Nicholas Jaber

ECE 590, Final Paper

Sunday, November 21st

Key Technologies for Scaling Superconducting Quantum Computers

In 2000, as the framework for quantum computers began to take shape, David P. DiVincenzo, of IBM, proposed a set of criteria which would be required to build a useful quantum computer. This fundamental definition of goals is the generally agreed upon criteria needed to build a Noisy Intermediate Scale Quantum (NISQ) computer, which would enable a quantum speedup. That is, an exponential speedup over classical supercomputers using quantum computers. The definition purposed has been dubbed the "DiVincenzo Criteria" and are as follows: scalable system with well characterized qubits, the ability to initialize the state of qubits, coherence times being much longer than gate operation time, a universal set of gates and qubit specific measurement capability [2]. A scalable system means a modular design, which is small enough to be repeatedly added to the quantum chip. Well characterized qubits are qubits which do not have frequency collisions and are anharmonic systems. Frequency collisions are when a qubit oscillates at the same frequency as another qubit causing degeneracies, which increases cross talk and eventually leads to gate errors [12]. Anharmonicity is the ratio of the transition energy from the ground state to state 1 and the transition energy from state 1 to state 2 [2]. A more anharmonic system will be less likely to enter state 2 and as result is less likely to destroy the quantum information. Initialization is the ability to reliably set a starting state of the qubit. Coherence times come in two main forms T_1 and T_2 coherence times. T_1 measures changes in qubit energy independent of gate operation. This excitation or relaxation of qubits leads to an irrecoverable loss of quantum information. T_2 measures the dephasing; through dynamic decoupling codes, this information can be recovered, if gate speeds are fast relative to decay times. A universal set of quantum gates can be generated from CNOT, Hadamard and phase gates [24]. Finally, one must be able to reliably measure qubits, without significantly modifying the state of the quantum computer known as non-demolition measurement [1].

Of the many quantum computation systems, only superconducting qubit and trapped ion systems have made significant progress towards the stated goals of the DiVincenzo Criteria. In this paper we will be focusing on superconducting quantum systems, because they are arguably the most advanced technology in terms of the DiVincenzo Criteria. This is because superconducting quantum systems have an established scaling process, very short gate times, minimal crosstalk and minimal frequency collisions [18]. The main drawbacks of superconducting systems are two-level systems forming in and on the boundaries of the dielectrics and the short coherence times [1,21]. As the materials, design and post-processing advance, researchers are able to increasingly suppress these types of errors [21].

Of the many superconducting quantum system designs, there are three main archetypes, which are defined by their E_J/E_C ratio. E_J/E_C is the ratio of energy stored in the qubit's josephson junction divided by the energy stored in the qubit's capacitor. The three archetypes are charge qubits (defined by $E_J/E_C << 1$), flux qubits (defined by $100 > E_J/E_C >> 1$) and phase qubits (defined by $E_J/E_C >> 1$) [18]. There are two major considerations when choosing an E_J/E_C ratio: anharmonicity and charge noise.

Anharmonicity increases with E_J/E_C ratio; by increasing anharmonicity one incidentally increases sensitivity to low frequency, 1/f charge, noise [5]. The current dominant superconducting qubit design is the transmon, this is a capacitively shunted flux qubit design, which exponentially reduces sensitivity to 1/f charge noise, while linearly decreasing anharmonicity [1,18].

There are four leading transmon designs: traditional transmons, xmons, gmons and three-dimensional transmons [18]. The "sweet spot" for the optimal configuration of anharmonicity and charge noise of traditional transmons is $E_J/E_C \sim 100$ [18]. A diagram of a traditional transmon is shown in **figure 1a**. Traditional transmons have a fixed frequency and always on coupling between qubits. This is particularly advantageous, because of their low charge dispersion meaning that T_1 and T_2 coherence times can exceed 100 microseconds [22]. Xmons are similar in design and functionality to traditional transmons, the main difference is owed to the independent XY and Z control lines, as shown by the red and green lines respectively in **figure 1b** [18]. Independent control lines means faster gate speeds and as a result of dynamic decouple longer T_2 coherence times. The drawbacks of this design are increased thermal dissipation through control lines, leading to shorter T_1 coherence times, and increased design complexity, which can introduce unforseen crosstalk while scaling [23]. Gmons have the same independent XY and Z controls as xmons, but additionally implement inductors to modulate qubit coupling strength, as seen in **figure 1d** [4,18]. This means that gmons can limit frequency crowding and increase computation complexity by selectively coupling for a short period [4]. Unfortunately, this comes at a cost to the qubit's coherence time [4]. Three-dimensional transmons utilize the traditionally wasted space above and below the qubits to serve as a waveguide cavity, which means qubits can have a high resistance to charge and magnetic field noise, as shown in **figure 1c** [7,18]. Because three-dimensional transmons have a higher resistance to noise, qubits can be packed closer together. The main drawback to this design is the untraditional manufacturing process [7]. All of these transmon designs are used in commercially viable systems. There is still a general disagreement amongst the large superconducting qubit companies, about which transmon design is best suited for NISQ computers.

