nature portfolio

https://doi.org/10.1038/s41928-021-00687-6

Supplementary information

A cryo-CMOS chip that integrates silicon quantum dots and multiplexed dispersive readout electronics

In the format provided by the authors and unedited

Supplementary Information

1. Statistical characterization of quantum devices

At cryogenic temperatures, QDs can be induced in the standard 40 nm MOS transistors used in this work when low source voltages are applied. Supplementary Fig. 1 shows Coulomb diamond diagrams of each QD transistor Q_{ij}. From the Coulomb diamond pattern, such as in Supplementary Fig. 1i, the gate lever arm, representing the ratio of applied source voltage (V_S) to the width of the Coulomb diamond (ΔV_G) at V_S, can be derived as:

$$\alpha = \frac{V_S}{\Delta V_G} = \frac{C_G}{C_\Sigma}, \tag{2}$$

where C_G and C_Σ are the gate-QD capacitance and the total capacitance to the QD, respectively. The total capacitance is C_Σ=C_G+C_S+C_D, where C_S and C_D are the source-QD and QD-drain capacitances, respectively (Supplementary Fig. 2a shows a cross-section schematic of a device). C_S and C_D can be further derived from the Coulomb diamond boundaries as:

$$C_S = -\frac{C_G}{m_1} \quad \text{and} \quad C_D = -\frac{C_G(1-m_2)}{m_2}, \tag{3}$$

thus

$$\frac{C_S}{C_D} = -\frac{m_2}{m_1(1-m_2)}, \tag{4}$$

where m_1 and m_2 are the slopes of the Coulomb diamond boundaries as in Supplementary Fig. 1i. From the ratio of C_S/C_D, the location of QDs can be estimated, since in general the capacitance is inversely proportional to the separation between two conductors. The charging energy E_c can be estimated as:

$$E_c = \frac{e^2}{C_\Sigma} = e \times \Delta V_S, \tag{5}$$

where e is the electron charge and ΔV_S is the height of the Coulomb diamond. We find E_c=21.4 meV in Supplementary Fig. 1i. However, some of the quantum devices in this work have multiple QDs induced in the channel, as in Supplementary Fig. 1g. The non-monotonic Coulomb diamond pattern suggests a multi-QD system. In the case of a multi-QD system, the charging energy of each quantum dot is hard to estimate from the diamond pattern, because the parameter ΔV_S can not be well defined.

We perform a statistical analysis for the d.c. transport parameters, namely (1) gate lever arm α, (2) C_S/C_D, and (3) $V_{G,1st}$, the gate voltage where the first observable Coulomb-blockade peak occurs (as shown in Supplementary Fig. 2b). The results are reported in Supplementary Fig. 2c-e, respectively. The data in Supplementary Fig. 2c-e are collected from two chips, namely Chip 1 (used in this manuscript) and Chip 2 (another instance not discussed in the main text). Each of these chips contains a 3×3 array of quantum devices with nominally identical physical dimensions. Supplementary Table 1 shows the summary of the electrical transport parameters. We conclude that: (i) all DUTs in this work can induce QDs; (ii) the lever arm $\alpha \approx 0.66$ eV/V in these quantum devices is considerably larger than in the planar quantum dot devices in [53, 54], and just below those in 3D nanowire QD transistors [41], importantly, since the strong gate-QD coupling is essential for dispersive readout [41][55, 56]; (iii) the ratio of $C_S/C_D \approx 1.1$ indicates that the locations of QDs are well centered in the channel, suggesting that the QDs are formed due to charge carrier accumulation by applying gate bias, rather than due to the dopant diffusion from source and/or drain electrodes; (iv) the consistent $V_{G,1st}$=0.463±0.035 V suggests a small variation from DUT to DUT in this work, and approaches the requirement to be able to load a single electron in multiple transistors with a common gate bias [16], the condition being $E_c/\alpha > \sigma(V_{G,1st})$, where σ indicates the standard deviation.

2. Retention time study

We characterize the charge retention time for an individual cell at 50 mK. A simple equivalent circuit model is shown in Supplementary Fig. 3a [30]. To extract the retention time, we apply the sequence (i) charging and (ii) discharging. For (i) charging: the cell is firstly charged by applying V_{WL}=1.49 V, much higher than the threshold voltage of the access transistor, while V_{DL}=0.8 V, as in Supplementary Fig. 3b, setting the access transistor well in the on state. This is then followed by (ii) discharging: V_{WL} is reduced to 0.5 V, where the access transistor is highly resistive. The effective voltage on the QD transistor gate $V_{DL,eff}$ as a function of time can be expressed as:

