

# Architectures for Multinode Superconducting Quantum Computers

James Ang,<sup>1</sup> Gabriella Carini,<sup>2</sup> Yanzhu Chen,<sup>3</sup> Isaac Chuang,<sup>4</sup> Michael Austin DeMarco,<sup>2,4,\*</sup> Sophia E. Economou,<sup>3</sup> Alec Eickbusch,<sup>5,6</sup> Andrei Faraon,<sup>7</sup> Kai-Mei Fu,<sup>8</sup> Steven M. Girvin,<sup>5</sup> Michael Hatridge,<sup>9</sup> Andrew Houck,<sup>10</sup> Paul Hilaire,<sup>3</sup> Kevin Krsulich,<sup>11</sup> Ang Li,<sup>1</sup> Chenxu Liu,<sup>3</sup> Yuan Liu,<sup>4</sup> Margaret Martonosi,<sup>12</sup> David C. McKay,<sup>11</sup> James Misewich,<sup>2</sup> Mark Ritter,<sup>11</sup> Robert J. Schoelkopf,<sup>6</sup> Samuel A. Stein,<sup>1</sup> Sara Sussman,<sup>10</sup> Hong X. Tang,<sup>13,5</sup> Wei Tang,<sup>12</sup> Teague Tomesh,<sup>12</sup> Norm M. Tubman,<sup>14</sup> Chen Wang,<sup>15</sup> Nathan Wiebe,<sup>1,16</sup> Yong-Xin Yao,<sup>17</sup> Dillon C. Yost,<sup>14,18</sup> and Yiyu Zhou<sup>5,13</sup>

<sup>1</sup>*Pacific Northwest National Laboratory, Richland, WA 99354, USA*

<sup>2</sup>*Brookhaven National Laboratory, Upton, NY 11973, USA*

<sup>3</sup>*Department of Physics, Virginia Tech, Blacksburg, VA 24061, USA*

<sup>4</sup>*Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

<sup>5</sup>*Yale Quantum Institute, Yale University, New Haven, CT 06511, USA*

<sup>6</sup>*Departments of Applied Physics and Physics, Yale University, New Haven, CT 06520, USA*

<sup>7</sup>*Thomas J. Watson, Sr., Laboratory of Applied Physics, California Institute of Technology, Pasadena, CA, USA*

<sup>8</sup>*Department of Physics, University of Washington, Seattle WA 98195, USA*

<sup>9</sup>*Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA 15260, USA*

<sup>10</sup>*Department of Electrical Engineering, Princeton University, Princeton, USA*

<sup>11</sup>*IBM Quantum, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA*

<sup>12</sup>*Department of Computer Science, Princeton University, Princeton, NJ 08544, USA*

<sup>13</sup>*Departments of Electrical Engineering, Yale University, New Haven, CT 06511, USA*

<sup>14</sup>*Quantum Artificial Intelligence Laboratory (QuAIL), Exploration Technology Directorate, NASA Ames Research Center, Moffett Field, California 94035, USA*

<sup>15</sup>*Department of Physics, University of Massachusetts-Amherst, Amherst, MA, 01003, USA*

<sup>16</sup>*Department of Computer Science, University of Toronto, Toronto, ON M5S1J7, Canada*

<sup>17</sup>*Ames National Laboratory and Iowa State University, Ames, Iowa 50011, USA*

<sup>18</sup>*KBR, 601 Jefferson St., Houston, Texas 77002, USA*

Many proposals to scale quantum technology rely on modular or distributed designs where individual quantum processors, called nodes, are linked together to form one large multinode quantum computer (MNQC). One scalable method to construct an MNQC is using superconducting quantum systems with optical interconnects. However, a key limiting factor of these machines will be internode gates, which may be two to three orders of magnitude noisier and slower than local operations. Surmounting the limitations of internode gates will require a range of techniques, including improvements in entanglement generation, the use of entanglement distillation, and optimized software and compilers, and it remains unclear how improvements to these components interact to affect overall system performance, what performance from each is required, or even how to quantify the performance of a component. In this paper, we employ a ‘co-design’ inspired approach to quantify overall MNQC performance in terms of hardware models of internode links, entanglement distillation, and local architecture. In the particular case of superconducting MNQCs with microwave-to-optical interconnects, we uncover a tradeoff between entanglement generation and distillation that threatens to degrade MNQC performance. We show how to navigate this tradeoff in the context of algorithm performance, layout how compilers and software should optimize the balance between local gates and internode gates, and discuss when noisy quantum internode links have an advantage over purely classical links. Using these results, we introduce a research roadmap for the realization of early MNQCs, which illustrates potential improvements to the hardware and software of MNQCs and outlines criteria for evaluating the improvement landscape, from progress in entanglement generation to the use of quantum memory in entanglement distillation and dedicated algorithms such as distributed quantum phase estimation. While we focus on superconducting devices with optical interconnects, our approach is general across MNQC implementations.

## I. INTRODUCTION

Modular, distributed, or multinode quantum computers (MNQCs) [1–8], wherein smaller devices or “nodes” are networked together [9] to make a unified multinode quantum computer, are considered a leading approach

to building large scale systems [10] without the associated difficulties of producing large monolithic devices [5]. Leading platforms include ion-trap computers with multiple traps [11–13], solid-state systems [14–16], atomic systems [17–19], and superconducting devices [2, 7, 20–27].

In superconducting devices, which we focus on in this paper, a motivation for MNQCs is not only the complexities associated with building larger devices, but the limitations set by the individual capacity of the cryogenic di-

\* mdemarco@bnl.gov

lution refrigerator required to cool the device [28]. Building links between devices in different refrigerators is thus a key capability [29]. Early-stage MNQCs with cryogenic links between refrigerators have been demonstrated [23], and when cryogenic links can be feasibly constructed they are a leading candidate for small MNQCs [30]. On the other hand, future large quantum systems may involve many nodes distributed over tens or even hundreds of meters, at which scale both serviceability requirements [31] and cable loss [23, 30, 32–36] become an issue. Rather than using cryogenic links, a system composed of transmon array devices [37, 38] housed in separate refrigerators with room-temperature microwave-to-optical (M2O) quantum internode links [39–45] between them is a more scalable proposal for building future MNQCs.

