Interfacing spin qubits in quantum dots and donors—hot, dense, and coherent

Vandersypen, Lieven; Bluhm, H; Clarke, J. S.; Dzurak, A. S.; Ishihara, Ryoichi; Morello, A.; Reilly, D. J.; Schreiber, L. R.; Veldhorst, Menno

DOI
10.1038/s41534-017-0038-y

Publication date
2017

Document Version
Final published version

Published in
NPJ Quantum Information

Citation (APA)
Vandersypen, L. M. K., Bluhm, H., Clarke, J. S., Dzurak, A. S., Ishihara, R., Morello, A., ... Veldhorst, M. (2017). Interfacing spin qubits in quantum dots and donors—hot, dense, and coherent. NPJ Quantum Information, 3(1), 34. https://doi.org/10.1038/s41534-017-0038-y

Important note
To cite this publication, please use the final published version (if applicable). Please check the document version above.
Interfacing spin qubits in quantum dots and donors—hot, dense, and coherent

L. M. K. Vandersypen¹,², H. Bluhm³, J. S. Clarke², A. S. Dzurak⁴, R. Ishihara¹, A. Morello⁴, D. J. Reilly⁵, L. R. Schreiber³ and M. Veldhorst¹

Semiconductor spins are one of the few qubit realizations that remain a serious candidate for the implementation of large-scale quantum circuits. Excellent scalability is often argued for spin qubits defined by lithography and controlled via electrical signals, based on the success of conventional semiconductor integrated circuits. However, the wiring and interconnect requirements for quantum circuits are completely different from those for classical circuits, as individual direct current, pulsed and in some cases microwave control signals need to be routed from external sources to every qubit. This is further complicated by the requirement that these spin qubits currently operate at temperatures below 100 mK. Here, we review several strategies that are considered to address this crucial challenge in scaling quantum circuits based on electron spin qubits. Key assets of spin qubits include the potential to operate at 1 to 4 K, the high density of quantum dots or donors combined with possibilities to space them apart as needed, the extremely long-spin coherence times, and the rich options for integration with classical electronics based on the same technology.

npj Quantum Information (2017)3:34; doi:10.1038/s41534-017-0038-y

INTRODUCTION
The quantum devices in which quantum bits are stored and processed will form the lowest layer of a complex multi-layer system.¹–³ The system also includes classical electronics to measure and control the qubits, and a conventional computer to control and program these electronics. Increasingly, some of the important challenges involved in these intermediate layers and how they interact have become clear, and there is a strong need for forming a picture of how these challenges can be addressed.

Focusing on the interface between the two lowest layers of a quantum computer, each of the quantum bits must receive a long sequence of externally generated control signals that translate to the steps in the computation. Furthermore, given the fragile nature of quantum states, large numbers of quantum bits must be read out periodically to check whether errors occurred along the way, and to correct them.⁴ Such error correction is possible provided the probability of error per operation is below the accuracy threshold, which is around 1% for the so-called surface code. While many qubit architectures and strategies for scaling have been proposed,¹³–¹⁷ the structure of these transistors bears a lot of resemblance with that of promising semiconductor-based qubits.⁵, ¹⁰ However, an important difference is that conventional processor chips have only ≈10⁵ input-output connections (IO’s), for instance Intel’s land grid array 2011 socket has 2011 pins that contact the backside of the processor (http://www.intel.nl/content/www/nl/nl/processors/core/core-i7-iga-2011-datasheet-vol-1.html). This brings the transistor-to-IO ratio over 10⁵. This scaling of the number of pins with the number of devices is empirically described by Rent’s rule.¹¹, ¹² In the absence of multiplexing or on-chip control logic, the limit for the qubit count is probably similar to the pin-limit of the package, which is currently around 10⁶ (http://www.intel.nl/content/www/nl/nl/processors/core/core-i7-iga-2011-datasheet-vol-1.html).

Therefore, the notion that semiconductor quantum bits that are manufactured by complementary-metal-oxide-semiconductor (CMOS)-compatible technology are easily scalable, is too simplistic. While many qubit architectures and strategies for scaling have been proposed,¹³–⁴⁰ a completely worked out pathway to create qubit systems that can be expanded to a large-scale quantum processor yet has to be defined and a key step is the design of a scalable classical-quantum interface.

Here, we focus on the quantum-classical interface requirements and possible solutions for qubits encoded in electron spins in semiconductor quantum dots and donors.⁹, ¹⁰ We thereby consider specifically quantum dots that are probed and controlled using electrical signals, referring to ref. ¹ for a discussion of optically addressed quantum dots. Electrically controlled quantum...
dots and donors are two promising qubit realizations that have much in common both conceptually and in terms of qubit specifications and hardware requirements. There is significant scope to make these realizations compatible with industrial CMOS technology, which is optimized for high-yield, reproducibility and cleanliness. Indeed, there is a lot of effort in this direction and qubits that are partly fabricated with industrial technology have already been realized.41

We begin with a brief summary of electron spin qubits in quantum dots and donors, then derive the control signal requirements and challenges, and next present possible solutions to overcome these challenges. These focus on dense 2D tunnel coupled spin qubit arrays, sparse arrays with coherent links between them, and on the possibility of operating spin qubits at 1 or 4 K, allowing for more complex electronics to be integrated with the qubits.

ELECTRON SPIN QUBITS IN QUANTUM DOTS OR DONORS
We first briefly introduce electron spin qubits in electrically detected quantum dots and donors as a starting point for discussing the control and interfacing requirements (for more extensive reviews, see refs 9, 10).

