

# Scaling Superconducting Quantum Computers with Chiplet Architectures

Kaitlin N. Smith  
University of Chicago  
Chicago, USA  
kns@uchicago.edu

Gokul Subramanian Ravi  
University of Chicago  
Chicago, USA  
gravi@uchicago.edu

Jonathan M. Baker  
University of Chicago  
Chicago, USA  
jmbaker@uchicago.edu

Frederic T. Chong  
University of Chicago  
Chicago, USA  
chong@cs.uchicago.edu

**Abstract**—Fixed-frequency transmon quantum computers (QCs) have advanced in coherence times, addressability, and gate fidelities. Unfortunately, these devices are restricted by the number of on-chip qubits, capping processing power and slowing progress toward fault-tolerance. Although emerging transmon devices feature over 100 qubits, building QCs large enough for meaningful demonstrations of quantum advantage requires overcoming many design challenges. For example, today’s transmon qubits suffer from significant variation due to limited precision in fabrication. As a result, barring significant improvements in current fabrication techniques, scaling QCs by building ever larger individual chips with more qubits is hampered by device variation. Severe device variation that degrades QC performance is referred to as a defect. Here, we focus on a specific defect known as a frequency collision.

When transmon frequencies collide, their difference falls within a range that limits two-qubit gate fidelity. Frequency collisions occur with greater probability on larger QCs, causing collision-free yields to decline as the number of on-chip qubits increases. As a solution, we propose exploiting the higher yields associated with smaller QCs by integrating quantum chiplets within quantum multi-chip modules (MCMs). Yield, gate performance, and application-based analysis show the feasibility of QC scaling through modularity. Our results demonstrate that chiplet architectures, relative to monolithic designs, benefit from average yield improvements ranging from  $9.6 - 92.6 \times$  for  $\lesssim 500$  qubit machines. In addition, our simulations explore the design space of chiplet systems and discover configurations that demonstrate average two-qubit gate infidelity reductions that are at best  $0.815 \times$  their monolithic counterpart. Finally, we observe that carefully-selected modular systems achieve fidelity improvements on a range of benchmark circuits.

**Keywords**—Quantum computing; quantum architecture; superconducting quantum computers

## I. INTRODUCTION

The proposed benefits of quantum information processing has garnered much interest from industry, academia, and government alike. Specifically, quantum computers (QCs) are forecast to accelerate the solving of some of today’s most complex problems through the careful use of quantum superposition, interference, and entanglement. Fault-tolerant QCs have been proposed for applications in cryptography [76], big data [30] chemistry [37], optimization [53], and machine learning [13]. Although current QCs are noisy prototypes without error correction, there exist early, yet exciting, demonstrations of quantum advantage [9], [87].

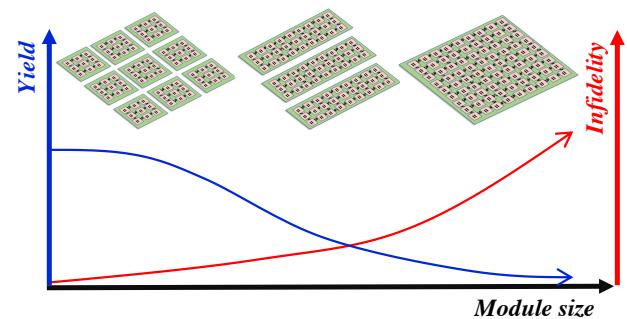


Figure 1. Infidelity and yield trade-off vs. individual QC module size in qubits.

The distinction between classical and quantum processors will become more apparent as scaled QCs emerge that can access exponentially larger computational spaces. While QCs with over 100 physical qubits have been built [2], [7], far more qubits are required to solve meaningful, real-world problems at rates that outperform or even compete with existing classical hardware. Additionally, near-term QCs, referred to as noisy intermediate scale quantum (NISQ) machines [68], have high gate infidelity and low coherence times that prohibit error correction. Large-scale quantum algorithms such as Shor’s Factoring [75] and Grover’s Search [31] will require millions of fault-tolerant qubits for computation [64].

Scaling QCs requires confronting many challenges including, but not limited to, qubit quality, external noise, power consumption, control routines, device packaging, hardware footprint, and system cooling [59]. In the near term, especially on QCs based on superconducting (SC) circuits, limitations associated with qubit fabrication are among the most debilitating on the path to QC scaling. The physical processing that attempts to realize targeted, per-qubit specifications suffers from unavoidable variation. This fabrication imprecision can inject random defects into QCs [25], [32], severely impacting QC fidelity. As fabricated chips increase in size, imprecision-related defects become statistically more likely. Thus, intuition expects that average device infidelity correlates with device size, as illustrated on the right axis in Fig. 1. Eventually, infidelity surpasses usefulness thresholds for a QC, causing fabrication yields to

decline, as shown on the left of Fig. 1.

Fixed-frequency transmon QCs, like those in the IBM Quantum systems [5], are considered promising for scaling. Transmons are constructed of superconducting (SC) circuits, and they take advantage of fabrication materials and processes refined for classical integrated circuits. Transmons also have fairly good coherence times that reliably extend to tens of microseconds [5], [18], [33] along with gate fidelities that are approaching fault-tolerance thresholds [73]. Although they have many benefits, transmons are not exempt from fabrication-related scaling challenges. In this work we primarily focus on frequency collisions.

Frequency collisions are a specific type of defect commonly seen in fixed-frequency transmon QCs. A frequency collision is caused by a qubit-qubit frequency detuning that falls in a range that degrades two-qubit gate fidelity [15]. Stochastic variation in QC fabrication influences the operational characteristics of Josephson Junctions (JJs), or the essential components of transmons that realize qubit state. Intrinsic variation alters both JJ dimension and critical current, and as a result, actual qubit frequency drifts from its intended value [41]. This causes collision conditions that restrict QC operation because qubits lose their ability to participate in high-quality multi-qubit entangling operations.

Larger QCs with more qubits experience greater variation during fabrication. This variation accumulates and introduces error channels in the form of QC defects. In the case of frequency collisions, the probability of a collision occurring on chip increases with QC size [32], [86], and thus, the collision-free yield decreases rapidly. This insight motivates our work. As a solution, we propose techniques for transmon QC scaling that boost device yields and performance through modularity. At time of writing, scaling through modularity has gained traction within concurrently-released quantum roadmaps [8], and in our study, we provide a novel analysis of quantum chiplets and quantum multi-chip modules (MCMs) that considers the physical effects limiting yield and application fidelity. MCMs can be thought of in a similar manner as multi-core systems and are designed to exploit the larger yields of smaller quantum chiplets. In some extremes, our simulations show that MCM architectures can provide fidelity gains and construct QCs of dimensions that result in zero monolithic yield. Our contributions are as follows:

① We describe current fidelity trends of today's QCs based on real-machine data, showing that smaller QCs are favored in terms of minimized variation and gate error.

② We investigate prior work related to classical scaling via modular systems. We draw parallels to transmon QC technology and current challenges that exist with scaling.

③ We identify frequency collision conditions and critical parameters in state-of-art device fabrication. We explore the frequency allocation and fabrication precision design space, highlighting their influence on collision-free yield.

④ Motivated by frequency collision defects, we propose

a flexible chiplet design for integration within large-scale transmon QCs. We keep eventual fault-tolerance in mind with chip bonding that preserves heavy-hex connectivity.

⑤ We introduce the customizable, quantum multi-chip module (MCM) and propose emerging link technology for chip unification.

⑥ We develop a simulation framework that models monolithic and MCM QCs for yield, performance, and application-based analysis. This framework will be open sourced and is highly parameterized for adaptability: simulations and QCs can be programmed to model future improvements in fabrication and devices.

⑦ We provide simulation results that show significant collision-free yield improvements for chiplets and, thus, MCMs. Analysis using real quantum benchmarks and state-of-art error metrics showcase situations where MCM properties and performance surpass monolithic counterparts on top of yield advantages, showing points of crossover in terms of architecture of choice.

The models presented here are extremely timely: modularity is gaining widespread acceptance as the best route to quantum scaling [8], [70]. In our study, we hope to guide the development of modular QCs, showing the feasibility of chiplet-based MCMs via yield, performance, and application-based analysis. Our results show that chiplet architectures, relative to monolithic designs, benefit from average yield improvements ranging from  $9.6 - 92.6 \times$  for  $\lesssim 500$  qubit machines. In addition, our simulations explore the design space of chiplet systems and discover configurations that demonstrate average two-qubit gate infidelity reductions  $0.949 - 0.815 \times$  of their monolithic counterparts. Finally, we observe that carefully-selected modular systems achieve fidelity improvements on a range of benchmark circuits.

## II. BACKGROUND

### A. Quantum Information

The qubit can hold a superposition of the basis states  $|0\rangle$  and  $|1\rangle$  until measurement where the state  $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$  collapses into either  $|0\rangle$  or  $|1\rangle$ . CX along with single-qubit rotation operations are frequently supported on near-term QCs, and quantum circuits are cascades of these gates.

Limitations of today's quantum hardware in terms of coherence and noise force a one-to-one mapping between logical qubits of an algorithm to physical qubits on a QC. A goal in quantum hardware design is to achieve error rates below the thresholds required for fault-tolerance via error correcting codes (ECCs). ECCs encode a logical qubit on many physical qubits, smoothing occasional gate and decoherence errors on individual physical qubits. Prior work estimates that at minimum, millions of high-quality physical qubits will be required for fault-tolerant (FT) quantum computation [63]. Once FT QCs emerge, disruptive quantum applications in areas such as cryptography [76] and

big data [30] will be feasible. More quantum computing fundamentals can be found in [61].

### B. Transmon Qubits

Fixed-frequency transmon devices provide an especially viable platform for realizing physical qubits. Compared to alternative technologies, transmon QCs show great promise due to recent breakthroughs in device coherence, operation fidelity, and addressability [35], [40], [52]. Transmon devices are also favorable to scaling as they take advantage of fabrication techniques refined by classical electronics. By employing Josephson Junctions (JJs) within SC circuits, transmons realize quantum information by acting as mesoscopic-scale, artificial atoms characterized by an anharmonic energy spectrum. Transmon qubits have many properties that must be properly characterized for successful control.

The spectrum associated with a transmon qubit contains energy levels beyond the (lowest) two that are used for information encoding. A transmon's operation frequency,  $f_i$ , is set by the transition between the qubit  $|1\rangle$  and  $|0\rangle$  state,  $f_i = f_{i,01} = f_{|1\rangle} - f_{|0\rangle}$ . Operation frequency is determined by the anharmonic energy levels of the qubit. Transmon anharmonicity,  $\alpha$  is fixed by the energy level difference between the  $|1\rangle \rightarrow |2\rangle$  transition and the  $|0\rangle \rightarrow |1\rangle$  transition,  $\alpha = f_{i,12} - f_{i,01}$ . Today's state-of-art transmon devices target  $f_i$  values in the neighborhood of  $\sim 5$  GHz while qubit anharmonicity is  $\alpha \approx -0.330$  GHz [86].

The IBM transmon QCs are up to 127 qubits in size [7]. Microwave drives implement single- [20] and two-qubit gates [19] on these devices with fidelities that are moving toward FT thresholds. While all-to-all physical qubit connectivity is desirable, design constraints associated with limited space on-chip restricts near-term transmon qubits to communication with their nearest neighbors. It might be assumed that QC topologies would be designed for the highest qubit degree possible, but highly-connected qubits have an elevated risk of crosstalk, or unintended classical or quantum coupling, that degrades computation [17]. As a solution, IBM QCs implement heavy-hexagon, or heavy-hex, topologies to enable sufficient physical qubit-qubit communication while minimizing cost in terms of crosstalk and hardware requirements [6]. An example of three IBM QCs that use the heavy-hex qubit layout can be found in Fig. 3(a). The heavy-hexagon lattice is targeted for the hybrid surface/Bacon-Shor ECC with an error threshold of 0.45% [17].

