scientific reports

Check for updates

OPEN

A tri-linear quantum dot architecture for semiconductor spin qubits

R. Li[1]✉, V. Levajac[1,3], C. Godfrin[1], S. Kubicek[1], G. Simion[1], B. Raes[1], S. Beyne[1], I. Fattal[1,2], A. Loenders[1,2], W. De Roeck[3], M. Mongillo[1], D. Wan[1] & K. De Greve[1,2]

Semiconductor quantum dot spin qubits hold significant potential for scaling to millions of qubits for practical quantum computing applications, as their structure highly resembles the structure of conventional transistors. Since classical semiconductor manufacturing technology has reached an unprecedented level of maturity, reliably mass-producing CMOS chips with hundreds of billions or even trillions of components, conventional wisdom dictates that leveraging CMOS technologies for quantum dot qubits can result in upscaled quantum processors with thousands or even millions of interconnected qubits. However, the interconnect requirements for quantum circuits are very different from those for classical circuits, where for each qubit individual control and readout wiring could be needed. Although significant developments have been demonstrated on small scale spin qubit systems, qubit numbers remain limited, to a large extent due to the lack of scalable qubit interconnect schemes. Here, we present a tri-linear quantum dot array that is simple in physical layout while allowing individual wiring to each quantum dot. By means of electron shuttling, the tri-linear qubit architecture provides qubit connectivity that is equivalent to or even surpasses that of 2D square lattice qubit arrays. Assuming the current qubit fidelities of small-scale qubit devices can be extrapolated to large-scale arrays, even medium-length shuttling arrays on the order of tens of microns in length would already allow the realization of million-scale qubit systems, while maintaining manageable overheads. Beyond the qubit architecture, we also present a scalable control scheme, where the qubit chip is 3D-integrated with a low-power switch-based cryoCMOS circuit for parallel qubit operation with limited control inputs. As our tri-linear quantum dot array is fully compatible with existing semiconductor technologies, this qubit architecture represents one possible framework for future research and development of large-scale spin qubit systems.

Fault tolerant quantum computers and practical quantum computation algorithms require the number of physical qubits to be in the million scale[1]. Controlling and interfacing such large quantum processors is a giant system-level challenge. A quantum computer is therefore expected to require a multi-layer architecture[2,3], a hierarchical structure not that dissimilar from classical computers. Along with this analogy, and in view of the complexity needed, both qubits and the lowest level control machinery of a quantum computer would benefit a lot if the unprecedented sophistication of classical semiconductor technology could be leveraged. This technology, and particularly CMOS processing, enabled scaling of the number of transistors on a single integrated chip from millions back in the 1990s to hundreds of billions nowadays[4].

Among different qubit technologies, semiconductor quantum dot spin qubits most closely resemble CMOS transistor technology[5–8]. From a qubit perspective, operation fidelities beyond the quantum error correction threshold have been demonstrated[9–16] in small scale quantum dot spin qubit systems in academic laboratories, and proof-of-principle long-range interactions have also been demonstrated – providing an existence proof for upscaling considerations[17–24]. Recently, also quantum dot spin qubits fabricated by fully CMOS-foundry-compatible methods have begun to show high device yield, good uniformity, and performance comparable to or even beyond those of the best academic-lab results[25–27]. Despite this progress, a clear roadmap to upscaling quantum dot qubits has been found lacking: in spite of extensive efforts, results on quantum dot spin qubit arrays have so far been limited to systems with qubit numbers on the order of 10, far below the desired million-scales[28–31]. These limitations can be partially explained by the small size of quantum dot spin qubits, which has the effect of a double edged sword on qubit scaling – while the small size allows for high density integration, the wiring interconnection at very tight and sustained pitches is challenging and quickly becomes the bottleneck[8,32]. This is

[1]IMEC, Leuven, Belgium. [2]Department of Electrical Engineering, KU Leuven, Leuven, Belgium. [3]Department of Physics, KU Leuven, Leuven, Belgium. ✉email: ruoyu.li@outlook.com

additionally complicated by the fact that - unlike transistor circuits, where inputs and outputs can be cascaded – each qubit requires individual control and readout. Addressing those scaling challenges is not straightforward, as wiring fanout and crosstalk correction favor simple one-dimensional qubit arrays, while quantum algorithms and error correction rely on more complex connectivity that is inherent to two-dimensional qubit grids[33,34] or beyond. There are many studies trying to address the wiring challenges from different aspects[8,35–41]. However, one or more elements in those proposed architectures generally require technologies well beyond the current state-of-the-art CMOS fabrication capabilities, such as compact co-integration of qubits with on-site control circuits, or high qubit uniformity at the spin qubit level[35,36]. One solution could be separating qubits (or qubit arrays) with long range qubit couplers to open spaces for wiring interconnects[8,37–39]. However, long range couplers, either shuttling or spin-photon coupling based, also require a certain level of uniformity, and the quantum error correction schemes need further developments to consider the reduced connectivity of these couplers. For the above reasons, no complete error correction scheme has been implemented on quantum dot spin qubit systems so far.

