Cryogenic CMOS for Qubit Control and Readout

Pellerano, Stefano; Subramanian, Sushil; Park, Jong-Seok; Patra, Bishnu; Xue, Xiao; Vandersypen, Lieven M.K.; Babaie, Masoud; Charbon, Edoardo; Sebastiano, Fabio

DOI
10.1109/CICC53496.2022.9772841

Publication date
2022

Document Version
Final published version

Published in
2022 IEEE Custom Integrated Circuits Conference, CICC 2022 - Proceedings

Citation (APA)

Important note
To cite this publication, please use the final published version (if applicable). Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.
Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

https://www.openaccess.nl/en/you-share-we-take-care

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.
Cryogenic CMOS for Qubit Control and Readout

Stefano Pellerano¹, Sushil Subramanian³, Jong-Seok Park¹, Bishnu Patra¹, Todor Mladenov¹, Xiao Xue², Lieven M. K. Vandersypen², Masoud Babaie², Edoardo Charbon³, Fabio Sebastiani²

¹Intel Corporation, OR, USA, ³Delft University of Technology, The Netherlands

Introduction

Quantum computers have been heralded as a novel paradigm for the solution of today’s intractable problems, whereas the core principles of quantum computation are superposition, entanglement and interference. Three fundamental properties of quantum mechanics [1]. A quantum computer generally comprises a quantum processor, made of an array of quantum bits or qubits, and a classical controller, which is used to control and read out the qubits. Quantum algorithms are generally mapped onto a circuit of quantum gates that operate on multiple qubits. Unlike conventional digital bits, qubits can take a coherent state ranging from |0⟩ to |1⟩ on a continuous sphere, known as the Bloch Sphere and they are implemented based on several mechanisms. While many solid-state implementations of qubits exist, an exhaustive description of available technologies is beyond the scope of this paper [2][3].

To execute a quantum algorithm, analog/RF signals are required to manipulate the state of individual qubits and entangle multiple qubits. Additional control signals are required to perform qubit state readout. These signals are generated by the classical controller, which is generally operated at room temperature (RT), while qubits must be cooled to cryogenic temperature (CT), usually from few tens of milli-Kelvin to a few Kelvin, to preserve their coherence. The resulting large temperature gradient must be handled to enable a direct electrical interconnection, while preventing energy transfer from RT to CT. To keep phonon flow to a minimum, while ensuring fast and compact electrical connections, multiple thermalization stages are used, where each stage attenuates control or amplifies readout signals. The goal is to match the impedance of each signal generator and only generate the noise equivalent to that temperature. An impedance of 50Ω is typically used to easily adapt existing RT instrumentation. This approach has worked well with <100 qubits. However, with an increasing size of quantum processors, there has been a need of moving classical control to lower temperatures, close, or ideally identical, to those of the qubits, to enable scalable, compact, and reliable machines in the future. An additional concern is the capability of such a system to be quickly debugged and upgraded, something hard to implement in a RT approach.

To address these issues, in 2016 we have proposed the use of cryogenic electronics, and in particular, cryogenic CMOS, or cryo-CMOS, as a good candidate to accomplish the tasks of a generic classical controller [4]. With the lack of thermalization requirements, cryo-CMOS electronics can generate control signals at an amplitude far lower than if they were generated at RT. In addition, readout circuits can also benefit from low thermal noise, while the proximity in space and temperature to qubits can reduce the complexity of interconnect, where superconductors can be used to maximize thermal isolation with virtually zero electrical losses. With high-temperature qubits [5]-[7], this configuration could be simplified further, eliminating the requirement for advanced refrigeration and milli-Kelvin temperatures.

The main limitation of the cryo-CMOS controller is the power of classical circuits, which must be within the limits of thermal absorption by the refrigeration unit. Non-CMOS devices, such as high-electron-mobility transistors (HEMTs), SiGe heterojunction bipolar transistors (HBTs), GaAs logic, and rapid single-flux quantum (RFQ) circuits have been proposed to achieve high-performance circuit functionality of the type needed in the classical controller [8]-[10]. However, recent technological advances, e.g., high-speed HBTs, can take advantage of 60 years of industrial development aimed at scalability, like CMOS [5][11][12]. An advanced CMOS process, Intel 22nm FinFET technology, was selected in this work to implement complex functionality at low power.

The classical control systems proposed in this paper are part of a larger context and require therefore specific interfaces to the quantum stack that supports quantum algorithms. The quantum stack in Fig. 1 is the ensemble of architectural components designed to translate such algorithms to quantum circuits and finally to quantum control via assembly languages, like QASM. Such architectures are complex and significant development is ongoing [13].

In this paper, we first give a brief overview of the current state of the art in cryogenic electronics for qubit control and readout. We then describe the architecture, circuit design and measurement results of two cryogenic SoCs, code-named “Horse Ridge”, designed to operate at 4K and implemented in Intel 22nm FinFET technology. Horse Ridge 1 is designed for state manipulation of spin qubits and transmons. Coherent qubit control, randomized benchmarking and two-qubit algorithm are demonstrated with a 2-qubit quantum processor. Horse Ridge 2 is targeted at spin qubits and can implement all the required control functions of driving, readout and gate bias pulsing. Its performance is fully characterized at 4K. We then conclude the paper with some considerations on future directions.