### C. Modularity in Classical Scaling

Classical computing systems have drastically improved since their emergence in the mid 1900s. Thus, we are motivated to examine classical methodologies to find quantum parallels that could lead to accelerated QC scaling.

Monolithic electronics have benefits that stem from the simplicity of being entirely self-contained as a system on a

chip (SoC). SoCs, however, have added costs associated with yields [43]. Additionally, monolithic SoCs often have steep resource requirements [26] and demonstrate inflexibility in implementing highly customized or specialized hardware as the SoC footprint increases [34]. Fortunately, chiplet-based architectures can alleviate some of these problems, pushing forward architectural innovations that allow new computing milestones [24], [34], [39], [65], [85]. In a chiplet design, multiple smaller and locally-connected chips replace a larger, monolithic SoC. Chiplet architectures are often referred to as multi-chip modules (MCMs).

A few benefits of chiplet are as follows. First, chiplet architectures take advantage of well-established electronic manufacturing techniques because each chiplet, like its monolithic counterpart, is an integrated circuit (IC). Second, as seen in Fig. 2, the smaller-footprint chiplet has lower fabrication cost that results in a higher yield of functional devices since significantly more can be fabricated simultaneously on a substrate. Third, because chips contain fewer components, the complexity of design verification and post-fabrication test is reduced. Fourth, chiplets enable mixed-process integration within a single package, potentially allowing significant processing gains. Because individual chiplets contain less logic, application-specific components can utilize more optimized technology libraries and acceleration that would not provide advantage to more generic logic. Finally, since chiplet designs are modular, failures can be isolated and removed without the need to discard the system as a whole.

Chiplet-based classical computation offers many advantages that promote scaling. However, communication bottlenecks, often in the form of latency in link hardware, must be properly evaluated as a trade-off. MCMs sometimes demonstrate performance loss as compared to their monolithic counterparts [39], but the benefits associated with modularity are often overwhelmingly worth the cost.

## III. TRANSMON QC CHALLENGES

This section focuses on three interconnected transmon constraints that currently impede progress toward realizing larger transmon QCs: Cross-Resonance gate error, architecture rigidity, and fabrication imprecision.

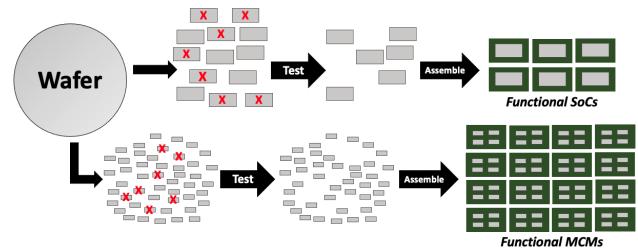


Figure 2. Monolithic (top) vs. chiplet (bottom) architecture yields with seven faulty devices in each batch.

Type	Criteria	Threshold	Note	Qubit Topology
1	$f_i = f_j$	$\pm 0.017 \text{ GHz}$	Nearest neighbors $Q_i$ and $Q_j$	
2	$f_i + \alpha/2 = f_j$	$\pm 0.004 \text{ GHz}$	Nearest neighbors control $Q_i$ and target $Q_j$	
3	$f_i = f_j + \alpha$	$\pm 0.03 \text{ GHz}$	Nearest neighbors $Q_i$ and $Q_j$	
4	$f_j < f_i + \alpha$ <i>or</i> $f_i < f_j$	-	Nearest neighbors control $Q_i$ and target $Q_j$	
5	$f_j = f_k$	$\pm 0.017 \text{ GHz}$	$Q_i$ is control to nearest neighbors $Q_j$ and/or $Q_k$	
6	$f_j = f_k + \alpha$ <i>or</i> $f_j + \alpha = f_k$	$\pm 0.025 \text{ GHz}$	$Q_i$ is control to nearest neighbors $Q_j$ and/or $Q_k$	
7	$2f_i + \alpha = f_j + f_k$	$\pm 0.017 \text{ GHz}$	$Q_i$ is control to nearest neighbors $Q_j$ and/or $Q_k$	

Table I  
FIXED-FREQUENCY TRANSMON COLLISION CRITERIA TO BOUND GATE ERROR FROM FREQUENCY-RELATED NOISE TO  $\lesssim 1\%$ .

### A. The Cross-Resonance Gate

A hallmark of a promising quantum implementation is strong qubit-qubit entanglement. Transmons can become entangled after executing the Cross-Resonance (*CR*) gate on a pair of qubits connected by a resonator [47], [66], [69], [83]. This gate operates by applying a microwave drive to one qubit, the control, at the frequency of a second, the target. The ideal *CR* effect is a *ZX* interaction between the participating qubits. This interaction can then be combined with single-qubit operations to achieve more standard two-qubit gates used in circuits, such as the *CX* gate.

*CR* error rates lower than  $10^{-2}$  have been reported [35], [36], [73], but *CR* gate implementation must improve to reach FT QCs. *CR* error rates can theoretically reach levels as low as  $10^{-4}$  [49]. Reaching this lower bound is necessary to improve QC scaling prospects. However, *CR* fidelity is limited if system noise is inaccurately characterized, leading to coarse models used during *CR* optimization. Despite advances made in qubit optimal control at the pulse-level, unexpected environmental coupling or irregular QC hardware properties stemming from manufacturing inconsistency or defects can significantly influence *CR* error.

### B. Architectural Constraints

Fixed-frequency transmon devices must be carefully designed in terms of their frequency assignment to 1) allow for selective addressability among qubits and 2) to enable high-quality *CR* interactions between nearest neighbors. Table I lists seven conditions that include numerically defined definitions and bounds from [32], [47] that are likely to lead to  $\gtrsim 1\%$  gate error due to transmon qubit *frequency collisions*, or conditions where qubit frequencies “collide,” degrading *CR* performance. Frequency collisions are a leading challenge in scaling transmon devices [86]. As a note, current transmon systems on-average demonstrate  $\sim 1\text{-}2\%$  infidelity contributed by various noise sources unrelated to error caused by frequency [51], [78].

The criteria of Table I define fabrication targets that minimize error stemming from *CR* interactions between qubits with improper frequency spacing. Violation of Table I criteria amplifies gate infidelity considerably as the

frequency collision becomes the dominant source of *CR* error. Careful frequency assignment in terms of targeted individual frequency and nearest neighbor detuning during the QC design process helps prevent collisions. Unfortunately, current limitations in fabrication precision inject random variation, causing actual QCs to have features that deviate from what is theoretically expected. More information about fabrication variation can be found in Section III-C.

**Ideal Frequency Assignment:** To avoid collisions, each qubit’s frequency must be distinguishable from its neighbor. It is also desirable to minimize the number of target frequencies,  $F_i$ , on chip to simplify control. Frequency collision conditions can be avoided using three frequencies in the heavy-hex lattice where  $F_0 < F_1 < F_2$  [32]. Here, the highest frequency qubits,  $F_2$ , act as control qubits in *CR* interactions and never have a degree greater than two. As a note, unidirectional two-qubit interactions are natively supported by hardware, but the direction of two-qubit operations can be reversed with single-qubit rotation operations.

The ideal heavy-hex frequency assignment is pictured in the lower half of Table I, column 5. As the pictured QC topology scales, the frequency assignment pattern remains constant. To avoid collisions, nearest neighbor frequencies should never be equivalent, and all  $F_2$  qubits should be surrounded by  $F_0$  and  $F_1$  but never by two of the same kind.

**Frequency Detuning:** The frequency difference of qubits must fall within a “straddling regime” to avoid collisions: too small of a frequency detuning results in inaddressability and too far prevents entanglement from being created [47]. Second, the relationship between the frequencies of next-nearest neighbors is also of importance. For example, consider the three qubits in the top half of Table I, column 5. If a control transmon,  $f_i = F_2$ , is connected to two qubits of almost equivalent frequencies,  $f_j = F_1 \approx f_k = F_0$ , applying a drive of  $f_j$  to the control would cause a frequency collision since a high-noise *CR* interaction results if  $f_j$  is near-resonant to  $f_k$ . This is an example of a “near-null,” frequency collision, or a Type 1 or 5 collision in Table I.

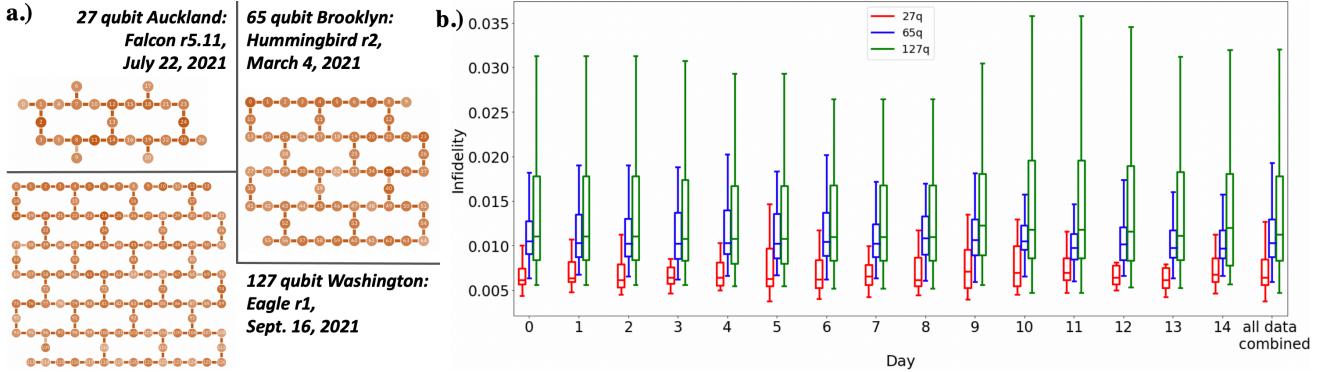


Figure 3. (a) Device name, processor type, date online, and connectivity graph for three IBM QCs. (b) Box plot comparing of 15 days of CX infidelity statistics for three generations of IBM processors of different sizes.

### C. Device Fabrication Variation

Precise control over qubit frequency is critical to produce high-fidelity transmon QCs. In reality, this can be extremely challenging due to the small feature size, approximately  $100 \times 100$  nm, of the transmon JJ [32]. Small imperfections appear in JJ positioning and dimensions during processing, influencing the frequency at which the fixed-frequency transmon operates [41]. Fabrication variation that prohibits or severely limits a transmon qubit's functionality is known as a defect. Here, we will focus on defects that deviate a qubit's frequency from its ideal, resulting in spectral overlaps that cause frequency collisions. This variation is stochastic, causing the final frequency profile of each chip to be unique. The distribution that describes fabrication precision is characterized by the standard deviation  $\sigma_f$ , where a lower  $\sigma_f$  indicates higher precision characterized by actual device frequencies landing near design targets. Shortcomings in fabrication accuracy currently result in devices with a frequency spread of  $\sim 0.1$  GHz around targeted frequencies [86]. This large amount of variation causes qubit-qubit frequency detunings that satisfy the Table I criteria and drive down the number of functional QC devices in the collision-free yield. With such high variation, yields reach zero well before devices reach  $10^2$  qubits. For example, a real-world fabrication variation of  $\sigma_f = 0.1323$  GHz [32] is seen as the blue curve in the plots of Fig. 4. At this poor precision, there is little hope of creating high-yield quantum chips containing more than 20 qubits. Fortunately, transmon qubit frequency can be characterized and laser tuned after fabrication, enabling a much higher level of precision characterized by  $\sigma_f \approx 0.014$  GHz, boosting yield [32]. Laser tuning has been applied to improve  $< 100$  qubit QC yields by  $15 \times$  [86]. Further improvements to  $\sigma_f$  are required to maintain yield while increasing the number of qubits manufactured on a monolithic QC.