The three most advanced superconducting quantum computer companies are Google, IBM and Rigetti. Of these Google and IBM are the current front runners and they have diametrically opposed design philosophies. Google's quantum computation research began in earnest through funding and eventual employment of Dr. Martinis, a professor at University of California, Santa Barbra. Google built on Martinis' design using a fixed frequency xmon architecture, starting with their 22 qubit Foxtail architecture in 2017 and the 72 qubit Bristlecone architecture in 2018 [10]. In 2019 Google switched to the 54 qubit tunable frequency gmon design, because they were running into many issues with frequency collision and unwanted coupling. This is likely a result of high connectivity inherent in their architecture, which consisted of a dense grid of x shaped qubits, as shown in **figure 2a** [18]. Google is able to specifically control each qubit to qubit coupling strength with 140 individually tunable zones [10,18]. This allowed Google to arguably declare that they had achieved quantum supremacy (developed a NISQ that could perform a computation infeasible on a classical supercomputer) in 2019 [10,14,15]. This has been disputed by IBM, because the problem Google solved was designed by Google and was of no real world use [10,14,15]. In fact, it was designed to be particularly hard for classical computers and easy for quantum computers, but was still completed by IBM in ~2 days on their supercomputer [10,14,15].

IBM has been working on superconducting quantum computers for over 2 decades and as a result have developed many processors with 5 main design architectures [8,13]. Over time the number of interconnections per qubit has dropped with each successive architecture [13]. Similar to Google, IBM found fixed frequency transmons in highly interconnected qubit systems have far too many unwanted frequency collisions. IBM however found that simply decreasing the number of interconnects and increasing the space between qubits allowed for a dramatic increase in fidelity, without veering from traditional qubit designs [8,9]. Eliminating the transition to flux tuned coupling allows IBM to maintain low "crosstalk" by preventing magnetic fields from triggering adjacent qubits unintentionally. IBM has decided that the "heavy-hex" design is the greatest compromise between computational complexity and gate fidelity [8,9,13]. This heavy-hex design is a series of hexagons stacked together sharing one side with each of their neighbors and has an additional qubit in the middle of each side, meaning that each hexagon contains 18 qubits, as shown in **figure 2b** [9]. IBM has used this design since their 27 qubit Canary architecture in 2019 and has been utilized on the 65 qubit Hummingbird architecture in 2020 [11]. Most recently IBM released the 127 qubit Eagle architecture in 2021 [11]. IBM claims that this Eagle architecture has fully surpassed the quantum supremacy milestone, but has yet to release the research paper or additional data on their Eagle processor.

Finally, Rigetti, a quantum computer hardware startup has combined the two main architectures implemented by Google and IBM for a hybrid fixed and variable frequency multi-die design [17,23]. This has yielded the most theoretically powerful superconducting design, but has introduced significant obstacles to practical implementation. Rigetti developed octagonal and enneadecagon (19 sided polygon) rings of alternating fixed and variable frequency qubits [16,17,20]. These rings are printed on small dies. These dies are then capacitively coupled through a plate which they all sit above [20,23]. The octagonal version of this model is shown in **figure 2c** [23]. This idea of highly modular and reconfigurable designs can rapidly scale and reconfigure quantum circuits to optimize hardware for specific types of quantum algorithms [23]. Rigetti is currently attempting to scale their multi-die ring structures to a 128 qubit system [16]. Given their financial limitations and the complicated nature of multi-die systems, inter-die gates currently have highly variable fidelities. Despite this, hardware optimization of quantum processors could prove to be a powerful addition to this field.

There are many technologies that have dramatically improved superconducting qubit performance, but I will focus on three technologies which have the potential to radically improve NISQ computer performance in the next few years. These technologies are as follows: Laser Annealing of Stochastically Imparied Qubits (LASIQ) published in 2020, Quantum Integrated Circuits (QuICs) published in 2021 and Josephson Quantum Filter (JFQ) published in 2020 [19,22,23]. I chose these three technologies, because they were each published in the last 2 years and each solves a major limitation of current superconducting quantum computers.