State of the Art Review

To demonstrate the reliability and performance of CMOS technology for realizing complex cryogenic integrated circuits, several building blocks have been published over the past few years [14]-[32]. However, this section only focuses on the past art that demonstrated promising performance at several hundreds of millikelvin or the ability to directly control and readout qubits.

The first integrated implementation of a cryogenic pulse modulator was presented in [15], where a pair of current-mode digital-to-analog converters (DAC) controlled by a waveform memory and a single-sideband upconverter are employed to manipulate the state of superconducting qubits. Although the pulse modulator consumes <2mW/qubit, its scalability is limited by the need for one high-frequency local oscillator (LO) per qubit, and the use of fixed-length symmetric pulses limits its compatibility with other qubit technologies. To alleviate these issues, [29] proposed a rotational phase accumulator in the digital domain to generate a microwave pulse with any arbitrary duration, phase, and frequency. This work also proposed a sinusoid shaping nonlinear DAC followed by a linear interpolating DAC to improve the controller power efficiency. However, the scalability of this structure is limited since it can only track the phase of one qubit at a time.

The work in [25] reported a switched-capacitor-based Cryo-CMOS controller in a 28nm FDSOI CMOS technology that operates at 100mK and can generate static and dynamic voltages for the parallel control of qubits. To generate the required static voltages of qubits electrodes, the output of a room-temperature DAC is multiplexed by using capacitive sample-and-hold circuits periodically refreshed to counteract the charge leakage of the hold capacitors. The required dynamic voltage pulses are then implemented by toggling the potential of the bottom plate of the hold capacitors with two fixed external voltage sources. The reported average power consumption for generating voltage pulses of 100mV is about 18nW per cell. Consequently, this structure can potentially be scaled up to generate voltage for a few tens of thousands of electrodes in a commercially available dilution refrigerator.

To demonstrate the viability of various functions down to 110mK, [20] presented a Quantum Integrated Circuit (QIC) in a 28nm FDSOI...
CMOS technology. The QIC chip comprises a double quantum dot (QDUT) with digital and analog blocks for QDUT biasing, manipulation, and read-out. This work successfully demonstrated a quantum effect called “charge pumping” by applying a 2.8GHz signal from an on-chip modulator to the QDUT and measuring its current through an on-chip trans-impedance amplifier (TIA) while sweeping the voltages of the plunger gates.

In [26], a fully-integrated cryo-CMOS I/Q receiver is proposed integrating frequency synthesis, low-noise amplifiers, Gilbert cell active mixers, and polyphase filters to perform phase-sensitive readout of qubit signals. The receiver, operating at frequencies of 4.5 to 8.5GHz, is based on a non-zero IF down-conversion with >20dB image-rejection ratio (IRR). At a temperature of 3.5K, it achieves a conversion gain up to 70dB with an effective noise equivalent temperature of 39.2K, IIP3>-72dBm, and P1dBIn>-85dBm. The proposed receiver supports up to 70 qubits in frequency-division multiple-access (FDMA), dissipating 1.5mW/qubit.

The work in [27] proposed a cryogenic receiver for gate-based RF readout of spin qubits, achieving a 58dB gain and a 45K effective noise temperature at 4.2K while consuming 0.17mW/qubit. Compared to the prior-art cryogenic DC readout IC in [20], this work supports FDMA for scalability, 1000x faster readout, with 5x lower energy consumption. However, in order not to degrade the inherent readout of qubit signals, a superconducting parametric amplifier with a near quantum-limit noise temperature (Tn~0.4K) or a HEMT LNA (Tn~2.2K) must be placed prior to the receiver to amplify the signal by >20dB.

The work in [31] proposed a single-pole-4-throw (SP4T) reflective cryo-CMOS multiplexer using standard 28nm CMOS technology to reduce the number of required output RF lines from the base temperature stage of a dilution refrigerator to the control/readout electronics placed at 4K stage. The multiplexer is controlled with a parallel or serial interface and exhibits DC-to-10GHz bandwidth, 1.6dB insertion loss and 34dB isolation at 6GHz, with 2ns switching time and 36.2μW heat dissipation while operating at 32mK. This work can improve cryogenic characterization throughput and alleviate the wiring problem in future large-scale quantum computing systems.

In [33], dispersive readout of the charge state is demonstrated on an array of 3x3 quantum dots, accessed by a combination of TDMA with digital and analog blocks for QDUT biasing, manipulation, and read-out. This work successfully demonstrated a quantum effect called “charge pumping” by applying a 2.8GHz signal from an on-chip modulator to the QDUT and measuring its current through an on-chip trans-impedance amplifier (TIA) while sweeping the voltages of the plunger gates.