## IV. MOTIVATING QUANTUM CHIPLETS

### A. Recent Trends in Quantum Hardware

Calibration data for three IBM machines of different processor generations was gathered to analyze relationships between CX gate performance and processor size. The machines included Auckland, a 27-qubit Falcon, Brooklyn, a 65-qubit Hummingbird, and Washington, a 127 qubit Eagle. Details of these machines, including date online, revision, and heavy-hexagon connectivity graphs, are featured in Fig. 3(a). All QCs under analysis were released in 2021 and their size correlates to processor generation: larger devices with more qubits feature newer hardware designs [4]. Statistics detailing CX infidelity for each device over the course of 15 days, i.e. 15 different calibration cycles, are found in Fig. 3(b).

Newer IBM processors feature technological advancements for improved scaling. For example, hardware upgrades enable computational benefits such as longer qubit coherence, reduced gate and readout times, and smaller footprint with innovations such as “direct couplers” [4], [7]. However, Fig. 3(b) shows that median CX infidelity directly correlates with chip size such that larger devices demonstrate higher two-qubit gate error rates over the 15 calibration cycles. Moreover, the CX infidelity of the larger devices generally exhibits a larger distribution of CX quality over the investigation period. These results suggest that greater variation exists as QCs increase in size, degrading overall performance. As described in Section III, fabrication variation exists in near-term QCs, and the extent of heterogeneity of qubits becomes more extreme as devices increase in size. We postulate that as fabrication precision stays constant, larger devices inadvertently have the potential to contain more qubit-qubit frequency detunings that push them within or close to frequency collision regions. Thus, we are motivated to pursue greater consistency and performance in machines through modularity.

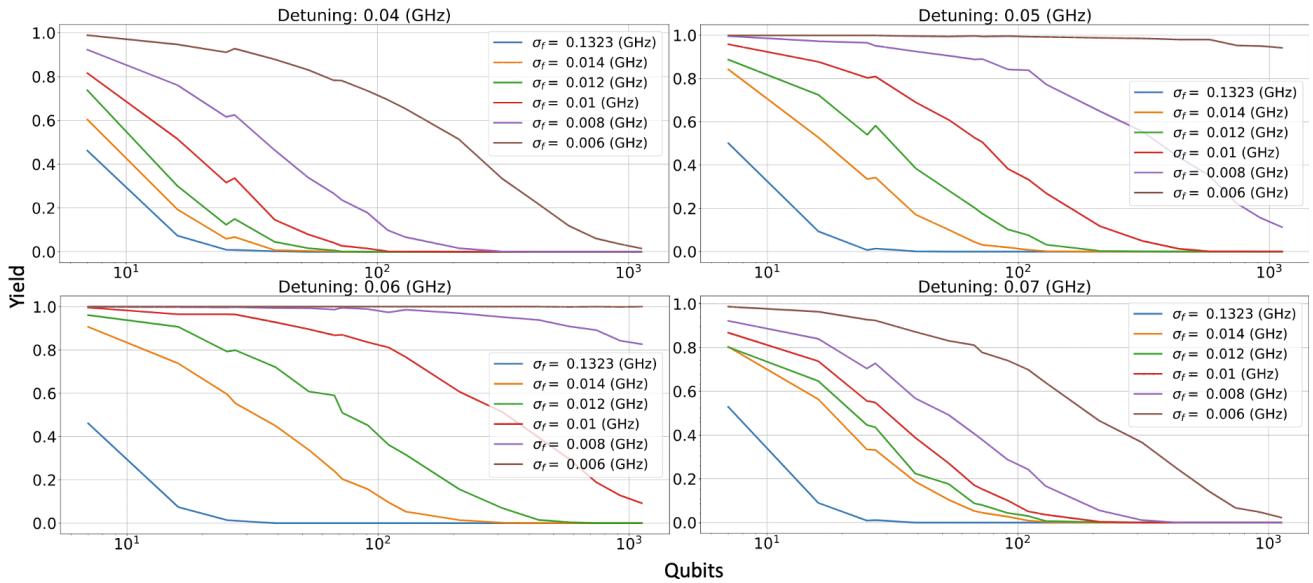


Figure 4. Collision-free yield vs. qubits for detuning values ranging 0.04–0.07 GHz between ideal frequencies,  $F_{0,1,2}$ .

### B. Modeling Yield vs. Qubits

To gain a better understanding of the relationship between chip size and collision-free yield, we employ a Monte Carlo (MC) model to quantify collision events as heavy-hex devices increase in qubit count. The heavy-hex layout was chosen for analysis because of its use in state-of-art IBM transmon QCAs, and current gate error rates are gradually approaching the thresholds required for ECCs on heavy-hex devices. More information can be found in Section II-B.

Yield simulation involved virtually constructing heavy-hex QCAs up to  $\sim 10^3$  qubits in size. To emulate variation associated with fabrication in our modeling, qubit frequency was assigned by randomly sampling values drawn from normal distributions characterized by both the qubit's ideal frequency,  $F_{0,1,2}$  (Section III-B), and fabrication precision,  $\sigma_f$  (Section III-C). Next, QCAs were evaluated for collisions by checking the criteria in Table I for satisfiability. If all seven criteria return `false`, a QC is categorized as collision-free. A batch size of  $10^3$  devices was used in MC simulation.

The first step in simulation required determining the optimum frequency detuning for  $F_{0,1,2}$  that results in highest yield. Transmon qubits are designed to operate at  $\sim 5$  GHz [35], [49], therefore  $F_0$  was set as 5 GHz. It should be noted that although detuning between frequencies is important, absolute values are not [32]. In this study we assume step size to be equal between  $F_{0,1,2}$  as done in prior work [32], [45]; exploring the impact of varying the distance between ideal frequencies could be an area for future work.

Fig. 4 shows yield simulation results where each plot employs a different step size ranging from 0.04 to 0.07 GHz between  $F_{0,1,2}$ . In these plots, the values for  $\sigma_f$  were chosen based on the results of [32], where  $\sigma_f = 0.1323$  GHz

was the qubit frequency spread around targeted frequencies directly after fabrication,  $\sigma_f = 0.014$  GHz was the fabrication precision obtained via qubit laser tuning, and  $\sigma_f = 0.006$  GHz, if physically obtainable, is proposed as the threshold for variation to achieve QCAs larger than  $10^3$  qubits using current collision criteria found in Table I. The lower left plot of Fig. 4 shows that a target qubit-qubit detuning of 0.06 GHz produces the highest yields during simulation for each level of precision, closely recreating the projections from [32]. This match helps validate our implemented model. Thus,  $F_{0,1,2}$  is set as 5.0, 5.06, and 5.12 GHz, respectively, in further analysis as it provides the most optimum frequency step to target during fabrication when considering the Table I criteria. Fig. 4 demonstrates that yield decreases as frequency step size moves away from 0.06 GHz. The supporting intuition is that if step size is too small, the likelihood increases for post-fabrication frequency detuning to be near-null, producing a Type 1 or 5 collision in Table I. If step size is too large, however, the likelihood increases of encountering a near half-anharmonicity or near anharmonicity detuning, resulting in Type 3, 4, 6, and 7 collisions.

The key takeaway from Fig. 4 is that machine variation becomes detrimental as QCAs increase in size. When  $\sigma_f > 0.006$  GHz and ideal frequency detuning is constant, larger chips experience lower yields due to the increased probability of a frequency collision. Today's state-of-art transmon processing has resulted in precision values of 0.0185 GHz [86] and 0.014 GHz [32]. Thus, significant improvements are needed to reach the ideal  $\sigma_f < 0.006$  GHz that will assist QC scaling beyond  $10^3$  qubits if current collision criteria is considered. Notably, a lower bound for a physically possible  $\sigma_f$  will likely be reached as research in the area continues. At this time, the

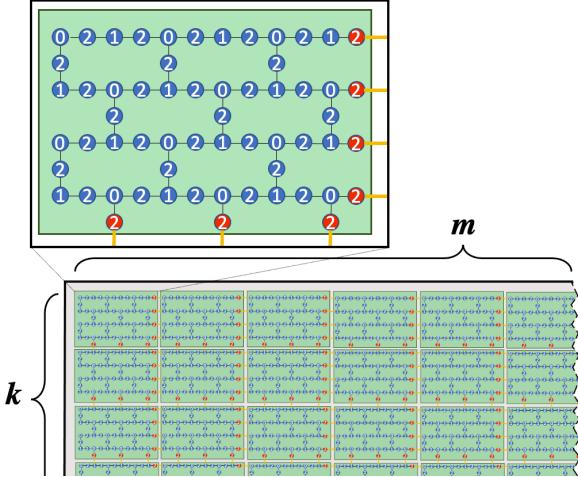


Figure 5. 60-qubit chiplet within a  $k \times m$  MCM. Links connecting chiplets are indicated in yellow.

$\sigma_f$  lower bound is unknown and difficult to speculate.

In the interest of keeping our modeling forward-focused without establishing overly aggressive projections, we set  $\sigma_f = 0.014$  GHz, the current state-of-art [32], in our QC system modeling. Some amount of variation in QC fabrication will likely be unavoidable moving into the future, having detrimental impact on the overall fidelity of monolithic QCs. As a result, we are motivated to explore chiplet-based architectures to produce larger machines with greater success.

## V. ARCHITECTING QUANTUM MULTI-CHIP MODULES

Transmon QC yield modeling of Section IV-B revealed challenges centered around fabrication variation. The key takeaway is that when monolithic QCs are designed to contain more qubits, the probability of encountering a frequency collision increases. As a solution, we investigate quantum chiplets that achieve higher yields by containing fewer qubits. Chiplets are interconnected with a host chip, creating multi-chip modules (MCMs) as a monolithic alternative.

It should be noted that an inadvertent bonus of MCM architecture, in addition to aforementioned benefits relating to yield, is that failures within a sub-component have the opportunity to be isolated, preventing widespread errors in computation. For example, the sensitivity of current QCs forces them to suffer from large amounts of correlated error when exposed to random external stimulus such as stray radiation or cosmic rays [50]. Since each chiplet has some buffer from the active components of its neighbors, large-scale qubit corruption from electromagnetic contamination can be avoided.

### A. Chiplet-based Design

**Chiplet Design:** A chiplet containing 20 qubits was first introduced in Table I, column 5. This device was designed to include a single, complete heavy-hex honeycomb – other

chiplet designs include a partial heavy-hex ring or multiple. Each chiplet has edge qubits, indicated in gray, that support coupling to qubits on adjacent chips. The heavy-hex three-frequency pattern is described in the Table I, column 5 drawing. When linked to other devices using inter-chip links, the design of the chiplet allows the heavy-hex lattice and frequency allocation pattern to be preserved in the resulting MCM, keeping future error correction with the hybrid surface/Bacon-Shor code in mind.

An example of a 60-qubit chiplet is pictured within a  $k \times m$  MCM in Fig. 5. This device's structure closely matches that of the 20-qubit chiplet in Table I but with two additional dense rows (i.e. rows containing qubits labeled as 0, 1, and 2) holding four extra qubits each and two additional sparse rows (i.e. rows that only contain qubits labeled as 2) holding one additional qubit each. Larger heavy-hex chiplets require additional rows and columns that are added similarly.

**MCM Interconnects:** Recently, SC qubits on separate chips were coupled with high fidelity using microwave link technology [48], [88]. In another demonstration highly relevant to this work, SC qubits on two separate physical modules were flip-chip bonded to a larger carrier chip, or interposer [28]. Inter-chip coupling rates and entanglement quality comparable to intra-chip qubit-qubit coupling were accomplished. Although either linking technique could be applied toward networking an MCM, here, we focus on the latter work that makes progress in flip-chip architectures.