$$V_{DL,eff} = V_0\,[1 + \frac{R_{acc}}{R_G}\exp(-\frac{t}{\tau})]\,, \tag{6}$$

where $V_0 = \frac{V_{DL}R_G}{R_{acc}+R_G}$ is the equilibrium voltage at the QD transistor gate at $t \to \infty$, and R_G and R_{acc} are the gate leakage resistance of the QD transistor and the channel resistance of the access transistor, respectively. $\tau = \frac{C_{cell}R_G R_{acc}}{R_G+R_{acc}}$ is the circuit time constant, i.e. retention time, where C_{cell} is the parallel sum of the QD transistor gate capacitance and the storage capacitance C_C in Fig. 1b. By monitoring I_S after V_{WL} is switched from 1.49 V to 0.5 V, Coulomb oscillations are observed as a function of time due to the decay of $V_{DL,eff}$ in time, as shown in Supplementary Fig. 3b. The observed Coulomb peaks in the time domain, marked as P_1, P_2, P_3, and P_4, have their counterparts in the voltage domain, as shown in Supplementary Fig. 3c. Combining the marked Coulomb peaks in time and voltage domains, we fit the data points to (6) and find the time constant $\tau \approx 207$ ms in this

cell, as shown in Supplementary Fig. 3d. From Fig. 1d in the main text, we deduce that at V_{WL}=0.5 V, $R_{\mathrm{acc}} > R_{\mathrm{G}}$ and leakage occurs primarily through the gate resistance of the QD device. Considering $C_{\mathrm{cell}} \sim C_{\mathrm{C}}$=200 fF, we obtain $R_{\mathrm{G}} \sim$ 1 TΩ. Further improvements in the gate voltage retention time will require increasing R_{G}.

It is also worth mentioning that limited charge injection is visible in Supplementary Fig. 3b after switching. The design only uses a single nMOS pass transistor, instead of a transmission gate (nMOS and pMOS) for the access transistor, but additional dummy single-finger transistors have been included in the layout next to the access transistor to absorb and minimize the charge injected by switching.

3. Resonator characterization at room temperature

We characterize the integrated *LC* resonators by measuring the frequency spectrum, as shown in Supplementary Fig. 4. We perform the measurement at 300 K by directly connecting a VNA to the microwave port on the PCB (Setup A). We perform numerical fittings to the data by initially estimating the resonant frequencies (f_{res}) where the reflection coefficient has local minima, that is around 6.8 GHz for $\mathrm{Resonator}_1$, 7.4 GHz for $\mathrm{Resonator}_2$, 7.9 GHz for $\mathrm{Resonator}_3$, respectively. The resonant frequencies and the reflection coefficients at the resonant frequencies (ΔS_{11}) are then extracted from the fitted results (red curves in Supplementary Fig. 4). Furthermore, the quality factors (Q) can be derived from $Q = f_{\mathrm{res}}/\Delta f$, where Δf is the full width at half maximum of the fit. The characteristics of the three integrated resonators are listed in Supplementary Table 2.

It is worth mentioning that the resonators on chip are, instead of more conventional L-shaped *LC* matching networks, π *CLC* matching networks (C_{S}, L, C_{P}), since the former would impose a fixed *matching network* quality factor determined by the ratio of source (50 Ω) and load impedance (the gate of the QD device), while the latter introduces an additional degree of freedom, thus allowing to independently determine the *matching network* quality factor. Therefore, all components in the matching network are functional, not due to parasitics. Finally, however, the measured quality factor is in any case determined by the lowest *component* quality factor, in this case mostly by inductors (with a nominal $Q \sim$ 15-20).

4. Resonator characterization at deep-cryogenic temperature

We characterize the frequency spectrum of the integrated *LC* resonators at 50 mK, as shown in Supplementary Fig. 5. We first characterize the effect of the access transistor on the resonators by applying voltages to each word line individually, as shown in Supplementary Fig. 5a. At V_{WL1}=1 V, V_{WL2}=0 V, V_{WL3}=0 V, the access transistors T_{i1} are activated (grey curve in Supplementary Fig. 5a). Therefore, all three resonators show changes in S_{11} with respect to the initial state (blue curve in Supplementary Fig. 5a). Likewise, in the case when only V_{WL2} or V_{WL3} are set to 1 V, then only T_{i2} or T_{i3} are activated (red and green curves in Supplementary Fig. 5a, respectively). Because all the access transistors T_{ij} are nominally identical, then the frequency spectra for each $V_{\mathrm{WL}j}$=1 V curve overlap.