Horse Ridge 1: Qubit Drive
Horse Ridge 1 (HR1) is our first-generation control SoC, which implements the function of qubit drive: the generation of RF amplitude- and phase-modulated pulses to manipulate the state of qubits [34]. The controller integrates four identical transmitters, each capable of controlling 32 frequency-multiplexed single-electron spin qubits or transmons in the frequency range from 2 to 20GHz, for a total of 128 qubits over 4 RF cables. Each transmitter has been designed following the methodology presented in [35] to achieve a 99.999% fidelity, with both an SNR (25MHz bandwidth) and a spurious-free dynamic range (SFDR) better than 44dB. Figure 2 shows the simplified block diagram of one of the transmitters, integrating an instruction controller, a multi-channel direct digital synthesizer (DDS) and an analog/RF front-end.

The digital is clocked at 1GHz and implements two polar modulators (bank 1 & 2). The 32 digitally controlled oscillators (NCOs) are used for coherent control of 32 qubits with a single LO. The 22-bit frequency control word enables 0.2kHz resolution and a 12-bit phase word is used to adjust the NCO phase and implement Z-axis qubit rotations [35]. Phase and amplitude are modulated by a 10-bit and 8-bit sequence, respectively. These sequences are stored in the on-chip memory and calibrated for each qubit to implement specific instruction sets. Sin(φ) and Cos(φ) functions are generated through look-up tables (LUT) to reduce power consumption. Signal bit widths have been defined by system simulations to target 99.9999% fidelity, 10dB better than the overall controller target [35]. Digital IQ and offset compensation networks (α, β and γ in Fig. 2) are implemented for image and LO leakage suppression. When used at the same time, the two identical banks allow for simultaneous control of two qubits.

Two high-speed 10-bit (5-bit thermometer + 5-bit binary) current-steering DACs followed by a 2nd order GmC reconstruction filter convert the IQ digital modulated pulses into analog signals. Current bleeding in the DAC unit cell [36] and current-mode filtering enables high-linearity operation. The VGA is implemented as a 4-bit programmable current mirror, with a single-stage amplifier to boost the current mirror loop gain for increased linearity. A double-balanced Gilbert cell IQ mixer with current bleeding is used to upconvert the baseband pulse to the qubit Larmor frequency. Two

**Fig. 3. (a) HR1 die and (b) FCBGA package assembled on cryogenic PCB.**
operating modes are used to cover the 2-20GHz RF output bandwidth, selected by a mux embedded in the mixer output with cascode devices. In the low-band mode (2-15GHz), the fundamental harmonic of the LO is used with the mux steering the mixer output current into a resistive load. For the high-band (15-20GHz), the 3rd harmonic of the LO is used and the mixer current is steered into a resonant load. Two separate class-A amplifiers with wideband transformer-based matching networks are used to drive the 50Ω cables to the qubits. With 15 switchable unit cells, ~24dB of power control is achieved.

The 2-15GHz LO signal is provided externally. Wilkinson power dividers and discrete wire-bonded quadrature hybrids are used on the cryogenic PCB to generate multiple single-ended IQ signals. Single-ended to differential LO drivers with 20dB voltage gain and 15GHz bandwidth are integrated on-chip to deliver the required voltage swing to the mixers. The target phase noise of -116dBc/Hz at 1MHz offset [35] is achieved over the entire frequency range with ~7mW power for both IQ paths. The 1GHz clock for the DACs and digital baseband is also provided externally. This is shared across the 4 transmitters, each using a 200Ω on-chip termination. The analog/RF front-end receives the external clock and distributes it to the DACs and digital baseband after proper synchronization. All bias currents are generated by a constant-gm block, tunable from 0.5× to 2× the nominal value to account for the shift in biasing points between 300K and 4K. An external bias current can also be applied directly.

HR1 die is fabricated in Intel 22nm FinFET process and occupies an area of 16mm² for 4 transmitters [Fig. 3(a)]. A custom-made annealed and gold-plated copper enclosure shown in Fig. 3(b) is used for efficient thermalization when the PCB is installed in the dilution refrigerator.

The measured peak output power at 4K is shown in Fig. 4(a). Flatness is limited by additional ground inductance erroneously introduced in the layout. The SFDR at peak output power is >45dB, limited by HD2 and IRR [Fig. 4(b)]. The LO rejection ratio (LORR) after calibration is 36dB. This is not a limitation for qubit control since the LO leakage can be placed far away from any qubit by proper choice of LO and NCO frequencies. Measured IM3 and SNR (25MHz bandwidth) are better than 50dB and 48dB respectively [34].