In [28], the average coherence limited fidelity, or the maximum achievable fidelity predicted from the measured relaxation and dephasing times, of the two-qubit operations across module links was 92.5% while the median was 94.4%. These statistics show that the cross-module gates are more error-prone than the on-chip operations of the example IBM machines in Fig. 3(a). The difference in fidelity, however is not drastic, demonstrating that flip-chip bonding holds more than just theoretical viability when applied to QCs.

Inter-chip links of the Fig. 5 MCM are indicated in yellow. Since the right-most and bottom-most qubits in our chiplet design always have a  $F_2$  assignment, they will act as the control in inter-chiplet CR interactions.

### B. Customization and Reconfigurability

One of the most notable advantages of a chiplet architecture is flexibility. For instance, the dimensions of the MCM can be chosen ad hoc. When the MCM increases in total number of chiplets selected from the collision-free yield, the amount of possible system configurations grows at a factorial rate. This relationship of configurations vs. MCM size is seen on the left axis in Fig. 6. This plot shows configuration totals based on  $\sim 69.4\%$  yield of identically-designed, 20-qubit chiplets, resulting from a state-of-art  $\sigma_f = 0.014$  GHz and a QC batch size of  $10^5$  units. The influence of a fabrication precision of  $\sigma_f = 0.014$  GHz on the collision-free yield of QCs up to  $10^3$  qubits is shown by the orange curves in Fig. 4.

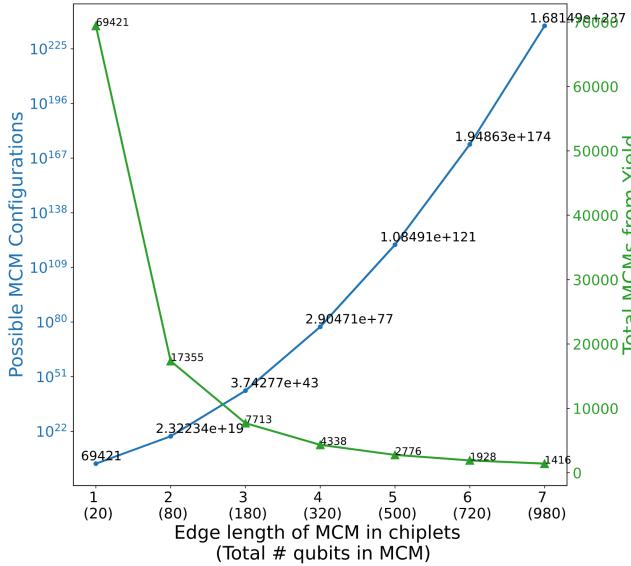


Figure 6. Potential configurations for a single MCM (left axis) and upper bound of number of assembled MCMs (right axis) vs. MCM size. Square MCMs ( $m \times m$ ) use 20-qubit chiplets. Calculations based on  $\sim 69.4\%$  yield with  $\sigma_f = 0.014$  GHz and a batch size of  $10^5$  units.

The right half of plot in Fig. 6 describes the total number of MCMs that result from the collision-free yield of 69,421/100,000 chiplets. As expected, the total number of MCMs resulting from the collision-free subset decreases as dimension increases, but there still exists the opportunity for a large amount of customization. We assume use of the industry-standard known good-die (KGD) testing techniques [38] where individual chips are tested before MCM assembly. Thus, QC chiplets are sorted in a process similar to speed-binning: devices with similar fidelity properties can be post-selected and linked, resulting in systems with uniform and high-fidelity quantum computation.

Fig. 6 considers MCMs composed of chiplets from the same batch. However, innovation in quantum processing, control, and carrier-chip methods could result in further MCM customization (e.g., mixed integration of heterogeneous physical qubit technologies). Highly specialized systems comprised of multiple types of physical qubits could present new opportunities for application-specific quantum accelerators.

### C. Monolithic vs. MCM Fabrication Output

Section IV-B demonstrates the collision-free yield benefits of smaller transmon QCs. Here, we explore how yield improvements can translate into significantly more fabrication output, and thus greater cumulative compute power, via MCM systems. We build an analytical model that supports the simple argument for chiplets presented in Fig. 2. The guiding principle is that chiplets exploit the ability to process more devices at once since their die takes less area on a wafer.

To build our model, the collision-free yield associated with

a  $q_m$ -qubit monolithic QC is defined as  $Y_m$  out of a batch size of  $B$  die. Conversely,  $Y_c$  is the yield associated with smaller,  $q_c$ -qubit chiplets. If the chiplets are designed for integration within an  $k \times m$  MCM, the total number of resulting MCMs,  $N$ , is calculated as

$$N = \frac{Y_c \times (B \times \frac{q_m}{q_c})}{k \times m} \quad (1)$$

where the die area ratio between the monolithic and chiplet QCs is approximated using their respective qubit capacity,  $\frac{q_m}{q_c}$ .

The yield simulation of Fig. 4 is applied to show how the upper bound of assembled MCMs compares to the yield of similarly-sized monolithic systems. Consider the experiment where  $\sigma_f = 0.014$  GHz and the frequency detuning is 0.06 GHz in the lower left plot of Fig. 4. With these parameters, a  $q_m = 100$  QC has a yield of approximately  $Y_m = 0.11$  for  $B = 1000$  chips. On the other hand, a  $q_c = 10$  chiplet is characterized by approximately  $Y_c = 0.85$ . Using the same wafer area as the monolithic device,  $B \times \frac{q_m}{q_c} = 1000 \times \frac{100}{10} = 10,000$  chiplets theoretically could be fabricated, where we would expect 8,500 as collision-free. MCMs of dimension  $2 \times 5$  enable ten-chiplet QCs that contain the same number of qubits as the monolithic devices, thus an  $N = 850$  upper bound of 100-qubit MCMs would result. As compared to the  $Y_m \times B = 0.11 \times 1000 = 110$  device yield from 100-qubit monolithic production, MCMs result in  $\sim 7.7 \times$  gain in manufactured QCs.

### D. Practical Constraints of MCMs

Section V-C describes the upper bound for MCM fabrication output as compared to monolithic devices. This provides an estimate for potential gains in fabrication output when taking a modular QC approach, but reaching this upper bound is unrealistic. In reality, some MCMs will be lost during carrier-chip bonding procedures due to link faults or other defects that prevent practical use. Thus, it is important that assembly yield is considered. Our proposed architecture unites quantum chiplets through flip-chip bonding to a larger, carrier chip, or interposer. Here, we will include MCM assembly yield to account for inevitable failures during assembly. Our model is based on silicon interposer defect rates in [38]: we estimate the success rate for the placement of a controlled collapse chip connection (C4) bump bond on a passive carrier chip. Fabrication information from [28] provides insight about the number of bump bonds required for each inter-chip linked qubit. Future work will refine the MCM assembly yield as data for fabricated quantum devices becomes available.

## VI. FRAMEWORK FOR QUANTUM MCM MODELING

Potential benefits of scaling via quantum MCMs include increased yield, more homogeneous computation, and added flexibility. The practicality of these systems, however, must

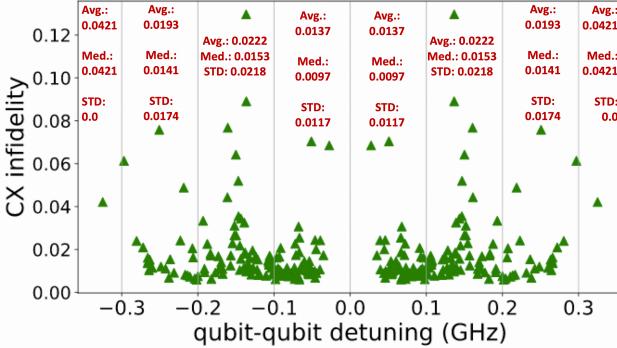


Figure 7. Plot of CX infidelity vs. qubit-qubit detuning for 127-qubit IBM Washington, an Eagle-class processor. Each datapoint represents an average of characterization data for a qubit-qubit connection over 15 calibration cycles. Median of plotted points is 0.012; average is 0.018.

be evaluated. In this section, we present our MCM modeling framework. A corresponding tool integrates these methods for simulating QC yield and estimating device and workload performance. As a note, while parameters in our analysis are based on state-of-art transmon [5] and flip-chip technology [28], properties involving device coupling, frequency assignment, fabrication precision, and inter- and intra-chip gate error are parameterized to accommodate future QC improvements.

#### A. On-chip Fidelity Assignment

Section IV-B describes collision-free yield. By avoiding Table I conditions, QCs theoretically have  $\lesssim 1\%$  frequency-related error contributing to gate infidelity. Real quantum systems, however, are characterized by diverse noise profiles that also contribute to gate error. Examples of two-qubit gate infidelity statistics for three QCs are found in Fig. 3(b). To better capture the complexities of real quantum operations in our modeling, we employ real QC data in our analysis.

The CR gate, Section III-A, along with single-qubit rotations generate the CX operation [21]. CX fidelilities and qubit frequencies are available in the IBM QC backend properties. Our model uses this information to develop an emperical relationship between qubit-qubit detuning and two-qubit operator fidelity. In our study, the QC under investigation is IBM Washington, a state-of-art Eagle-class processor that accomplished an quantum-industry first of surpassing the 100-qubit milestone [7]. IBM Washington's innovation in transmon control and design represents the future goals of quantum technology, so its properties were used to assign two-qubit gate fidelity values on-chip.

Real QC noise fluctuates over time [23], thus we gathered and averaged CX infidelity over 15 calibration cycles and correlated the averages with the frequency detuning of the involved qubits. Fig. 7 plots the relationship of average CX infidelity vs. detuning for IBM Washington. The median of plotted data is  $0.012 \rightarrow 1.2\%$  while average is  $0.018 \rightarrow$

$1.8\%$ . As illustrated in Fig. 7, data was binned according to detuning intervals of step-size 0.1 GHz. This interval was inspired by the observed fabrication-induced frequency spread described in [86], but the parameterized nature of the presented modeling framework allows the sampling bounds to be adjusted. After qubit-qubit detuning characterization, gate fidelity is assigned by sampling from the distribution of the corresponding bin.

#### B. Inter-chip Link Fidelity Assignment

An important consideration for MCM architectures is the trade-off between chiplet yield and inter-chip communication. In our MCM architecture, chiplets are proposed to be flip-chip, C4 bump bonded to an interposer. Successful two-qubit operations spanning QC modules have been experimentally demonstrated with fidelity averaging 92.5% (Section V-A). Our model assigns link infidelity based on the distribution of [28]. As a note, the link vs. on-chip average error ratio is  $e_{\text{link}}/e_{\text{chip}} = 7.5\% / 1.8\% \approx 4.17$ .

## VII. ARCHITECTURE EVALUATION

This section examines MCM and monolithic trade-offs in terms yield, average two-qubit gate infidelity, and application performance. Evaluation will be completed using devices within collision-free yields that are modeled using properties of state-of-art transmon QCs and quantum link hardware. In the absence of quantum chip interposer data, MCM assembly yield will be based upon silicon interposer defect rates [38] due to the parallels between classical and emerging quantum flip-chip bonding techniques [28].

#### A. Benchmarks

Seven benchmarks were used to evaluate the MCM and monolithic systems. The circuits were chosen to adequately cover the application space of realistic QC workloads [81], [82]. Circuits were designed for 80% system qubit utilization to allocate ancilla for compiler mapping and optimization.

**Bernstein-Vazirani (BV):** demonstrates a quantum advantage over classical programming by guaranteeing the return of the bitwise product of some input with a hidden string [12].

**Quantum Approximate Optimization Algorithm (QAOA):** a hybrid quantum-classical algorithm for applications solving NP-Hard combinatorial optimization problems [27].