Despite many promising experimental platforms, internode links in these systems are likely to be much noisier and slower than local gates and thus threaten the viability of MNQCs [46, 47]. These weak internode links hamper performance both by directly causing errors and by creating a computational bottleneck which allows decoherence [37, 38] to degrade quantum information. While true across platforms, this problem is particularly pronounced in superconducting devices with M2O links, where the conversion faces serious limitations due to the weakness of the nonlinear conversion process, fiber-to-chip coupling, thermal added noise, and other hardware difficulties [39–44]. In order to be viable, systems with quantum internode links must outperform not only monolithic quantum systems but also systems with only classical ‘circuit cutting’ links between nodes [48, 49].

To guide the development of M2O MNQCs, we must quantify the available performances of internode links using M2O hardware, determine how these performances affect algorithm execution performance, and determine how hardware and software should jointly navigate design space tradeoffs.

However, evaluating the performance of MNQCs becomes complicated because MNQCs must balance three expensive resources: local two-qubit gates, internode gates, and classical circuit cutting links. In particular, any internode gate may always be cut and replaced with a circuit cutting link, thereby exchanging the noise and time of an internode link for the multiple executions required by circuit cutting [50–53]. On the other hand, compiler and algorithm design can often trade an internode gate for increased local computation, thus exchanging an internode link for a longer and deeper circuit [54, 55]. Thus, in order to understand the attainable performance of an MNQC, we must not only understand the computational cost of internode gates, but also the relative costs of local computation and circuit cutting vis-a-vis internode gates.

Evaluation of internode gate performance itself presents a challenge of complexity and scale. Rather than directly transmitting quantum information between nodes, MNQCs can use heralded protocols to distribute entangled pairs (EPs) [46, 56, 57] between nodes, distill

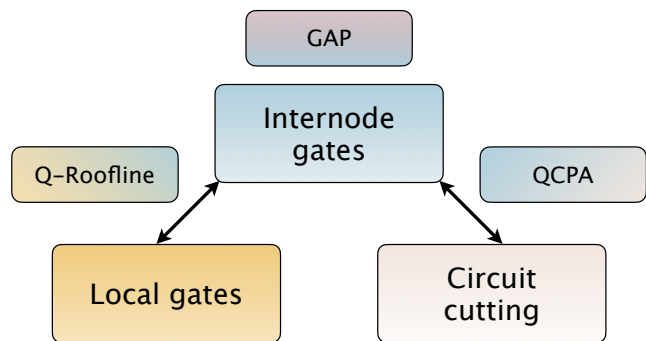


FIG. 1. Multinode Quantum Computers (MNQCs) must balance three key resources: internode gates, local computation, and circuit cutting gates. To evaluate the costs of these resources, we introduce three models. The Gate-Algorithm Performance (GAP) model determines the overall performance of internode gates in terms of detailed models of the internode hardware and local distillation and compares this to the demands of algorithms. Building on this analysis, the Quantum Roofline (Q-Roofline) model compares the relative costs of internode gates and local computation and guides compiler balancing between the two. Similarly, the Quantum vs. Classical Performance Analysis (QCPA) evaluates the relative cost of internode gates and circuit cutting gates to optimize the usage of both.

them [30, 58, 59], and execute teleported gates [60]. However, modeling this as a unified system quickly becomes untenable: modeling the internode link including the entanglement generation [60] and up to just six rounds of distillation requiring  $2^6$  EPs quickly scales to require a minimum of 18 qubits just in the internode link, rendering the simulation of a just a 10-qubit MNQC equivalent to the simulation of a  $2^{28} \times 2^{28}$  density matrix, well beyond what can be simulated today [61, 62]. Furthermore, understanding the relative costs of internode gates, circuit cutting, and local computation will require integrating the already large internode link performance models with an understanding of compiler frameworks [54, 55, 63–65] and network architectures that are able to optimize around the weak links and scale with system sizes. Thus to address the scaling behavior of large MNQCs, we must determine simple and scalable ways to evaluate circuit cutting and local computation against internode gate performance. Without a clear understanding of how all the components of an MNQC, from entanglement generation and distillation to compiler and algorithm design, interact to affect system performance, future improvements in hardware and software may be incompatible and lead to reduced or even no improvements in MNQC performance [66].

Classical computing has navigated similarly complex design constraints to build high-performance multi-node systems [67–69]. Early multicomputers including the ALEWIFE [70, 71] and BEOWULF [72] systems utilized existing state-of-the-art hardware to lay the foundations of classical multinode networks with distributed mem-

ory and communication systems that evolved into the interconnection architectures such as the Infiniband and Slingshot networks [73–75] used by contemporary high-performance computing systems. Physical hardware and software constraints were key to building modern interconnection architectures, from ‘fat-tree’ networks [76] that balance network bandwidth against the size of an architecture to adaptive routing [77] and complex network architectures [78] that maximize system performance while minimizing wiring overhead. From an architectural perspective, these tradeoffs are captured in ‘Roofline’ models [79] which quantifies the relative burden of local computation and memory communication. Taken together, this approach of navigating tradeoffs by designing hardware and software jointly came to be called ‘co-design’ [80] and has played a significant role in the design of modern high-performance and exascale computing [81, 82].

On the other hand, considerable research has been directed towards the design of networked quantum systems, of which MNQCs would be a subset. Building on early proposals for quantum networks [83] and quantum internet [84], recent works have elaborated a vision for the development of a truly distributed quantum ecosystem [1, 85, 86], although hardware which is capable of delivering the requisite performance largely remains to be developed [87, 88]. Layered link protocols [89–91] focused on the preparation of nonlocal entanglement [89, 92] modeled from the classical internet have also been elaborated. However, how these will interact with highly constrained platforms has only begun to be understood, with progress on routing optimization [93] and dedicated compilers and frameworks [94–98]. With substantial progress envisioned in the realization of high-performance quantum interlinks [9], joint co-design of hardware and software will be key to enabling quantum networks [99].

