubits, such as III–V-type semiconductor [111] or donor-implanted [112], [113] qubit types, are not discussed here.

QDs can be realized in CMOS technologies by accumulating one or a few charges in the silicon substrate beneath high-$k$ gates [see Fig. 12(a)]. The charge behavior in these devices can be understood classically, via the lumped element picture shown in Fig. 12(b). The QD is represented as a metallic island of area $A$ with a total capacitance to ground $C$ and a potential voltage $V$ referred to ground [114]. The dot is capacitively coupled to the gate with a capacitor $C_{g,\text{dot}}$, permitting control of the dot potential $V = \alpha V_g$, where $V_g$ is the gate voltage and $\alpha = (1 + C/C_{g,\text{dot}})^{-1}$ is the gate-dot coupling. For typical nanometric oxides of modern CMOS nodes, $\alpha$ is usually between 0.3 and 0.6 [115].

Bringing a single charge onto the QD requires changing the dot potential by $\pm e/C$, which corresponds to a gate voltage change of $\pm e/\alpha C$, where $e$ is the charge of a single electron and the sign depends on the charge type (a plus
FIGURE 12. (a) Semiconductor QD array. A doped region, named reservoir, provides charges to the single QD. Charges are localized at the semiconductor–oxide interface by adjusting the electrostatic landscape via the application of voltages on gates. Plunger gates are commonly used to adjust dot–dot coupling. (b) Lumped element representation of a single QD. The application of dc gate voltage $V_g$ (top electrode) increases or decreases the island potential resulting in the tunneling in or out of single electrons or holes from nearby charge reservoir (usually highly doped regions). To maintain a charge in the dot, the thermal energy $kT_{dot}$ has to remain well below the charging energy $e^2/C$ related to the QD size, as shown in (c).

Movement of charges forms the basis of computation of charge qubits, which were used in some of the earliest demonstrations of coherent control in QD structures [121], [122], [123], [124], [125]. A charge qubit is composed of a single charge that is associated with two well-defined positions in a double QD. The naturally strong coupling of charges to the environment enables fast manipulation on the order of 10–100 ps for a typical relaxation time $T_1$ of a fraction of a nanosecond. Fast square voltage detuning pulses between the dot potentials bring the left- and right-position energy levels in resonance, activating the fast coherent exchange. Recent articles envision a universal quantum gate set with baseband-only control by adjusting the left- and right-state interaction with square pulses of controlled length [117]. Alternatively, using a single charge equally shared between two QDs decreases the charge qubit sensitivity to charge noise and increases the relaxation time above the nanosecond level [126]. Manipulation of such a qubit is identical to microwave-driven qubits and requires drive that is resonant with the few-gigahertz qubit energy level spacing. The state of a charge qubit can be measured using dispersive readout of a charge sensor, in a similar fashion as for superconducting qubits (see Section III-B). However, whereas phase shifts of ±90° impressed on carriers in the 5-GHz range are typical for dispersive readout when applied to superconducting qubit technology, for charge qubits, readout is typically carried out in the range of 300 MHz to a few gigahertz, and typical phase shifts range from a few millidegrees to a fraction of a degree [127], [128]. Although charge qubits have been less popular than spin qubits due to the latter’s lower sensitivity to environmental noise, they have been gaining attention recently, thanks to the development of modern commercial nanometer silicon-on-insulator (SOI) CMOS technologies, whose base technologies support the formation and fast control of charge qubits [120], [129], [130], [131], [132].

A single spin state is a logical and widely studied candidate for use in a quantum computer [133]. An energy splitting can be introduced between the spin-up and spin-down state of a single charge through the application of an external magnetic field. Spin rotation can be induced via ESR with an ac magnetic field that is on-resonance with this energy splitting [134]. Spins are easily driven with an oscillating magnetic field on top of the dc component and remain insensitive to charge noise, allowing fast gates and long coherence times. Because spin qubits are biased and driven using magnetic fields, they are sensitive to any stray magnetic
fields. Removing all magnetic contribution from the qubit environment by purifying the silicon substrate to keep the spin-free 28Si isotope, coherence times exceeding a second have been demonstrated [135].

Spin qubits can also be manipulated via electric dipole spin resonance (EDSR) using an oscillating electric field by converting the charge movement to an effective oscillating magnetic field on the spins. This can be accomplished in multiple ways, for example, by leveraging the naturally high spin–orbit coupling of holes in a silicon crystal [136] or with electrons in asymmetric wells [137]. An alternative approach is to generate a nonuniform magnetic field in the vicinity of the spin qubit, which can be accomplished using micro-magnets [138], for instance. Despite increasing the coupling of the qubit to charge noise via EDSR, this solution allows using a single gate to generate local electrical fields for better single-qubit addressability, reducing crosstalk issues associated with ESR-based control approaches. Single-qubit gates with fidelities above 99% have been achieved in silicon, with typical gate times from 100 ns to a few microseconds [139].