One major issue with manufacturing fixed frequency qubits is the wide tolerances. Typical transmon frequency tolerances are about 1-2% [22]. This wide tolerance is generated by a 2-4% variance in the tunnel junction resistance [22]. This tolerance can be vastly narrowed by using the LASIQ method to tune each qubit independently. LASIQ is the process of irradiating quantum processors in steps to tune fixed frequency qubits to a desired frequency [22]. This is done by selectively irradiating individual qubits with visible light (532 nm) laser [12,22]. This laser is 1.6-2 W and is engaged for 2-80 seconds depending on the needed change in frequency [12]. Transmon tunnel junction resistance is measured and then irradiated in incremental steps [22]. This works because irradiating the circuitry for an unknown reason, increases tunnel junction resistance. Because tunnel junction resistance and frequency are linearly related, one can continuously measure the resistance and then re-irradiate until the resistance has changed such that the desired frequency is met [22]. Using these methods researchers at IBM were able to precisely tune 390 fixed frequency qubits to within .3% (~10MHz) of their desired target, with a success rate of 89.5% [22]. This radical reshaping of qubit frequency is shown in **figure 3c**. The resistance of some qubits has been shifted by as much as 14.2%, as shown in **figure 3b** [22]. Frequency tuning is most accurate when changes in tunnel junction resistance are towards the middle of the viable 1%-14% change in resistance range [12]. All of these facts together means that a large number of widely variable fixed frequency qubits can be independently tuned over a large range of frequencies to a desired frequency within a tight tolerance at a high reliability. Ideally, the reliability of this shift will be fine tuned such that it can approach 100% in the soon future. This will certainly be attainable, if manufacturing can bring its tolerances towards the center the 1-14% range. Even with current limitations of this technology, the collision free yield of qubits studied is predicted to be increased from less than 5% to a predicted greater than 90%, as shown in **figure 3a** [22]. LASIQ was able to do all of this without negatively affecting the long qubit coherence times native to fixed frequency transmon systems [22]. This all means that creating high collision free yield NISQ computers is in the near future. This technology was likely implemented on IBM's latest Eagle processor. Hopefully future developments of LASIQ and other post-processing technologies will allow fixed frequency transmons to continually increase their performance and scalability [12].

One major limitation to conventional scaling architectures is the rigidity of their design. Rigetti hopes to change this and to upset the status quo by implementing their QuICs system. This system would allow Rigetti to easily design a number of chip architectures each optimized for different types of quantum algorithms [23]. For a wafer of 220 QuIC dies there are over 2.2 billion unique possible die organizations [23]. The current generation of these dies are 8 qubit dies with alternating fixed and flux modulating frequency qubits [17,20]. The idea is to build dies with an octagonal qubit orientation on each die where each qubit is coupled to its adjacent qubit and to the nearest qubit on the next die, as shown in **figure 2c** [16,20]. By maintaining the die orientation, one can ensure that a fixed frequency qubit is only ever coupled to a flux tunable qubit, as shown in **figure 2c**. This is very advantageous, because one has a much greater distance between tunable qubits minimizing crosstalk [23]. Additionally, coupling a tunable qubit only to fixed qubits means the possibility of frequency collisions is minimized. The main downside of this design is the complexity. More types of qubits means more control infrastructure and a more complicated troubleshooting process. Currently, fixed and tunable qubits have a T_1 coherence time of 73 and 18 microseconds and a T_2 coherence time of 43 and 15 microseconds respectively [23]. The very short tunable qubit coherence times are a critical factor Rigetti is hoping to improve on. This is believed to come from unwanted interaction between capacitive coupling and flux modulation [23]. Notably this is not caused by inter-die interactions, as shown by **figure 4a** [23]. In Rigetti's latest QuICs design, the 8 qubit dies are elevated above the main plate by a series of indium "bumps" and are capacitively coupled to the main plate, as shown in **figure 4c** [23]. This bump design originated in classical computing flip-chip design, where classical chips are linked together [3]. Indium was used in classical flip-chip design, because of its flexibility and resistance to oxidation [3]. These indium bumps are cylinders of indium deposited on the surface of the main plate [23]. These bumps have a radius of 20 microns and a height of 6.5 microns before die compression and 1.5-4 microns after bonding [23]. It is extremely challenging to maintain a sub micron coplanar surface between the die and the plate, because of die curvature, plate curvature, thermal expansions and contractions during bonding and while lowering assembly temperature from room temperature to millikelvin [23]. The wide variety of bump heights means that the strength of inter-die coupling will be highly variable leading to a divergence between designed and measured qubit frequency, as shown in **figure 4b [23]**. Currently, fidelities of 99 - 99.5% can be achieved in the best case despite the wide variety of post bonding indium bump heights [23]. In order to significantly increase the fidelity Rigetti must reduce the spread of heights and modify capacitor geometry to minimize sensitivity to height [23]. If Rigetti is able to solve these engineering limitations, they will be able to optimize computation complexity and stability based on the required quantum algorithm, vastly increasing the quantum speedup for certain quantum algorithms.

The final major limitation to scaling superconducting qubits that we will cover is dissipation of qubit thermal energy into control lines. The current standard to diminish this effect is implementation of a Purcell filter [19]. Purcell filters act as a band pass filter and thus become ineffective when the radiative frequency is near the control frequency [19]. To significantly increase performance one must further decouple control lines and in turn increase control signal power, however increasing the control signal power can lead to uneven heating and crosstalk [19]. This places a practical limit on coherence times; this limit is removed by the addition of a JQF to the control line. The introduction of a JQF allows researchers to reflect relaxation energy from flowing onto the control lines as a result of T_1 qubit decoherence, without disrupting control signals, as shown in **figure 5c** [19]. This JQF acts as a nonlinear mirror, by reflecting the single photon decay energy, but becoming inert when many photon control signals are sent [6]. This incredible advance is shown by an increase of the thermal population from 2.8% to 16.2% [19]. This does not represent additional thermal noise, but shows that the qubit is now decoupled from the control lines [19]. As evidenced by the much longer T_1 and T_2 coherence times, when the JQF is near resonance with the qubit, as seen in **figure 5b** [19]. The main limitations of the JQF design are the requirements that the JQF needs to be one half of the rabi wavelength away from the qubit on the control line and that the JQF needs to be flux tuned, as seen in **figure 5a** [19]. This means that more complicated designs need to be implemented, so that the JQF is one half of a rabi wavelength from the qubit, without having the flux needed to tune the JQF cause unwanted interference with neighboring qubits [19]. The vast benefit in coherence times provided by JQF will likely have a profound impact on NISQ computer's evolution from experiment to tool.