Next, we characterize individual resonators by activating each data line when V_{WL1}=1 V, V_{WL2}=0 V, V_{WL3}=0 V, as shown in Supplementary Fig. 5b. When V_{DL1} increases to 0.7 V, and V_{DL2} and V_{DL3} remain at 0 V (gold curve in Supplementary Fig. 5c), then T_{11} is turned off due to the high source voltage. The reflected power at the resonant frequency f_1 of $\mathrm{Resonator}_1$ then returns to its original state (blue curve in Supplementary Fig. 5c). By increasing each $V_{\mathrm{DL}i}$, only the corresponding resonator shows response in its frequency spectrum, and hence each resonator can be identified, as shown in Supplementary Fig. 5c-e.

5. Power consumption analysis

In the proposed row-column architecture, for N quantum devices, there are $2\sqrt{N}$ lines ($V_{\mathrm{DL}i}$ and $V_{\mathrm{WL}j}$) and $\sqrt{N}$ resonators to read $\sqrt{N}$ devices in parallel (the ones that can be frequency-multiplexed). During operation, power is dissipated due to switching, and this can increase the operating temperature and limit the number of quantum devices that can be read with this approach. The power dissipation per operation cycle is given by the power required to charge/discharge both the access transistor gates and the quantum device gates. When switching, the column to be read is activated, while the previously read column is deactivated, so, for every clock cycle, two word lines with $\sqrt{N}$ access transistors each are charged/discharged. At the same time, all data lines are swept and all $\sqrt{N}$ quantum device gates on the activated column are charged, while the others are disconnected by the access transistors in the off state. The overall power dissipation per operation cycle is then:

$$\begin{aligned} P_{\mathrm{diss,DRAM}} &= \sqrt{N} \times (2C_{\mathrm{A}}\Delta V_{\mathrm{WL}}^2 + C_{\mathrm{cell}}\Delta V_{\mathrm{DL}}^2) f_{\mathrm{a}} \\ &= \sqrt{N}\beta/t_{\mathrm{r}}, \end{aligned} \tag{7}$$

where N is the number of quantum devices, C_{A} is the access transistor capacitance to be charged/discharged by a voltage $\Delta V_{\mathrm{WL}} = V_{\mathrm{WL}}^{\mathrm{High}} - V_{\mathrm{WL}}^{\mathrm{Low}}$, while C_{cell} is the cell capacitance (containing the storage capacitance C_{C} and also the much smaller quantum device capacitance) to be charged/discharged with a voltage $\Delta V_{\mathrm{DL}} = V_{\mathrm{DL}}^{\mathrm{High}} - V_{\mathrm{DL}}^{\mathrm{Low}}$, β is a technology parameter, $f_{\mathrm{a}} = 1/t_{\mathrm{r}}$ is the single-column activation frequency and t_{r} is the single quantum device readout time. Multiple operations are needed for a full readout cycle, namely the total readout time is:

$$t_{\mathrm{tot}} = \sqrt{N} \times t_{\mathrm{r}}. \tag{8}$$

In the case of qubits, one needs to make sure that the total readout time is shorter than the qubit relaxation/decoherence time, depending on the quantum computing application, because the state should be preserved until it is read

out again.
By replacing (8) into (7) one obtains:

$$P_{\text{diss,DRAM}} = N \times (2C_{\text{A}}\Delta V_{\text{WL}}^2 + C_{\text{cell}}\Delta V_{\text{DL}}^2)/t_{\text{tot}} \\ = N\beta/t_{\text{tot}}. \tag{9}$$

The access transistor capacitance C_{A} can be estimated to be 15 fF from simulations and the cell capacitance C_{cell} to be 200 fF from circuit component values. The word-line voltage change can be set to ΔV_{WL}=1 V to completely turn on/off the access transistor, as in our measurements, with a data-line voltage change ΔV_{DL}=10 mV, to perform a projective spin readout across a voltage range a few times the full width at half maximum of a Coulomb-blockade peak. The total readout time t_{tot} can be set to 25 ms, as in our experiment. Under these assumptions, the overall power dissipation per QD-access transistor cell becomes 1.2 pW/qubit (1.2 pW$\times N$ in total).
At this point, assuming a maximum power dissipation allowed in the dilution refrigerator at the sample stage of 400 μW, which would make the mixing chamber temperature rise up to 100 mK, the calculated power consumption per qubit would allow to read out in excess of 3.3×10^8 devices in the row-column architecture from the mere power consumption standpoint.

The power consumption expression in (7) is plotted in Supplementary Fig. 6 for a given number of qubits N and as a function of single-qubit readout time t_{r}. This expression, for $P_{\text{diss,DRAM}}$=400 μW, determines a boundary (blue line in Supplementary Fig. 6) that separates the plot in a region where operation is allowed, since the cooling budget is met, and a region where operation is prohibited, because either the excessive number of qubits or the short readout times (fast operating frequency) would result in excessive power consumption.