In this paper, we use a co-design layer architecture to simplify the problem of internode links and thus quantify the full range of available internode performance with present technology; we then integrate these results into models that determine the relative costs of local, internode, and circuit cutting links to model algorithm performance and map out tradeoffs; and finally we lay out a research roadmap that proposes improvements and quantifies their possible effects in terms of these models. Our analysis reveals a key tradeoff between the time of execution and the fidelity of internode gates and we show how to navigate this using M2O pump power settings, entanglement distillation, and error mitigation techniques. We present simulation results on sample hardware demonstrating this tradeoff and quantify the available performance of internode links. Using these results, we evaluate algorithm execution performance using internode links using a ‘Gate-Algorithm Performance’ (GAP) model, introduce a ‘Quantum Roofline’ (Q-Roofline) model which determines the relative costs of internode and local computation and guides compiler resource balancing, and perform a ‘Quantum-Classical

Performance Analysis’ (QCPA) to demonstrate the relative costs of error-mitigated internode links against circuit cutting links. Finally, we discuss proposed improvements to each layer of the MNQC, from entanglement generation and distillation to algorithm and compiler design, and use the GAP, Q-Roofline, and QCPA models to discuss their effects and relative tradeoffs. While we focus on superconducting systems with M2O interlinks in this paper, our approach is generic to any physical MNQC implementation which uses entanglement generation to execute remote operations or with links that may be characterized by the time and fidelity of operations, including quantum networks [85, 86] and cryogenic microwave links [7, 20, 30, 100].

The next section presents an overview of the specific platform we consider, namely superconducting transmon devices with M2O interlinks. In Section III, we discuss the co-design architecture that allows the problem of internode links to be drastically simplified by splitting the internode link into distinct layers. Section IV then presents and analyzes models of each of the layers. Section V unifies these models into a full stack model, and presents the GAP, Q-Roofline, and QCPA analyses. In Section VI, we present a research roadmap for the development of highly performant MNQCs and discuss advances in light of the MNQC architecture and analyses. Finally, Section VII discusses potential applications of our methodology to other quantum platforms.

## II. SUPERCONDUCTING DEVICES WITH M2O INTERLINKS

Over the past two decades, the superconducting circuit has become an established platform for large-scale quantum information processing. While systems with several hundred superconducting qubits have been built, and systems with thousands are planned [2, 101], scaling remains a serious challenge. Available cryogenic capacity and qubit control infrastructure are two major limitations for achieving devices at the scale required for cutting-edge applications. Furthermore, today’s large superconducting devices are monolithic so smaller subsystems cannot be tested in isolation or individually replaced to improve system performance. Hence we consider a multinode superconducting quantum computer with each ‘node’ comprised of a refrigerator holding a superconducting circuit which handles intermediate-scale computation tasks via local state preparation, gate operation, and measurement. For simplicity, we assume the classical communication between nodes to be fast, reliable and well-synchronized to a single clock. The key to creating an effective multinode architecture is then establishing quantum links between nodes. While short cryogenic microwave interlinks have been built [23], the challenges of scaling such links between distant refrigerators lead us to consider room temperature links based on M2O transduction. In this section, we review the

progression of remote entanglement distribution experiments done with superconducting qubits, and discuss the use of M2O protocols to link them. In particular, we discuss how the transmon decoherence rate sets a lower bound on the M2O transduction rate, which will be an engineering challenge for MNQCs.

Although various superconducting processor designs are being prototyped [101, 102], the most widely used architecture in both academic and industry settings is a two-dimensional lattice of nearest-neighbor coupled transmons [103, 104] cooled to milli-Kelvin temperatures in dilution refrigerators. Transmons are robust quantum computing building blocks, fortified by recent engineering breakthroughs in high-fidelity two-qubit gates [105–107] and individual qubit coherence, which can approach 0.5 ms [37, 38]. Transmon processors with as many as 433 qubits are available [108] and steady progress is being made towards the goal of fault-tolerant quantum computing using stabilizer codes with 49 qubits [104], 23 qubits [109], and 17 qubits [110, 111].

Remote entanglement distribution experiments with transmon-qubit-based processors connected by cold microwave links have evolved after a period of intensive engineering. Almost a decade ago, a series of heralding-based probabilistic entanglement distribution experiments were performed between qubits on separate chips within a fridge [24, 112, 113]. Then, deterministic entanglement distribution experiments were done between chips separated by 1-5 m of cable, one of which housed the chips in separate fridges [23], and the resulting fidelities were largely limited by cable loss [23, 32–35]. The next generation of deterministic entanglement distribution experiments took care to minimize cable loss so transfer and process infidelities were instead limited by qubit loss [114, 115]. Heralded deterministic entanglement experiments have also been done [116] where sophisticated techniques were developed to mitigate cable loss [36]. Nonetheless, today’s state-of-the-art deterministic entanglement distribution experiments are again limited by cable loss [30, 117]. In all of these experiments, the quantum links were cryogenic and their length was on the order of a few meters, which poses a challenge for scaling to a many node system distributing entanglement across tens or hundreds of meters.