The amplitude and phase modulation information for pulse shaping is stored in on-chip SRAM, allowing envelopes up to 41μs, which are referenced by a look-up table (LUT) that can define 8 instructions per qubit, each with a phase offset to efficiently reuse envelopes for different rotation axes. Figure 4(c) demonstrates the versatility of this design by showing the measurement of an instruction sequence containing 5 unique waveforms targeting 5 different qubit frequencies. At 1GHz clock frequency, the power consumption is 330mW and 54mW for the digital and analog/RF, respectively. The digital power consumption can be substantially reduced by design optimization and clock-gating. Nevertheless, thanks to the frequency multiplexing architecture, the total power consumption of 12mW/qubit would still allow the control of >320 qubits in a state-of-the-art dilution refrigerator over just 10 RF cables.

Extensive testing of HR1 with spin qubits is conducted [37]. We used the 2-qubit quantum processor shown in Fig. 5(a). First, to demonstrate coherent qubit control, Rabi oscillations are measured. As shown in Fig. 5(b), the qubit is initialized to ground state and then an RF pulse of duration t is generated by HR1 at the qubit Larmor frequency (13.4GHz for this experiment) and applied to the gate MW in Fig. 5(a). The qubit state is then measured. Figure 5(c) shows that as the pulse duration t is increased the measured state oscillates between |0⟩ and |1⟩, as the qubit state is rotated along the Bloch sphere, demonstrating coherent qubit control. The limited amplitude of the measured Rabi oscillations shown in Fig. 5(c) is due to the non-zero error probability in the readout process. To quantify the fidelity of a single-qubit gate operation and compare with RT control electronics, a randomized benchmarking experiment is performed [37]. A random set of qubit rotations ["Clifford Operations" in Fig. 5(d)] is executed and the error in the final qubit state is measured as a function of the number of consecutive operations. Results show that within the measurement error margin, HR1 achieves the same fidelity...
as the RT control electronics at 99.7%, limited by the qubit sample. The two qubits can be simultaneously controlled by HR1 through frequency multiplexing. As shown in Fig. 5(e), simultaneous coherent control is demonstrated by using two tones spaced by about 100MHz around a 13.5GHz LO frequency, corresponding to the two qubit resonance frequencies. Two-qubit operations can also be implemented. Once the potential barrier between the two qubits is lowered [2Q-B gate in Fig. 5(a)], due to reciprocal interaction each qubit resonance frequency splits in two (by ~10MHz for this experiment), depending on the state of the other qubit. The controlled-NOT (CNOT) gate is the quantum equivalent of an XOR gate: the target qubit state is flipped only if the control qubit state is |1⟩. If f₀ and f₁ correspond to the resonance frequencies of the target qubit when the control qubit state is |0⟩ or |1⟩ respectively, then by pulsing at f₁ a CNOT is implemented (with some residual phase rotation on the control qubit), since the pulse would be at the target qubit resonance and hence able to rotate it only if the control qubit state is |1⟩. To perform a single-qubit operation on the target qubit, two consecutive pulses at f₀ and f₁ are generated, so that the target qubit is rotated regardless of the state of the control qubit. Leveraging these single- and two-qubit operations, HR1 is programmed to execute the 2-qubit Deutsch-Josza algorithm, which determines if a function gives a constant or balanced outcome [38]. Figure 5(f) shows the pulse sequence (Q1 is the input qubit and Q2 is the output qubit) and the measured results for the 4 different implementations of the function U. Q2 final state correctly predicts if U is constant or balanced with about 80% probability. This experiment highlights the ability of HR1 to perform arbitrary sequences of operations leveraging frequency multiplexing and DDS-based frequency generation.

Horse Ridge 2: Qubit Drive, Readout and Gate Pulsing

Horse Ridge 2 (HR2) is the second-generation cryogenic controller SoC. It extends the capabilities of HR1 by adding qubit state readout and gate voltage pulsing functionality [39]. Figure 6 shows HR2 system architecture. The qubit drive is similar to HR1: 16 NCOs shared across two polar modulators allow for a frequency multiplexed control of up to 16 qubits, with simultaneous pulsing of any pair. To reduce memory usage, this implementation shares a single envelope template that is scaled by 2¹² 16-bit coefficients for each qubit to implement different rotations, each with 2¹² additional 22-bit phase settings for each NCO. This corresponds to 2¹⁹ programmable codewords per qubit. As for HR1, NCO phase update is also used for rotation along the Z-axis for correction of unwanted Z-rotations due to AC-Stark shift induced on adjacent qubits by frequency multiplexing [35]. To further improve fidelity, a digital FIR is integrated to create two notches in each pulse power spectral density at programmable frequencies for each qubit to suppress energy leakage to adjacent qubits. The SoC also integrates a microcontroller for increased flexibility in implementing the instruction set. The digital and DACs speed has been increased to 2.5GHz with respect to HR1 to extend the complex signal bandwidth to 2GHz. For the 4K characterization, a clock of 1.6GHz has been used to reduce power consumption.