**Greenberger–Horne–Zeilinger (GHZ):** prepares large-scale entanglement that is required by many complex quantum algorithms and communication protocols for quantum-enabled computational speedups [29].

**Ripple-Carry Adder:** a critical subroutine in quantum algorithms such as Shor's quantum factoring. We include the implementation described in [22].

**Quantum Primacy:** generates random quantum circuits similar to those proposed for and used to demonstrate quantum primacy, or advantage [9], [14].

**Bit Code:** implements a syndrome measurement in a bit-flip ECC.

**Hamiltonian:** constructs circuits that simulate 1D Transverse Field Ising Models (TFIM) used to discover static properties of quantum systems, such as ground-state energies of molecules [10].

### B. Methodology

Variation during fabrication, on-chip gate error, and link quality were set using state-of-art QC data. We anticipate future improvements will scale down infidelity as qubits scale up. Here, our focus is on the near term, and our analysis includes QCs comprising of several hundred qubits.

We considered chiplets with 10, 20, 40, 60, 90, 120, 160, 200, and 250 qubits. We evaluated a total of 102 MCMs against their monolithic counterparts. MCM dimensions of  $k \times m$  were chosen so that each MCM in a chiplet category had a unique size  $\leq 500$  qubits. Since duplicates were avoided in terms of number of qubits, MCM dimensions that were more ‘square’ were prioritized to reduce topology graph diameter. For example, a 40-qubit MCM of dimension  $2 \times 2$  with 10-qubit chiplets was included in our analysis whereas a  $4 \times 1$  configuration with 10-qubit chiplets was omitted.

Yield information resulted from MC simulation with  $F_{0,1,2}$  set to 5.0, 5.06, and 5.12 GHz (Section IV-B),  $\sigma_f = 0.014$  GHz (fabrication precision of [32]), and an original batch size of 10,000 units. After Table I criteria evaluation, collision-free chiplets were grouped for MCM assembly.

MCM assembly creates as many high-quality MCMs as possible. The exploration space for the best chiplets to use within an MCM is exponential, and rather than search for optimums or use random selection from the collision-free bin, infidelity and frequency information gathered from prior KGD characterization guides MCM assembly. Chiplet stitching procedures use the chiplets with the lowest error rates first and the highest error rates last from the collision-free bin. As many complete and collision-free MCMs possible are created following the principle that once a chiplet is within an MCM, it is removed from the collision-free bin. If a frequency collision between adjacent chiplets is found with a particular MCM configuration, chiplet placement is shuffled within the MCM. If a collision-free MCM is not discovered according to time-out criteria selected during runtime (100 maximum reconfigurations), chiplets are returned back to the bin and MCM assembly continues with a new subset of chiplets from the sorted, collision-free bin.

Yield analysis includes loss due to the inability to place chiplets within collision-free MCMs and linking faults. Linking faults during assembly describe the failure to properly bond linking qubits on the chiplet edge to the carrier chip so that they can be coupled inter-chip. We used data involving silicon interposer defect rates from [38] to derive a success probability estimation for each C4 bump bond on a passive carrier chip of  $s_l = 99.999960642\%$ . Fabrication information

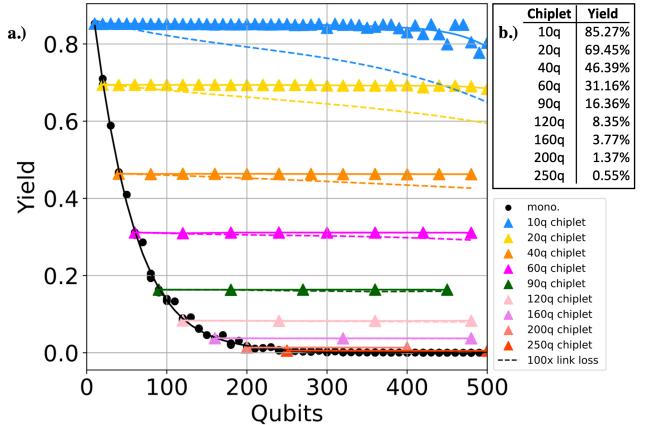


Figure 8. (a) Plot of yield vs. qubits. Curves represent different architecture types: monolithic and MCM. MCM yield includes yield losses from assembly and linking. Yield vs. qubits relationship where link qubit bonding failure is  $100 \times$  indicated with dashed line. (b) Chiplet yields.

from [28] inspired the allocation 25 bump bonds for each linked qubit on a chiplet’s edge. Thus, each qubit linked inter-chip has a  $s_l^{25}$  probability of a successful bond.

During infidelity analysis, two-qubit gate infidelity for collision-free devices was assigned using real-machine data (Section VI-A). Average infidelity averaged across every qubit pair is assumed as known and used to rank chiplets by performance potential to prioritize the use of the best chiplets within MCM construction (Section V-D). Once assembled and flip-chip bonded, we assigned each MCM’s two-qubit gate error over inter-module links (Section VI-B). We studied the influence of MCM architectures on overall system average two-qubit operation infidelity.

For application analysis, the benchmarks described in Section VII-A were generated to target 80% of qubits on the MCM and monolithic devices. Application evaluation used fidelity product of gates, a figure of merit based on the estimated probability of success (ESP) metric frequently used in prior work to predict quantum program success on hardware [46], [57], [62], [80]. Since our simulation methods include the assignment of two-qubit infidelity between intra- and inter-chip connected qubits post fabrication and assembly, respectively, our fidelity product that estimates benchmark success is calculated by multiplying all two-qubit operator fidelities. We note that some higher-dimension MCMs lack a monolithic counterpart during evaluation due to 0% yield during MC simulation.

### C. Results

In this section, yield, gate performance, and application-based analysis show the feasibility of QC scaling via modular design. Data used to evaluate quantum MCM and monolithic architectures can be found in [77].

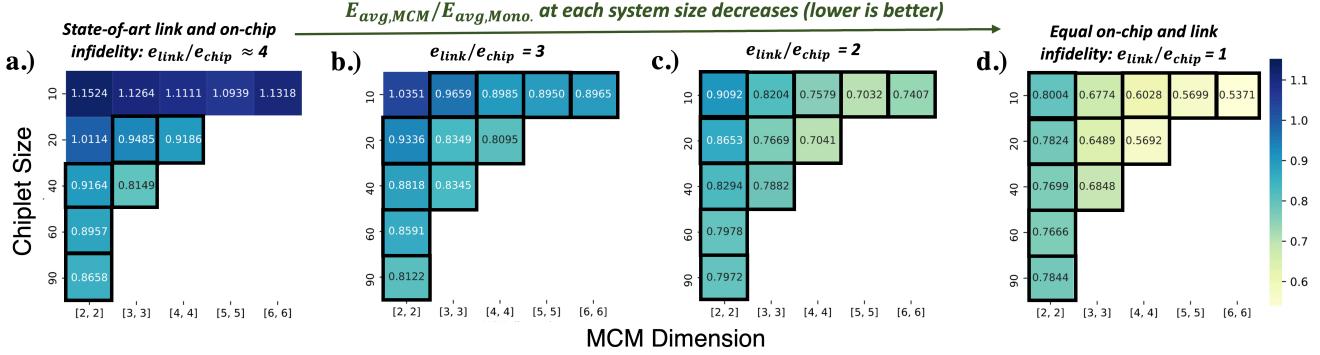


Figure 9. Heatmaps showing the MCM vs. monolithic ratio of infidelity averaged across every qubit pair,  $E_{avg,MCM}/E_{avg,Mono}$  using the scaled collision-free yields. A ratio less than one indicates that the MCM systems have lower average two-qubit gate infidelity. Modeling includes (a) state-of-art link and on-chip infidelity,  $e_{link}/e_{chip} \approx 4$ , along with cases where links in MCMs have improved error: (b)  $e_{link}/e_{chip} = 3$ , (c)  $e_{link}/e_{chip} = 2$ , and (d)  $e_{link}/e_{chip} = 1$  (on-chip and link two-qubit gate infidelity equal).

**1) Yield Evaluation:** We first evaluate MCM vs. monolithic QC architectures in terms of yield. Yield is an important consideration in QC design: although it might be possible to fabricate a high-quality, monolithic QC of a certain qubit size, the trade-off of scrapped chips with critical defects (i.e. frequency collisions) produced along with one good chip might be prohibitively expensive. Yield analysis results are found in Fig. 8. A plot of yield vs. system qubits for monolithic QCs and MCM QCs constructed with varied chiplet size is featured in Fig. 8(a). It should be noted that the yield data for each system type is not strictly monotonic due to the stochastic nature of MC simulation and the slight influence that QC dimension has on the occurrence of frequency collisions (i.e. more “square” devices have more qubit-qubit communication) and thus yield. The MCM curves consider loss during assembly caused by 1) chiplets that cannot be placed in a complete, collision-free MCM and 2) link failures. Post-assembly yield is calculated as the ratio of total chiplets used to assemble complete, collision-free MCMs vs. total chiplets in the original fabrication batch (i.e. 10,000) multiplied by  $(s_l^{25})^L$ , where L is the total number of edge qubits used for inter-chip MCM linking. To show sensitivity to link qubit loss during assembly, dashed lines in 8 show adjusted yield when the probability of bonding failure is 100×. Regardless of the amplified failure during assembly, MCM systems show considerable benefits in terms of yield. Monolithic systems, however, show the most dramatic drop in yield, dropping to ~10% at 120 qubits, alluding that monolithic architectures are the most costly in terms of wasted resources during fabrication. Chiplet yields are explicitly stated in Fig. 8(b). These serve as the upper bound of MCM yield if all elements of the collision-free chiplet bins can be successfully placed within a collision-free MCM. The assembly and linking contributions to yield cause the plotted points of the MCM curves to fall below the yields of Fig. 8(b). We note, however, that our results find that the assembly and linking loss only slightly impact yield benefits of the

MCM sizes under investigation.

A key takeaway from Fig. 8 is that with  $\sigma_f = 0.014$  GHz, it is unfeasible to create a monolithic device  $\gtrsim 400$  qubits because of prohibitively low yield. As devices approach 500 qubits, MCM architectures have slightly lower yields than their base-chiplet yields described in Fig. 8. This is due to the increased likelihood of encountering a defect during assembly as more inter-chiplet links are required. Additionally, a greater number of leftover chiplets result that cannot be placed in a complete, collision-free MCM of larger dimension. Regardless, MCM yields are still typically higher than that of the monolithic QCs. All configurations considered, MCM groups based on the 10-, 20-, 40-, 60-, 90-, 120-, 160-, and 250-qubit chiplets demonstrate 9.58×, 11.34×, 13.45×, 16.53×, 23.03×, 30.44×, 92.61×, and 52.99× average yield improvement relative to their monolithic counterparts, respectively. The 200-qubit chiplet was excluded from this analysis since it only creates a 400-qubit MCM that had a corresponding monolithic device with 0% yield during MC simulation. Theoretically, MCM yield improvement is infinite when monolithic yields are 0%.

**2) Infidelity Evaluation:** We compare the influence of chiplet size on average infidelity averaged across every qubit pair,  $E_{avg.}$ , for the MCM and monolithic systems. We also evaluate the relationship between  $E_{avg,MCM}$  and  $E_{avg,Mono}$  as average link error,  $e_{link}$ , improves.

To focus this analysis, we evaluate a subset of our experimental systems where the MCMs are  $n \times n$  (i.e. chiplets are assembled in a ‘square’). Square structures are favorable because they reduce the graph diameter of the system topology, minimizing worst-case communication distance between qubits. Fig. 9 shows heatmaps describing the MCM vs. monolithic average infidelity averaged across every qubit pair,  $E_{avg,MCM}/E_{avg,Mono}$  for various link quality conditions. A ratio less than one (magnitude indicated by cell brightness) indicates that the MCM has lower average error rates as compared to its monolithic counterpart. These areas within

the heatmap are highlighted with a thick, black border. It is important to reemphasize that yield benefits of smaller chiplets often enable the production many more MCMs than monolithic devices of equivalent size. Thus, the yield relationship scales the number of MCMs used in architecture comparison: we compare the devices in the collision-free monolithic yield to the MCMs resulting from the chiplets in the scaled, collision-free bin.