Superconducting quantum computers are currently the most advanced form of quantum computation. With the diverse set of approaches to solve the remaining design limitation by academic researchers, as well as major superconducting qubit companies such as Google, IBM and Rigetti, we will soon develop useful superconducting NISQ computers. While following fixed frequency, flux tunable frequency and hybrid systems simultaneously is not the most efficient way to solve this problem, it is advantageous to work on every feasible platform as many revolutionary technologies can only be applied to specific platforms. For example, LASIQ can only be applied to fixed frequency transmons in quantum processors like IBM's Eagle. Other technologies such as QuICs are more broadly applicable and can be an enormous benefit, only if manufacturing can significantly narrow the range of post bonding indium bump heights [21]. The last type of upcoming technology is those like JQF which have broad applicability and can cause a major transformation of the superconducting quantum computer industry right now. Given that superconducting quantum computers currently have the greatest computation yield and a number of highly impactful technologies poised to revolutionize the quantum industry, superconducting qubits will remain the dominant quantum computation technology for the foreseeable future.

Figure 1

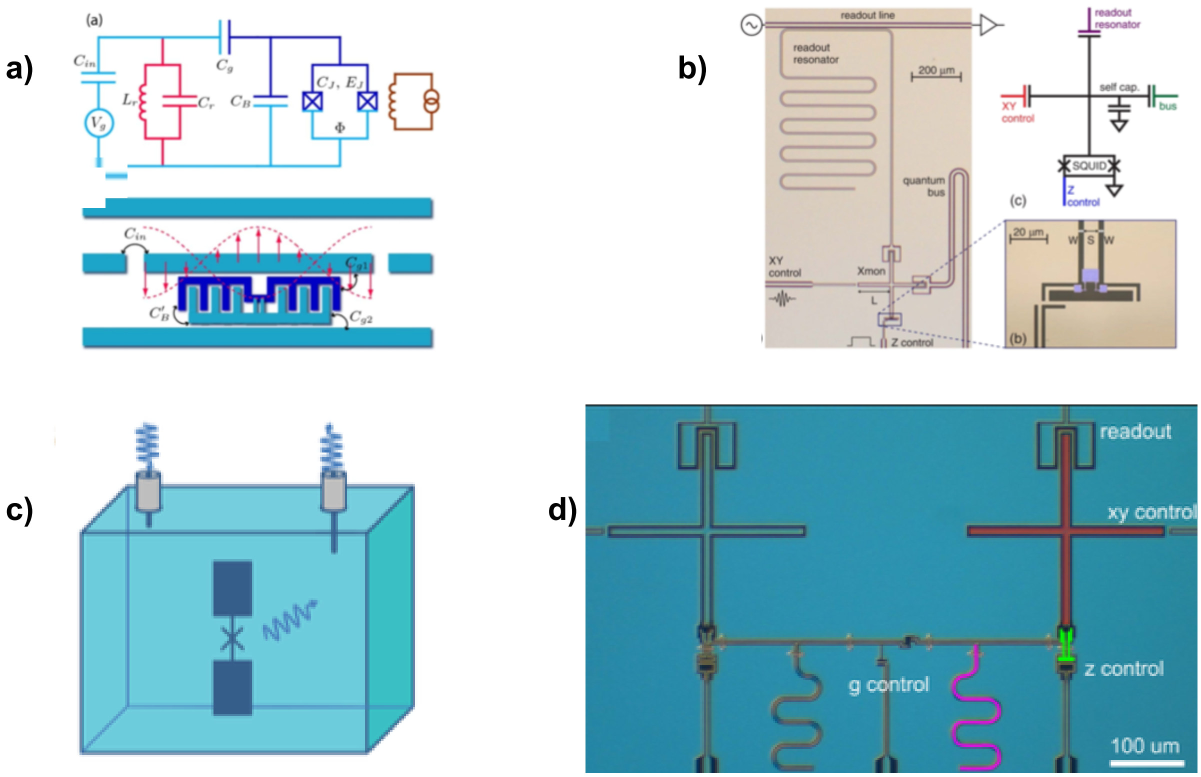

Figure 1 a) Traditional transmon circuit diagram E_J/E_C ~ 100 [18]. **b)** Xmon circuit diagram and lithography with independent (red) XY and (green) Z control lines [18]. **c)** Three-dimensional transmon diagram of a transmon with waveguide cavity and incoming noise as shown by blue wave [18]. **d)** Gmon diagram with modulated flux coupling (purple) to rapidly turn on and off qubit coupling [3,18].