A schematic of a prototypical quantum dot device is shown in Fig. 1a. A combination of bandgap offsets and electrostatic gates are used to confine one or more free electrons (or holes)41–47; for brevity we will refer to electrons throughout the text) in a small space in a semiconductor, typically a few tens of nm in diameter. For qubit experiments, the gate voltages are usually tuned so the quantum dots contain exactly one electron each, although for certain initialization and read-out protocols, an electron is pushed off a dot or onto a neighboring dot. Figure 1b shows a schematic of a donor-based device. Donor atoms such as phosphorous in silicon have one excess electron compared to the atoms in the surrounding lattice, and at low temperatures this electron is bound to the donor atom (acceptors with one excess hole can be used as well; we will just refer to donors for brevity). With a gate voltage, this electron can be pushed off the donor or a second electron can be bound to the donor, provided the required electric fields are below values that result in population of the excited state.49 The quantum-point-contact (QPC) or single-electron-transistor (SET) is used to probe the number of charges on the dots. They could potentially be avoided via gate-based dispersive read-out57.

Regardless of the chosen qubit encoding, one generally requires the ability to individually rotate every qubit about two
different axes in the corresponding qubit Bloch sphere, and to entangle neighboring qubits with each other; see Fig. 2 for rotation axes of different qubit encodings. Altogether, this forms a universal set of quantum gates, which can be used to perform arbitrary logic.25 Both single-qubit and two-qubit gates can be accomplished in one of two modes: (1) fast gate voltage pulses that rapidly switch the Hamiltonian so that the qubit(s) will start evolving around a new axis (in Hilbert space) or (2) radio-frequency or microwave-frequency electric or magnetic fields resonant with the energy difference between specific single- or two-qubit states. Gate durations vary from sub-ns to microsecond timescales.9, 10

Spin states are hard to detect directly, but can be converted to charge states via a sequence whereby a charge movement between dots or from a dot to a nearby electron reservoir is made to be spin state dependent, via “spin-to-charge conversion”.53, 54 Simultaneous single-charge detection then reveals what the spin state was before the measurement. Real-time single-charge detection can be accomplished in several ways. In the first method, the conductance through a nearby charge detector is probed, either at baseband55 or via radio frequency (RF) modulation.95 The charge detector can be a narrow channel called quantum point contact (QPC) or a small island that itself is capacitively coupled to the quantum dot or donor. In either case, the conductance through it directly depends on the charge occupation of the dot or donor (see Fig. 1a, b). Alternatively, the ability of charges to move back and forth in response to an oscillating excitation can be probed. This amounts to an electrical susceptibility measurement, which is commonly implemented by looking at the reflection of an RF signal applied to one of the quantum dot gates56 or reservoirs.58 Single-shot measurement times down to 200 ns have been achieved in specific settings,59 and read-out fidelities as high as 99.8% have been reported.60 Qubit reset or initialization could be achieved by thermalization to the ground state, but that would be very slow given that spin relaxation times are often in the millisecond to second range.9, 10 Faster approaches include initialization by measurement52 and spin-selective tunneling from an electron reservoir or dot to a dot or donor.54, 61, 62

Finally, we note that microscopic variations in the semiconductor substrate and non-uniformities in the gate patterns lead to substantial variations from site to site in a realistic device. While progress has been made and high-quality double quantum dots have been reported, an attractive but challenging solution would be to reach a uniformity level where a common (set of) DC voltage(s) would suffice to place each of several quantum dots in the desired configuration; for example, systematically having a dot-to-dot variation in required gate voltage for single electron occupancy smaller than the charging energy. Donor fabrication introduces more challenges, but the strong confining potential can have specific advantages here due to the intrinsic large energy scales. Fabrication based on scanning-tunneling-microscopy64 as compared to ion implantation has the further advantage that uncertainties in donor placement and capacitive coupling to nearby stray donors are significantly reduced. However, a systematic study on the relevant variations for a large array is missing. Furthermore, nominally identical operations currently require DC gate voltages, gate voltage pulses, and
microwave control signals that all differ in amplitude or duration from qubit to qubit.

**CONTROL SIGNAL REQUIREMENTS**

The discussion of electron spin qubits in quantum dots or donors leads us to the following commonly recurring requirements for the control signals. As can be seen from Fig. 2, not all requirements apply to each of the encodings, and this can be a criterion for comparing the merits of different encodings with each other.

1. an independently calibrated and tuned DC gate voltage on every site (typically up to ±1 V)
2. independently calibrated and tuned gate voltage pulses on every site (typically up to tens of mV and with sub-ns rise times)
3. independently calibrated and tuned microwave magnetic or electric fields at every site (typically –40 to –20 dBm, 1–50 GHz bursts of 10 ns to 1 µs duration)
4. a high precision of each of the control signals to achieve error rates comfortably below the 1% accuracy threshold
5. initialization, operations and read-out on timescales short compared to the relevant decoherence time.

We now examine some of these requirements in more detail, and in particular consider which requirements can be relaxed. In the next section, we will present some general guidance for meeting the necessary requirements.