To read the qubit state, spin-to-charge conversion is adopted [40]. As shown in Fig. 7, the qubit dot plunger gate potential Vₚ is adjusted such that the Fermi energy level (E_f) of the nearby electron reservoir (metal contact) is between the qubit spin-up and spin-down energy levels. Only if the qubit is in the spin-up state, the electron energy is sufficient to tunnel through the barrier into the reservoir and a spin-down electron successively tunnels back to the dot. This charge...
To suppress spurs, an 8th-order adjustable low-pass filter with VGA is implemented after the double-balanced IQ mixer. Fully-differential op-amps with dual-differential inputs and feedforward compensation are used in both filters and VGAs (Fig. 8). The baseband IQ signal is then sampled by two 7.5-bit 400MS/s SAR ADCs. The receiver bandwidth can be adjusted from 60 to 200MHz with 40 to 90dB total gain. The total power consumption of the receiver is 38mW. For a linear array of 7 qubits and respective SETs, 22 DACs are integrated to control all the gate potentials. Each DAC is composed of two 11-bit pMOS and nMOS current source banks to support both positive and negative pulses while saving power when idling (Fig. 9). To reduce area, a <2 radix is used in each DAC with a look-up table (LUT) to map the 11-bit code to the redundant 12-bit input. A 50Ω output driver for each DAC is used to absorb pulse reflections from the open termination at the qubit gates. Since we use square pulses, each DAC code is only updated twice for each pulse. This also saves power in the LUTs. All 22 DACs can be pulsed simultaneously with different amplitudes, which is critical to cancel crosstalk due to capacitive coupling among the gates on the qubit chip. Each DAC static power consumption is 2.3mW, limited by the 50Ω output impedance of the driver.

For a linear array of 7 qubits and respective SETs, 22 DACs are integrated to control all the gate potentials. Each DAC is composed of two 11-bit pMOS and nMOS current source banks to support both positive and negative pulses while saving power when idling (Fig. 9). To reduce area, a <2 radix is used in each DAC with a look-up table (LUT) to map the 11-bit code to the redundant 12-bit input. A 50Ω output driver for each DAC is used to absorb pulse reflections from the open termination at the qubit gates. Since we use square pulses, each DAC code is only updated twice for each pulse. This also saves power in the LUTs. All 22 DACs can be pulsed simultaneously with different amplitudes, which is critical to cancel crosstalk due to capacitive coupling among the gates on the qubit chip. Each DAC static power consumption is 2.3mW, limited by the 50Ω output impedance of the driver.

For a linear array of 7 qubits and respective SETs, 22 DACs are integrated to control all the gate potentials. Each DAC is composed of two 11-bit pMOS and nMOS current source banks to support both positive and negative pulses while saving power when idling (Fig. 9). To reduce area, a <2 radix is used in each DAC with a look-up table (LUT) to map the 11-bit code to the redundant 12-bit input. A 50Ω output driver for each DAC is used to absorb pulse reflections from the open termination at the qubit gates. Since we use square pulses, each DAC code is only updated twice for each pulse. This also saves power in the LUTs. All 22 DACs can be pulsed simultaneously with different amplitudes, which is critical to cancel crosstalk due to capacitive coupling among the gates on the qubit chip. Each DAC static power consumption is 2.3mW, limited by the 50Ω output impedance of the driver.

For a linear array of 7 qubits and respective SETs, 22 DACs are integrated to control all the gate potentials. Each DAC is composed of two 11-bit pMOS and nMOS current source banks to support both positive and negative pulses while saving power when idling (Fig. 9). To reduce area, a <2 radix is used in each DAC with a look-up table (LUT) to map the 11-bit code to the redundant 12-bit input. A 50Ω output driver for each DAC is used to absorb pulse reflections from the open termination at the qubit gates. Since we use square pulses, each DAC code is only updated twice for each pulse. This also saves power in the LUTs. All 22 DACs can be pulsed simultaneously with different amplitudes, which is critical to cancel crosstalk due to capacitive coupling among the gates on the qubit chip. Each DAC static power consumption is 2.3mW, limited by the 50Ω output impedance of the driver.

For a linear array of 7 qubits and respective SETs, 22 DACs are integrated to control all the gate potentials. Each DAC is composed of two 11-bit pMOS and nMOS current source banks to support both positive and negative pulses while saving power when idling (Fig. 9). To reduce area, a <2 radix is used in each DAC with a look-up table (LUT) to map the 11-bit code to the redundant 12-bit input. A 50Ω output driver for each DAC is used to absorb pulse reflections from the open termination at the qubit gates. Since we use square pulses, each DAC code is only updated twice for each pulse. This also saves power in the LUTs. All 22 DACs can be pulsed simultaneously with different amplitudes, which is critical to cancel crosstalk due to capacitive coupling among the gates on the qubit chip. Each DAC static power consumption is 2.3mW, limited by the 50Ω output impedance of the driver.