Figure 2

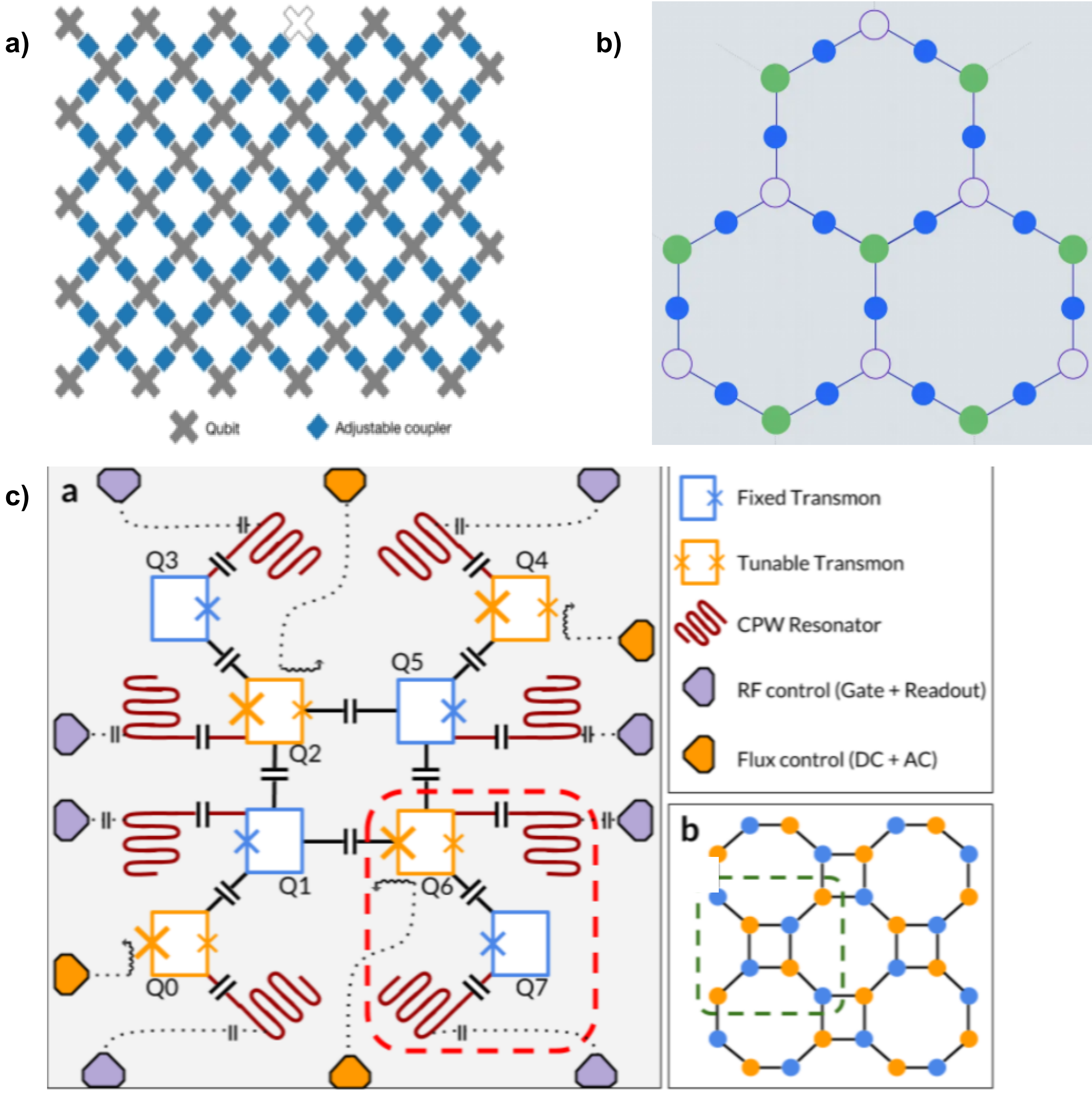

Figure 2 a) Google's current 54 gmon qubit Sycamore architecture released in 2019 [19]. **b)** IBM's heavy-hex design implemented on it's 127 fixed frequency transmon qubit Eagle architecture released in 2021 [11]. **c)** Rigetti's modular multi-die design with a hybrid fixed and variable frequency qubit architecture [20].