Here, we present a linear spin qubit architecture that allows realistic individual qubit wiring, while having a qubit connectivity equivalent to that of a 2D lattice grid. The core of our proposed architecture is based on earlier work, where we and others proposed how a 2D qubit lattice grid could be mapped onto a bi-linear qubit array for both superconducting and silicon spin qubits, with overlapping resonators providing the qubit connectivity between the two linear qubit arrays[42,43]. However, for silicon quantum dot spin qubits, resonator-mediated two qubit gates tend to be challenging due to the inherently limited coupling strength, resulting in limited fidelities thus far[20]. Also, the required dense arrays of resonators with well-matched qubit frequencies impose multiple, stringent engineering challenges. In this work, we therefore propose introducing instead a third linear array of quantum dots, in-between the two linear qubit arrays, to enable a similar qubit connectivity – but this time enabled via qubit shuttling. Unlike conventional shuttling-based long-range connections that rely on shared gate control to free up the space for interconnects, this tri-linear architecture allows for connecting each individual quantum dot in the shuttling array and applying different voltages separately, in turn allowing tuning each in the array for best shuttling performance[18]. We discuss the physical implementation of our tri-linear architecture for different system sizes, while also addressing the wiring interconnect bottleneck with 3D integration and cryoCMOS. We further discuss how to bypass potentially malfunctioning quantum dot sites in the proposed architecture, and how to achieve even higher order qubit connectivity - beyond the nearest-neighbor interactions of conventional 2D lattice grids.

Tri-linear quantum dot array

Figure 1 shows the concept of the mapping of a 2D lattice grid onto a tri-linear array. For illustration purposes, we use a 4×4 lattice as shown in Fig. 1a. Each spin qubit is defined by a quantum dot, as denoted by the filled blue circles, and is connected to its nearest neighbors, as denoted by the blue lines. Figure 1b shows the tri-linear array as a direct mapping of the 2D lattice, where each odd (even) row from the 2D lattice is placed into the upper (lower) 1D array[43]. Two-qubit interactions within the row direction (green line in Fig. 1a) remain in the tri-linear array and can be performed directly, as represented by the green line between the corresponding sites in Fig. 1b. For two-qubit interactions along the column direction in the 2D lattice (purple line in Fig. 1a), interactions between the upper and lower 1D array are required in the tri-linear array. We implement such an operation by shuttling one qubit from the lower or upper 1D array to the middle array, which is composed of empty quantum dots - as shown by the empty blue circles in Fig. 1b. Then the qubit is shuttled along the middle

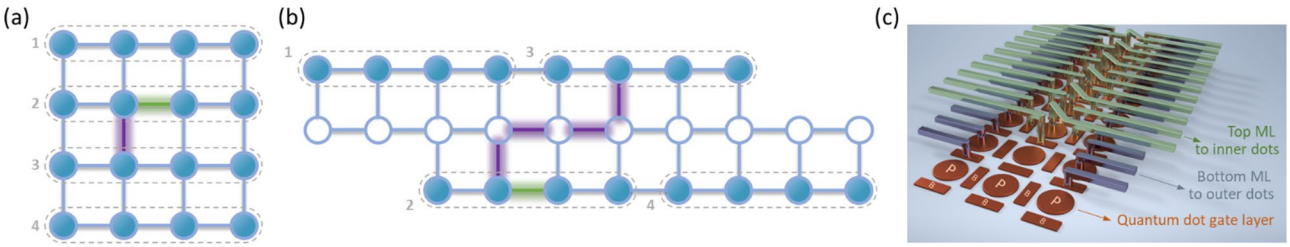

Fig. 1. Design of the linear quantum dot architecture. **(a)** Schematic representation of the 2D quantum dot array with square lattice and nearest neighbor connectivity. **(b)** The tri-linear array with equivalent connectivity as (**a**). The upper (lower) 1D array in the tri-linear architecture are equivalent to row 1 and 3 (2 and 4) of the 2D array as in (**a**). The middle of the tri-linear array consists of empty quantum dots for qubit shuttling. From a qubit connectivity perspective, interactions within the same row maintain the same operation scheme for the 2D and the tri-linear array, as noted by the green shaded lines connecting two qubit sites in (a) and (b) as an example. Interactions along the vertical directions between different rows require more steps on the tri-linear array than the 2D array, which needs moving the qubit, say, from the top 1D array to the middle and shuttle to the targeting site above the bottom array for two qubit operation, and then shuttle back to the original site. An example operation in the tri-linear array, and the equivalent one in the 2D array, are shown by the purple shaded lines in (**a,b**). **(c)** Three-dimensional model of the array gate structure. Two metal layers (MLs) allow individual connection to the plunger (P) and barrier (B) gates of each quantum dots. Some metal layers at the bottom left of the figure are removed for clarification.

array to the desired site (still in the middle array, next to the other qubit) to perform the two-qubit operation and shuttle back. This operation procedure is represented by the purple line in Fig. 1b.

The tri-linear architecture could simplify both the physical qubit structure as well as their operation schemes[26,40,41,44–46]. The physical structure can be straightforwardly implemented with current CMOS technology as the gate pitch and critical dimensions are in the few tens of nanometer range – a range currently achieved for small-scale qubit demonstrators[26,47,48]. As shown in Fig. 1c, two wiring layers (blue and green) would suffice for fanout, allowing full control of the chemical potential of the quantum dots (via the circular plunger gates) and their coupling to the neighbouring dots (via the rectangular barrier gate) for each quantum dot. For simplicity of the discussion, here we omit the control and readout periphery which can be straightforwardly added to the scheme (see Supplementary Information S1). A key advantage of the linear nature of the proposed architecture layout is that such and other (linearly scaling) physical structures can be readily added to the layout. For example, an array of nanomagnets can be fabricated on top of the quantum dots for qubit addressability in the case of Si-based electron spin qubits (SiMOS and Si/SiGe heterostructures). For Ge-based hole spin qubits (Ge/SiGe heterostructures), strong spin-orbit interaction can be exploited for qubit addressability and control, so no nanomagnet array is needed. Next, an array of single electron transistors can be added on each side along the tri-linear array for the qubit readout. These could be implemented in the same gate layer as the quantum dot gates and are readily available with the current CMOS technology[45,49]. The separate wiring enables individual qubit fidelity optimization for every step of the quantum operations, including single and two-qubit gates and shuttling. Additionally, the device tune-up and crosstalk compensation will be more straightforward here than in 2D lattices[29,50–52]. Moreover, individual control also enables parallel operations at different qubit sites, and multiple qubits can be shuttled simultaneously in the middle quantum dot array[45].