In the case of Noisy Intermediate Scale Quantum (NISQ) algorithms, the requirement is that the state of the qubit is preserved until read at the end of the computation. Considering a single-electron spin qubit implementation in silicon, where T_1 times approaching 10 s have been measured [57, 58], and considering that for high fidelity readout the measurement time needs to be $T_1/25$ [59], the total readout time should not exceed 0.4 s (red dashed line in Supplementary Fig. 6). By referring to the previously derived expression (9), such readout time would result in 75 fW/qubit (75 fW$\times N$ in total) power consumption with the row-column architecture. If this is compared now to the typical power budget of a dilution refrigerator at the sample stage of 400 μW, it would allow to read 5.3×10^9 devices in the row-column architecture.

For the quantum-error-correction era, the requirement should be to measure faster than the coherence time of the qubit $T_2 \sim 28$ ms [4], so a qubit measurement should occur approximately every 1 ms (orange dashed line in Supplementary Fig. 6). In this case, by replacing a total readout time t_{tot}=1 ms in the power dissipation expression (9), the power consumption per cell would become 30 pW/qubit (30 pW$\times N$ in total). Consequently, assuming the same limit of 400 μW for the sample stage of modern dilution refrigerators, this would allow to read out 1.3×10^7 qubits with the presented row-column architecture. Note that here we consider that dynamical decoupling techniques are used between cycles.

Finally, for fast fault-tolerant quantum computing (QC), the requirement should be to be able to read the array faster than approximately 10× the intrinsic gate time (roughly the number of operations in an error correction cycle [14]). For electron spins in silicon, considering an intrinsic gate time around 100 ns, one would need a total readout time of such an architecture in the order of 1 μs (green dashed line in Supplementary Fig. 6). In this case, the power consumption per cell would become 30 nW/qubit (30 nW$\times N$ in total) leading to a total of 1.3×10^4 qubits that can be read with this architecture. Note that this would require a single-qubit readout time of $\sim$ 10 ns, which is below the fastest high-fidelity electrical readout demonstrated of 300 ns [60].

Therefore, one can conclude that the proposed architecture gives a clear advantage in terms of wiring with respect to brute-force scaling, while keeping the power consumption below the required target both for present and near-term quantum systems.

Supplementary References

[53] Simmons, C. B. et al. Tunable spin loading and T_1 of a silicon spin qubit measured by single-shot readout. *Phys. Rev. Lett.* **106**, 156804 (2011).

[54] Petit, L. et al. Spin lifetime and charge noise in hot silicon quantum dot qubits. *Phys. Rev. Lett.* **121**, 076801 (2018).

[55] Betz, A. C. et al. Dispersively detected Pauli spin-blockade in a silicon nanowire field-effect transistor. *Nano Lett.* **15**, 4622–4627 (2015).

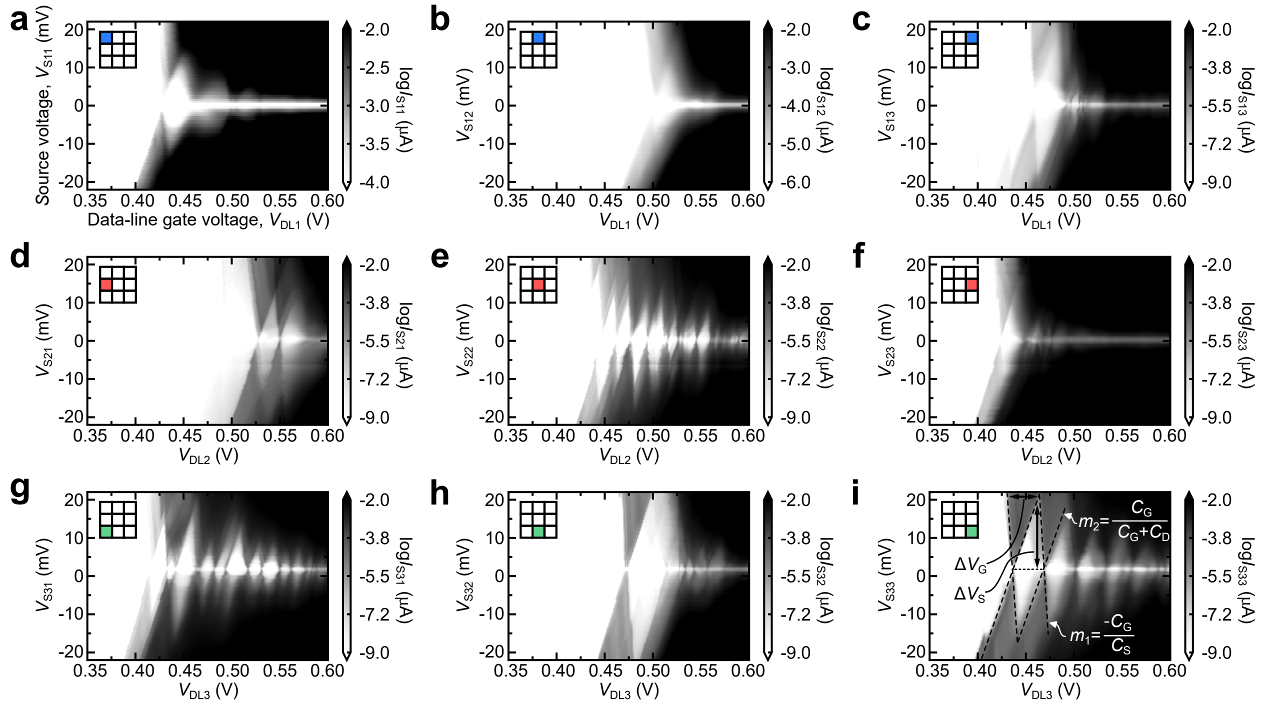
[56] Crippa, A. et al. Gate-reflectometry dispersive readout and coherent control of a spin qubit in silicon. *Nat. Commun.* **10**, 2776 (2019).