For the pulsed control signals, often only one of the two pulse stages requires precise tuning. For instance, the precise strength of the exchange interaction is important when the exchange is turned “on”, but the exchange strength in the “off” state merely needs to be below some threshold, which is a much more relaxed constraint. The exchange is commonly controlled by changing gate voltages along the so-called detuning axis that controls the relative potential of neighboring dots (see Fig. 2). To reach the “off” state, it suffices to pulse gate voltages to a sufficiently far detuned condition. Similarly, when performing exchange gates at the so-called symmetry point (see the caption of Fig. 2), it suffices to pulse the gate voltages to any condition where the residual exchange is sufficiently suppressed, though this may require larger voltage amplitudes than when pulses the detuning. Similarly, accurate level alignment is needed during read-out of a single spin based on spin-selective tunneling to a reservoir, but when not reading out it suffices to stay in the regime with one electron per site. Spin read-out of two-electron spin states is typically even more forgiving, as it suffices to pulse from somewhere deep in the regime with one electron on each dot, to somewhere in the so-called pulse triangle with two electrons on one of the dots. Therefore, one could imagine that voltage pulses to, say, control exchange gates or initiate read-out can be made uniform across multiple (all) dots, by fine-tuning the exact qubit operating points via DC bias voltages. The main assumption in these examples is that the qubit is not sensitive to the exact DC gate voltage while in the “off” state. As the qubit transition frequency may in fact vary with DC gate voltage, unintentional single-qubit z-rotations could occur and these must be tracked or corrected separately for every qubit.

For microwave control signals, we need to separately consider the microwave frequency vs. amplitude and duration. The simplest approach is to assume that all qubits will need to be resonant with either a single frequency or a small number of frequencies. This can be achieved by $g$-factor control or Stark shifting, through either DC or pulsed control voltages, to bring qubits on specific sites in or out of resonance with the excitation. For conventional electron spin resonance (ESR) whereby a global microwave magnetic field is applied, the same microwave can be used to achieve the same angle of rotation on multiple qubits provided the amplitude variations are sufficiently small and the resonance frequency of all qubits resonant with the excitation is sufficiently uniform. Uncontrollable spin-orbit coupling renormalizing the $g$-factor can give qubit-to-qubit variations in the resonance frequency of order 10 to 100 MHz at $B = 1.5$ T (Ref. 67). A possible strategy to overcome such variations is operating at significantly lower magnetic field. Globally applied alternating current (ac) magnetic fields could give rise to excessive dissipation and heating, and the magnetic field profile may suffer from distortions due to all the metal interconnects. A strategy could be to integrate local microwave lines that are close to the qubits and only address subsections of the larger qubit array. Superconducting lines could further reduce dissipation.

For electric-dipole spin resonance, whether based on intrinsic spin–orbit interaction or on local magnetic field gradients to allow electric fields to drive spin transitions, or on local magnetic field gradients to allow electric fields to drive spin transitions, dot-to-dot variations in the confining potential may impose different microwave amplitudes for every qubit. All-electrical control is often argued to be beneficial because of fast and local control. Essential in the design will be the interconnection between the microwave source and the individual qubits. Power dissipation will be significantly reduced compared to ESR, but avoiding cross-talk will be challenging. A solution for cross-talk could include to spatially separate qubits with equal resonance frequency.

The main message from this technical discussion is that even though some requirements can be relaxed, especially if the quantum dot properties are homogeneous, at least a subset of signals (DC, pulsed, or microwave) will need to be independently calibrated and delivered to each and every qubit.

**CONTROL SIGNAL WIRING SOLUTIONS**

How can we route qubit-specific classical control signals to a large number of quantum dot or donor qubits? The common understanding in the field is that directly connecting via wires or coax lines say $10^5$ sub-100 mK qubits to room temperature voltage sources, pulse sources, and microwave sources, is impractical for several reasons. At the qubit chip level, it conflicts with Rent’s rule in classical systems and practical limits to the number of pins on a chip. At the level of the transmission lines from room temperature to the chip, heat transport causes a heat load of a few mW on the 4 K plate. For comparison, cooling powers of currently used pulse tube systems are in the range of a few W at 4 K. Below 4 K, superconducting lines can be used, which are poor thermal conductors and thus minimize heat load, but power dissipated in the attenuators can heat up the coldest parts of the dilution refrigerator. A common view is that instead a combination of two ingredients will be required:

1. Multiplexing strategies
2. A first layer of classical electronics residing next to the qubits and commensurate with the inter-qubit spacing

Other layers of classical electronics may reside farther away from the qubit plane and at higher temperature, as the data rates between layers higher up in the quantum computer architecture are orders of magnitude smaller than those between the physical qubits and the first control layer.

Within this framework, important choices include

1. What qubit density to work with?
2. At what temperature do the qubits reside? Is operation at 1.5 or 4 K possible?
3. What is the functionality of the first electronics layer?
4. What specifications must the electronics meet (clock speed, noise, resolution, frequency range, memory, power dissipation)?
These questions are interrelated, for instance the qubit density and the cooling power (which depends on the temperature) impact the functionality and specifications of the electronics that can be achieved. We next discuss platforms based on a dense qubit array or a sparse qubit array, and an operation temperature ranging from 100 mK to 4 K.

Dense qubit array and cross-bar addressing

The most widely used mechanism for two-qubit gates using quantum dots is based on the exchange interaction. This interaction couples the spin states of two electrons when their respective wave functions overlap, i.e., when the respective dots are tunnel coupled. The two-qubit exchange gate is very fast: it can be operated on sub-ns timescales, limited in practice by the bandwidth of the control electronics rather than by the underlying physics. In the absence of nuclear spin noise that is mostly relevant in III–V quantum dots, the fidelity is often limited by electrical noise, usually charge noise from the amorphous materials and interfaces, and electrical noise on the gates. Coherent spin exchange between neighboring spins has been extensively realized in double dots as well as in linear arrays of three dots, and scaling up a linear array to larger sizes is relatively straightforward.