The  $E_{avg,MCM}/E_{avg,Mono}$  ratios of Fig. 9(a) analyze systems that use state-of-art inter-chip link and on-chip infidelity. As described in Section VI-B,  $e_{link}/e_{chip} = 7.5\% / 1.8\% \approx 4$ , prohibiting MCMs with 10-qubit chiplets from showing an advantage in terms of average infidelity, even if they have considerable yield benefits. Small chiplets are typically characterized by lower two-qubit error rates, especially those that rank higher in the sorted, collision-free bin. However, the number of links that the smaller-chiplet MCMs require inflate average error rates. With chiplets of 20 qubits and larger, we begin to see MCMs surpass the performance of the monolithic systems. For example, the  $3 \times 3$  MCM with a 20 qubit chiplet (180 qubits) has  $E_{avg,MCM}/E_{avg,Mono} = 0.9485$  and the  $2 \times 2$  MCM with a 40 qubit chiplet (160 qubits) has  $E_{avg,MCM}/E_{avg,Mono} = 0.9164$ . Our MCM modeling that uses state-of-art error values show improvements where MCM average two-qubit gate infidelity is reduced to be about 81.5% of its monolithic counterpart for the case of the 40-qubit chiplet arranged in a  $3 \times 3$  MCM.

In the future, it is projected that the quality of communication between separate chiplets coupled on an carrier interposer a will improve as engineering challenges are overcome [28]. In the extreme case where average link error rates are equivalent to on-chip error rates, pictured in Fig. 9(d), MCMs outperform monolithic devices in terms of  $E_{avg}$  in 100% of considered system configurations. However, achieving equivalent on-chip and link qubit-qubit coupling is unnecessary for MCM overall system infidelity, in terms of  $E_{avg}$ , to be less than its monolithic counterpart. We find at  $e_{link} = 2e_{chip}$ , all  $E_{avg,MCM}/E_{avg,Mono}$  ratios are  $< 1$ .

3) *Application Evaluation:* Fig. 10 shows MCM and monolithic QC performance when BV, QAOA, GHZ, adder, quantum primacy, bit-code, and Hamiltonian simulation workloads (Section VII-A) were mapped to the systems under investigation. Since our study analyzed hundreds of compiled benchmarks, details for select benchmarks, those used in the Fig. 9(a), column 1 systems, were included in Table II to provide insight about the gate composition and critical paths of example circuits. The structures we evaluate surpass the capacity of today's most powerful quantum simulators, so our figure of merit is the fidelity product of all two-qubit gates (Section VII-B). Since we are interested in comparing fidelity of the two architectures, the y-axis of the 14 plots in Fig. 10 measures the ratio between MCM and monolithic fidelity. Fig. 10(a) includes all considered MCM/monolithic devices while Fig. 10(b) only includes the ‘square’ systems featured

Chip.	Dim.	Qubits	Bench:1q / 2q / 2q critical
10q	2x2	40	bv: 192 / 437 / 228 g: 3 / 31 / 31 q: 34 / 62 / 33 a: 390 / 1339 / 750 p: 226 / 315 / 74 bc: 16 / 30 / 30 h: 191 / 62 / 62
20q	2x2	80	bv: 384 / 1193 / 604 g: 3 / 63 / 63 q: 66 / 126 / 65 a: 806 / 3741 / 1881 p: 461 / 890 / 88 bc: 32 / 62 / 62 h: 383 / 126 / 126
40q	2x2	160	bv: 768 / 2251 / 1135 g: 3 / 2170 / 1147 q: 130 / 3691 / 1263 a: 1638 / 6915 / 3695 p: 927 / 1552 / 114 bc: 64 / 2557 / 1313 h: 767 / 2320 / 1267
60q	2x2	240	bv: 1152 / 5865 / 2952 g: 3 / 6081 / 3105 q: 195 / 9320 / 3245 a: 2470 / 16164 / 7654 p: 1380 / 4690 / 300 bc: 96 / 3804 / 1973 h: 1151 / 6241 / 3235
90q	2x2	360	bv: 1728 / 6393 / 3282 g: 3 / 5977 / 3112 q: 291 / 10410 / 3413 a: 3717 / 19682 / 9818 p: 2073 / 4382 / 205 bc: 144 / 6016 / 3133 h: 1727 / 6455 / 3480

Table II  
DETAILS OF SELECT COMPILED BENCHMARKS.

in the average infidelity analysis described in Section VII-C2, Fig. 9(a). A value above one indicates a MCM advantage whereas a value below one indicates a monolithic advantage. The distance away from one provides insight about the extent that one architecture outperforms the other. A red “X” (11 total on each Fig. 10(a) plot) indicates where monolithic devices encounter a 0% collision-free yield. Here, the MCM device offers the only solution for implementing the selected benchmark, theoretically resulting in an infinite MCM/monolithic fidelity ratio. It should be highlighted that the extreme values seen on the y-axis are caused by on-chip and inter-chip infidelity of magnitude  $10^{-2}$ .

A significant takeaway from Fig. 10 is that MCM advantage appears in select cases – chiplet size and MCM dimension likely influence function. For example, as devices increase in size, the amount of variation also increases, negatively impacting overall device performance. Postselected chiplets can create higher-fidelity clusters within MCMs, but there are trade-offs between chiplet size and link infidelity due to communication needed between MCM modules. It is possible, however, to select MCM configurations such

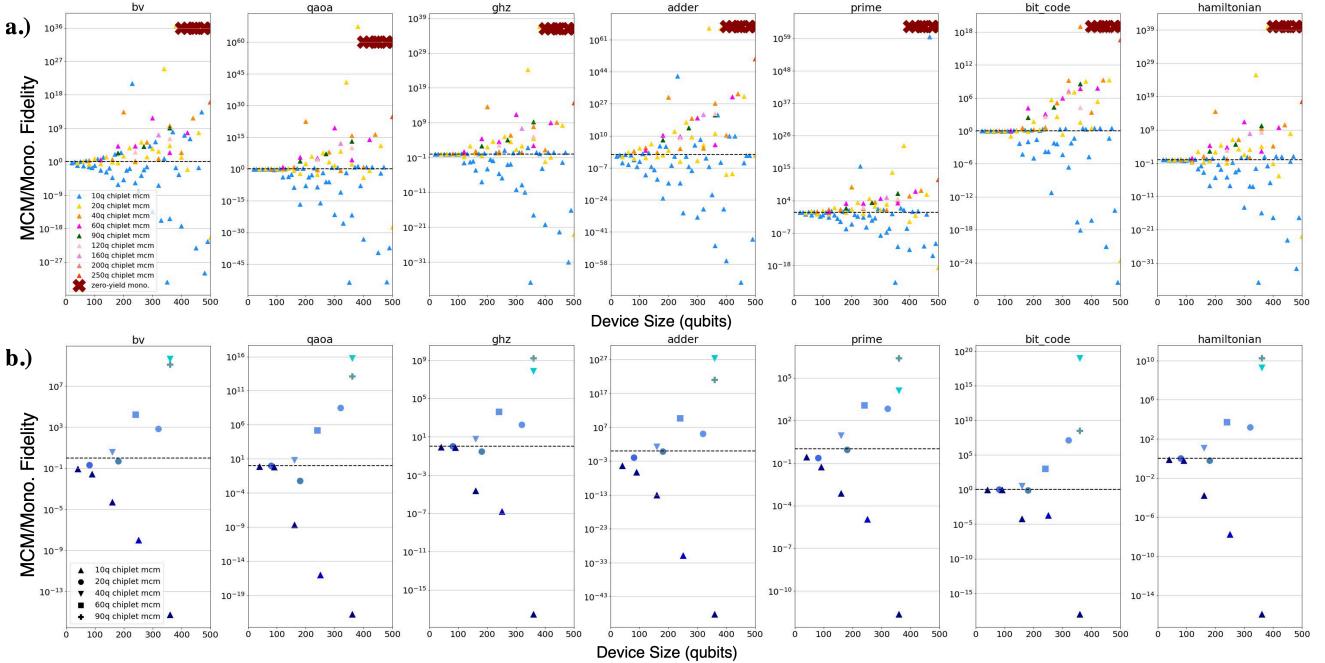


Figure 10. MCM and Monolithic QC benchmark evaluation. (a) Results for all evaluated systems and (b) Results for ‘square’ systems featured in Fig. 9(a)

that higher link errors are balanced and their impact on benchmark performance is reduced. While Fig. 10 shows the viability of MCMs, the MCM is not always the highest-performing architecture, especially in smaller systems. This outcome is not unique to our study; past work in classical chiplets have also reported performance losses in modular systems, but often these losses are heavily outweighed by the benefits that manufacturing chiplets over monolithic devices brings [39]. However, we would like to emphasize that our results show the potential for guiding future quantum architecture design, especially with the results of Fig. 10(b). Here, the points are colored according to their assignment within the heatmap in Fig. 9(a), and it is shown that MCMs with lower  $E_{avg.,MCM}/E_{avg.,Mono}$  tend to have better benchmark performance as compared to their monolithic counterparts. We observe a relationship between the MCM vs. monolithic infidelity (lower better) and the MCM vs. monolithic application fidelity (higher better). In Fig. 10(b) all of the 40-, 60-, and 90-qubit chiplets show advantage, and their values in Fig. 9(a) are  $< 1$ .

### VIII. FUTURE SCALING CONSIDERATIONS

As modular quantum architectures mature, chiplet and MCM specific compilers, perhaps ones based on circuit cutting approaches [79], will be essential for informed use of modular systems through awareness of error rate distributions. Intelligent compilation routines that consider links would allow improved global mapping and optimization.

The proposed chiplets maintain heavy-hex connectivity,

supporting eventual adoption of the hybrid surface/Bacon-Shor ECC. However, errors associated with the flip-chip linking technology should be characterized to see if alternative ECCs should be targeted. Dynamic ECC compilation could be explored to allow for adaptive code distances across lower fidelity or more varied sections of the MCM network.

Last, we hope to improve our modeling by applying first principles [49] to better capture the relationship between frequency detuning and operator fidelity during analysis. Additionally, we hope to refine the modeling of quantum MCM assembly as data becomes available.

### IX. RELATED WORK

Advances in quantum hardware [1], [2], [17], algorithm compilation [44], [56], [58], [74], and control mechanisms [11], [16] boost the performance of near-term QCs, improving their viability for scaling. Methods to better characterize and reduce quantum defects that impact coherence [60], [67], [84] are also relevant as they aim to make qubits better-suited for future, large-scale computation. Studies are emerging that model linked quantum devices to better understand distributed QC performance in SC-based systems [42] and trapped-ion devices [54]. Finally, work exists that outlines architecture considerations for teleportation-based quantum multi-core systems [70] and explores latency and scaling trade-offs needed in a technology-agnostic manner for effective communication and computation in such systems [71], [72]. In this section, we describe how our contributions differ from prior art related

to this study.

The frequency allocation problem associated with transmon QCs along with the yield impact of frequency collisions appeared in prior work [15], [55]. The work of [32] provides a solution to minimize collision conditions by presenting a technique that adjusts transmon frequencies post-fabrication with laser annealing to significantly boost transmon QC yield. Afterwards, laser annealing was successfully applied on real IBM QCs [86] currently deployed on the IBM Quantum Cloud [3]. This prior work, along with the fabrication data of the experimental paper [28], collectively provide theoretical foundations that guided parameter selection in our simulations. We extended the ideas of these papers into the space of modular quantum systems, demonstrating potential yield and performance benefits that MCMs offer.