Two-qubit gates are realized for spin qubits by turning on and off the electromagnetic coupling between two neighboring spins via a square voltage pulse on a plunger gate. Such interactions can be used to implement controlled-phase gates [140]. Alternatively, by resonantly driving the $|10\rangle$ → $|11\rangle$ transition, a controlled-NOT can be realized [141]. The readout of the spin state is done after performing a spin-to-charge conversion, converting spin states to different charge movements thanks to Pauli spin blockade. Spin-dependent charge movement to a reservoir or to a filled dot is then sensed with a charge sensor (ideally to charge qubits) [142] or by directly sensing the gate capacitance of the dot [127].

Silicon spin qubits have already demonstrated key metrics for quantum computing with >99.9% fidelity single-qubit gates [143], 98% fidelity two-qubit gates [134], and relaxation times $\tau_s$ of 1 s [142]. The demonstrated ability to coherently shuttle qubits between neighboring dots in 1-D [116] and 2-D [118] array increases the connectivity between qubits. Strong efforts are being pursued to bring all these metrics into one device, with promising devices made out of standard industrial fabrication flow in, e.g., fully depleted silicon on insulator (FDSOI) [136] or FinFET [144], and the realization of early-scale QPs from 2 to 6 qubits [145], [146], [147].

The nanoscale dimension of the qubits is common to all of the semiconductor-QD-based platforms discussed above. While the ability to densely pack qubits is an advantage of the technology, the nanoscale dimensions also create a challenge in scaling, as fanout might become a limiting factor. Several large-scale architectures based on arrays of thousands of QDs have been proposed and research is currently underway to prepare for the availability of large-scale QPs, which could exploit these architectures [130], [148], [149], [150], [151]. The holy grail of QD-based quantum computing is to integrate the control and measurement electronics on the same die or 3-D assembled with the QP to solve the interconnection challenge when increasing the number of qubits. As presented in this article, many of the groups active in CMOS circuit design related to semiconductor quantum computing have this ultimate goal in mind. For this to work, the qubits must operate at temperatures high enough that efficient heat removal is feasible, as the control and measurement circuitry will dissipate significant power and cryogenic cooling fundamentally becomes less and less efficient as the physical temperature is reduced. While “hot” operation of single-dot silicon spin qubits at temperatures as high as 4.2 K has recently been reported [127], [152], [153]—albeit at degraded fidelities with respect to deep cryogenic operation—thermalization is still a major challenge, and intense research efforts are currently focused not only on the design of cryogenic control circuitry but also on the optimization of the overall system architecture, balancing the classical and quantum design.

1) RESEARCH AT CEA-LETI

CEA-Leti aims to accelerate the development of industrial quantum computing by developing key skills and knowledge for the advent of quantum computers. CEA-Leti fabricates spin qubits from a 300-mm wafer industrial FDSOI CMOS platform. Since the demonstration of the first spin qubit built using an industrial CMOS platform in 2016 [136], the CEA-Leti team has expanded their scope, which now also covers key fields such as fabrication, modeling, testing, integration, and circuit design. Fig. 13 depicts the envisioned cryogenic electrical interface chain for silicon qubits [154] and relates what the CEA-Leti group has already published [155], [156], [157], [158], [159], [160], [161], [162], [163], [164].

The CEA-Leti team leverages 28-nm FDSOI technology for their qubit research. In comparison to bulk-CMOS technology, FDSOI has several key advantages: 1) the increase in transistor threshold voltage $V_{TH}$ that typically occurs with cryogenic cooling can be compensated using the back-gate terminal [155]; 2) kink effects and hysteresis at cryogenic temperatures do not occur [156], [157]; 3) the leakage currents are kept low; and 4) since the body is undoped, there is a natural immunity to random dopant fluctuations [155].

The absence of accurate modeling of the cryogenic behavior of transistors hinders the development of large integrated circuits achieving the stringent requirement for quantum computing. As a starting point in understanding the cryogenic performance of their technology, the CEA-Leti IC team designed an integrated on-chip matrix of 1000 individually addressable transistors to massively characterize their CMOS 28-nm FDSOI technology from room temperature to qubit temperature (i.e., from 300 to 0.1 K) [158], [159]. The proposed approach enabled the early collection of electrical current–voltage characteristics of a wide representative set of devices without the need for a cryogenic probe station, which are usually limited to operation above a few kelvin. The packaged IC was installed at the coldest stage of a dilution
fridge to extract dc characteristics, such as transconductance, subthreshold slope, threshold voltage, and mismatch [158], [159]. The extracted data fueled the analysis of the cryogenic behavior for better modeling such as in [163], where residual traps in the silicon bandgap were found to cause the subthreshold slope to saturate at 4.2 K. All these contributions converged to the development of a first design-kit for IC designers based on the Leti-UTSOI model [164].

To better understand the subtleties associated with the implementation of cryogenic ICs meeting the performance and power requirements of quantum computing, the team of researchers is also exploring the implementation of critical analog and digital blocks at cryogenic temperatures [155], [160], [161], [162]. First, the team showed that, by adjusting the threshold voltage via the back gate, that ring oscillators could be operated down to 0.325 V, drastically improving the energy–delay product (EDP) [155]. Another example of a proof-of-concept circuit designed by the team is a differential 8-bit current-steering DAC [162]. The DAC is aimed to bias the gate network of a matrix of qubits by generating signals with a precision of about 26 μV [162].