Scaling to a large two-dimensional array of tunnel coupled dots presents a great opportunity for realizing highly dense qubit arrays, but also presents practical challenges to wire up all the qubits, given the geometric constraints. To make things concrete, in order to have sufficient tunnel coupling between neighbors in the array, the center-to-center distance between dots must be no more than a few 100 nm in GaAs, <100 nm for qubits defined in silicon (a result of the three times larger effective mass in Si), and <20 nm for donors (due to the strong confinement potential). Taking a 100 nm qubit-to-qubit pitch as an example, a 2D array of 1 mm$^2$ would allow space for a massive 10$^8$ qubits. However, with at least one control line per qubit, fan-out of the control lines requires multiple layers of interconnects. If the interconnects and control lines can come from all sides, but have the same width as the quantum dots, a 4 × 4 array requires two layers (the first layer to contact the outer dots, the second to contact the inner dots), five layers are needed for a 10 × 10 array and 50 layers for a 100 × 100 array. There is clearly a practical limit to the size of a monolithic array that can be wired up in this way.

A possible strategy to partly overcome this fan-out problem is to borrow concepts from dynamic random access memory (DRAM). Rather than connecting every gate continuously to a voltage source, an individual gate is connected to a capacitor that stores the desired voltage. The voltages can be set efficiently via a cross-bar addressing scheme (Fig. 3). Given that a single electron charge $e$ is the smallest amount of charge that can be added to the capacitor, the capacitance required to achieve a gate voltage resolution $\Delta V$ must be $C > |e|/\Delta V$. For $\Delta V = 1 \text{ mV}$, this gives $C > 160 \text{ fF}$. Furthermore, thermal noise in the circuit when the switch is closed translates to an uncertainty in the gate voltage given by $V_{\text{rms}} = \sqrt{kT/C}$, which is a function of the capacitance but independent of the circuit resistance. Reaching an uncertainty $V_{\text{rms}} = 1 \text{ mV}$ at a temperature of 50 mK would require a capacitance larger than 800 fF. One challenge with this approach is the chip area required for such relatively large capacitances. Conventional planar approaches with 10 fF/μm$^2$ are not compatible with the envisioned small dot spacing so that other solutions such as concentric pillar capacitors will be needed.

These charge-storage electrodes may have to be periodically refreshed, due to leakage or variations in the capacitive coupling to nearby structures. Such refreshing is routinely done in classical electronics. For instance, a typical refreshing interval time of DRAM is 64 ms where a refresh cycle is performed within 30 ns. If the 1% weakest electrodes can be excluded, the interval time can be extended to a second. While the tolerances of quantum dot voltages are much more stringent, leakage is strongly reduced at a few Kelvin or below, so such an approach might be feasible. Experimental drifts of approximately one Coulomb per hour ~8 mV/h have already been observed in charge-storage electrodes integrated with quantum devices. However, more research is needed to demonstrate these drifts using electrodes that have a size comparable to the quantum dots and to minimize possible leakage pathways.

Globally controlling these floating electrodes could be done via an efficient cross-bar addressing scheme, using horizontal and vertical control lines that each have a spacing corresponding to the dot-to-dot distance. Assuming a dot-to-dot pitch of 50 nm, consistent with requirements for quantum dots, would imply an interconnect pitch of 50 nm, which is similar to what is possible with 14 nm node technology, the most advanced that is commercially available today (http://www.intel.com/content/www/us/en/silicon-innovations/intel-14nm-technology.html).

Furthermore, 50 nm is below the 70 nm transistor gate pitch for the 14 nm node. Therefore, unless dot dimensions can be kept slightly larger, integrating a single transistor above every quantum dot requires continued scaling of conventional CMOS devices, dictated by Moore’s Law.

A cross-bar approach can also provide a relatively economical avenue for qubit control. For instance, we can apply a voltage pulse on one of the vertical lines (combined with the DC voltage required by that site via a bias-tee) and use the horizontal line to select to which qubit the pulse is applied. As discussed in the section of control signal requirements, it should be possible to allow the same pulse amplitude to induce an exchange gate or initiate read-out across multiple dots. In this case, parallel addressing of multiple dots will be possible, as well as addressing for instance all dots or half of the dots (any combination of dots compatible with cross-bar selectivity is possible). It has indeed been shown that the cross-bar approach can be used to run the surface code, both in donor and dot platforms. It was also shown that surface code variations can be implemented with reduced local control.

Initiating parallel read-out is possible with a cross-bar approach as well, with vertical lines used to select the set of qubits underneath and horizontal lines used to carry the corresponding read-out signals. It may be possible to re-use the same cross-bar that is used for control, also for read-out, for instance using dispersive gate read-out. An RF signal is then applied to a vertical line (again added to a DC gate voltage) and the horizontal lines select the qubit that is read out. This procedure comes at a
cost. In its simplest form, an array of \( N \) qubits requires \( \sqrt{N} \) repetitions of this read-out protocol to measure all the qubits. This slow-down has two sides. First, it requires that probability of error of a qubit during \( \sqrt{N} \) read-out cycles stays far below the accuracy threshold. Here, the extremely long memory times of spin qubits under dynamical decoupling, of order one second\(^{40, 60} \) are crucial. Second, it slows down the net clock cycle of the surface code operation by a factor \( \sqrt{N} \). Here, we note that it is not clear what the optimal effective clock cycle is. Too slow is not good since it slows down the computation. Too fast is not good either, since then the classical processors cannot keep up processing the massive data streams produced by the surface code syndrome measurements, and this will pose a hard boundary. This flexibility in choosing the clock cycle of the classical computer may turn out to be an important advantage of electron spin qubits over, e.g., superconducting qubits.