Application-specific architecture design for SC QCs was explored in [45] with a bottom-up approach. This prior work contributes a methodology that analyzes benchmarks for connectivity and optimizes for communication and qubit frequency allocation when mapping to transmon QCs. Resulting QC architectures of 20 qubits or less were analyzed in terms of both yield and mapped algorithm gate counts. While [45] motivates researching the balance between QC performance and hardware yield, our work differs by taking a more top-down approach to explore the practicality of modular quantum systems, quantum MCMs, as QCs scale to hundreds of qubits. We develop a MCM simulation tool, currently parameterized with state-of-art fabrication and real machine details, but adaptable to future QC improvements. We target a wide range of architecture sizes during yield, device property, and performance analysis.

## X. CONCLUSION

Quantum machines have rapidly improved in fidelity and qubit count, but they face significant scaling challenges. Here, we propose a modular approach as a strategy to overcome the obstacles associated with fabricating larger QCs. Our work presents a first step toward modeling the yield, fidelity, and performance properties of fixed-frequency, transmon-based quantum chiplet architectures, exploring tradeoffs in system design. Our results show that chiplet architectures, relative to monolithic designs, benefit from average yield improvements ranging from approximately  $9.6 - 92.6 \times$  for  $\lesssim 500$  qubit machines. In addition, we discover MCM configurations that demonstrate average two-qubit gate infidelity reductions of approximately  $0.949 - 0.815 \times$  compared to monolithic counterparts. We hope that this work leads to greater exploration of quantum chiplet design.

## ACKNOWLEDGMENT

The authors would like to thank Malcolm Carroll, Nathan Earnest, Nikos Hardavellas, Jared Hertzberg, Ali Javadi-Abhari, and Jason Orcutt for their helpful discussions and comments during the preparation of this manuscript. This

work is funded in part by EPiQC, an NSF Expedition in Computing, under award CCF-1730449; in part by STAQ under award NSF Phy-1818914; in part by NSF award 2110860; in part by the US Department of Energy Office of Advanced Scientific Computing Research, Accelerated Research for Quantum Computing Program; in part by the NSF Quantum Leap Challenge Institute for Hybrid Quantum Architectures and Networks (NSF Award 2016136); and in part based upon work supported by the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers. This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. KNS is supported by IBM as a Postdoctoral Scholar at the University of Chicago and the Chicago Quantum Exchange. GSR is supported as a Computing Innovation Fellow at the University of Chicago. This material is based upon work supported by the National Science Foundation under Grant # 2030859 to the Computing Research Association for the CIFellows Project. FTC is Chief Scientist for Quantum Software at ColdQuanta and an advisor to Quantum Circuits, Inc.

## REFERENCES

- [1] “Hardware-aware approach for fault-tolerant quantum computation,” <https://www.ibm.com/blogs/research/2020/09/hardware-aware-quantum/>, accessed: 2022-04-01.
- [2] “Ibm quantum breaks the 100-qubit processor barrier,” <https://research.ibm.com/blog/127-qubit-quantum-processor-eagle>.
- [3] “Ibm quantum experience,” <https://quantum-computing.ibm.com>, accessed: 2020-11-18.
- [4] “Ibm quantum processor types,” <https://quantum-computing.ibm.com/services/docs/services/manage/systems/processors>, accessed: 2022-03-31.
- [5] “Ibm quantum systems,” <https://quantum-computing.ibm.com/services?systems=all>, accessed: 2020-11-18.
- [6] “The ibm quantum heavy hex lattice,” <https://www.research.ibm.com/blog/heavy-hex-lattice>, 2021.
- [7] “Eagle’s quantum performance progress,” <https://research.ibm.com/blog/eagle-quantum-processor-performance>, 2022.
- [8] “Expanding the ibm quantum roadmap to anticipate the future of quantum-centric supercomputing,” <https://research.ibm.com/blog/ibm-quantum-roadmap-2025>, 2022.
- [9] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell *et al.*, “Quantum supremacy using a programmable superconducting processor,” *Nature*, vol. 574, no. 7779, pp. 505–510, 2019.
- [10] L. Bassman, K. Liu, A. Krishnamoorthy, T. Linker, Y. Geng, D. Shebib, S. Fukushima, F. Shimojo, R. K. Kalila, A. Nakano *et al.*, “Towards simulation of the dynamics of materials on quantum computers,” *Physical Review B*, vol. 101, no. 18, p. 184305, 2020.

- [11] Y. Baum, M. Amico, S. Howell, M. Hush, M. Liuzzi, P. Mundada, T. Merkh, A. R. Carvalho, and M. J. Biercuk, “Experimental deep reinforcement learning for error-robust gate-set design on a superconducting quantum computer,” *PRX Quantum*, vol. 2, no. 4, p. 040324, 2021.
- [12] E. Bernstein and U. Vazirani, “Quantum complexity theory,” *SIAM Journal on computing*, vol. 26, no. 5, pp. 1411–1473, 1997.
- [13] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, “Quantum machine learning,” *Nature*, vol. 549, no. 7671, pp. 195–202, 2017.
- [14] S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, “Characterizing quantum supremacy in near-term devices,” *Nature Physics*, vol. 14, no. 6, pp. 595–600, 2018.
- [15] M. Brink, J. M. Chow, J. Hertzberg, E. Magesan, and S. Rosenblatt, “Device challenges for near term superconducting quantum processors: frequency collisions,” in *2018 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2018, pp. 6–1.
- [16] A. R. Carvalho, H. Ball, M. J. Biercuk, M. R. Hush, and F. Thomsen, “Error-robust quantum logic optimization using a cloud quantum computer interface,” *Physical Review Applied*, vol. 15, no. 6, p. 064054, 2021.
- [17] C. Chamberland, G. Zhu, T. J. Yoder, J. B. Hertzberg, and A. W. Cross, “Topological and subsystem codes on low-degree graphs with flag qubits,” *Physical Review X*, vol. 10, no. 1, p. 011022, 2020.
- [18] J. B. Chang, M. R. Vissers, A. D. Córcoles, M. Sandberg, J. Gao, D. W. Abraham, J. M. Chow, J. M. Gambetta, M. Beth Rothwell, G. A. Keefe *et al.*, “Improved superconducting qubit coherence using titanium nitride,” *Applied Physics Letters*, vol. 103, no. 1, p. 012602, 2013.
- [19] J. M. Chow, A. D. Córcoles, J. M. Gambetta, C. Rigetti, B. R. Johnson, J. A. Smolin, J. R. Rozen, G. A. Keefe, M. B. Rothwell, M. B. Ketchen *et al.*, “Simple all-microwave entangling gate for fixed-frequency superconducting qubits,” *Physical review letters*, vol. 107, no. 8, p. 080502, 2011.
- [20] J. M. Chow, L. DiCarlo, J. M. Gambetta, F. Motzoi, L. Frunzio, S. M. Girvin, and R. J. Schoelkopf, “Optimized driving of superconducting artificial atoms for improved single-qubit gates,” *Physical Review A*, vol. 82, no. 4, p. 040305, 2010.
- [21] A. D. Córcoles, J. M. Gambetta, J. M. Chow, J. A. Smolin, M. Ware, J. Strand, B. L. Plourde, and M. Steffen, “Process verification of two-qubit quantum gates by randomized benchmarking,” *Physical Review A*, vol. 87, no. 3, p. 030301, 2013.
- [22] S. A. Cuccaro, T. G. Draper, S. A. Kutin, and D. P. Moulton, “A new quantum ripple-carry addition circuit,” *arXiv preprint quant-ph/0410184*, 2004.
- [23] S. Dasgupta and T. S. Humble, “Stability of noisy quantum computing devices,” *arXiv preprint arXiv:2105.09472*, 2021.
- [24] Y. Demir, Y. Pan, S. Song, N. Hardavellas, J. Kim, and G. Memik, “Galaxy: A high-performance energy-efficient multi-chip architecture using photonic interconnects,” in *Proceedings of the 28th ACM International Conference on Supercomputing*, 2014, pp. 303–312.
- [25] M. H. Devoret, A. Wallraff, and J. M. Martinis, “Superconducting qubits: A short review,” *arXiv preprint cond-mat/0411174*, 2004.
- [26] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger, “Dark silicon and the end of multicore scaling,” in *2011 38th Annual international symposium on computer architecture (ISCA)*. IEEE, 2011, pp. 365–376.
- [27] E. Farhi, J. Goldstone, and S. Gutmann, “A quantum approximate optimization algorithm,” *arXiv preprint arXiv:1411.4028*, 2014.
- [28] A. Gold, J. Paquette, A. Stockklauser, M. J. Reagor, M. S. Alam, A. Bestwick, N. Didier, A. Nersisyan, F. Oruc, A. Razavi *et al.*, “Entanglement across separate silicon dies in a modular superconducting qubit device,” *npj Quantum Information*, vol. 7, no. 1, pp. 1–10, 2021.
- [29] D. M. Greenberger, M. A. Horne, and A. Zeilinger, “Going beyond bell’s theorem,” in *Bell’s theorem, quantum theory and conceptions of the universe*. Springer, 1989, pp. 69–72.
- [30] L. K. Grover, “A fast quantum mechanical algorithm for database search,” in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, 1996, pp. 212–219.
- [31] L. K. Grover, “A fast quantum mechanical algorithm for database search,” in *ANNUAL ACM SYMPOSIUM ON THEORY OF COMPUTING*. ACM, 1996, pp. 212–219.
- [32] J. B. Hertzberg, E. J. Zhang, S. Rosenblatt, E. Magesan, J. A. Smolin, J.-B. Yau, V. P. Adiga, M. Sandberg, M. Brink, J. M. Chow *et al.*, “Laser-annealing josephson junctions for yielding scaled-up superconducting quantum processors,” *npj Quantum Information*, vol. 7, no. 1, pp. 1–8, 2021.
- [33] A. Houck, J. Schreier, B. Johnson, J. Chow, J. Koch, J. Gambetta, D. Schuster, L. Frunzio, M. Devoret, S. Girvin *et al.*, “Controlling the spontaneous emission of a superconducting transmon qubit,” *Physical review letters*, vol. 101, no. 8, p. 080502, 2008.
- [34] N. E. Jerger, “Chiplet-based systems,” <https://www.sigarch.org/chiplet-based-systems/>, accessed: 2022-03-30.
- [35] P. Jurcevic, A. Javadi-Abhari, L. S. Bishop, I. Lauer, D. Borgorin, M. Brink, L. Capelluto, O. Gunluk, T. Itoko, N. Kanazawa *et al.*, “Demonstration of quantum volume 64 on a superconducting quantum computing system,” *Quantum Science and Technology*, 2021.
- [36] A. Kandala, K. Wei, S. Srinivasan, E. Magesan, S. Carnevale, G. Keefe, D. Klaus, O. Dial, and D. McKay, “Demonstration of a high-fidelity cnot gate for fixed-frequency transmons with engineered z z suppression,” *Physical Review Letters*, vol. 127, no. 13, p. 130501, 2021.