In comparison to a 2D lattice, the tri-linear qubit structure requires extra shuttling for two-qubit operations along the column directions. As shown in Fig. 1b, the adjacent 2D odd and even rows are at the upper and lower 1D qubit array, respectively, with one shifted horizontally by half of the row size. This means that the shuttling length in the middle array is half of the row size. Since qubits are shuttled forward and backward, the total number of shuttling steps corresponds to the number of quantum dots within the row, which is $\sqrt{N}$ for N qubits in the full array. For upscaling, the critical consideration is the shuttling length against the qubit array size. In Fig. 2, we plot the required shuttling length against the total number of qubits N in the tri-linear array. The shuttling length is obtained by multiplying the total number of shuttling steps $\sqrt{N}$ and the quantum dot pitch (assumed here to be 100 nm). By considering shuttling lengths in the few-micron range, we obtain that our architecture can easily host a few thousands of qubits. For million-scale system sizes, the shuttling length is in the range of tens of microns, which is still comparable to the designed length in current shuttling-connected sparse architectures[37,38]. Also, a recent experiment has demonstrated shuttling over an accumulated distance of 10 µm with shuttling fidelity exceeding 99% and shuttling speed reaching 50 m/s[18]. However, for systems beyond the million scale, the extensive shuttling length likely requires very high shuttling fidelity and qubit coherence, which could be very challenging to realize.

For extremely large systems, the single qubit rows in the top and bottom of the tri-linear array could be replaced by multiple rows and, ultimately, by semi-2D arrays, as illustrated by the gray inset in Fig. 2 – at the expense of more complex interconnection layers (more back-end-of-line layers to sustain the semi-2D arrays). In the ultimate limit, this would result in $\sqrt{N}$ square 2D sub-arrays of $\sqrt{N}$ quantum dots each, which would

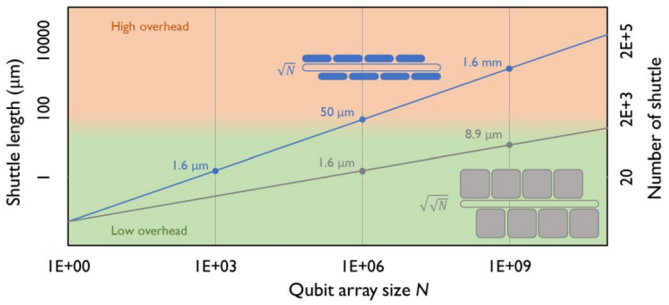

Fig. 2. Shuttle length vs. different qubit array sizes for the corresponding 2D lattice grid with nearest neighbor connectivity. In comparison to a 2D lattice, the linear qubit structure requires extra shuttling overhead for two qubit operations along the vertical directions. The blue line shows the shuttle length in the tri-linear array, where the number of shuttles needed is $\sqrt{N}$ – scaling with the number of rows in a 2D grid with N qubits; each of those shuttles also requires on the order of sqrt(N) empty quantum dots to be able to connect neighboring rows in the 2D grid. For thousands of qubits, the shuttle length is in the few microns scale, which is well within the low overhead regime. For million scale qubit systems, the shuttling length reaches the tens of micron range, which is still within the reasonable overhead region but could require excessively high shuttling fidelity and speed. For very large-scale systems, the 1D qubit array at the top and bottom of the tri-linear structure can be replaced by semi-2D arrays, as illustrated by the gray inset at the bottom right of the plot. Both the number of shuttles as well as the length of each shuttle then further reduces to $\sqrt{\sqrt{N}}$, and the shuttle length can be in the range of few microns even for billion scale systems. A quantum dot pitch of 100 nm is used for the shuttle length calculation.

then each require shuttling over their sub-array row length of $\sqrt{\sqrt{N}}$. so that the shuttling length can be reduced to a few microns – even for billion scale systems as shown by the gray line in Fig. 2. Interconnecting a semi-2D qubit array would require multi-layer wiring, and transferring qubits across the semi-2D array would require multiple operations[47,48], so this would need to be carefully analyzed in terms of potential tradeoffs vs. shuttling length. As intermediate steps towards such semi-2D array, the 1D qubit arrays (single rows, aka arrays of size $1 \times \sqrt{N}$) at each side of the linear architecture can already be replaced by multiple-row qubit arrays (M rows, aka arrays of size $M \times \sqrt{N}/M$). This will require a limited increase in the number of interconnect wiring layers, while reducing the shuttling length to ($\sqrt{N}/\mathrm{M}$). Along the same lines, a few more 1D arrays can also be added to include more shuttling paths to enable more system operation protocols and avoid a single shuttling array resulting in bottlenecks. One will need to carefully analyze the tradeoffs between the benefits of shorter shuttling distances and more parallelized shuttling path on one side, and the challenges of fabricating and crosstalk on the other. Of course, developing useful quantum computers with millions of interconnected physical qubits is a gargantuan task, and it may in practice take a large-scale effort, for which this architecture could provide a blueprint, to be augmented by dedicated and significant developments on materials, devices and the wiring layers themselves for the detailed implementation.