Figure 3

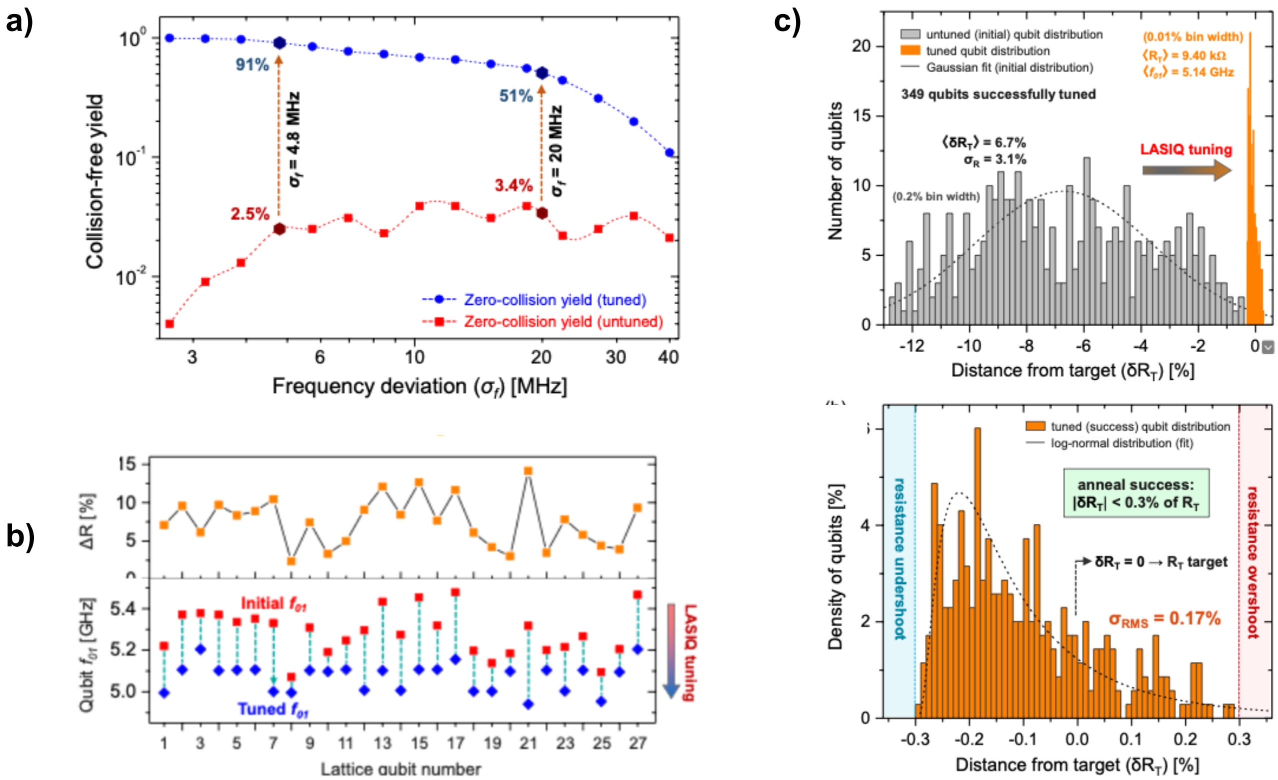

Figure 3 Laser Annealing of Stochastically Impaired Qubits (LASIQ) **a)** Using LASIQ frequency collisions in fixed frequency qubits can be dramatically reduced by precisely modifying transmon frequency [22]. **b)** The upper graph indicates the change in tunnel junction resistance through the transmon while pulses of light thermally load the qubit. The lower graph indicates the wide variety of transitions in frequency. The magnitude of this frequency transition is directly related to the change in resistance [22]. **c)** 349 qubits with a wide variety of qubit frequencies can be focused into a very narrow region (within .3% of the target frequency) [22].