- [37] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, “Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets,” *Nature*, vol. 549, no. 7671, pp. 242–246, 2017.
- [38] A. Kannan, N. E. Jerger, and G. H. Loh, “Enabling interposer-based disintegration of multi-core processors,” in *2015 48th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2015, pp. 546–558.
- [39] M. M. Kim, M. Mehrara, M. Oskin, and T. Austin, “Architectural implications of brick and mortar silicon manufacturing,” in *Proceedings of the 34th annual international symposium on Computer architecture*, 2007, pp. 244–253.
- [40] M. Kjaergaard, M. E. Schwartz, J. Braumüller, P. Krantz, J. I.-J. Wang, S. Gustavsson, and W. D. Oliver, “Superconducting qubits: Current state of play,” *Annual Review of Condensed Matter Physics*, vol. 11, pp. 369–395, 2020.
- [41] J. Kreikebaum, K. O’Brien, A. Morvan, and I. Siddiqi, “Improving wafer-scale josephson junction resistance variation in superconducting quantum coherent circuits,” *Superconductor Science and Technology*, vol. 33, no. 6, p. 06LT02, 2020.
- [42] N. LaRacuente, K. N. Smith, P. Imany, K. L. Silverman, and F. T. Chong, “Short-range microwave networks to scale superconducting quantum computation,” *arXiv preprint arXiv:2201.08825*, 2022.
- [43] R. C. Leachman, “Yield modeling and analysis prof. robert c. leachman ieor 130, methods of manufacturing improvement fall, 2020,” 2020.
- [44] G. Li, Y. Ding, and Y. Xie, “Tackling the qubit mapping problem for nisq-era quantum devices,” in *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2019, pp. 1001–1014.
- [45] G. Li, Y. Ding, and Y. Xie, “Towards efficient superconducting quantum processor architecture design,” in *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2020, pp. 1031–1045.
- [46] G. Li, A. Wu, Y. Shi, A. Javadi-Abhari, Y. Ding, and Y. Xie, “Paulihedral: a generalized block-wise compiler optimization framework for quantum simulation kernels,” in *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 2022, pp. 554–569.
- [47] E. Magesan and J. M. Gambetta, “Effective hamiltonian models of the cross-resonance gate,” *Physical Review A*, vol. 101, no. 5, p. 052308, 2020.
- [48] P. Magnard, S. Storz, P. Kurpiers, J. Schär, F. Marxer, J. Lütolf, T. Walter, J.-C. Besse, M. Gabureac, K. Reuer *et al.*, “Microwave quantum link between superconducting circuits housed in spatially separated cryogenic systems,” *Physical Review Letters*, vol. 125, no. 26, p. 260502, 2020.
- [49] M. Malekakhlagh, E. Magesan, and D. C. McKay, “First-principles analysis of cross-resonance gate operation,” *Physical Review A*, vol. 102, no. 4, p. 042605, 2020.
- [50] J. M. Martinis, “Saving superconducting quantum processors from decay and correlated errors generated by gamma and cosmic rays,” *npj Quantum Information*, vol. 7, no. 1, pp. 1–9, 2021.
- [51] D. C. McKay, S. Sheldon, J. A. Smolin, J. M. Chow, and J. M. Gambetta, “Three-qubit randomized benchmarking,” *Physical review letters*, vol. 122, no. 20, p. 200502, 2019.
- [52] D. C. McKay, C. J. Wood, S. Sheldon, J. M. Chow, and J. M. Gambetta, “Efficient z gates for quantum computing,” *Physical Review A*, vol. 96, no. 2, p. 022330, 2017.
- [53] N. Moll, P. Barkoutsos, L. S. Bishop, J. M. Chow, A. Cross, D. J. Egger, S. Filipp, A. Fuhrer, J. M. Gambetta, M. Ganzhorn *et al.*, “Quantum optimization using variational algorithms on near-term quantum devices,” *Quantum Science and Technology*, vol. 3, no. 3, p. 030503, 2018.
- [54] C. Monroe, R. Raussendorf, A. Ruthven, K. Brown, P. Maunz, L.-M. Duan, and J. Kim, “Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects,” *Physical Review A*, vol. 89, no. 2, p. 022317, 2014.
- [55] A. Morvan, L. Chen, J. M. Larson, D. I. Santiago, and I. Siddiqi, “Optimizing frequency allocation for fixed-frequency superconducting quantum processors,” *Physical Review Research*, vol. 4, no. 2, p. 023079, 2022.
- [56] P. Murali, J. M. Baker, A. Javadi-Abhari, F. T. Chong, and M. Martonosi, “Noise-adaptive compiler mappings for noisy intermediate-scale quantum computers,” in *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2019, pp. 1015–1029.
- [57] P. Murali, D. M. Debroy, K. R. Brown, and M. Martonosi, “Architecting noisy intermediate-scale trapped ion quantum computers,” in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2020, pp. 529–542.
- [58] P. Murali, D. C. McKay, M. Martonosi, and A. Javadi-Abhari, “Software mitigation of crosstalk on noisy intermediate-scale quantum computers,” in *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2020, pp. 1001–1016.
- [59] National Academies of Sciences, Engineering, and Medicine, *Quantum computing: progress and prospects*. National Academies Press, 2019.
- [60] A. Nersisyan, S. Poletto, N. Alidoust, R. Manenti, R. Renzas, C.-V. Bui, K. Vu, T. Whyland, Y. Mohan, E. A. Sete *et al.*, “Manufacturing low dissipation superconducting quantum processors,” in *2019 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2019, pp. 31–1.
- [61] M. A. Nielsen and I. Chuang, *Quantum computation and quantum information*. Cambridge University Press, 2010.
- [62] S. Nishio, Y. Pan, T. Satoh, H. Amano, and R. V. Meter, “Extracting success from ibm’s 20-qubit machines using error-aware compilation,” *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 16, no. 3, pp. 1–25, 2020.

- [63] J. O’Gorman and E. T. Campbell, “Quantum computation with realistic magic-state factories,” *Physical Review A*, vol. 95, no. 3, p. 032338, 2017.
- [64] J. O’Gorman and E. T. Campbell, “Quantum computation with realistic magic-state factories,” *Physical Review A*, vol. 95, no. 3, Mar 2017. [Online]. Available: <http://dx.doi.org/10.1103/PhysRevA.95.032338>
- [65] S. Pal, J. Liu, I. Alam, N. Cebry, H. Suhail, S. Bu, S. S. Iyer, S. Pamarti, R. Kumar, and P. Gupta, “Designing a 2048-chiplet, 14336-core waferscale processor,” in *2021 58th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 2021, pp. 1183–1188.
- [66] G. Paraoanu, “Microwave-induced coupling of superconducting qubits,” *Physical Review B*, vol. 74, no. 14, p. 140504, 2006.
- [67] A. P. Place, L. V. Rodgers, P. Mundada, B. M. Smitham, M. Fitzpatrick, Z. Leng, A. Premkumar, J. Bryon, A. Vrajitoarea, S. Sussman *et al.*, “New material platform for superconducting transmon qubits with coherence times exceeding 0.3 milliseconds,” *Nature communications*, vol. 12, no. 1, pp. 1–6, 2021.
- [68] J. Preskill, “Quantum computing in the nisq era and beyond,” *Quantum*, vol. 2, p. 79, 2018.
- [69] C. Rigetti and M. Devoret, “Fully microwave-tunable universal gates in superconducting qubits with linear couplings and fixed transition frequencies,” *Physical Review B*, vol. 81, no. 13, p. 134507, 2010.
- [70] S. Rodrigo, S. Abadal, E. Alarcón, and C. G. Almudever, “Will quantum computers scale without inter-chip comms? a structured design exploration to the monolithic vs distributed architectures quest,” in *2020 XXXV Conference on Design of Circuits and Integrated Systems (DCIS)*. IEEE, 2020, pp. 1–6.
- [71] S. Rodrigo, S. Abadal, E. Alarcón, M. Bandic, H. Van Someren, and C. G. Almudéver, “On double full-stack communication-enabled architectures for multicore quantum computers,” *IEEE micro*, vol. 41, no. 5, pp. 48–56, 2021.
- [72] S. Rodrigo, S. Abadal, C. G. Almudéver, and E. Alarcón, “Modelling short-range quantum teleportation for scalable multi-core quantum computing architectures,” in *Proceedings of the Eight Annual ACM International Conference on Nanoscale Computing and Communication*, 2021, pp. 1–7.
- [73] S. Sheldon, E. Magesan, J. M. Chow, and J. M. Gambetta, “Procedure for systematically tuning up cross-talk in the cross-resonance gate,” *Physical Review A*, vol. 93, no. 6, p. 060302, 2016.
- [74] Y. Shi, N. Leung, P. Gokhale, Z. Rossi, D. I. Schuster, H. Hoffmann, and F. T. Chong, “Optimized compilation of aggregated instructions for realistic quantum computers,” in *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2019, pp. 1031–1044.
- [75] P. W. Shor, “Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer,” *SIAM Journal on Computing*, vol. 26, no. 5, p. 1484–1509, Oct 1997. [Online]. Available: <http://dx.doi.org/10.1137/S0097539795293172>
- [76] P. W. Shor, “Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer,” *SIAM review*, vol. 41, no. 2, pp. 303–332, 1999.
- [77] K. Smith, “chiplets-for-qc-scaling,” <https://github.com/knsmith/chiplets-for-QC-scaling>, 2022.
- [78] M. Takita, A. W. Cross, A. Córcoles, J. M. Chow, and J. M. Gambetta, “Experimental demonstration of fault-tolerant state preparation with superconducting qubits,” *Physical review letters*, vol. 119, no. 18, p. 180501, 2017.
- [79] W. Tang, T. Tomesh, M. Suchara, J. Larson, and M. Martonosi, “Cutqc: using small quantum computers for large quantum circuit evaluations,” in *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 2021, pp. 473–486.
- [80] S. S. Tannu and M. Qureshi, “Ensemble of diverse mappings: Improving reliability of quantum computers by orchestrating dissimilar mistakes,” in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO ’52. New York, NY, USA: Association for Computing Machinery, 2019, p. 253–265. [Online]. Available: <https://doi.org/10.1145/3352460.3358257>
- [81] T. Tomesh, “quantum-circuit-generator,” [https://github.com/teague томеш/quantum\\_circuit\\_generator](https://github.com/teague томеш/quantum_circuit_generator), 2021.
- [82] T. Tomesh, P. Gokhale, V. Omole, G. S. Ravi, K. N. Smith, J. Viszlai, X.-C. Wu, N. Hardavellas, M. R. Martonosi, and F. T. Chong, “Supermarq: A scalable quantum benchmark suite,” in *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2022, pp. 587–603.
- [83] V. Tripathi, M. Khezri, and A. N. Korotkov, “Operation and intrinsic error budget of a two-qubit cross-resonance gate,” *Physical Review A*, vol. 100, no. 1, p. 012301, 2019.
- [84] J. I. Wang, M. A. Yamoah, Q. Li, A. H. Karamlou, T. Dinh, B. Kannan, J. Braumueler, D. Kim, A. J. Melville, S. E. Muschinske *et al.*, “Hexagonal boron nitride as a low-loss dielectric for superconducting quantum circuits and qubits,” *Nature Materials*, pp. 1–6, 2022.
- [85] J. Yin, Z. Lin, O. Kayiran, M. Poremba, M. S. B. Altaf, N. E. Jerger, and G. H. Loh, “Modular routing design for chiplet-based systems,” in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2018, pp. 726–738.
- [86] E. J. Zhang, S. Srinivasan, N. Sundaresan, D. F. Bogorin, Y. Martin, J. B. Hertzberg, J. Timmerwilke, E. J. Pritchett, J.-B. Yau, C. Wang *et al.*, “High-fidelity superconducting quantum processors via laser-annealing of transmon qubits,” *arXiv preprint arXiv:2012.08475*, 2020.
- [87] H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu *et al.*, “Quantum computational advantage using photons,” *Science*, vol. 370, no. 6523, pp. 1460–1463, 2020.

- [88] Y. Zhong, H.-S. Chang, A. Bienfait, É. Dumur, M.-H. Chou, C. R. Conner, J. Grebel, R. G. Povey, H. Yan, D. I. Schuster *et al.*, “Deterministic multi-qubit entanglement in a quantum network,” *Nature*, vol. 590, no. 7847, pp. 571–575, 2021.