For a linear array of 7 qubits and respective SETs, 22 DACs are integrated to control all the gate potentials. Each DAC is composed of two 11-bit pMOS and nMOS current source banks to support both positive and negative pulses while saving power when idling (Fig. 9). To reduce area, a <2 radix is used in each DAC with a look-up table (LUT) to map the 11-bit code to the redundant 12-bit input. A 50Ω output driver for each DAC is used to absorb pulse reflections from the open termination at the qubit gates. Since we use square pulses, each DAC code is only updated twice for each pulse. This also saves power in the LUTs. All 22 DACs can be pulsed simultaneously with different amplitudes, which is critical to cancel crosstalk due to capacitive coupling among the gates on the qubit chip. Each DAC static power consumption is 2.3mW, limited by the 50Ω output impedance of the driver.

For a linear array of 7 qubits and respective SETs, 22 DACs are integrated to control all the gate potentials. Each DAC is composed of two 11-bit pMOS and nMOS current source banks to support both positive and negative pulses while saving power when idling (Fig. 9). To reduce area, a <2 radix is used in each DAC with a look-up table (LUT) to map the 11-bit code to the redundant 12-bit input. A 50Ω output driver for each DAC is used to absorb pulse reflections from the open termination at the qubit gates. Since we use square pulses, each DAC code is only updated twice for each pulse. This also saves power in the LUTs. All 22 DACs can be pulsed simultaneously with different amplitudes, which is critical to cancel crosstalk due to capacitive coupling among the gates on the qubit chip. Each DAC static power consumption is 2.3mW, limited by the 50Ω output impedance of the driver.

For a linear array of 7 qubits and respective SETs, 22 DACs are integrated to control all the gate potentials. Each DAC is composed of two 11-bit pMOS and nMOS current source banks to support both positive and negative pulses while saving power when idling (Fig. 9). To reduce area, a <2 radix is used in each DAC with a look-up table (LUT) to map the 11-bit code to the redundant 12-bit input. A 50Ω output driver for each DAC is used to absorb pulse reflections from the open termination at the qubit gates. Since we use square pulses, each DAC code is only updated twice for each pulse. This also saves power in the LUTs. All 22 DACs can be pulsed simultaneously with different amplitudes, which is critical to cancel crosstalk due to capacitive coupling among the gates on the qubit chip. Each DAC static power consumption is 2.3mW, limited by the 50Ω output impedance of the driver.

For a linear array of 7 qubits and respective SETs, 22 DACs are integrated to control all the gate potentials. Each DAC is composed of two 11-bit pMOS and nMOS current source banks to support both positive and negative pulses while saving power when idling (Fig. 9). To reduce area, a <2 radix is used in each DAC with a look-up table (LUT) to map the 11-bit code to the redundant 12-bit input. A 50Ω output driver for each DAC is used to absorb pulse reflections from the open termination at the qubit gates. Since we use square pulses, each DAC code is only updated twice for each pulse. This also saves power in the LUTs. All 22 DACs can be pulsed simultaneously with different amplitudes, which is critical to cancel crosstalk due to capacitive coupling among the gates on the qubit chip. Each DAC static power consumption is 2.3mW, limited by the 50Ω output impedance of the driver.

For a linear array of 7 qubits and respective SETs, 22 DACs are integrated to control all the gate potentials. Each DAC is composed of two 11-bit pMOS and nMOS current source banks to support both positive and negative pulses while saving power when idling (Fig. 9). To reduce area, a <2 radix is used in each DAC with a look-up table (LUT) to map the 11-bit code to the redundant 12-bit input. A 50Ω output driver for each DAC is used to absorb pulse reflections from the open termination at the qubit gates. Since we use square pulses, each DAC code is only updated twice for each pulse. This also saves power in the LUTs. All 22 DACs can be pulsed simultaneously with different amplitudes, which is critical to cancel crosstalk due to capacitive coupling among the gates on the qubit chip. Each DAC static power consumption is 2.3mW, limited by the 50Ω output impedance of the driver.

For a linear array of 7 qubits and respective SETs, 22 DACs are integrated to control all the gate potentials. Each DAC is composed of two 11-bit pMOS and nMOS current source banks to support both positive and negative pulses while saving power when idling (Fig. 9). To reduce area, a <2 radix is used in each DAC with a look-up table (LUT) to map the 11-bit code to the redundant 12-bit input. A 50Ω output driver for each DAC is used to absorb pulse reflections from the open termination at the qubit gates. Since we use square pulses, each DAC code is only updated twice for each pulse. This also saves power in the LUTs. All 22 DACs can be pulsed simultaneously with different amplitudes, which is critical to cancel crosstalk due to capacitive coupling among the gates on the qubit chip. Each DAC static power consumption is 2.3mW, limited by the 50Ω output impedance of the driver.