The team also demonstrated a transimpedance amplifier (TIA) only consuming 1 μW and used it to measure the current flowing through a QD [160], [161]. Importantly, by integrating the TIA used for measurement on the same die as the quantum device, they were able to speed up the dot characterization by removing the loading of long cables. In this case, keeping the TIA power low was essential in allowing the measurement of a millimeter-away QD structure while preventing heating that would otherwise perturb the quantum state, even for physical temperatures as low as 10 mK [161].

The IC presented in [160] demonstrated a hybrid circuit combining silicon QD devices with classical digital and analog circuits. The circuit combines a silicon-based electron double QD with biasing thanks to on-chip bias tees and integrated generation of gigahertz-range millivolt-level signals in order to expose quantum dynamic phenomena such as charge pumping, along with in situ nanoampere current measurement capabilities. Sustainable power consumption (<300 μW) has been reported to maintain the IC operations down to 110 mK.

The team also reported a cryogenic circuit incorporating a CMOS-based active inductor (AI) to enable the compact readout of semiconductor spin qubits [25]. As shown in Fig. 14, the impedance measurements achieves a 10-aF sensitivity. As opposed to commonly used schemes based on dispersive RF reflectometry, which require millimeter-scale passive inductors, it allows for a markedly reduced footprint (50 μm × 60 μm), facilitating multichannel readout integration with reduced cross coupling. The frequency and quality factor of the formed resonator (qubit+AI) can be digitally adjusted for selecting the best tradeoff between readout sensitivity and duration.

These early results [25], [160], [161], [162] are an important step toward monolithic integration of the quantum core along with qubit manipulation, characterization, and readout electronics. To overcome the limitations of classical readout topologies, the CEA-Leti team explores disruptive approaches to explore the limit of seamless readout with limited cooling power such as the on-chip impedance measurement of a CMOS active resonator capacitively coupled to a QD [25].

2) RESEARCH AT THE UNIVERSITY OF TORONTO

As noted earlier, the operating temperature of a QP is an important parameter since the dissipated power a cryostat can remove increases exponentially with temperature. Given this constraint, researchers at the University of Toronto are studying the operation of high-temperature (2–12 K) QPs based on heterostructure SiGe hole-spin qubits cointegrated with control and readout electronics in a commercial 22-nm FDSOI CMOS technology. However, elevated-temperature qubits require millimeter-wave (mm-Wave) spin control electronics, the design of which is challenging because, not
FIGURE 14. Compact gate-based readout of multiplexed quantum devices with a cryogenic CMOS active inductor.

FIGURE 15. Proposed architecture for a QP with a linear FDSOI SiGe hole-spin qubit array monolithically integrated with direct modulation mm-wave spin manipulation circuits, capacitively coupled (through shallow trench isolation) SET+TIA readout electronics, and SRAM for control pulse sequence and result storage. The PLL, which consumes 300 mW [165], can optionally be left off-chip to reduce cryostat thermal lift requirements.

least, of the lack of foundry transistor and passive models that are valid at cryogenic temperatures.

The proposed architecture in Fig. 15 integrates 80-GHz spin manipulation electronics, SRAM, qubits, and readout circuitry on the same die in production 22-nm FDSOI CMOS [166]. This technology allows for short 20-ps control pulses, 1000 times faster than in [3], and for spin manipulation amplifiers and switches beyond 200 GHz [167]. The controller, based on minimum size CMOS inverters and switches, without inductors, minimizes power consumption and layout footprint, while still achieving over 80-GHz bandwidth [168]. Unlike current classical electronic controllers based on IQ upconversion for spin manipulation [3], [110], [169], which are challenging to implement at mm-wave frequencies with low power consumption, direct amplitude and phase modulation of a single or frequency-division-multiplexed low-phase-noise mm-wave carriers [165] are employed. The QP core consists of linear arrays of elevated-temperature coupled QD hole-spin qubits [166], sketched in Fig. 16, each consuming under 5 pW (assuming $V_{DS} = 1$ mV and $|I_{DS,peak}| \leq 5$ nA like in Fig. 17) under nominal bias conditions.