As a more sophisticated and efficient read-out variant, it may be possible to combine the cross-bar approach with frequency multiplexing\(^{60} \) when using RF techniques for read-out.\(^{57, 58} \) In this case, each horizontal line can carry multiple read-out signals simultaneously. The demonstrated on-chip resonators\(^{80} \) will be challenging to fit locally into a dense array. However, frequency multiplexing could also be achieved by clever crossbar operation. For example, if the gates that control the interdot tunnel coupling are connected to vertical lines, these can be frequency-modulated so that each vertical line has a different modulation frequency. The resonance frequency of the readout circuits, measured along the horizontal lines, will then shift corresponding to the respective modulation frequency. This frequency multiplexing enables simultaneous read-out along horizontal lines. Global simultaneous read-out is then obtained by connecting each horizontal line to a separate circuit or by frequency multiplexing each horizontal line. If \( k \) frequencies can be simultaneously read out, \( k/\sqrt{N} \) qubits can be read out in parallel. This gives further design flexibility and room for optimization.

An important consideration for any of the above uses of cross-bar addressing is whether power dissipation in the switching circuits is compatible with dilution refrigerator temperatures. The desired functionality of the control circuits will determine the number of required active components, the total power dissipation, and the minimum operation temperature. Dynamic power dissipation is a major source of power dissipation in classical electronics, here a single switch contributes \( P = CV^2f \), with \( C \) is the circuit capacitance, \( V \) the applied voltage, and \( f \) the activity factor relative to clock cycle with frequency \( f \). For example, if refreshing the voltage on the floating gate would involve compensating a 10 \( \mu \)V drift at 1 V gate voltage and at a conservative refresh rate of 1 MHz per qubit, dissipation would amount to 8 pW for a capacitance of 800 fF, the lowest capacitance that can give 1 \( \mu \)V noise and resolution, as discussed above. The additional power needed to drive the switching of the transistors could be dissipated at higher temperature stages. Large dilution refrigerators are now capable of providing cooling power beyond 1 mW at 100 mK. Therefore, many millions of transistors could potentially operate in combination with floating gates at the lowest temperature stage, provided they can be interconnected to higher temperature stages with dissipationless (superconducting) lines. Simple functionalities such as multiplexing strategies could become compatible with the discussion here and research to find the optimal hybrid, with essential electronics operating at the lowest temperature and all other electronics at higher temperature stages is, therefore, key to scaling spin qubits.

Sparse qubit arrays and local electronics

Several alternative spin qubit coupling mechanisms exist besides direct exchange coupling, that allow the building of two-dimensional spin qubit arrays without the need for direct tunnel coupling between neighboring qubits in four directions (north, south, east, west). Many of these mechanisms have in common that they allow the separation of the qubits by larger distances, varying from roughly 1 \( \mu \)m to roughly 1 mm. Proposals for coupling spin qubits at a distance rely on the use of superconducting resonators,\(^{53} \) ferromagnets,\(^{28} \) superconductors,\(^{27, 31} \) intermediate dots or dot arrays,\(^{19, 20, 53} \) or surface acoustic wave cavities.\(^{39} \) An alternative approach consists of shuttling electrons across the chip between distant quantum dots, where the electrons are propelled by surface acoustic waves\(^{21, 22, 81} \) or time-varying gate voltages.\(^{6, 35, 82, 83} \) Whereas with enough motivation, any of these platforms could be realized in industry cleanrooms, those that only require additional gate metal are most easily integrated with CMOS technology.

When combining coupling mechanisms at a distance with local registers of tunnel coupled qubits, a modular structure arises as illustrated in Fig. 4. Modular architectures are currently considered across a wide variety of platforms, from trapped ions to superconducting qubits to impurity spins of NV centers in diamond.\(^{86} \) Quantum error correction schemes such as the surface code can be naturally implemented on modular or distributed quantum computers. For instance by moving two logical qubits onto the same local register, two-qubit logical gates can be performed with known methods.\(^{85} \)

Widely spaced qubit arrays can alleviate fan-out and wiring problems, simply by allowing more space for routing as also seen in Fig. 4. Yet, even if this allows space to connect each qubit to one or more control lines running off the chip, we mentioned before that connecting individual qubits to sources and generators a large distance away is not viable. Therefore, the more important advantage of space between the qubits may be that it allows a first layer of control electronics that is commensurate with the inter-qubit spacing to be placed directly above or in the qubit layer. If placed above the qubit layer, this classical layer can be interfaced with the qubit layer via an interposer, flip-chip (C4) technology or similar methods. Thermal isolation between the quantum and classical chips could be provided by using superconducting vias for connection. In this way, heating of the qubits by thermal dissipation in the classical circuitry is minimized. When transistors are realized in the same plane as the qubit layer, they could be integrated directly with traditional CMOS fabrication.