While microwave photon loss poses a significant obstacle to scaling MNQCs, optical photons at the telecommunication wavelengths are promising candidates for mediating information exchange due to the extremely low loss and negligible thermal photon noise of optical fibers at room temperature. For medium or long-distance quantum communication between superconducting chips, it is more promising to transduce quantum information from the microwave regime to optical wavelengths and generate entanglement through heralded schemes. Recently, electro-optomechanical transducers were integrated into transmon qubit systems and used for qubit readout [118–120], but these converters are not yet efficient and broadband enough for use in a remote entanglement distri-

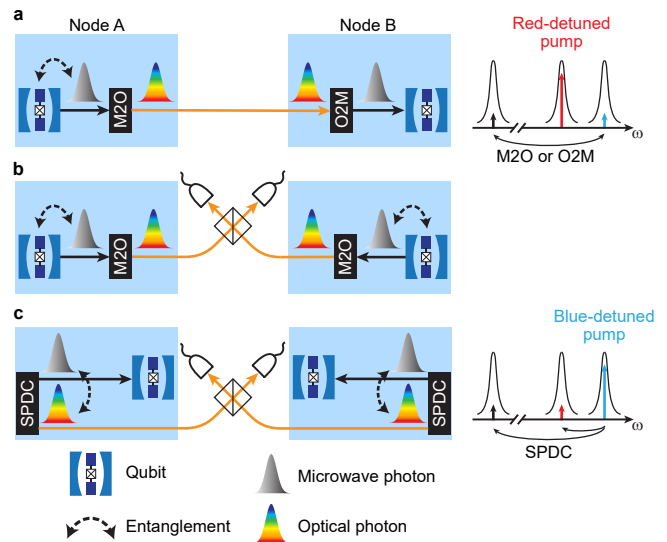


FIG. 2. Schemes for entanglement generation between remote nodes. (a), entanglement between a qubit and a microwave photon is prepared at node A. By applying an M2O and a subsequent O2M converter (see the right panel), the state of the microwave photon can be transferred to the qubit at node B. (b), the direct conversion heralded scheme. Entangled qubit and microwave photon states  $|g0\rangle + |e1\rangle$  are prepared at two nodes. The microwave photons are up-converted to optical photons and interfere at a beamsplitter. The single-photon detector click heralds the generation of entangled qubits  $|ge\rangle \pm |eg\rangle$  at two nodes. (c), the SPDC heralded scheme. Entangled microwave-optical photon pairs are generated at two nodes when an M2O converter is used as a SPDC source by pumping at the blue-detuned resonance frequency (see the right panel). By erasing the which-path information with a beamsplitter, a click from the detectors heralds the generation of microwave entangled Bell state  $|10\rangle \pm |01\rangle$  at two nodes. The microwave photon states are transferred to qubits, which leads to entangled qubit state  $|ge\rangle \pm |eg\rangle$ .

bution experiment. A high-fidelity M2O converter will be an essential component for realizing large-scale distributed superconducting quantum computing.

While an ideal M2O converter should have unity quantum conversion efficiency, in practice the intrinsic weakness of optical nonlinearity poses an extreme challenge for high-efficiency M2O conversion. Various schemes have been proposed and experimentally demonstrated, including cavity electro-optics [57, 121–128], opto-magnonics [129–132], electro-optomechanics [118, 119, 133–141], cold atoms [142–144] and rare-earth ions [145–149]. Reviews of recent experimental advances in M2O conversion can be found in Refs. [39–44]. The conversion efficiency achieved in state-of-the-art experiments, however, remains far less than unity. Despite the relatively high on-chip conversion efficiency, the total efficiency can be significantly lower due to the inevitable fiber-to-chip coupling loss and the optical-pump-rejection filtering loss. In addition, because a high-power optical pump is needed to boost the conversion efficiency, ther-

mal microwave photons generated by the optical-pump-induced heat can be combined with transduced signal in the optical output channel as ‘added noise’. The performances of state-of-the-art M2O converters are summarized in Table V of Appendix B.

In order to generate entanglement between separated refrigerators, one straightforward method is to locally generate entangled qubit-microwave photon pair at one node and subsequently apply an M2O and an O2M converter to deliver the microwave photon to another node as shown in Fig. 2(a). However, this scheme is sensitive to the low M2O conversion efficiency and thus suffers from a low entanglement fidelity. Alternatively, direct M2O conversion could be used in a heralded scheme. Analogous to the optical photon heralded schemes [150–152], the superconducting qubit is first entangled with a microwave photon at each node as  $|g0\rangle + |e1\rangle$ . The microwave photons at both nodes then undergo direct M2O conversion and the optical photons are then routed and detected as shown in Fig. 2(b) (referred to as the direct conversion heralded scheme). The optical photons from both nodes interfere at a beamsplitter, and a click from the optical detector heralds the generation of entangled qubit state  $|ge\rangle \pm |eg\rangle$ . In addition, a remote entanglement generation scheme previously developed for atomic ensembles [150, 153–157] is another option for superconducting platforms [60] to obtain high-fidelity entanglement generation in the presence of low M2O conversion efficiency. As shown in Fig. 2(c), an M2O converter can be pumped at the blue-detuned resonance frequency and thus be used as a spontaneous parametric down conversion (SPDC) source to generate entangled microwave-optical photon pairs (referred to as the SPDC heralded scheme). The optical photons generated at two fridges interfere in a beamsplitter to erase the which-path information, and the click from the optical single-photon detectors heralds the generation of an entangled microwave Bell state  $|01\rangle \pm |10\rangle$  between the two fridges. This scheme has the benefit that the entanglement fidelity is less sensitive to the M2O conversion efficiency, while the entanglement generation rate depends on the conversion efficiency as well as the bandwidth.

Ultimately, the performance of MNQCs made of superconducting qubits and M2O converters will depend strongly on the conversion efficiency and bandwidth of the M2O converters, which is very slow and noisy compared to the local operations. We estimate that the on-chip entangled pair generation rate of the state-of-the-art M2O converters is of the order 1 MHz with infidelity of 0.2 (see Sec. IV). In comparison, transmon two-qubit gate infidelity has already been engineered down to less than 0.002 [107], with two-qubit gate times on the order of 100 ns. Hence we see that the internode operations are the major limitation on MNQC performance, and future MNQCs will need to marshal a range of techniques to surmount the weak internode links.