Wiring interconnectivity and 3D integration

Small qubit arrays allow separate wiring from room-temperature electronics to each qubit to demonstrate proof-of-principle quantum information processing[53,54]. For large qubit arrays, it is impractical to connect each gate electrode to a separate analog circuit – meaning that certain levels of multiplexing at the qubit-to-analogue-control interface should be implemented to address the interconnect bottleneck[32]. The proposed tri-linear array architecture, in view of its accommodation of reduced effective pitches, lends itself well to a natural implementation that brings the control electronics to the qubit chip without room-temperature wiring. Our proposed approach encompasses two steps. First, we propose to contact each quantum dot gate through 3D integration, and second, to address those gates via multiplexed gate control involving cryoCMOS switch circuits.

In Fig. 3a, we show the qubit wiring scheme with the proposed 3D integration solution. The linear qubit architecture permits the fanout of all quantum dot gates to take place on a single chip plane, and the large space outside the linear array grants opportunities for contacting the qubits by staggered contact points with very

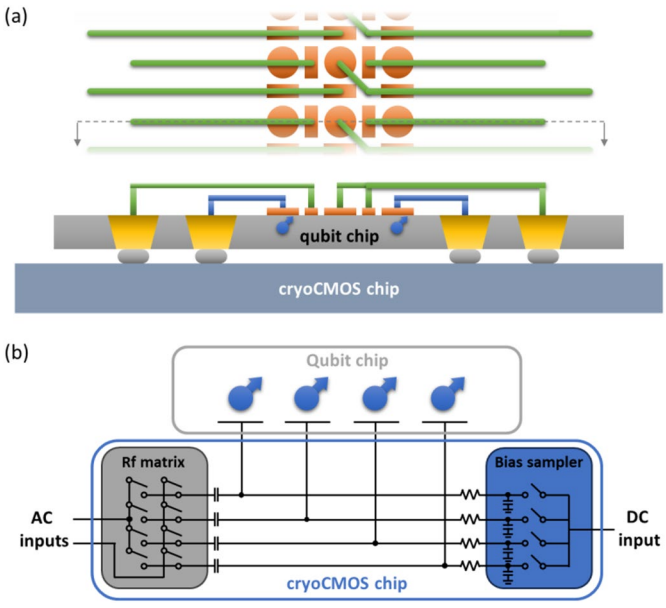

Fig. 3. 3D integrated qubit and cryoCMOS chips for interconnectivity. **(a)** Schematic of the 3D integrated qubit system. Outside the tri-linear array, there is ample space to interconnect each qubit gate and ensure that realistic wiring pitches can be used. By interleaving the wiring ending points, each quantum dot gate can be routed to interconnect structures having dimensions much larger than the pitch of the quantum dots. In this schematic, we use TSVs (through-silicon-vias) to bring the interconnect to the back of the qubit chip. A similar approach with flip-chip bonding to interposers or control chips on the front side of the chip can also be considered (not shown). The qubit chip is subsequently 3D integrated to a cryoCMOS chip for qubit control. **(b)** The scheme for scalable interconnection to each quantum dot gate with the cryoCMOS chip. Each gate needs two inputs: DC to set the gate bias point and AC for shuttling and qubit operations. For DC, different gate biases are stored with a sampling circuit so that a single input can provide DC bias voltages for multiple gates in a time-based multiplexing manner. For AC, an RF switch matrix can be used to route different sets of AC pulses to different sets of gates simultaneously, allowing for parallel qubit operations with a limited number of AC inputs. The exact number of AC inputs in this scheme depends on the degree of homogeneity achieved after tuning each dot carefully using the DC biasing scheme.

relaxed pitches. The contact points can therefore easily reach an effective pitch in the micron range, unlike the 100 nm or below linear quantum dot gate pitch, which is well compatible with existing 3D integration techniques, offering in turn space for separate CMOS control of each quantum dot. Rather than cointegrating the qubits and cryoCMOS control electronics on the same chip, here we separate the qubit chip and the cryoCMOS chip. The quantum dot gate wiring is routed to through-silicon-vias (TSVs) in the qubit chip, and then directly 3D-bonded to the cryoCMOS chip. Such separation of the two modules allows for their individual optimization and, in addition, offers possibilities for thermal isolation between the two chips – the latter is important in view of the expected thermal load from large scale cryoCMOS chips and the need to keep electron temperatures of the quantum dot chips well below 1 K. Alternatively, a flip-chip method can be used on the qubit chip to eliminate the TSV process[55], and an interposer chip or a redistribution layer can be added depending on the integration requirements[56].

In Fig. 3b, we illustrate the multiplexing scheme with cryoCMOS switch circuits. For each quantum dot gate, we split the voltage input based on the signal bandwidth and functionality. We use a low frequency (quasi-static DC) input for biasing static quantum dot working points, while a high frequency input (here referred to as AC, in contrast to the quasi-static DC control) serves for applying pulses for fast control and readout. These two inputs are ultimately combined by a bias-tee circuit and the sum of the DC and AC signals is applied to each quantum dot gate. Due to nonuniformity among different quantum dot sites[8,35], different DC biases may be required for the quantum dot plunger gates to reach the last-electron occupation in different quantum dots. Similarly, different barrier gates may need different DC biases to tune different dots into the same tunnel coupling range[57]. For these reasons, providing multiplexing capabilities for individual gate tuning is highly desired. We can enable this by using a sampling circuit, where charges stored on a capacitor maintain the voltage even after the input has been disconnected by opening the transistor switch[35]. This type of floating gate control is very similar to the storage mode in DRAM, but with the crucial difference that long holding times have been demonstrated experimentally at cryogenic temperatures due to the low-leakage rate of cryogenic CMOS – potentially allowing for very high multiplexing factors and low charge refresh rates[58,59].