For a linear array of 7 qubits and respective SETs, 22 DACs are integrated to control all the gate potentials. Each DAC is composed of two 11-bit pMOS and nMOS current source banks to support both positive and negative pulses while saving power when idling (Fig. 9). To reduce area, a <2 radix is used in each DAC with a look-up table (LUT) to map the 11-bit code to the redundant 12-bit input. A 50Ω output driver for each DAC is used to absorb pulse reflections from the open termination at the qubit gates. Since we use square pulses, each DAC code is only updated twice for each pulse. This also saves power in the LUTs. All 22 DACs can be pulsed simultaneously with different amplitudes, which is critical to cancel crosstalk due to capacitive coupling among the gates on the qubit chip. Each DAC static power consumption is 2.3mW, limited by the 50Ω output impedance of the drive.
HR2 is also implemented in Intel 22nm FinFET technology. Figure 10 shows the 16mm² die micrograph integrating more than 10M transistors. The PCB hosting the FCBBGA packaged SoC is enclosed in a two-piece copper enclosure for thermalization at 4K. Thermal resistance is minimized by a large number of ground bumps and BGA balls and an array of indium springs for direct back-side silicon contact. Figure 11 shows HR2 assembled in the copper enclosure and installed on the 4K plate of the dilution refrigerator. The SoC has been fully characterized at 4K inside the fridge, clocked at 1.6GHz [39]. All the following measurements are reported at 4K, unless noted otherwise. As shown in Fig. 12, RF drive achieves >10dBm peak power and >-15dBm power for IM3<-50dBc over 11-17GHz. Measured SNR over 25MHz bandwidth is >44dB, meeting the 99.99% single-qubit gate fidelity target [35]. Receiver linearity has been tested with a 6-tone signal (Fig. 13), showing intermodulation products <30dBc, sufficient for a 99.9% readout fidelity [39]. The measured receiver noise temperature is 44K, which we believe might be limited by device self-heating [42]. Analysis of readout fidelity based on receiver noise figure and other parameters can be found in [39]. A loop-back configuration is used to test qubit readout (Fig. 14). The TX generates a pulse at 375MHz for 6µs while the RX receives for 12µs. The TX signal is first routed outside the fridge, attenuated and fed back to the RX with -85dBm input power to match the level expected for RF reflectometry. After amplification, IQ downconversion with a 400MHz LO and further amplification and filtering at baseband, the signal is sampled by the ADCs and processed by the digital demodulator (Fig. 14). One of the 6 channels is used to downconvert the complex tone to DC by digital mixing. Each IQ sample is then filtered with a moving average and binned into a 32-32 Re/Im plane. As can be seen in Fig. 14, the cluster of points corresponding to the -85dBm signal is clearly distinguishable. By setting a total count threshold for a specific region of interest (ROI in Fig. 14), the qubit state corresponding to signal reflection can be detected.

The table in Fig. 15 summarizes the performance and compares features of HR1 and HR2 SoCs. Substantial improvement can be seen in HR2 digital power consumption, thanks to an optimized architecture and implementation of clock gating.

**Future Directions and Conclusion**

As state-of-the-art quantum processors comprise only a few qubits (<100), each qubit can still be individually wired from the cryogenic chamber to the room-temperature electronics. Current setups even allow for a few cables per qubit to provide different signals and functionalities, such as biasing, microwave drive for single-qubit operations, and fast pulsing for 2-qubit gates [43]. Consequently, the use of cryogenic electronics is minimal in most quantum-computer prototypes and typically limited to a few low-noise cryogenic amplifiers that exploit the low-temperature to ensure the signal-to-noise ratio in the read-out [43][44]. Not requiring the development of integrated circuits, this approach offers fast deployment and flexibility. However, the complexity of room-temperature electronics has already forced the shift from general-purpose off-the-shelf equipment to specialized discrete electronics implementations. Furthermore, ensuring the reliable operation of hundreds of cables through the multiple stages of a dilution refrigerator from room-temperature down to 100mK is an engineering feat in itself, and it is arguable that we are already very close to the limits imposed by refrigerator size, wiring heat load, and reliability.

A first approach to overcome the wiring bottleneck is re-using each cable for multiple functions, e.g., exploiting frequency multiplexing on a single line to drive single-qubit operations on multiple qubits, as proposed in the Horse Ridge SoCs. Although this can alleviate the issue in the short term, it is unlikely to allow scaling from today’s ~100 qubits to the required millions of qubits, due to inherent multiplexing limits, e.g., crosstalk and limited bandwidth for frequency multiplexing. A more promising alternative is arranging N² qubits in a 2D array and addressing them using a cross-bar scheme requiring ~N bit and word lines, similar to standard semiconductor memories [45]. While cross-bar addressing can bring more aggressive scalability of the control signals, it imposes stringent requirements on the uniformity of the quantum devices, which is extremely challenging in itself.

As a consequence, an increasing number of researchers are developing cryogenic interfaces, as previously discussed, with more and more functionalities of the electronic control interface being
integrated for operation at cryogenic temperature. Those ICs are typically operated around 4K, a temperature at which existing refrigerators offer a cooling capability of a few Watts. Although serving thousands of qubits or more will surpass such capability, the cooling power of the cooling infrastructure may be enormously improved at 4K by adopting custom-made refrigerators [46]. However, operating the electronics above 1K due to the cooling requirements also forces the electronics to be far from the qubits, which typically operate at 20–100mK on a different temperature stage, thus still resulting in a wiring bottleneck. The straightforward solution would be to design “hot” qubits operating at the same temperature of the electronics [47], and encouraging results in this direction have been shown, demonstrating silicon spin qubits operating above 1K [5][6]. As an alternative, an electronic front-end with simple functionality, and hence with power requirements compatible with the mK cryogenic stage (typically < 1mW at 100mK), can be operated close to the sub-K qubits to just multiplex the electronic signals, thus alleviating the wiring bottleneck. An example is shown in [25].