### III. MULTINODE QUANTUM COMPUTING ARCHITECTURE

A central task in evaluating MNQC performance is to understand the available performance of internode gates. However, the evaluation of internode gates in multinode systems is itself the evaluation of a multipart quantum system: MNQCs using superconducting devices with M2O links will need to compensate for weak internode links using a combination of entanglement generation [46, 56, 57] settings, entanglement distillation [30, 58, 59], and compiler optimization [54, 55, 63–65]. One direct approach might be to simply conduct a simulation of the full system, treating M2O conversion, entanglement distillation, remote gate execution, local operations, and measurement in one large analysis. However, this calculation quickly grows too large even for relatively simple MNQCs. With current density matrix simulations limited to  $\mathcal{O}(20)$  qubits [61, 62], allotting just a few qubits for entanglement distillation, measurement ancillae, and treating M2O conversion with a quantum framework costs approximately 18 qubits per internode link. This quickly limits the system to algorithms performed on a single-digit-number of qubits even with the best simulation algorithms. What is needed is a framework for organizing these components and their interactions into a structure that can be treated quantitatively. This framework must simplify the complex interactions between components into a few quantities that describe the relevant interaction while identifying key tradeoffs in the operation of components.

Classical computing network architecture has long faced similarly complex systems and developed approaches to tackle them. A foundational example is the Open Systems Interconnection (OSI) model which tackles the complex problem of networks and distributed systems by splitting the system into ‘layers’ in a ‘stack’. Each layer has its role in the system, often referred to as the ‘service’ it provides to layers above it in the stack. While the OSI model is foundational, more contemporary classical analogs of MNQCs, including the architectures underlying the ARES and InfiniBand network systems which directly provide network services for multinode computers [73–75], often use a 5 layer network stack comprising a Physical layer which transmits signals, a Link layer which manages packet transmission, a Network layer which provides routing and network management, a Transport layer which is responsible for the reliable transition of data, and Upper (or Application) layers where the users operate (for a detailed discussion of these layers, see [67]). A key concept behind the operation of multinode systems is *transparency* [69]: modern parallel classical platforms seek to offer users a seamless transition between single-node and multinode operations, with the multinode system appearing to the user as a single unified system. The chief service of the network stack is then to manage the execution of internode operations for the compiler in a transparent way.

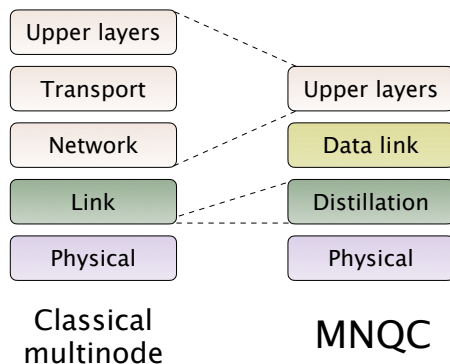


FIG. 3. Comparison of the layer stack for classical multinode architectures [67, 74] and our MNQC architecture. Weak, noisy internode link merits a dedicated Distillation layer, while the low-level access of quantum compilers and high cost of network resources breaks the abstraction between the Upper layers and the Transport and Network layers.

However, there are several key differences between quantum MNQCs and their classical counterparts (see Figure 3). First, quantum internode communication suffers from far higher error rates than those in comparable classical architectures, and MNQCs will need dedicated resources to compensate for an extreme noise environment. Connected to the problem of internode noise is the presence of within-node noise that accumulates with time. Executing operations more slowly is not sufficient to improve performance, as the time of execution is itself a source of noise that will need to be accounted for. Furthermore, the prevalence of noise has led modern quantum compilers to operate at the level of gates or even pulse sequences, a much lower level than their classical counterparts. We expect links to be largely optimized to generate fast and high fidelity entanglement between just two nodes, independent of other links (Figure 4a). Any switching will be costly and is to occur within the local compute nodes, and the compiler is directly issuing commands for two-qubit, two node gates, and any non-nearest-neighbor gates will need to be transpiled directly into nearest-neighbor gates. Finally, the principle of transparency of quantum systems means that the function of the quantum network stack should be to offer internode links to the compiler in the same way as local links, simply with longer gate times, lower fidelities, and probabilistic success.

An efficient method for the execution of remote gates is to use EPs produced from the M2O SPDC process, local operations, and classical internode communication to execute remote gates [158, 159]. However, the low rate and high infidelity of EPs may lead to low fidelity of internode gates. This performance may be improved by using entanglement distillation [160], which consumes raw (not distilled) EPs to produce distilled EPs, which may then be used for remote gates. The function of the network stack is then to produce raw EPs, distill them, and manage the execution of internode gates, offering

internode gates as a resource to the upper layers while abstracting away the details of their execution.

Taken together, raw EPs, distilled EPs, and internode gates form a chain of key resources for remote gates, each produced from the previous. We may use these key resources to construct an MNQC network stack in analogy with the classical approaches by devoting a ‘layer’ of the system to the production of each key resource. At the ‘bottom’ of the stack, M2O SPDC hardware produces raw EPs. In analogy with the classical approach, we call this the ‘Physical layer’. Next, a ‘Distillation layer’ converts raw EPs into distilled EPs at a lower generation rate. Finally, a ‘Data layer’ manages the execution of internode gates, and exposes them as a resource to the compiler. A comparison of this MNQC architecture with that from modern classical interconnection architectures is given in Fig. 3. Similar models for network architectures have been proposed in several pioneering works [89–91, 161]. These papers lay out criteria for quantum networks, and also find that a stack based on classical interconnection architectures, with the added function of distillation, is an effective way to structure the network. While these are general studies, a hardware-focused model has been proposed for networks using NV centers [90]. An excellent general overview of planned progress in quantum interlink technologies may be found in [9]; here we focus on a particular technology (M2O interlinks) and provide detailed studies of MNQC algorithm performance. To our knowledge, this paper is the first to present a detailed hardware-based model of multinode (or networked) architectures using superconducting devices with M2O interconnects.

An important property of the network stack is that each layer interfaces only with the layers above and below it in the stack. For example, all raw M2O EPs are passed to the Distillation layer, and there is no need for the Data layer to interact with Physical M2O generation. Similarly, entanglement distillation is hidden from the compiler. Instead, layers only interact by passing instructions and success flags as indicated in Figure 4. Hence this layer architecture can simplify the conception and operation of MNQCs by reducing the potential complexities of component interactions to a linear layer and stack structure.