[57] Morello, A. et al. Single-shot readout of an electron spin in silicon. *Nature* **467**, 687–691 (2010).

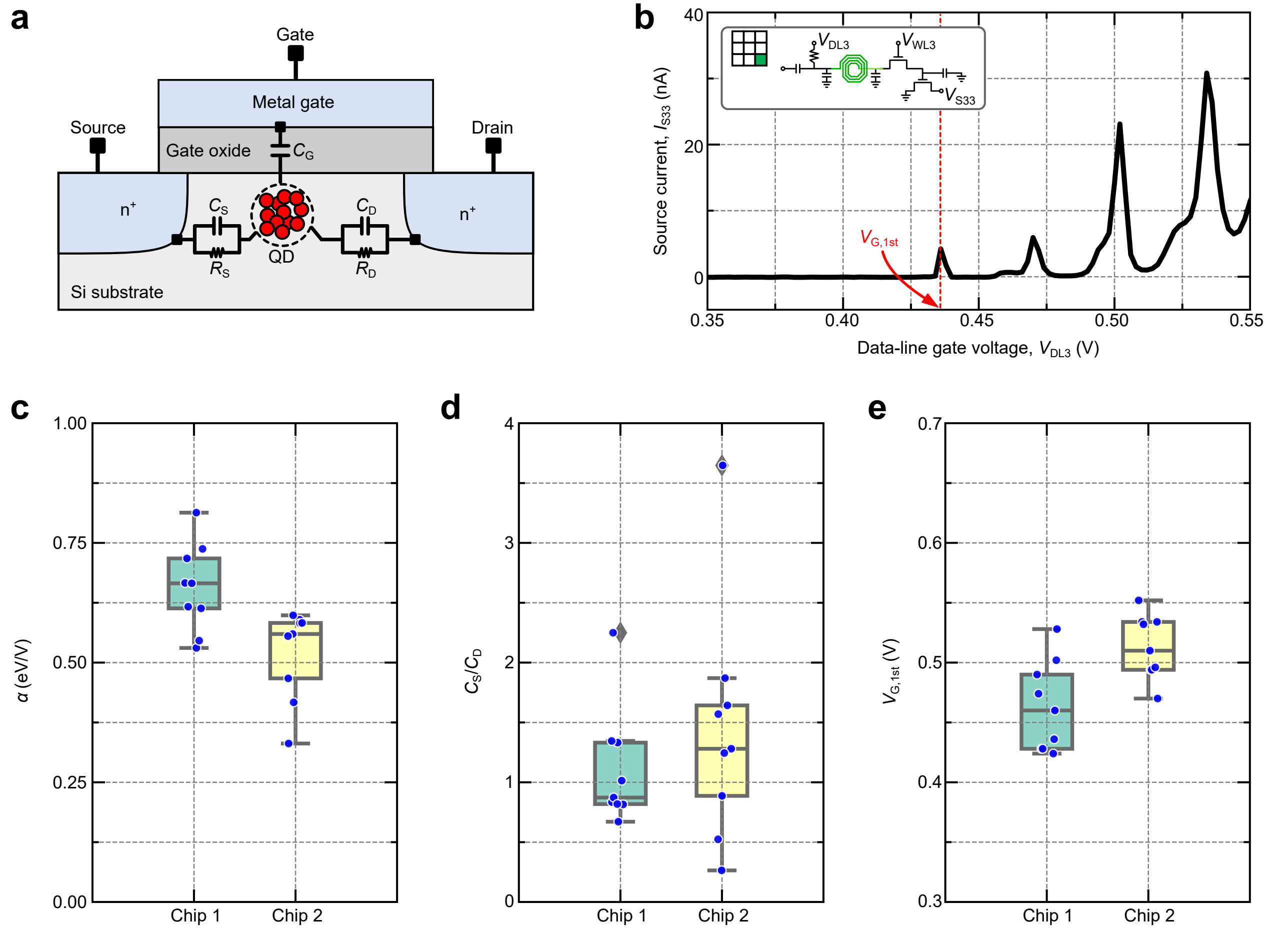
[58] Ciriano-Tejel, V. N. et al. Spin readout of a CMOS quantum dot by gate reflectometry and spin-dependent tunneling. *PRX Quantum* **2**, 010353 (2021).