Even with the electronic interface and the quantum processor operating at the same temperature, the challenge remains on how to connect the classical and the quantum section. 3D packaging of the electronics in separate dies to exploit processes optimized for the electronics and high-speed low-power performance coupled with a mature EDA infrastructure and the potential for qubit co-integration, thus representing a suitable candidate to support future large-scale quantum computers.

Acknowledgments

References

<table>
<thead>
<tr>
<th>Qubit platform</th>
<th>HR1</th>
<th>HR2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transmons + spin qubits</td>
<td>Spin qubits</td>
<td></td>
</tr>
<tr>
<td>Controller capability</td>
<td>Drive</td>
<td>Drive, Readout with Digital Detector, Gate Pulsing, µ-Controller</td>
</tr>
<tr>
<td>Power</td>
<td>Analog</td>
<td>Drive: 5.2mW/qubit²</td>
</tr>
<tr>
<td>Digital</td>
<td>Digital: 330mW @ 1GHz clock</td>
<td>Readout TX/RX: 1.3/9.3mW/qubit², Gate pulsing: 2.9mW/channel</td>
</tr>
<tr>
<td>Chip area</td>
<td>4mm²</td>
<td>16mm²</td>
</tr>
<tr>
<td>Technology</td>
<td>22nm FinFET CMOS</td>
<td>22nm FinFET CMOS</td>
</tr>
</tbody>
</table>

Fig. 15. HR1 and HR2 performance summary and comparison.

### Table 1.

<table>
<thead>
<tr>
<th># Channels</th>
<th>22 (simultaneous)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq. range</td>
<td>2 – 20GHz @ &gt;-45dBm</td>
</tr>
<tr>
<td>Freq. multiplexing</td>
<td>11 – 17GHz @ &gt;10dBm</td>
</tr>
<tr>
<td>Sampling rate</td>
<td>32, 16 qubits</td>
</tr>
<tr>
<td>IM3</td>
<td>1G/s, up to 2.5G/s</td>
</tr>
<tr>
<td>Data bandwidth</td>
<td>&lt;50dBc @ -18dBm (6.25GHz)</td>
</tr>
<tr>
<td>Digital FIR</td>
<td>&lt;50dBc @ &gt;17dBm (11 to 17GHz)</td>
</tr>
<tr>
<td>Envelope size</td>
<td>Yes, 2 notches per qubit</td>
</tr>
<tr>
<td>Instruction set</td>
<td>1GHz</td>
</tr>
<tr>
<td>Output impedance</td>
<td>50Ω</td>
</tr>
<tr>
<td>LO IQ generation</td>
<td>External off-chip hybrid (PCB)</td>
</tr>
<tr>
<td>Output rms noise</td>
<td>200μV rms</td>
</tr>
<tr>
<td>Output impedance</td>
<td>50Ω</td>
</tr>
<tr>
<td>TX output freq.</td>
<td>400MHz LO, 100MHz baseband</td>
</tr>
<tr>
<td>TX output power</td>
<td>14pF load</td>
</tr>
<tr>
<td>RX RF freq.</td>
<td>1GHz up to 2GHz</td>
</tr>
<tr>
<td>RX noise temperature</td>
<td>4K</td>
</tr>
<tr>
<td>RX baseband BW</td>
<td>60 – 200MHz</td>
</tr>
<tr>
<td>Filter order</td>
<td>8th</td>
</tr>
<tr>
<td>ADC</td>
<td>SAR, 7.5 bits, 400MS/s</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Qubit state detector</th>
<th>Integrated</th>
</tr>
</thead>
</table>

### Table 2.

<table>
<thead>
<tr>
<th>Freq. multiplexing</th>
<th>Yes, 8 qubits</th>
</tr>
</thead>
<tbody>
<tr>
<td>TX output freq.</td>
<td>DC – 0.6GHz</td>
</tr>
<tr>
<td>TX output power</td>
<td>-70 to -400mW</td>
</tr>
<tr>
<td>RX RF freq.</td>
<td>200 – 600MHz</td>
</tr>
<tr>
<td>RX gain</td>
<td>40 – 90dB</td>
</tr>
<tr>
<td>RX noise temperature</td>
<td>4K</td>
</tr>
<tr>
<td>RX baseband BW</td>
<td>60 – 200MHz</td>
</tr>
<tr>
<td>Filter order</td>
<td>8th</td>
</tr>
<tr>
<td>ADC</td>
<td>SAR, 7.5 bits, 400MS/s</td>
</tr>
</tbody>
</table>

### Table 3.