For our present purposes, the key property of this MNQC architecture is that it also simplifies the simulation and benchmarking of an MNQC. Because the layers interact only with their neighbors in the stack, we can quantify the operation of each layer in terms of the production quality of key resources. Beginning from the bottom of the stack, the Physical layer works to produce raw M2O EPs, which are quantified in terms of the heralded rate of production and the density matrix of produced EPs. The Distillation layer then is responsible for taking these raw EPs and producing distilled EPs, which are quantified by the minimum time to produce a distilled pair, the density matrix of the produced pair, and the success probability of the operation, each as a

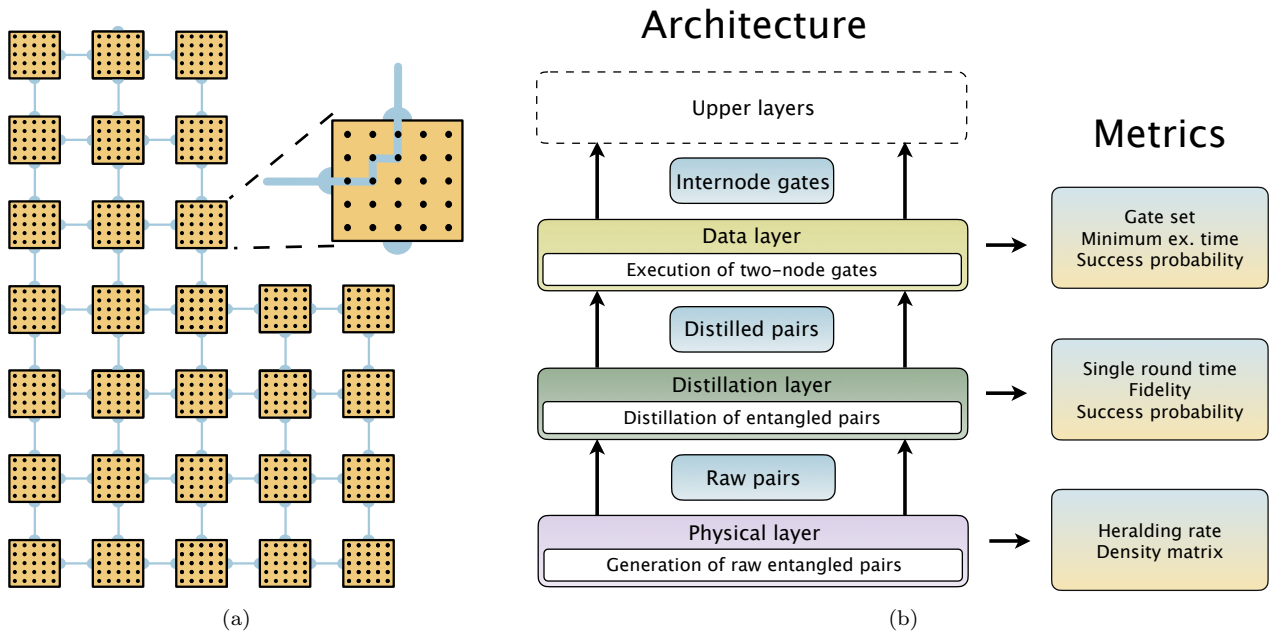


FIG. 4. (*color online*). (a): Schematic depiction of a multinode quantum system. Nodes (orange boxes) are connected with quantum links (blue lines) that allow internode gates between remote qubits (black dots). Internode gates in an MNQC will be slow and noisy, which leads us to focus on two-node, nearest-neighbor internode gates. This breaks the usual abstraction between the network and local hardware components. (b): Schematic description of the layers, functions, interfaces, and key metrics of the MNQC model. Our focus is on the MNQC network stack (the Physical, Distillation, and Data) layers and how their performance affects overall MNQC performance.

function of the number of rounds applied. At the top of the network stack, the Data layer uses distilled EPs to execute internode gates, which are quantified by the set of available gates as well as the minimum time and success probability of each gate. These metrics are denoted in Figure 4b.

Each layer then may be treated quantitatively by simulating the metrics of the resource it produces in terms of the incoming resource. In the next section, we delve into models of each layer in order to examine the available performance profiles and lay out a key tradeoff in joint hardware and software operation.

One significant simplification which we make in this paper is suppressing the probabilistic nature of the entangled pair generation and distillation processes, characterizing each by the average time of the process. This greatly simplifies our models, but the probabilistic processes should be treated fully in future, more detailed studies in order to reveal the complex interplay of real-time application execution.

#### IV. MODELS OF THE MNQC NETWORK LAYERS

The MNQC architecture organizes entangled pair generation, entanglement distillation, and remote gate execution into layers. Furthermore, because the key re-

sources between layers are specified, the operation and performance of each layer can be treated individually and later unified into a whole model. In this section, we examine the available performance of each layer of the MNQC stack using models of expected hardware and software performance. The next section will then unify these layer models into a model of overall MNQC performance.

Beginning from the bottom of the stack, our first task is to estimate the fidelity and generation rate of EPs created using M2O converters in the Physical layer. In the following, we focus on the direct conversion heralded scheme shown in Fig. 2(b), where the qubit-microwave photon pair is initialized in  $\sqrt{0.5}|g0\rangle + \sqrt{0.5}|e1\rangle$  at both nodes. The simulated entanglement infidelity and generation rate for current experimental platforms are shown in Fig. 5(a) (see details in Appendix B). Here, we use the parameter sets of three resonator-based M2O converters (see Appendix B Table VI) to perform the simulation. We assume the intrinsic decay rate of both microwave and optical resonators to be five times lower than the actual experiment values, which we expect to be achievable relatively soon. The best M2O conversion efficiency is obtained at the pump power that reaches a unity cooperativity  $C = 1$  [162], which consequently leads to the highest generation rate and lowest infidelity. For the No. 1 converter (the green curve), the entangled qubit generation rate can approach 1 MHz with an infidelity near 0.2. However, for the No. 2 (the red curve) and No. 3 (the blue