The key to monolithic integration of QPs and control electronics in FDSOI CMOS is the fact that minimum-size MOSFETs behave like QDs in the subthreshold triode region at very low $V_{DS}$ ($V_{DS} < 10$ mV), biased between the first and second current peaks of the transfer characteristics shown in Fig. 17, while large gate finger devices behave as classical transistors when biased in the active region. A 3-D quantum well (or QD) is formed in the channel, with the gate and buried oxides acting as fixed “infinite” potential barriers in two directions perpendicular to the channel and the gate oxide spacers forming tunable potential barriers between the channel and the source/drain regions. The latter act as electron or hole reservoirs. The tunable potential barriers between the QD and the charge reservoirs and between neighboring QDs can be voltage controlled by the top or back gate [166]. As sketched in Fig. 16(b), a spin qubit is formed by applying a dc magnetic field ($B_0 = 0.5–2.5$ T) perpendicular to the channel, which splits the degenerate ground level in the QD into spin-up and spin-down energies separated by $\hbar \omega_01 \propto B_0$, which must be larger than the thermal energy, $kT$, to improve qubit fidelity. For example, the spin resonance frequency, $f_{01} = \omega_01/2\pi$, is 240 GHz at an energy separation of 1 mV with the corresponding thermal noise temperature of 12 K. The spin qubit is
initialized by injecting a single electron/hole from the reservoir into the lowest energy level of the QD, i.e., the MOSFET is in the single-electron transistor (SET) or single-hole transistor (SHT) regime. The electron (hole) spin is manipulated (quantum gate operation) by applying a (much) smaller amplitude oscillating mm-wave magnetic/electric field $B_1$ ($E_1$) at $\omega_0$, perpendicular (parallel) to $B_0$, that determines the Rabi frequency, $\omega_R$, which acts as the clock frequency of the QP. The readout of the quantum gate operation is done by spin-to-charge conversion followed by charge sensing and current (see Fig. 15) [166], [170] or voltage [130] amplification. Alternatively, reflection-type readout with IQ downconversion [171], as in transmon processors [3], has also been demonstrated [127], [172] but requires much larger die area. In the architecture of Fig. 15, SET/SHTs are capacitively coupled to each qubit in the 1-D quantum core array as charge detectors, each SET/SHT exhibiting the measured $I$–$V$ characteristics shown in Fig. 17 with a threshold voltage shift proportional to the charge present in its paired QD and to the coupling capacitance. In 22-nm FDSOI, current peaks in the range of 10 pA to 10 nA with peak-to-valley current ratios as large as 100 (see Fig. 17) can be selected for detection. Fig. 18 shows the measured stability diagram of a 22-nm p-MOSFET with QD characteristics observable up to 50 K.

Full cryogenic characterization down to 2 K of the 22-nm FDSOI technology, including large and small transistors, passives, and TIAs, has been performed up to 65-GHz demonstrating improved performance over room temperature operation and tunability of the threshold voltage to reach the peak-$f_T$ and peak-$f_{MAX}$ current densities at all temperatures in the 2–400 K range (see Fig. 19). Although the threshold voltage changes by more than 200 mV, over this temperature range, the current densities corresponding to peak-$g_m$, peak-$f_T$, and peak-$f_{MAX}$ of p-MOSFETs and n-MOSFETs remain constant over temperature from 400 K down to 2 K, while the peak-$g_m$, peak-$f_T$, and peak-$f_{MAX}$ values themselves improve as the temperature decreases. This behavior, characteristic of all CMOS technologies irrespective of foundry and technology node, allows us to employ a constant current biasing technique to design circuits for operation at cryogenic temperatures based solely on standard room temperature foundry models. In addition, unlike bulk planar CMOS or FinFET CMOS technologies, in FDSOI CMOS technologies, the threshold voltage and bias current density of a transistor can be adjusted from the back-gate voltage, without having to change the supply voltage or drain–source
voltage, in order to compensate for the threshold voltage variation with temperature. By appropriately setting the back-gate voltages of p- and n-MOSFETs independently, it is possible to ensure that p-MOSFETs and n-MOSFETs are perfectly matched in all CMOS logic and analog circuits and simultaneously biased at the desired current density between 0.1 and 0.4 mA/μm, at all temperatures from 2 to 400 K [173]. The latter is critical for the design of cryogenic control electronics without compromising performance.

Using the above design methodology with standard design kit models, monolithic integration of the readout TIA and QDs has been demonstrated in this technology [166]. Unlike in other applications, this TIA was optimized to read qubit currents in the range of 10 pA and 10 nA with ultralow input capacitance to avoid overloading the qubits and to maximize the spin-readout bandwidth when driving a 50-Ω load. A TIA input referred noise of $\sim 1 \text{ pA}_{\text{rms}}/\sqrt{\text{Hz}}$ was measured at room temperature, as shown in Fig. 20 [170]. A 10× improvement in SNR was simulated at 12 K [170].

While the proposed architecture provides a scalable approach to the implementation of a fully integrated quantum computer, numerous challenges must be overcome before the realization of such a device can become reality. Although a quantum core with 1 billion qubits can be integrated within a 5-mW power budget in 22-nm FDSOI (using the 5 pW per QD qubit estimate noted previously), the development of large (>1000) QPs of the proposed architecture is limited by the power consumption and associated heat dissipation of the analog-mixed-signal control and readout electronics and by the challenge of interconnecting such a large number of qubits with the control electronics. Although readout SETs and electronics can be capacitively coupled in a 3-D wafer stack architecture, the 2-D tuneable coupling of these spin qubits, needed for a large-scale full 2-D processor architecture, as in [3], remains a challenge due to the strict manufacturing rules of nanoscale production CMOS technologies, which allow transistor layouts of only one orientation. It will require process changes that are not currently economically justified by the miniscule QC market. Alternatively, spin-to-THz photon coupling, as suggested in [175], may be feasible for coupling qubits in the second dimension using dielectric waveguides in the BEOL, not unlike in silicon photonics platforms.