Depending on the actual spacing between qubit arrays and on the power budget, the functionality of the classical layer can be more or less advanced.\(^{86} \) At the lowest level, simple multiplexing strategies based on switches can be implemented. What would have more impact is if analog to digital converters (ADC), digital to analog converters (DAC), and vector modulation could be implemented locally in the first classical layer. In this case, only digital signals must flow between the first classical layer and a second layer higher up in the control structure, potentially even at room temperature, where the digital data is processed. The required bandwidth of the communication channel between the classical layers is then much smaller, as per qubit or a few bits of information must be transmitted per clock cycle, instead of time traces containing a large number of analog data points. Even then, data rates to room temperature are substantial. For example, if \( 10^9 \) qubits are repeatedly read out at 1 \( \mu \)s intervals and each qubit measurement provides one bit of information, the data flow amounts to 100 Tb/s. Control will require a few bits and several operations per surface code cycle. Therefore, local error decoding would be highly attractive but also most demanding in terms of circuit complexity.

The feasibility of this approach hinges on a number of questions that each constitute a full research question, for which only an initial analysis has been performed to date. First estimates indicate that footprints on the order of (100 \( \mu \)m)\(^2 \) and a power budget in the microwatt range per qubit could be sufficient to implement...
reliable gate operations from the physics perspective range from footprint in perspective, convenient qubit spacings to allow (ref. 89), but a key question is whether the resulting temperature provide a cooling power of 144 kW at 4.5 K and 20 kW at 1.9 K Collider magnet system has eight cryocoolers that together operated at 1.8 K or higher (for comparison, the Large Hadron power itself would not be a severe limitation if electronics can be lead to a substantial reduction in the required footprint. The dissipation could potentially be reduced by several orders of magnitude if the transistors are fully optimized for low dissipation could potentially be reduced by several orders of magnitude if the transistors are fully optimized for low temperature operation by reducing the supply voltage.88 Cooling power itself would not be a severe limitation if electronics can be operated at 1.8 K or higher (for comparison, the Large Hadron Collider magnet system has eight cryocoolers that together provide a cooling power of 144 kW at 4.5 K and 20 kW at 1.9 K (ref. 89)), but a key question is whether the resulting temperature gradient between qubits and electronics can be maintained in conjunction with a sufficient interconnect density. Regarding error decoding, it has been shown that the best-performing surface code algorithm lends itself to parallelization,90 and that it requires about 2 μs per round and qubit on a current high performance CPU.91 Substantial improvement can be hoped for with an application-specific integrated circuit, but it remains to be seen if the resulting circuit complexity and power consumption will be acceptable. Alternatively, other decoders that are less computation-intensive may become an option,92, 93 including decoders based on neural networks.94, 95 To put the electronics footprint in perspective, convenient qubit spacings to allow reliable gate operations from the physics perspective range from 1–10 μm for capacitive couplers,23, 24, 36 from 1–100 μm for spin shuttles16, 35, 81–83 and from 100–1000 μm for superconducting resonators15, 18, 26, 29. A final consideration is that the constraints on power dissipation as well as the interconnects between electronics and qubits would be greatly reduced if spin qubits could be operated at higher temperature, without excessive compromises in the fidelity of initialization, coherent operations, memory time, and read-out. We explore this attractive possibility in more detail in the next section.

Hot qubits
Much would be gained by qubits that can operate at 1 to 4 K. At 4 K, the cooling power of a single commercial pulse tube cooler as used in qubit experiments today is 1–2 W. By comparison, powerful dilution refrigerators offer a cooling power of 1 mW at 100 mK. At T < 100 mK, we, therefore, expect that only very simple functionality can be realized without excessive heat dissipation. Superconducting classical circuits96 dissipate very little power, but are complex in design, lack the memory function, and have a large footprint. Operating spin qubits at 4 K, with a thousand-fold increase in available cooling power, makes the prospect of electronics commensurate with and right next to the qubit plane more realistic. An integrated quantum-classical structure would have multiple advantages in solving the fan-out problem, would simplify the RF wiring and reduce signal losses.

A major attraction of Si-MOS-based quantum dots and donor-based qubits is that they can have energy scales that are compatible with 1 to 4 K operation. Proper operation requires that the relevant energy scales are about five times larger than the thermal energy, which is 340 μeV at 4 K. Charging energies of donors and small quantum dots are easily in excess of 10 meV and orbital energies can be of order 10 meV as well,97 satisfying this requirement. However, in silicon there is also a valley degree of freedom. Silicon has a sixfold degeneracy due to crystal symmetry, which is broken at the interface leaving two relevant valley states. These lowest-energy valley states can be split via a sharp confinement potential, e.g., the silicon-SiO₂ or Si/SiGe interface, and a vertical electric field. In Si/SiGe devices, valley splittings are typically no more than 100 μeV in current devices.51 Possibly this energy scale can be significantly increased by reducing dot size or adopting novel growth approaches. Alternatively, the valley could be initialized using advanced methods such as a measurement-based active reset for high-temperature operation. By comparison,
in Si-MOS dots, the valley splitting has reached almost 1 meV, and could also be pushed further by reducing the device dimensions and increasing the electric field by confinement gates. This would allow initialization in the lowest-energy orbital and valley state.