[59] Laucht, A. et al. Roadmap on quantum nanotechnologies. *Nanotechnology* **32**, 162003 (2021).

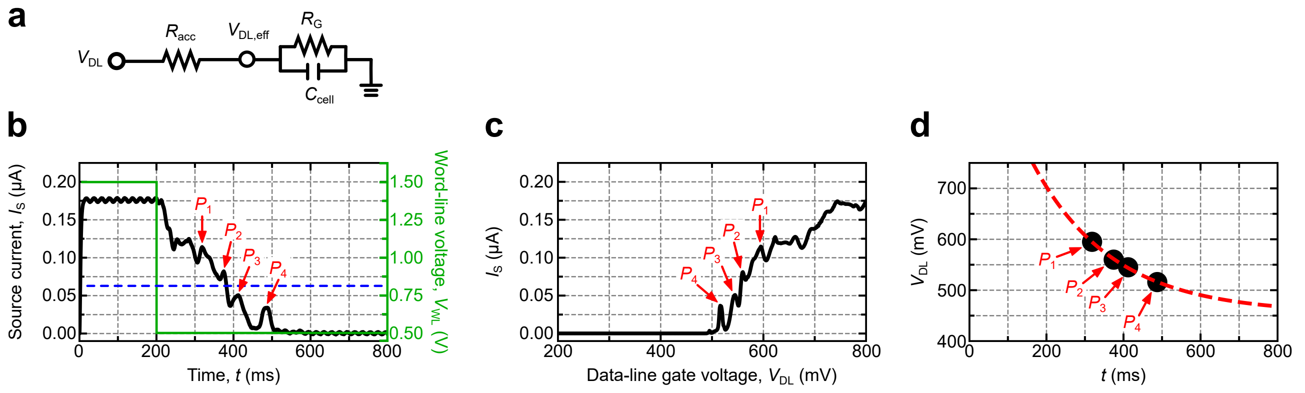
[60] Ibberson, D. J. et al. Large dispersive interaction between a CMOS double quantum dot and microwave photons. *PRX Quantum* **2**, 020315 (2021).



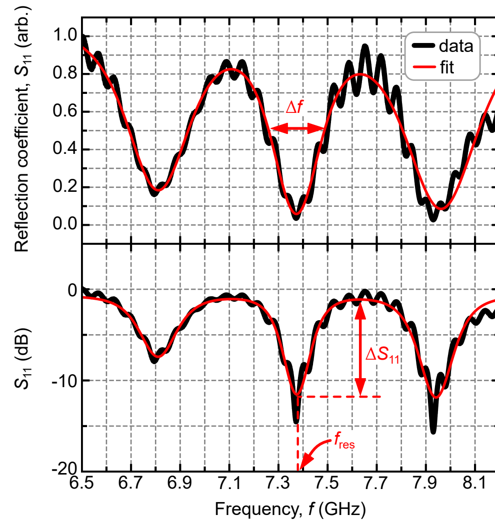
Supplementary Fig. 1 | Coulomb diamonds at 50 mK. a-**i**, Coulomb diamonds of quantum devices Q_{11} to Q_{33}, respectively. In these measurements, all word-line voltages are V_{WLj}=1.5 V. The insets indicate the corresponding quantum device in the 3×3 device array, with a different color for each row.