For AC multiplexing, experimental demonstrations of time-domain operations have been reported with both spin and superconducting qubits[59,60]. Here, we take a step further and discuss the possibility of parallel operations with a limited set of AC inputs. Importantly, as we compensate for the nonuniformities among qubits by using the DC floating gates introduced above, one might be able to use a limited, finite set of distinct AC pulses for all qubits – down to potentially a single, common AC pulse control level with sufficient DC compensation and uniformity. In this ideal case, a finite, fixed set of AC pulses, for exchange control, shuttling, readout, would be needed regardless of the qubit number. To achieve this regime, a careful design and highly uniform fabrication of nanomagnets (see Supplementary Information S1) are essential in order to ensure highly uniform Rabi frequencies for all the qubits. Furthermore, we propose to employ an RF switch matrix that simultaneously routes the same AC pulses to different quantum dots, and ultimately allows for parallel operations of different qubits. The inputs of the RF switch matrix consist of its digital switch controls, as well as the fixed set of AC pulses for qubit operation and shuttling, and their total number scales sub-linearly with the number of qubits. A detailed example of qubit operation – with the multiplexing scheme is discussed in Supplementary Information S2.

Practical operation considerations for a larger spin qubit array

When building larger quantum dot spin qubit arrays, another important challenge to tackle, besides the interconnect challenges discussed above, is the required high yield of quantum dots[26]. Disorder at the atomic scale could impact the confinement potential at multiple sites in the array – such that some quantum dots cannot be defined or the tunnel coupling between some dots cannot be controlled[57]. Moreover, in linear arrays that utilize shuttling, a single point of excessive disorder could break one system into two. In Fig. 4a, we show an example where one defective quantum dot site forbids the shuttling of qubits between its two sides – severely restricting the connectivity between the qubits on the left and on the right.

To mitigate this problem, the quantum dots surrounding the defective site can be reconfigured for shuttling at the cost of sacrificing a few discrete quantum dots to preserve the operation of the entire array, as shown in Fig. 4b. This reconfiguration is possible thanks to the individual wiring of each quantum dot, where the cryoCMOS circuits change the dot DC bias with the sampler and the AC pulses with the RF switch matrix. For larger qubit arrays, sacrificing a few qubits in this way will not strongly influence the overall performance of the array – as shuttling is preserved and the original connectivity can be achieved via only slightly longer shuttling distances for certain qubits (around 200 nm per a single occurring defective site). However, if there are multiple defective sites that happen to cut across the array, the above solution will not work. On the architectural level, multiple shuttling arrays or semi-2D qubit arrays, as shown in Fig. 2, can reduce the chance for such a scenario. Also, by considering advancements on the device level, high quantum dot yield rates can be achieved with advanced manufacturing, and further improvement is required for larger systems[26].

The connectivity between qubits is critical for efficient quantum algorithms and quantum error correction codes[61]. For example, nearest neighbor connectivity is essential for implementation of the surface code in 2D qubit grids. As discussed previously (Fig. 2), an N qubit 2D grid can be mapped onto our tri-linear architecture where the nearest neighbor connectivity can be preserved with a shuttling overhead of $\sqrt{N}$. For parallel operations, depending on how many spins are simultaneously shuttled in the middle array, the time overhead has a scaling that is in-between two limits, with $\sqrt{N}$ time overhead (maximally parallelized shuttling operations) and $N\sqrt{N}$ time overhead (fully sequential shuttling operations). Here, we highlight that higher orders of connectivity, between any two qubits in neighboring rows or columns, can be also achieved, with shuttling overheads of $3\sqrt{N}$ or less, as represented by the light blue lines in the upper 2D grid in Fig. 4c. One such example is highlighted by the purple line in the 2D grid, and its corresponding shuttling path is denoted

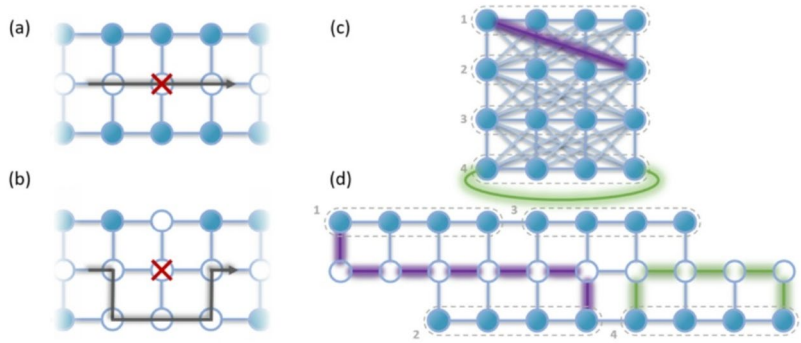

Fig. 4. Practical qubit array operation considerations. **(a,b)** Scheme to avoid defect quantum dots in the tri-linear array. The qubit quantum dots around the defective site can be configured as empty shuttling quantum dots to fix the connectivity. **(c,d)** Schemes for higher order of qubit connectivity. (**c**) A 2D qubit array with square lattice that has row-wise all to all connectivity (connectivity represented by the light blue lines connecting different qubits). In the tri-linear array having N qubits as shown by (**d**), such a connectivity can be achieved with a shuttling overhead of $3\sqrt{N}$. The purple and green shaded line shows the example of connecting qubits beyond the nearest neighbour at different rows and within the same row, respectively.