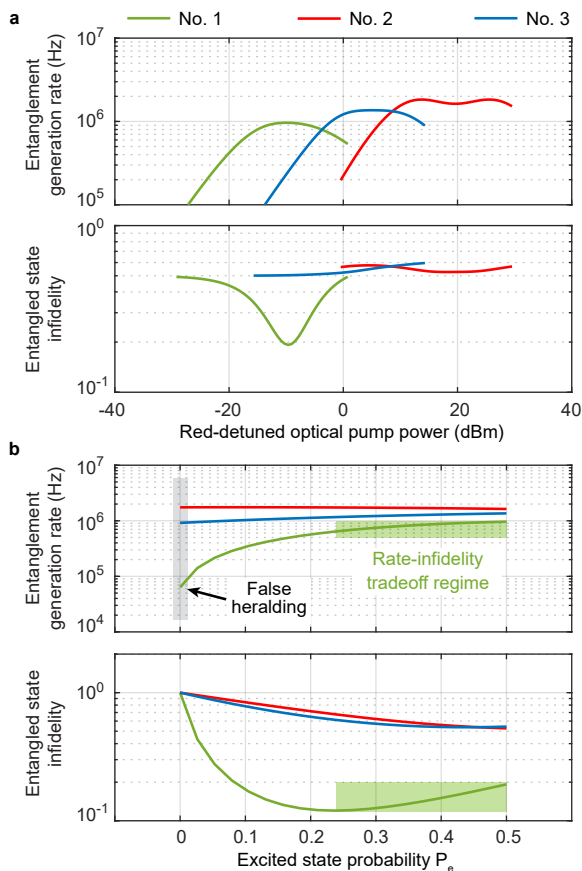


FIG. 5. The estimated Bell state generation rate and infidelity (log scales) for the scheme shown in Fig. 2(b). (a) the performance as a function of pump power with  $P_e$  fixed at 0.5. The simulation is performed using the parameter sets of the No. 1, 2, and 3 converters listed in Appendix B Table VI. We assume the intrinsic decay rate of both microwave and optical resonators to be five times lower than the current actual experiment values, which we expect to be available relatively soon. (b) the performance as a function of  $P_e$  with fixed pump power such that the cooperativity  $C = 1$ . The rate at  $P_e = 0$  is the false heralding rate triggered by thermal noise, and a rate-infidelity tradeoff regime (the green shaded area) can be identified for the No. 1 parameter set.

curve) converters, the infidelity remains  $>0.5$  because a high pump power is needed to achieve  $C = 1$ , and the microwave thermal added noise induced by the high pump power strongly limits the fidelity. We next set the initial qubit-microwave photon as  $\sqrt{1 - P_e} |g0\rangle + \sqrt{P_e} |e1\rangle$  where  $P_e$  is the probability of the excited qubit state and is experimentally tunable [24]. The pump power is fixed such that  $C = 1$ , and the results with  $P_e$  tuned from 0 to 0.5 are shown in Fig. 5(b). The entanglement generation rate at  $P_e = 0$  is thus the false heralding rate, which dominates for No. 2 and No. 3 parameter sets. The tuning of  $P_e$  also reveals a rate-infidelity tradeoff regime, which is highlighted as the green shaded area, where the rate increases but the infidelity also increases with an increasing  $P_e$ . In this regime, a larger  $P_e$  allows more optical

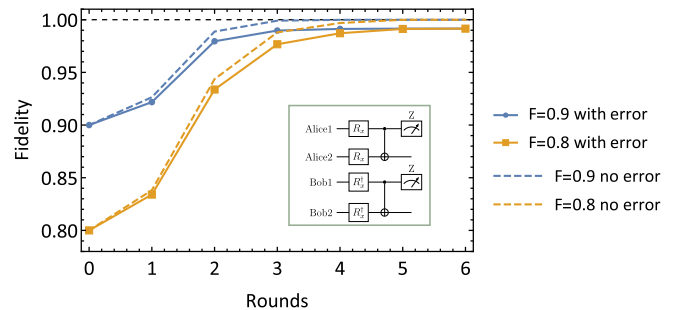


FIG. 6. The performance of entanglement distillation under the DEJMPS protocol [160] with and without imperfections from qubit decay and decoherence and quantum gates. We consider two cases, where the imperfect Bell states with fidelity  $F = 0.9$  (blue) and  $F = 0.8$  (orange) can be generated across superconducting chips. We perform entanglement purification using DEJMPS protocol. We consider all the imperfect Bell states have been generated at the beginning. The local gates and qubit measurements take  $1 \mu\text{s}$  for each purification step. We assume the local gates between superconducting qubits have a depolarizing error with probability 0.0001. The solid lines show the fidelity of the Bell states after  $n$  rounds of purification, while the qubits are suffered from the imperfections. The qubits are assumed to have  $T_1 = T_2 = 1 \text{ ms}$ . The dashed lines shows the corresponding state fidelity without decay and decoherence error. The inset is the quantum circuit for a single round of purification [160].

photons to be generated, but it also increases the error of having two nodes in the excited states simultaneously. In Section V, we will see how the demands of the MNQC must guide the entanglement generation in conjunction with entanglement distillation, to which we now turn.

The distillation layer faces a tradeoff between generation rate and fidelity of EPs as it explicitly consumes raw EPs to produce a smaller number of distilled EPs, thereby exchanging a higher fidelity for a lower generation rate. As detailed in Appendix C, we simulate the output of entanglement distillation using EPs produced from M2O conversion. We set  $T_1 = T_2 = 1 \text{ ms}$  and assume all local gates to take 100ns with a probability of depolarizing errors of .0001. In Fig. 6, we show the fidelity of the Bell state shared between remote superconducting chips after  $n$  rounds of recurrent entanglement purification performed using the DEJMPS protocol [160, 163]. Four or five rounds of entanglement distillation can significantly improve the fidelity of the generated EPs, likely leading to improved internode gate performance. However, the improvement quickly suffers diminishing returns, with further rounds yielding only modest increases. This is a significant problem, as every round decreases the rate of distilled entangled pair generation by a factor of 2, thereby slowing internode gates. This problem is further exacerbated by the presence of decoherence, which degrades partially distilled EPs as they wait for more raw EPs, and Figure 6 shows the performance of entanglement distillation with (solid line) decoherence or (dotted line) an ideal memory that prevents decoherence. How-



ever, it remains to see what this fidelity improvement can do at the level of internode gates.