3) RESEARCH AT JÜLICH

At the Institute of Electronic Systems (ZEA-2) of the Central Institute of Engineering, Electronics, and Analytics at Forschungzentrum Jülich, researchers are investigating approaches to scaling quantum computers with the ultimate goal of cointegrating control and measurement hardware as close as possible to the QP to minimize challenges associated with interconnects. In support of this goal, they are pursuing a research agenda following the “V-model” of systems engineering, performing both top-down system simulations and bottom-up implementations of CMOS circuits [176].

The Jülich team has carried out a systems study in which they analyzed the power and area requirements of realistic integrated circuits for the control of semiconductor spin qubits [177] to understand the tradeoff between the temperature at which a CMOS quantum controller is thermalized, the process parameters of that technology, and the scale of the QP. To maximize the insight provided by this study, a model was generated that encapsulated all the necessary functional components and allowed optimization at both the circuit and architecture levels. Their key result was that power consumption is the dominant obstacle in obtaining scalable circuits.
With the 1-mW cooling power available from a dilution refrigerator operating at 200 mK, they found that a controller implemented in 65-nm CMOS and operating from $V_{DD} = 1$ V could service 12 qubits. Furthermore, they found by moving to a 14-nm process optimized for $V_{DD} = 0.1$ V when operated at 200 mK, up to 328 qubits could be operated from the same power budget.

Based on the outputs of their model, the Jülich team determined that scaling to the roughly million qubits required for the implementation of a fault-tolerant quantum computer would not be feasible if CMOS electronics were thermalized with the QP at 200 mK. Instead, to meet this need, they conclude that the control electronics will have to be thermalized at an elevated temperature where more cooling power is available and advanced thermally isolating 3-D interconnection techniques must be developed to interface the QP to the electronics. They estimate that, with the electronics thermalized to 1.8 K, it should be feasible to control 130,000 qubits using electronics implemented in a 65-nm CMOS technology and dissipating 1 W [177]. The 130,000 qubits are modeled to be controlled by 300-MHz rectangular pulse sequence signals for singlet-triplet spin qubits. Low power dissipation is achieved through assumed microfabricated interconnects between the electronics and qubits with no 50-Ω matching needed. However, major challenges in this temperature distributed approach applied to spin qubits include matching the qubit footprint (currently in the micrometer range) and removing power from the electronics while keeping the qubits cold.

As part of the bottom-up implementation, the Jülich team has developed, implemented, and characterized a prototype chip in TSMC 65-nm bulk CMOS that contains all of the required building blocks for the control of a GaAs-based semiconductor qubit [178]. The chip contains a digital I2C interface for configuration and control, a 13-bit eight channel Bias-DAC with 1-V output range, and an 8-bit pulse DAC with a sampling rate of 250 MHz and ±4 mV voltage range. Furthermore, the chip contains test structures for future applications like a 20-GHz voltage-controlled oscillator (VCO), a digitally controlled oscillator (DCO) for clock generation, and single-device test structures for cryogenic device characterization.

The Bias-DAC is designed to provide up to eight individually set output voltages over eight pads, which can be bond-wired to the qubit chip metal electrodes for QD generation and tuning. Considering the need to minimize power consumption, a charge redistribution architecture was employed due to its negligible static power dissipation and low Johnson–Nyquist noise, which reduces proportional to ambient temperature. In addition, this architecture allows for a robust design in terms of cryogenic effects in CMOS processes, as transistors are used as transmission gates or digital control logic only, both being reported as mostly unaffected by cryogenic temperatures. Whereas leakage will be present, the effects of the cryogenic temperatures can be utilized in favor for this DAC design, resulting in an increased subthreshold slope and, hence, a reduced leakage current. A die photograph and the output curve of the Bias-DAC are shown in Fig. 21. The power consumption is less than 3 μW per channel. 99.5% of this power is dissipated in digital circuitry and scales well with CMOS technology node. Furthermore, it is not required to duplicate all of the digital blocks with each additional Bias-DAC, because timing signals controlling the DAC can be generated only once on-chip and distributed to multiple Bias-DACs. Only a few bytes of digital memory circuitry are required for each individual DAC. Fig. 21(b) shows the calibrated output curve of the Bias-DAC after the calibration procedure. The desired output voltage range of 1 V is achieved. A zoom-in reveals some singular nonmonotonic steps. However, qubit operation is unaffected by those as nonmonotonic steps are detectable when fine-tuning the qubit.

The pulse DAC is an 8-bit segmented current steering DAC. Due to the low output voltage amplitudes needed for the operation of a qubit, a 50-Ω resistance can be used to transfer the current into a voltage, which simplifies the measurements. The DAC consumes approximately 150 μW of analog power, which can be reduced by increasing the termination resistance, thereby reducing the current needed for the same output amplitude. One of the problems of a current-based DAC topology is the increased mismatch between current sources in a cryogenic environment. Therefore, each current cell can be independently calibrated.