in purple in the tri-linear array below in Fig. 4d. Moreover, shuttling overheads of $2\sqrt{N}$ can also enable the connectivity between the qubits at the two opposite vertical edges of the 2D grid as illustrated by the green lines in Fig. 4d, which is topologically equivalent to folding the 2D grid into a tube. For a longer tri-linear array, the two ends of the array could also be connected by arranging the array as a loop, where there is still enough space for wiring to the inner qubits, via additional interconnect layers, or through 3D integration. The loop could then turn the tube into a donut, as shown in Supplementary Information S3. Such effective donut connectivity would satisfy topological requirements for forms of error correction that are otherwise out of reach in finite 2D arrays[33]. Finally, we emphasize that the tri-linear architecture in principle allows for even higher orders of connectivity than those considered in Fig. 4c and d. We note that shuttling overheads for achieving such connectivity could exceed $\sqrt{\mathrm{N}}$ overhead in Fig. 2, and can even reach N if connecting the most separated qubits. We also note that the use of a shared shuttling bus may lead to temporally and spatially correlated errors that are detrimental for the performance of error correction and could requires dedicated error correction mechanisms. Although, a detailed discussion of possible algorithms and error correction codes is beyond the scope of this work, we point out that the architectural capabilities for flexible connectivity of the tri-linear architecture could enable novel and more efficient quantum computation and error correction protocols.

Conclusion and discussions

Qubit interconnect and wiring are arguably the central upscaling challenge for most quantum computing platforms. The small size of semiconductor quantum dot spin qubits offers great scalability on the qubit level with qubit pitches of around 100 nm[62]. However, it raises many challenges for the 2D scaling of the interconnect layer, due to difficulties of wiring each quantum dot in a densely packed 2D array. In this work, we have proposed a scalable solution in the form of a tri-linear quantum dot array in which the 2D connectivity is realized by qubit shuttling. This architecture allows for switch-based cryoCMOS modules for channel multiplexing and parallel qubit operation. Both the array structure and the control modules are compatible with the existing semiconductor (Si-based and Ge-based) qubit material platforms and 3D integration technologies. The key requirements for fault-tolerant quantum computation in our tri-linear architecture have so far been demonstrated in small scale systems. These include high quantum dot yield rate[26], single and two qubit gates beyond the fault-tolerant threshold[63], coherent shuttling[18] and operation of small-scale 1D arrays[64]. Next steps towards reaching such performance also in large qubit arrays would involve improving the uniformity of qubit devices by utilizing advanced semiconductor technology[25–27]. In the meanwhile, tri-linear quantum dot arrays with intermediate qubit numbers could be realized to examine the architecture performance on the system-level. In particular, assessing the fidelity of two-qubit gates mediated by shuttling is essential as it will be influenced by an interplay of various parameters such as shuttling speed, shuttling fidelity and qubit coherence.

The cost of the relatively simple layout of our tri-linear architecture is the overhead in the qubit operation coming from qubit shuttling. Recent experiments have shown fast, high fidelity shuttling in the 10-micron range[18], which would correspond to the shuttling length required for a million-qubit system size in the architecture presented here. By using the proposed cryoCMOS control, different gate voltages could be simultaneously optimized to further improve the qubit performance at different sites. Moreover, shuttling can be used to extend the qubit connectivity beyond the nearest neighbors, which opens hardware possibilities for more efficient error correction schemes and algorithms[65]. Also, this architecture is fully compatible with the qubit implementations beyond the Loss-DiVicenzo type of qubits. Namely, realizations of singlet-triplet or exchange-only qubits can also be achieved in the tri-linear architecture by utilizing its capability for selective control of individual quantum dots and pairs of them.

Finally, this work has focused on the general layout of the tri-linear architecture, and some more specific aspects of the functional qubit periphery have been omitted for clarity, as they scale linearly and are readily implementable in the proposed scheme. For completeness, we include in Supplementary Information S1 a detailed example of how the tri-linear array architecture can be equipped with global qubit control, nano-magnet based qubit addressability, and single electron transistors for qubit readout. Importantly, the additions proposed there are fully compatible with existing CMOS technologies. Nonetheless, we remark that all these different components bring specific requirements, and that building a full stack qubit architecture would need to be addressed as a system-level problem. We expect that the considered advantages of the tri-linear qubit architecture motivate system-level developments in the design, fabrication, and operation of future quantum processors.

Data availability

The data that support the findings of this study are available from the authors upon reasonable request.

Received: 15 September 2025; Accepted: 26 February 2026
Published online: 06 April 2026