Figure 4

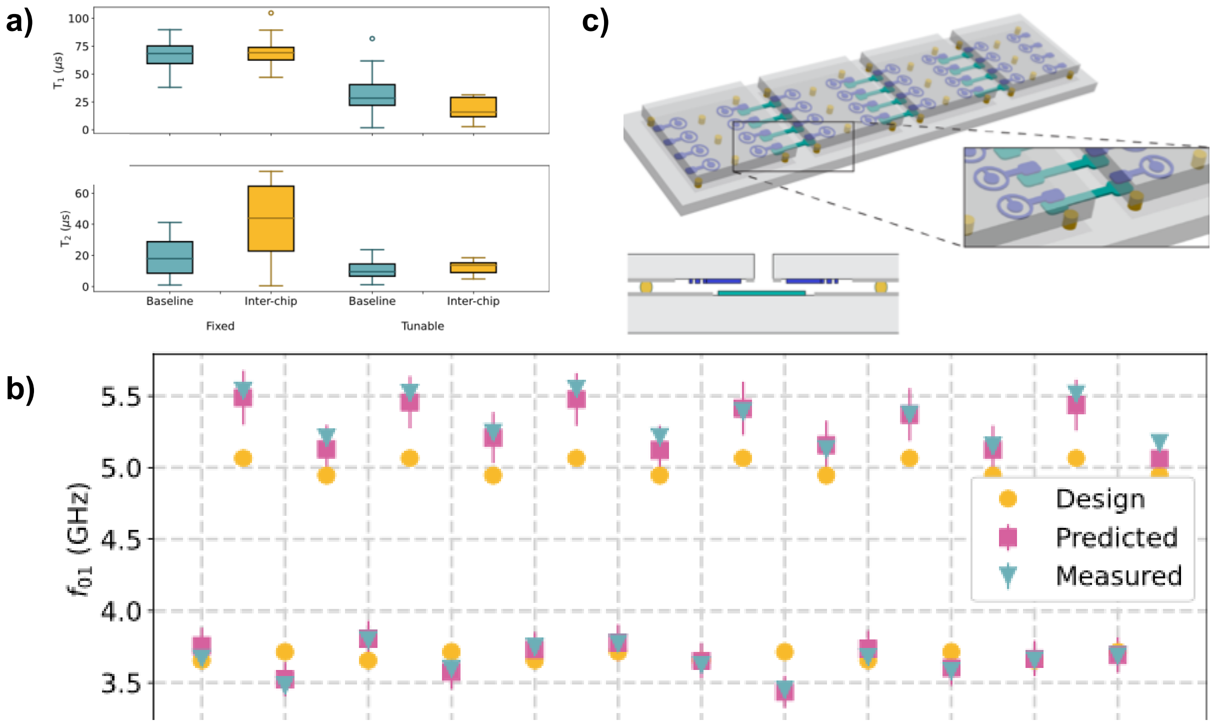

Figure 4 Quantum Integrated Circuits (QuICs) **a)** T_1 and T_2 coherence times are not significantly affected by inter-chip coupling [23]. **b)** Variable heights of indium bumps and unwanted interactions between capacitor and flux modulation causes discrepancy in qubit frequency [23]. **c)** Indium bumps raise the die above the main plate setting the height of the capacitor responsible for inter-die coupling [23].