Spin splittings in all spin qubit measurements to date are far below the thermal energy at 1–4 K. This would pose problems for conventional single-spin initialization and read-out schemes. Simply increasing the magnetic field and hence the spin splitting would imply impracticable qubit operation frequencies of (sub) THz and potentially too short relaxation times. Instead, high-fidelity initialization and read-out of spin states can make use of the single-dot singlet-triplet splitting, which is typically somewhat below the orbital or valley splitting (whichever is lower) due to the exchange interaction. For initialization, two electrons are loaded on the same dot, occupying the ground state valley and orbital state with the spins in a spin singlet configuration. One electron is then moved to the neighboring dot by adjusting the gate voltages, creating a state with one electron on each dot. If the movement is diabatic with respect to the difference in Zeeman energy between the dots, the spins will remain in their spin state and thus be initialized in the singlet state, which is a natural initial state for a $S>T_0$ qubit. When using $\{\uparrow\downarrow\}$ qubits, the spin singlet can be rotated to $|\uparrow\downarrow>$ via single- and two-qubit gates if desired. If the difference in Zeeman energy is large compared to the exchange energy, diabatic pulsing might not be an option. Instead, adiabatic transfer of one electron to the neighboring dot will directly result in the $|\uparrow\downarrow\rangle$ state (Fig. 2). Spin read-out at temperatures exceeding the spin splitting can be realized based on Pauli spin blockade, whereby two electrons can or cannot come together on the same dot depending on their relative spin states. The (relative) spin state can then be inferred from charge-sensing, as shown also in recent experiments using Si-MOS quantum dots.

With one well-initialized electron on each dot, qubit splittings can be chosen in a comfortable range, say 5–200 μeV, which corresponds to accessible microwave frequencies of 1–50 GHz. Hence by combining a large energy splitting for initialization and read-out with a lower level splitting during qubit manipulation, the frequencies for driving qubits do not have to be scaled up with the operating temperature.

The spin relaxation time $T_1$ will be reduced with higher temperature. Below 100 mK, $T_1$ is typically very long, especially in silicon, with measured $T_1$ times of over one second, see ref. 100 for a theoretical analysis on the limiting relaxation mechanisms. At low temperature, the temperature dependence of $T_1$ is dictated by one-phonon (direct) processes, and the relaxation rate will increase roughly linearly with temperature. However, the relaxation rate can have a much stronger temperature dependence at higher temperatures due to two-phonon transitions, such as $1/T_1 \propto T^{-9}$ (Raman) and/or $1/T_1 \propto e^{-\Delta E/k_B T}$ (Orbach), where $\Delta E$ is the energy to the first orbital state. For donors, the transition to the exponential temperature dependence due to Orbach transitions occurs at 6 K for phosphorus, 11 K for arsenic, 4 K for antimony, and 26 K for bismuth, all at a magnetic field of 0.3 T. The measured $T_1$ is above one second at 4 K in all cases. For silicon quantum dots, there are few experimental reports on the temperature dependence of $T_1$. Based on the large orbital splitting of order 10 meV that can be realized in silicon quantum dots, one would expect the transitions to two-phonon processes to occur at relatively high temperatures as well. However, imperfect interfaces give rise to spin-orbit coupling between the valley states, and this opens a new channel for relaxation as observed in experiment, which will have a strong sample-to-sample dependence. Nevertheless, long $T_1$ times have been achieved even in systems with very small valley splitting. This suggests that at least in this temperature range, multi-phonon processes do not dominate and more research on the temperature dependence is needed. Nonetheless, the long relaxation times leave a lot of margin, and we anticipate that it is possible to substantially increase the operating temperature of silicon spin qubits.

Decoherence from hyperfine interaction with nuclear spins in the substrate will be approximately temperature independent. An important question is to what extent both low-frequency and high-frequency charge noise will be enhanced by thermal excitations. Charge noise affects spin states most strongly during gates based on exchange or capacitive coupling (Fig. 2), but also a single spin is sensitive to electric fields through the Stark effect, and this sensitivity is higher if local magnetic field gradients are present. Established models indicate that low-frequency charge noise increases linearly with $T$, and such signatures are seen in recent experiments on SiGe and SiMOS dots. In GaAs, a quadratic dependence of high-frequency charge noise was observed between 50 and 250 mK. If silicon devices exhibit similar behavior, this would strongly impact two-qubit gate fidelities. Significant improvements in the quality of exchange oscillations (the basis for most two-qubit gates, and for single-qubit gates in some qubit representations) were recently obtained by keeping the qubits at all times at the so-called symmetry point (Fig. 2a). Altogether, we believe that potential 4 K operation of spin qubits is an attractive possibility.

CONCLUSIONS

Wiring up large qubit arrays is a common, central challenge across all qubit platforms. From the above discussion, we see that electron spin qubits in quantum dots or donors offer several particularly attractive features for overcoming this challenge. First, the sub-100 nm lateral dimensions of quantum dots or donors allow for highly dense qubit registers that nevertheless can be wired up with multiplexing and cross-bar approaches with charge-storage electrodes. The feasibility of such approaches strongly benefits from the extremely long coherence times of electron spins in nuclear-spin-free host materials such as isotopically purified $^{28}$Si, 105 which relax the requirements on parallel read-out and control that short-lived qubits must meet. Second, multiple ideas have been proposed for interconnecting qubit arrays over micron to mm distances. This leaves flexible space for interconnects and integrated electronics. Third, spin qubits on dots or donors may be operated at temperatures of 1–4 K, where the available cooling power is about 1000 times larger than below 4 K, the typical operating temperature today. This would greatly simplify the integration of a first layer of classical control electronics right next to the qubits, again strongly relaxing the interfacing challenges.