References

1. 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).
2. Jones, N. C. et al. Layered architecture for quantum computing. *Phys. Rev. X*. **2**, 031007 (2012).
3. Fu, X. et al. ACM, Como Italy,. A heterogeneous quantum computer architecture. In *Proceedings of the ACM International Conference on Computing Frontiers*. 323–330. https://doi.org/10.1145/2903150.2906827 (2016).
4. Bohr, M. A 30 year retrospective on Dennard's MOSFET scaling paper. *IEEE Solid-State Circuits Newsl.* **12**, 11–13 (2007).
5. 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).
6. Zwanenburg, F. A. et al. Silicon quantum electronics. *Rev. Mod. Phys.* **85**, 961–1019 (2013).
7. Burkard, G., Ladd, T. D., Pan, A., Nichol, J. M. & Petta, J. R. Semiconductor spin qubits. *Rev. Mod. Phys.* **95**, 025003 (2023).
8. Vandersypen, L. M. K. et al. Interfacing spin qubits in quantum dots and donors—hot, dense, and coherent. *Npj Quantum Inf.* **3**, 34 (2017).
9. Yoneda, J. et al. A quantum-dot spin qubit with coherence limited by charge noise and fidelity higher than 99.9%. *Nat. Nanotechnol.* **13**, 102–106 (2018).
10. Noiri, A. et al. Fast universal quantum gate above the fault-tolerance threshold in silicon. *Nature* **601**, 338–342 (2022).
11. Veldhorst, M. et al. An addressable quantum dot qubit with fault-tolerant control-fidelity. *Nat. Nanotechnol.* **9**, 981–985 (2014).
12. Lawrie, W. I. L. et al. Simultaneous single-qubit driving of semiconductor spin qubits at the fault-tolerant threshold. *Nat. Commun.* **14**, 3617 (2023).
13. Xue, X. et al. Quantum logic with spin qubits crossing the surface code threshold. *Nature* **601**, 343–347 (2022).
14. Yang, C. H. et al. Silicon qubit fidelities approaching incoherent noise limits via pulse engineering. *Nat. Electron.* **2**, 151–158 (2019).
15. Mills, A. R. et al. High-fidelity state preparation, quantum control, and readout of an isotopically enriched silicon spin qubit. *Phys. Rev. Appl.* **18**, 064028 (2022).
16. Tanttu, T. et al. Assessment of the errors of high-fidelity two-qubit gates in silicon quantum dots. *Nat. Phys.* **20**, 1804–1809 (2024).
17. Zwerver, A. M. J. et al. Shuttling an electron spin through a silicon quantum dot array. *PRX Quantum*. **4**, 030303 (2023).
18. De Smet, M., Matsumoto, Y., Zwerver, AM.J. et al. High-fidelity single-spin shuttling in silicon. *Nat. Nanotechnol.* **20**, 866 (2025)
19. Noiri, A. et al. A shuttling-based two-qubit logic gate for linking distant silicon quantum processors. *Nat. Commun.* **13**, 5740 (2022).
20. Dijkema, J. et al. Cavity-mediated iSWAP oscillations between distant spins. *Nat. Phys.* https://doi.org/10.1038/s41567-024-02694-8 (2024).
21. Van Riggelen-Doelman, F. et al. Coherent spin qubit shuttling through germanium quantum dots. *Nat. Commun.* **15**, 5716 (2024).
22. Yoneda, J. et al. Coherent spin qubit transport in silicon. *Nat. Commun.* **12**, 4114 (2021).
23. Xue, R. et al. Si/SiGe QuBus for single electron information-processing devices with memory and micron-scale connectivity function. *Nat. Commun.* **15**, 2296 (2024).
24. Struck, T. et al. Spin-EPR-pair separation by conveyor-mode single electron shuttling in Si/SiGe. *Nat. Commun.* **15**, 1325 (2024).
25. Elsayed, A. et al. Low charge noise quantum dots with industrial CMOS manufacturing. *Npj Quantum Inf.* **10**, 70 (2024).
26. Neyens, S. et al. Probing single electrons across 300-mm spin qubit wafers. *Nature* **629**, 80–85 (2024).
27. Steinacker, P., Dumoulin Stuyck, N., Lim, W.H. et al. Industry-compatible silicon spin-qubit unit cells exceeding 99% fidelity. *Nature* **646**, 81 (2025).
28. George, H. C. et al. 12-Spin-Qubit Arrays Fabricated on a 300 mm Semiconductor Manufacturing Line. *Nano Lett.* **25**, 793 (2025).
29. Mortemousque, P. A. et al. Coherent control of individual electron spins in a two-dimensional quantum dot array. *Nat. Nanotechnol.* **16**, 296–301 (2021).
30. Wang, C. A. et al. Operating semiconductor quantum processors with hopping spins. *Science* **385**, 447–452 (2024).
31. Zhang, X. et al. Universal control of four singlet–triplet qubits. *Nat. Nanotechnol.* https://doi.org/10.1038/s41565-024-01817-9 (2024).
32. Franke, D. P., Clarke, J. S., Vandersypen, L. M. K. & Veldhorst, M. Rent's rule and extensibility in quantum computing. *Microprocess Microsyst.* **67**, 1–7 (2019).
33. Bravyi, S. et al. High-threshold and low-overhead fault-tolerant quantum memory. *Nature* **627**, 778–782 (2024).
34. Xu, Q. et al. Constant-overhead fault-tolerant quantum computation with reconfigurable atom arrays. *Nat. Phys.* **20**, 1084–1090 (2024).
35. Veldhorst, M., Eenink, H. G. J., Yang, C. H. & Dzurak, A. S. Silicon CMOS architecture for a spin-based quantum computer. *Nat. Commun.* **8**, 1766 (2017).
36. Li, R. et al. A crossbar network for silicon quantum dot qubits. *Sci. Adv.* **4**, eaar3960 (2018).
37. Boter, J. M. et al. Spiderweb array: A sparse spin-qubit array. *Phys. Rev. Appl.* **18**, 024053 (2022).
38. Künne, M. et al. The SpinBus architecture for scaling spin qubits with electron shuttling. *Nat. Commun.* **15**, 4977 (2024).
39. Taylor, J. M. et al. Fault-tolerant architecture for quantum computation using electrically controlled semiconductor spins. *Nat. Phys.* **1**, 177–183 (2005).
40. Siegel, A., Strikis, A. & Fogarty, M. Towards early fault tolerance on a 2 × N array of qubits equipped with shuttling. *PRX Quantum*. **5**, 040328 (2024).
41. Paraskevopoulos, N. et al. Near-Term Spin-Qubit Architecture Design via Multipartite Maximally Entangled States. *PRX Quantum*. **6**, 020307 (2025).