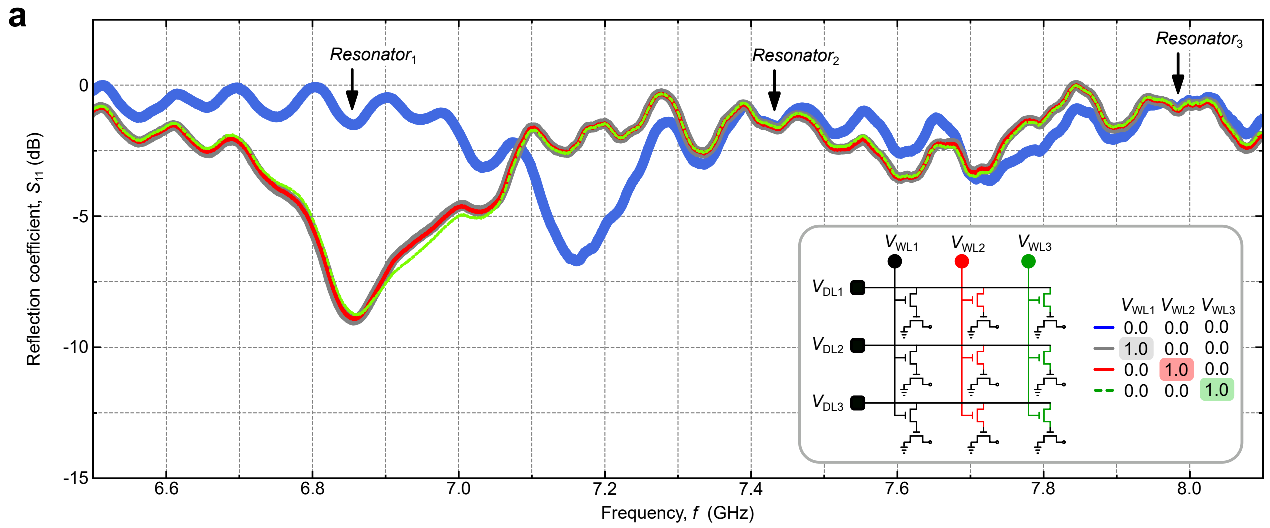
Supplementary Fig. 2 | Summary of the electrical transport parameters of DUTs in this work. a, Schematic of a MOS transistor cross-section view in this work. Red spheres represent charge carriers. **b**, Source current I_{S33} as a function of data-line gate voltage V_{DL3} for device Q_{33} at V_{WL3}=1.5 V, to highlight $V_{G,1st}$, the gate voltage where the first observable Coulomb-blockade peak occurs. Inset, circuit diagram of a single cell. **c**, Box plot for the gate lever arm parameter α. **d**, Box plot for the ratio of source-QD capacitance to QD-drain capacitance C_S/C_D. **e**, Box plot for the gate voltage of the first observable Coulomb-blockade peak $V_{G,1st}$. Chip 1 represents the DUT described in the main text, and consists of nine quantum devices. Chip 2 represents another chip, which is not discussed in the main text, with quantum devices having nominally identical physical dimensions as in Chip 1.