The DCO is a current starved ring oscillator whose bias current is controlled by a digital word and covers a frequency range from 450 to 550 MHz. The VCO is an LC tank NMOS and PMOS transistor cross-coupled pair at the top and bottom to ensure oscillation. The oscillation is centered around 19 GHz. Measurements show an increased oscillation frequency at cryogenic temperature as well as decreased far-out phase noise and increased close-in phase noise. Fig. 21(c) shows the output frequency of the VCO over the applied control voltage from 0 V to 1.2 V. With increasing control voltage, the output frequency of the VCO increases. At cryogenic temperature, the center frequency of the oscillator shifts from 18.12 to 19.15 GHz, which is an increase of about 5%.

Next to the prototype chip with functional blocks, the team implemented and tested chips with single devices in order to develop transistor models for cryogenic temperatures. The test chips also contain structures to evaluate local self-heating effects. Ongoing work by the Jülich team comprises the control electronics of silicon-based qubits, requiring operating frequencies up to 20 GHz, and implementation of integrated readout electronics. The CMOS technology of choice for upcoming implementations is 22-nm FDSOI due to its improved power consumption and further design freedom by back-gate control.

4) RESEARCH AT TU DELFT/INTEL
Charbon et al. [179] at TU Delft first proposed using cryogenic CMOS for quantum control in 2016, with the goal
of enabling easier scaling by relieving the interconnect bottleneck. The TU Delft team embarked in a comprehensive research program aimed at realizing complete solutions for the control and measurement of silicon spin qubits following the architecture shown in Fig. 22.

To begin, the academic members of the team fully characterized a 40-nm standard bulk-CMOS technology node to understand the degree to which critical design parameters change with cryogenic cooling. They then proceeded to develop a complete standard cell library named cooLib to permit digital synthesis [180]. Not only was this library able to capture cryogenic effects (e.g., through timing models), but the layout geometries and transistor dimensions were chosen to optimize cryogenic performance, mitigating effects such as latchup, which can become exacerbated at cryogenic temperatures.

To validate the library and highlight its performance advantages over standard PDK offerings, the test chip appearing in Fig. 23 was designed. The chip implemented an RISC-V processor that could operate at 4 K at reduced power supply, as low as 0.6 V, with an EDP better than at room temperature. In addition, a number of typical digital blocks (e.g.,
As a precursor to robust mixed-signal system-on-chip (SoC) designs, the academic members of the team have also developed a variety of analog/RF circuit blocks. One example is a reference voltage generator [181] that is able to provide a stable voltage at temperatures as low as 4 K. The resulting voltage reference is shown in Fig. 24. The figure shows plots of the voltage as a function of temperature for NMOS, PMOS, and DTMOS design styles. The NMOS, PMOS, and DTMOS circuits achieved voltage stability better than 10, 70, and 60 mV over a 4–300 K temperature range, respectively [181].

They also implemented analog/RF blocks that are essential to the realization of a complete quantum control and readout system for semiconductor spin qubits. For the readout, low-noise amplifiers are required, so a prototype device was designed and realized in a 160-nm standard CMOS technology and tested at 4 K. The device is based on a noise cancellation topology [182] to provide both input matching and low noise simultaneously, while programmable bias currents controllable externally by digital switches were also added to the design. The resulting noise figure yields a minimum noise temperature of 7 K at a physical temperature of 4 K [12].

In addition to components related to the receivers required for readout, the team has also demonstrated oscillators, recognizing that signal generation will be required for both control and readout of spin qubits. Fig. 25 shows a DCO implemented in 40-nm standard CMOS technology [183]. A digital calibration loop is introduced to automatically adjust the configuration of the differential-mode and common-mode capacitor banks to ensure that the oscillator always operates near its optimum performance, where the oscillator common-mode resonance is at twice the oscillation frequency. This technique suppresses the oscillator phase noise at 100-kHz offset frequency by >10 dB at 4 K. Though high-frequency signal and clocks can be provided externally, having an LO in situ is convenient to reduce complexity and to potentially achieve lower noise with relatively low effort, not to mention the possibility of generating many synchronized signals at practically any phase.

Finally, the TU Delft team, in collaboration with Intel Corp., designed an extensively digital controller for spin and superconductive qubits. The design, called Horse Ridge, is shown in Fig. 26. Unlike cryo-CMOS controllers, targeted for superconducting qubits [101], this controller was designed for spin qubits primarily, although it could be compatible with transmons, though it was not tested with them, and there might be some adjustments required.

It comprises an array of numerically controlled oscillators, which drive I and Q versions of a programmable
waveform (rect, cosine, raised cosine, triangle, and Gaussian envelopes). The waveforms are digitally synthesized and subsequently upconverted to a programmable frequency from 2 to 20 GHz. Image rejection is performed at this stage before being outputted through a balun [184], [185].