42. Mohiyaddin, F. A. et al. Large-scale 2D spin-based quantum processor with a bi-linear architecture. In *2021 IEEE International Electron Devices Meeting (IEDM)*. 27.5.1–27.5.4. https://doi.org/10.1109/IEDM19574.2021.9720606 (IEEE, 2021).
43. Mukai, H. et al. Pseudo-2D superconducting quantum computing circuit for the surface code: Proposal and preliminary tests. *New. J. Phys.* **22**, 043013 (2020).
44. Volk, C. et al. Loading a quantum-dot based Qubyte register. *Npj Quantum Inf.* **5**, 29 (2019).
45. Mills, A. R. et al. Shuttling a single charge across a one-dimensional array of silicon quantum dots. *Nat. Commun.* **10**, 1063 (2019).
46. Li, R. et al. A flexible 300 mm integrated Si MOS platform for electron- and hole-spin qubits exploration. In *2020 IEEE International Electron Devices Meeting (IEDM)*. 38.3.1–38.3.4. https://doi.org/10.1109/IEDM13553.2020.9371956 (IEEE, 2020).
47. Weinstein, A. J. et al. Universal logic with encoded spin qubits in silicon. *Nature* **615**, 817–822 (2023).
48. Ha, W. et al. A flexible design platform for Si/SiGe exchange-only qubits with low disorder. *Nano Lett.* **22**, 1443–1448 (2022).
49. Simion, G. et al. A scalable one dimensional silicon qubit array with nanomagnets. In *2020 IEEE International Electron Devices Meeting (IEDM)*. 30.2.1–30.2.4. https://doi.org/10.1109/IEDM13553.2020.9372067 (IEEE, 2020).
50. Unseld, F. K. et al. Baseband control of single-electron silicon spin qubits in two dimensions. *Nat Commun* **16**, 5605 (2025).
51. Lim, W. H. et al. A 2 × 2 Quantum Dot Array in Silicon with Fully Tunable Pairwise Interdot Coupling. *Nano Lett.* **25**, 10263 (2025)
52. Borsoi, F. et al. Shared control of a 16 semiconductor quantum dot crossbar array. *Nat. Nanotechnol.* **19**, 21–27 (2024).
53. Arute, F. et al. Quantum supremacy using a programmable superconducting processor. *Nature* **574**, 505–510 (2019).
54. Google Quantum, A. I. et al. Quantum error correction below the surface code threshold. *Nature* https://doi.org/10.1038/s41586-024-08449-y (2024).
55. Beyne, E. 3D system integration technologies. In *2006 International Symposium on VLSI Technology, Systems, and Applications*. 1–9. https://doi.org/10.1109/VTSA.2006.251113 (IEEE, 2006).
56. Beyne, E., Milojevic, D., Van Der Plas, G. & Beyer, G. 3D SoC integration, beyond 2.5D chiplets. In *2021 IEEE International Electron Devices Meeting (IEDM)*. 3.6.1–3.6.4. https://doi.org/10.1109/IEDM19574.2021.9720614 (IEEE, 2021).
57. Cifuentes, J. D. et al. Bounds to electron spin qubit variability for scalable CMOS architectures. *Nat. Commun.* **15**, 4299 (2024).
58. Li, R. et al. Stable floating-gate control of Si qubit devices with cryoCMOS circuits. In *Silicon Quantum Electronics Workshop (SiQEW) in Kyoto International Conference Center* (2023).
59. Pauka, S. J. et al. A cryogenic CMOS chip for generating control signals for multiple qubits. *Nat. Electron.* **4**, 64–70 (2021).
60. Acharya, R. et al. Multiplexed superconducting qubit control at millikelvin temperatures with a low-power cryo-CMOS multiplexer. *Nat. Electron.* **6**, 900–909 (2023).
61. Bravyi, S., Poulin, D. & Terhal, B. Tradeoffs for reliable quantum information storage in 2D systems. *Phys. Rev. Lett.* **104**, 050503 (2010).
62. Stuyck, N. D. et al. CMOS compatibility of semiconductor spin qubits. *Preprint.* https://doi.org/10.48550/ARXIV.2409.03993 (2024).
63. Stano, P. & Loss, D. Review of performance metrics of spin qubits in gated semiconducting nanostructures. *Nat. Rev. Phys.* **4**, 672–688 (2022).
64. Philips, S. G. J. et al. Universal control of a six-qubit quantum processor in silicon. *Nature* **609**, 919–924 (2022).
65. Breuckmann, N. P. & Eberhardt, J. N. Quantum low-density parity-check codes. *PRX Quantum*. **2**, 040101 (2021).

Acknowledgements

This work was performed as part of IMEC's Industrial Affiliation Program (IIAP) on Quantum Computing.

Author contributions

R.L. and C.G. developed the physical qubit chip layout. S.K and S.B verified qubit and 3D integration schemes. G.S. analyzed prospects for quantum error correction. R.L. conceived the project with support from all authors. R.L., M.M., D.W., and K.D.G. supervised the project. R.L., V.L., and K.D.G. wrote the manuscript with input from all authors.

Funding

The authors would like to acknowledge the support of the European Union Horizon Europe research and innovation program under grant agreement no. 101174557 (QLSI2). V.L., W. D. R., and K. D. G. acknowledge the support in part from the Excellence of Science (EOS) programme (FWO and F.R.S.-FNRS) through grant EOS G0H1122N EOS 40007526 CHEQS.

Declarations

Competing interests

R.L. C.G. S.K. G.S. B.R. are inventors on patent application related to this work, filed by IMEC (application no. EP 24166659.3; filing date, 27 March 2024). Additionally, R.L. C.G. S.K. G.S. D.W. K.D.G. are inventors on another patent application related to this work, also field by IMEC (application no. EP 25163893.8; filing date, 14 March 2025). The other authors declare that they have no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-42575-z.

Correspondence and requests for materials should be addressed to R.L.