Figure 5

a) **c)**

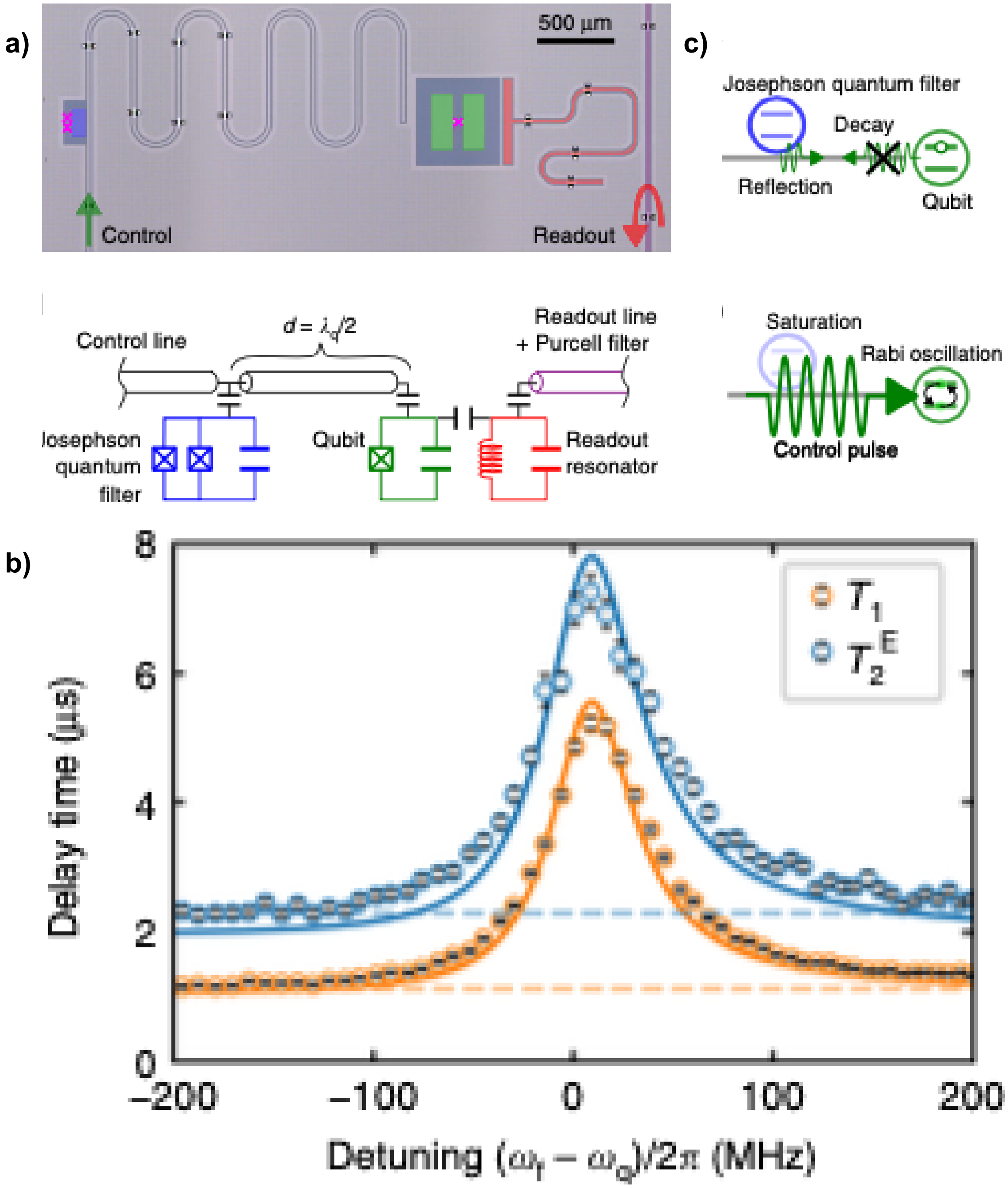

Figure 5 Josephson Quantum Filter (JQF) **a)** Diagram of a JQF on a fixed frequency transmon control line [19]. **b)** T_1 and T_2 coherence times are significantly increased by tuning JQF to near the same frequency as the transmon. These coherence times are far higher than the non-JQF system as indicated by the dashed lines [19]. **c)** Demonstration of decay reflection and inert to large amplitude control signals [19].

Works Cited

1. Krantz, P. *et al.* A Quantum Engineer's Guide to Superconducting Qubits. *Appl. Phys. Rev.* **6**, (2019).

2. DiVincenzo, D. P. The physical implementation of quantum computation. *Fortschritte der Phys.* **48**, 771–783 (2000).

3. Tizita Yeshitela. Flip-chip bonding by electroplated indium bumps. (2016).

4. Chen, Y. *et al.* Qubit architecture with high coherence and fast tunable coupling. *Phys. Rev. Lett.* **113**, 1–5 (2014).

5. Koch, J. *et al.* Charge-insensitive qubit design derived from the Cooper pair box. *Phys. Rev. A - At. Mol. Opt. Phys.* **76**, 1–21 (2007).

6. Koshino, K., Kono, S. & Nakamura, Y. Protection of a Qubit via Subradiance: A Josephson Quantum Filter. *Phys. Rev. Appl.* **13**, 1–13 (2020).

7. Rosenberg, D. *et al.* 3D integrated superconducting qubits. *npj Quantum Inf. 2017 31* **3**, 1–5 (2017).

8. The IBM Quantum heavy hex lattice | IBM Research Blog. Available at: https://research.ibm.com/blog/heavy-hex-lattice. (Accessed: 23rd November 2021)

9. The Quantum Aviary: October 2020. Available at: https://thequantumaviary.blogspot.com/2020/10/. (Accessed: 23rd November 2021)

10. Google AI Blog: Scaling Up Fundamental Quantum Chemistry Simulations on Quantum Hardware. Available at: https://ai.googleblog.com/2020/08/scaling-up-fundamental-quantum.html. (Accessed: 23rd November 2021)

11. IBM's roadmap for scaling quantum technology | IBM Research Blog. Available at: https://research.ibm.com/blog/ibm-quantum-roadmap. (Accessed: 23rd November 2021)

12. Hertzberg, J. B. *et al.* Laser-annealing Josephson junctions for yielding scaled-up superconducting quantum processors. *npj Quantum Inf. 2021 71* **7**, 1–8 (2021).

13. The IBM Quantum heavy hex lattice | IBM Research Blog. Available at: https://research.ibm.com/blog/heavy-hex-lattice. (Accessed: 23rd November 2021)

14. Oliver, W. D. Quantum computing takes flight. *Nat. 2021 5747779* **574**, 487–488 (2019).

15. Arute, F. *et al.* Quantum supremacy using a programmable superconducting processor. *Nat. 2019 5747779* **574**, 505–510 (2019).

16. The Rigetti 128-qubit chip and what it means for quantum | by Rigetti Computing | Rigetti | Medium. Available at: https://medium.com/rigetti/the-rigetti-128-qubit-chip-and-what-it-means-for-quantum-df757d1b71ea. (Accessed: 17th November 2021)

17. Caldwell, S. A., Didier, N., Ryan, C. A. & Sete, E. A. Parametrically Activated Entangling Gates Using Transmon Qubits. *Phys. Rev. Appl.* **10**, (2018).

18. Huang, H. L., Wu, D., Fan, D. & Zhu, X. Superconducting quantum computing: a review. *Sci. China Inf. Sci.* **63**, 1–21 (2020).

19. Kono, S. *et al.* Breaking the trade-off between fast control and long lifetime of a superconducting qubit. *Nat. Commun.* **11**, 1–6 (2020).

20. Hong, S. S. *et al.* Demonstration of a parametrically activated entangling gate protected from flux noise. *Phys. Rev. A* **101**, (2020).

21. Nersisyan, A. *et al.* Manufacturing low dissipation superconducting quantum processors. *Tech. Dig. - Int. Electron Devices Meet. IEDM* **2019**-**Decem**, (2019).

22. Zhang, E. J. *et al.* High-fidelity superconducting quantum processors via laser-annealing of transmon qubits. **01**, 1–9 (2020).

23. Gold, A. *et al.* Entanglement across separate silicon dies in a modular superconducting qubit device. *npj Quantum Inf.* **7**, (2021).

24. CNOT | Quantiki. Available at: https://www.quantiki.org/wiki/cnot. (Accessed: 23rd November 2021)