The figure also shows a typical quantum algorithm performed on a spin qubit by means of the controller. Extensive experimentation was conducted with Horse Ridge, leading to the characterization of the device with SiGe qubits, where single- and two-qubit gates were performed, with the single-qubit gates achieving over 99.7% fidelity. Randomized benchmarking was performed using 64 Clifford gates achieving an overall fidelity of 99.69%, compared to 99.71% fidelity using a conventional room temperature controller [186]. Horse Ridge is equipped with an instruction set, making it trivial to program a sequence of quantum operations. This feature, along with the chip’s overall performance, could pave the way to a new generation of scalable controllers operated more seamlessly in the quantum stack.

5) RESEARCH AT EQUAL1.LABS
In large agreement with the principles outlined above, the startup Equal1.Labs aims to realize a commercial application-specific quantum computer for neural networks. Their vision, as shown in Fig. 27, is to use a high-volume advanced CMOS process technology to monolithically integrate qubits with interface electronics on a single die and to operate it at 4 K using a custom-built portable cooler. In collaboration with University College Dublin (UCD) in Ireland, they have already taped out three generations of quantum SoC processors, each containing thousands of QD-based qubits and tens of millions of digital gates. Their initial employment of charge qubits was due to their relative simplicity over the spin qubits [189] and is justified due to: 1) no need to generate any magnetic fields, neither dc (permanent magnets) nor microwave; 2) straightforward control [injection into and movement inside of a QD array (QDA)] and detection of single electrons, which can be realized at very low power consumption and area; and 3) fast flip times of quantum gates (tens of picoseconds) in order to compensate for their shorter decoherence time. The currently measured decoherence time in their system is >50 ns, which is two to three orders of magnitude shorter than with the spin-based qubits. However, the cutoff frequency ($f_c$) of the 22FDX FD-SOI CMOS transistors used in their system is in the hundreds of gigahertz, allowing the realization of quantum gate operations <50 ps, which is two to three orders of magnitude faster than with the spin counterparts. This allows over 1000 gate operations per useful decoherence time. Therefore, based on the research carried at Equal1.Labs and UCD, the team has come to the conclusion that despite their apparent imperfections, when examined in isolation, the collectivity of charge qubits and their straightforward integration with the interfacing electronics appears the best choice for integrated large-scale quantum computing systems [117], [188], [190].

Moreover, their latest work includes an addition of hybrid qubits, which combine the benefits of charge and spin qubits. As indicated in Fig. 27, adding a permanent magnet in the cryo chamber is quite straightforward.

The realized quantum structures are either fully compliant with the fab’s design rule checker (DRC) or with benign DRC violators fully signed off by the fab. This makes it amenable to high-volume production of quantum SoC chips with potentially millions of qubits. The qubits are of lower individual quality than the other types described above, but they are targeted for applications and arrangements that favor quantum circuits of shallow depth but massive width.

Furthermore, they are incorporating a topological protection to charge qubits by arranging them in a special lattice as Su–Schrieffer–Heeger (SSH) rings [188]. The principle here is that a long chain of QDs, which can be arranged in a ring, can have an alternating pattern of “inner” and “outer” tunneling coefficients that could be readily controlled electrostatically. Their analysis indicates that under certain circumstances, the ring can behave as a topological insulator, i.e., insulating in the interior and conducting at the edges. The ring’s properties are protected and robust to noise due to topological attributes in its spectrum. This topological invariance is used to construct “static” noise-resistant charge qubit states. This is similar in effect to the popular attempt of performing QEC, which typically requires a qubit overhead in the range of well over 100–1000×, but here it is self-correcting and local in the sense that the “information” does not have to travel outside of the structure.

As mentioned above, the FDSOI process is of particular interest here. In contrast to bulk CMOS (or even the...
FIGURE 26. RF DAC controller for spin and superconductive qubits implemented in 22-nm CMOS technology with FinFETs. (Top) Micrograph of the chip, measuring 4 mm, and schematic. (Center) Quantum algorithm with Bloch sphere representation. (Bottom) Results of the measurement [184], [185].

FIGURE 27. (a) Vision of a single-chip quantum computer first proposed in [187]. Each qubit is based on a QD array. QDs can be arranged in a lattice as SSH rings for topological protection against noise [188]. (b) 3 × 3 mm² quantum SoC designed by Equal1.Labs and UCD housed in a commercial portable desktop-size 4 K cryogenic cooler consuming 1.5 kW. The unit is fully self-contained and air cooled. (c) Mechanical details of the portable cooler. The chip is placed at the bottom of the cool head, and the space underneath houses a magnet to support hybrid qubits.

more advanced FinFET technology), FDSOI provides a thin semiconductor layer isolated vertically from the substrate by a 20-nm buried oxide layer [see Fig. 28(a)]. Therefore, a quantum particle can be strictly confined inside the 5-nm-thick semiconductor film, where it precisely follows the gate control, and is isolated from the substrate impurities to further increase its decoherence time. This opens up new approaches for charge qubits that were popular but then abandoned over a decade ago [121], [122], [123], [124], [125].