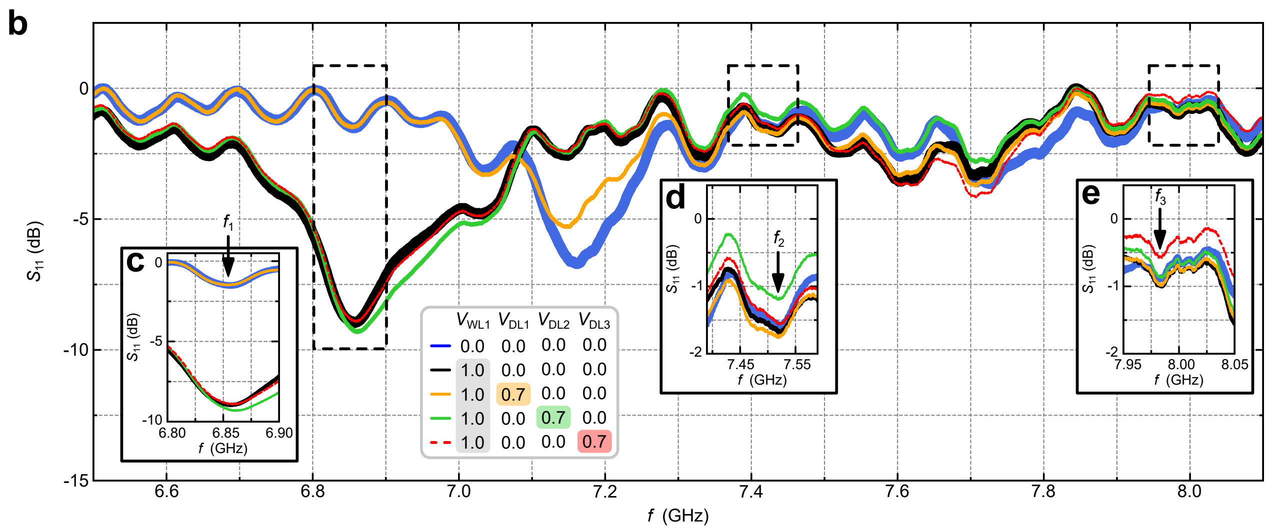


Supplementary Fig. 3 | Retention time experiment. **a**, Equivalent circuit of a single cell. **b**, Source current I_S as a function of time t with the application of word-line voltage V_{WL}^{High}=1.49 V for t=0 to 200 ms and V_{WL}^{Low}=0.5 V afterwards (green line), and a constant data-line voltage V_{DL}=0.8 V (blue dashed line). **c**, I_S as a function of V_{DL} at V_{WL}=1.49 V. **d**, Locations of Coulomb-blockade peaks P_1, P_2, P_3, and P_4 in time domain (in **b**), as a function of the corresponding Coulomb peaks in voltage domain (in **c**). The red dashed line is an exponential decay fit.

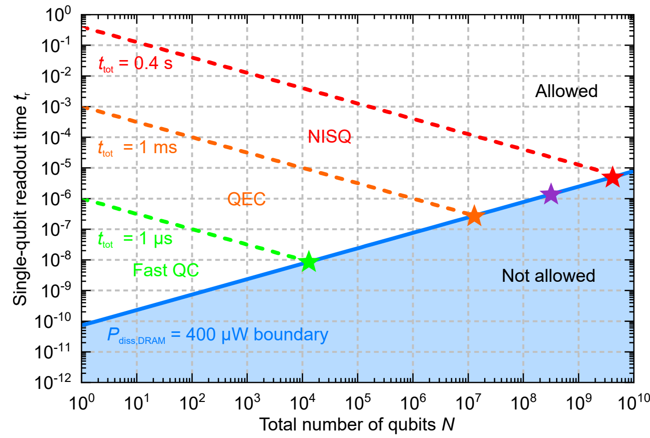


Supplementary Fig. 4 | Frequency spectrum analysis at 300 K. Frequency spectrum of the integrated *LC* resonators at room temperature, where the black traces are the experimental data in linear scale (upper panel) and decibel scale (bottom panel), and the red curves are squared Lorentzian fits. ΔS_{11}, f_{res}, and Δf are the dip amplitude, center of the dip, and the dip full width at half maximum from the fit, respectively.





Supplementary Fig. 5 | Frequency spectrum of the multi-resonator at 50 mK. a, Frequency spectrum at different word-line voltages. Data-line voltages are V_{DL1}=V_{DL2}=V_{DL3}=0 V. The black arrows indicate the locations of the resonant frequencies of Resonator$_1$, Resonator$_2$, and Resonator$_3$. **b**, Frequency spectrum at different data-line voltages. Word-line voltages are V_{WL1}=1 V (except for the blue curve) and V_{WL2}=V_{WL3}=0 V. **c**-**e**, Zoom-in of the black dashed regions, where the arrows indicate the resonant frequencies f_1, f_2, and f_3 corresponding to Resonator$_1$, Resonator$_2$, and Resonator$_3$, respectively.



Supplementary Fig. 6 | Power consumption analysis for different quantum computing scenarios. Boundary regions between different quantum computing scenarios as a function of total number of qubits N and single-qubit readout time t_r. The blue line indicates the cooling power boundary of a standard dilution refrigerator. Under this line the power consumed by the matrix exceeds the cooling budget and this area is not allowed. Above this line the power consumption is below the limit and this area is allowed. The red, orange and green dashed lines indicate regions of constant total readout time t_{tot} and bound the parameter space where noisy intermediate-scale quantum algorithms, quantum error correction and fast quantum computing will be possible.

Supplementary Table 1 | Benchmark of the electrical transport parameters. α, C_S/C_D, and $V_{G,1st}$ of Chip 1 and Chip 2 in Supplementary Fig. 2c-e, respectively.

	Chip 1	**Chip 2**
α (eV/V)	0.656±0.086	0.520±0.089
C_S/C_D	1.106±0.461	1.436±0.926
$V_{G,1st}$ (V)	0.463±0.035	0.513±0.025

Supplementary Table 2 | Characteristics of the integrated *LC* resonators at 300 K. Resonant frequency (f_{res}), reflection coefficient at the resonant frequency (ΔS_{11}), and quality factor (Q) from Supplementary Fig. 4.

Parameter	*Resonator*$_1$	*Resonator*$_2$	*Resonator*$_3$
f_{res} (GHz)	6.810	7.374	7.941
ΔS_{11} (dB)	-6.709	-10.830	-11.169
Q factor	14.478	20.111	14.850