These proposed solutions and approaches are not mutually exclusive. For instance, charge-storage electrodes can be beneficial also in sparse arrays, and a classical layer with (very) limited functionality could be incorporated with dense arrays. Furthermore, it is clear that there is still a big step to take from formulating general ideas as done here, to a complete proposal for an actual device, including device lay-outs, dimensions, power budgets, and so forth. Nevertheless, it is clear that spin qubits offer several particularly attractive possibilities in this direction. Finally, the continuous development of semiconductor technology provides further perspective that the wiring challenges can in fact be overcome, paving the way for the construction of a large-scale universal quantum computer.
ACKNOWLEDGEMENTS

We thank D.P. DiVincenzo and our respective group members for valuable discussions.

AUTHOR CONTRIBUTIONS

L.M.K.V. and M.V. wrote the manuscript. All authors discussed the work together and commented on the manuscript.

ADDITIONAL INFORMATION

Competing interests: The authors declare no competing financial interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Jones, N. C. et al. Layered architecture for quantum computing. Phys. Rev. X 2, 031007 (2012).
2. Metodi, T. S., Thaker, D. D. & Cross, A. W. A quantum logic array micro-architecture: scalable quantum data movement and computation. Proceedings of the 38th Annual IEEE/ACM International Symposium on Microarchitecture, 305–318 (IEEE Computer Society, 2005).
3. Hossain, R., Kouwenhoven, L. P., Petta, J. R., Tarucha, S. & Vandersypen, L. M. K. Surface charge architecture for donors and dots in silicon with imprecise and nonuniform qubit couplings. Phys. Rev. B 93, 035306 (2016).
4. Dickerson, B. E. A silicon-based nuclear spin quantum computer. Science 342, 853–858 (2013).
5. Raussendorf, R. & Harrington, J. Fault-tolerant quantum computation with high threshold in two dimensions. Phys. Rev. Lett. 98, 190504 (2007).
6. Fowler, A. G., Mariantoni, M., Martinis, J. M. & Cleland, A. N. Surface codes: towards practical large-scale quantum computation. Phys. Rev. A 86, 032324 (2012).
7. Wecker, D., Bauer, B., Clark, B. K., Hastings, M. B. & Troyer, M. Gate-count estimates for performing quantum chemistry on small quantum computers. Phys. Rev. A 90, 022305 (2014).
8. Van Meter, R. & Horsman, C. A blueprint for building a quantum computer. Commun. ACM 56, 84–93 (2013).
9. Hanson, R., Kouwenhoven, L. P., Petta, J. R., Tarucha, S. & Vandersypen, L. M. K. Spins in few-electron quantum dots. Rev. Mod. Phys. 79, 1217–1265 (2007).
10. Zwanenburg, F. A. et al. Silicon quantum electronics. Rev. Mod. Phys. 85, 961 (2013).
11. Lanzorotti, M. Y., Fiorenza, G. & Rand, R. A. Microminiature packaging and integrated circuitry: The work of E. F. Rent, with an application to on-chip interconnection requirements. IBM J. Res. Dev. 49, 4–5 (2005).
12. Rust, B. et al. Quantum control and process tomography of a semiconductor quantum dot hybrid qubit. Phys. Rev. A 1500022 (2015).
13. Kim, D. et al. Quantum control and process tomography of a semiconductor quantum dot hybrid qubit. Nature 511, 70–74 (2014).
14. Veldhorst, M. et al. An addressable quantum dot qubit in a silicon wire transistor. Nat. Nanotechnol. 9, 981 (2014).
15. Laucht, A. et al. Electrically controlling single spin qubits in a continuous microwave field. Sci. Adv. 3, e1500022 (2015).
16. Kawakami, E. et al. Electrical control of a long-lived spin qubit in a Si/Ge quantum dot. Nat. Nanotechnol. 9, 666–670 (2014).
17. Nielsen, M. A. & Chuang, I. L. Quantum Computation and Quantum Information (Cambridge University Press, 2000).
18. Oso, K., Austing, D. G., Tokura, Y. & Tarucha, S. Current rectification by Pauli exclusion in a weakly coupled double quantum dot system. Science 297, 1313 (2002).
19. Elzerman, J. M. et al. Single-shot read-out of an individual electron spin in a quantum dot. Nature 430, 431 (2004).
20. Vandersypen, L. M. K. et al. Real-time detection of single electron tunneling using a quantum point contact. Appl. Phys. Lett. 83, 4394–4396 (2003).
21. Barthel, C., Reilly, D. J., Marcus, C. M., Hanson, M. P. & Gossard, A. C. Rapid single-shot measurement of a single-electron qubit. Phys. Rev. Lett. 103, 160503 (2009).
22. Colles, J. I. et al. Dispersive readout of a few-electron double quantum dot with fast rf gate sensors. Phys. Rev. Lett. 110, 046805 (2013).
23. Gonzalez-zalba, M. F., Barraud, S., Ferguson, A. J. & Bezt, A. C. Probing the limits of gate-based charge sensing. Nat. Commun. 6, 6084 (2015).
24. Barthel, C. et al. Fast sensing of double-dot charge arrangement and spin state with a radio-frequency sensor quantum dot. Phys. Rev. B 81, 161308 (2010). (R).
25. Watson, T. F., Weber, B., House, M. G., Büch, H. & Simmons, M. Y. High-fidelity rapid initialization and read-out of an electron spin via the single donor D-charge state. Phys. Rev. Lett. 115, 166806 (2015).
26. Shulman, M. D. et al. Superseding qubit dephasing using real-time Hamiltonian estimation. Nat. Commun. 5, 5156 (2014).