Realization of a quantum computer entails quantum entanglement of multiple qubits. A linear arrangement of QDs [i.e., quantum shift register; see Fig. 28(c)] allows for the
individual electrons to travel within the structure [117], [175], [190]. For example, once the electron is injected into a QD, it can be transported to a neighbor QD through the $\pi/2$ phase shift [119]. It can then be transported to next QD through another $\pi/2$ phase shift. We can, thus, construct a 1-D topological array of QDs that move around an electron in its entirety (for $\pi/2$ phase shifts) or transfer around part of its wavefunction $\Psi_1$ (for phase shifts other than $\pi/2$). For example, $\pi/4$ implements a Hadamard gate (H). Other single-qubit gates (e.g., $R(\phi)$, rotation) can be obtained by adjusting the pulse duration of the imposer’s voltage. We refer the readers to [120] for further details regarding the implementation of a universal gate set via this topology.

As mentioned above, the quantum state of the QDA can be dynamically controlled by adjusting its potential barrier profile [see Fig. 28(b)]. This allows the splitting of the quantum particle wavefunction for the superposition of quantum states as well as the transfer of particles between QDs via tunneling to reach their intended entanglement stage [see Fig. 28(e)]. As shown in Fig. 28(f), this is accomplished through the tightly integrated DACs [131], [132] that control the tunneling barriers between the QDs. Due to the very light capacitive load of the qubit interface nodes, the DACs can be designed to consume very little power using a switched-capacitor approach. The final quantum state of the QDA can be read out by the single-electron detectors. The reset device engages once before the quantum experiment to ensure that no undesired electrons are present in the QDA. A 20-μm device exclusion region is maintained between the quantum structure and the interface circuits in order to minimize dopant induced decoherence of the quantum state. Only the reset switch and the first source follower of the detector path are placed in close proximity of the last QD (realized as a floating depleted well), since their loading capacitance impacts the readout charge-to-voltage ($Q$-to-$V$) conversion gain.

The capacitive DAC is shown in Fig. 29. It is 8-bit binary-controlled consisting of 255 identical weight units. Its key building blocks are the circuitry for clock gating, capacitor array, precharge (pedestal setting), and pulse shaping filter. The capacitors in the binary-controlled array are split into unit cells (UCs), one of which is shown on the right side of Fig. 29. The UC comprises logic gates driving the capacitive divider ($C_{u1}$ and $\sum C_{u3}$). A noteworthy feature inside this block is a parasitically coupled negative clock edge, through $C_{u2}$, to compensate for switch charge injection and clock
feedthrough at the DAC output. The clock gating reduces the dynamic power consumption by propagating the 2-GHz clock only to the enabled bits in the 8-bit DAC. The precharge circuit sets a dc voltage on N1 before the DAC is activated in order to establish the pedestal of the QD energy barrier levels. The 11-stage $RC$ pulse-shaping filter reduces fast transients at $V_{\text{out}}$. $R_c$ is realized with low metal layer Metal1 (M1) trace of 50-nm width, which still meets the DRC rules. At the time of design, it was not known that polysilicon resistors (without salicide) would work at 4 K [166], [170]. All the capacitors are designed with intermediate metal layers with a density of 5.08 fF per $\mu$m$^2$.

Fig. 30(a) details the readout circuitry for detecting the state of the quantum structure. It functions as a single-electron detector to observe a gain or loss of an individual electron within a window determined by $S_0$ and $S_1$ pulses from the floating diffusion (FD) interface node between the quantum and classic circuits. It consists of a double source follower $M_{3,4}$, a preamplifier (preamp) and a switched-capacitor correlated double sampler (CDS). The minimum-size $M_3$ was selected to maximize the $Q$-to-$V$ gain. However, such a choice brings up significant flicker noise from the device, but it is effectively rejected by the CDS scheme that samples the signal twice with $S_0$ and $S_1$ pulses within a short time interval, just before and after the electron is expected to be injected into or received from the floating depleted quantum well [labeled as “FD” in Fig. 28(f)].

IV. CONCLUSION

In this article, we have reviewed early work on CMOS integrated circuits for the QIS. While we have shown a broad range of circuits with applications ranging from NMR to quantum computing and covering many different quantum technology platforms, the astute reader may realize that the circuits described here are all far from what will be required once these quantum technologies have matured. As such, there is a large opportunity for circuit designers to contribute to the definition and demonstration of robust quantum sensing and computation systems. For instance, each of the different quantum computing platforms has a different control and measurement architecture, and the circuits described here leave considerable work in the demonstration of scalable systems. Some example areas where contributions are still required include low-power readout circuits meeting the stringent noise requirements of superconducting quantum computing (where receiver noise temperatures of about 2 K are required), quantum control circuits for both semiconductor and superconducting qubits with low enough power consumption that they can be cryogenically cooled at scale, compact biasing and readout circuits for superconducting or semiconducting photodetectors for state detection of trapped ions, and low-noise high-speed voltage sources compatible with the requirements for trapped-ion control (> 1.5 V) with low enough power dissipation for cryogenic operation at scale. For these reasons, we anticipate that the field of
integrated circuits for quantum applications will continue to grow over the coming years.

ACKNOWLEDGMENT

The authors would like to thank GlobalFoundries for providing silicon fabrication through the 22FDX university program.

REFERENCES