In order to translate the output of the Physical and Distillation layers into internode gates, we develop a model of the Data layer, which uses distilled EPs to execute remote internode gates. Since a CX gate provides computationally complete communication between nodes, we focus on the case of only internode CX gates. Gate teleportation of the CX gate can be accomplished via the consumption of one raw EP, two measurements, and two local CX gates [159]. Using our simulations of M2O conversion, we numerically calculate the production time and density matrices of the raw EPs from the M2O process. These outputs are fed into the next distillation layer to generate high-fidelity, purified EPs. The time for each round of distillation, the success probability, and the consequent density matrices of the purified EPs are used in the following data layer to simulate the performance of a single internode gate.

## V. FULL MNQC ANALYSIS

We now have models of each layer in the MNQC network stack, from the Physical layer with M2O generation models to the Data layer which manages the execution of internode gates. While understanding the available performance and tradeoffs of each of these layers is key to understanding MNQC performance, models of individual layers cannot tell us how the performances of each layer affects overall MNQC performance, how the layers together offer internode gates, or how to navigate tradeoffs that affect multiple layers. Most importantly, they cannot tell us the performance of internode gates, what algorithms require, or how to exchange internode gates with local computation or circuit cutting gates.

We need an overall model of how the layers in an MNQC interact to produce total system performance. In this section, we unite the models of the previous section into a simulation pipeline that models the full MNQC stack, which allows us to perform three quantitative studies of the system. First, we introduce a ‘Gate-Algorithm Performance’ (GAP) model which uses the output of the unified model to map out the available internode gate performance in terms of hardware models and compare this to the demands of algorithms. In doing so, we will see the effect of the tradeoffs in average internode gate execution time and fidelity and shows how to navigate them for a small MNQC. Next, we use the unified model output to develop a Quantum Roofline model (Q-Roofline) to show how the compiler can navigate the balance of internode and local computation at scale and identify the effects of hardware and software tradeoffs on internode communication bandwidth. Finally, we compare quantum links with error mitigation to classical circuit cutting links using a Quantum-Classical Performance Analysis (QCPA) to determine at what cost can internode links be exchanged for circuitry cutting links.

## A. Unification of Layers into an Overall MNQC Model

The unified model should allow us to quantify overall MNQC performance as a function of the performance of each layer. More specifically, it should accept as inputs hardware and software details of each layer, including the M2O drive strength and Hamiltonian, the entanglement distillation protocol, local operation fidelities and times, qubit  $T_1$  and  $T_2$ , a compiler, and the quantum application to be executed. To characterize the quantum network stack, it should return metrics of the key resource offered by the Data layer to the upper layers, namely the fidelity and average execution time of internode gates. Furthermore, to allow us to determine the needs of algorithms, the unified model should then allow us to study the behavior of the Application and Compilation layers, including performing a full density matrix simulation of algorithms running on the small systems, and providing data on compilation results and estimates of performance on large systems.

The MNQC layer architecture that we have used in the previous sections to organize our models plays a key role in enabling the unified model. As noted in Section III, an MNQC is a complex system, including M2O generation, entanglement distillation, and internode gate distillation in addition to the normal functioning of a quantum computer. Treating the total system at once would quickly outgrow available simulation capabilities. Because we have independent models of each layer, we can link the output of one layer to the input of the next. Roughly speaking, if each layer involves a Hilbert space of size  $N_i$ , then we must treat a series of  $N_i \times N_i$  density matrices. This should be compared to modeling the whole system with a density matrix of size  $\sum_i N_i \times \sum_i N_i$ . In practice, the layer architecture offers even further simplification by allowing us to use heuristic simplifications at the layer interfaces, for example by reducing error channels to the depolarizing channel or abstracting away probabilistically successful processes into an average execution time.

Let us unify the network stack first, which we can then connect with the Application and Compiler layers. At the top of the network stack, the Data layer supplies internode gates as a key resource to the Compiler and Application layers; our task is thus to quantify the fidelity and execution time of available gates as a function of the outputs of Distillation and Physical layers lower in the stack. Using our simulations of M2O conversion, we numerically calculate the production time and density matrices of the raw EPs from the M2O process. These outputs are fed into the next distillation layer to generate high-fidelity, purified EPs. The time for each round of distillation, the success probability, and the consequent density matrices of the purified EPs are then used in the Data layer simulation to evaluate the performance of a single internode gate.

Now we connect the MNQC network stack simulation

with the Application and Compiler layers to create an overall simulation of the MNQC. Naïvely, joining the Compiler and Application layers would involve a complex control issue. The compiler could be responsible for managing entanglement distillation and internode gate execution, each of which require multiple measurements, operations, and classical communication, in addition to its function managing local operations. However, the abstractions provided by the MNQC network stack, in particular the principle of transparency articulated in Section III, greatly simplify this task. To the compiler, an internode gate is presented in the same way as a local gate, albeit with a longer average execution time and lower fidelity. Transparency thus greatly simplifies the construction, as compilers designed for monolithic systems may be used at the top of the MNQC stack, though they may not be optimal.

We can construct the full simulation pipeline beginning from the M2O Physical layer simulation, Distillation, and Data layer simulations that we discussed in the previous section and unified in the gate model above to upper layers. Figure 7 shows an overview of this simulation pipeline with metrics for each interface. Once the average execution time and fidelity of the internode gate are simulated as in Section IV, they are used to evaluate the performance of the upper layers, which includes the compiler layer and the application layer, leading to the full pipeline simulation shown in Figure 7.

## B. Gate-Algorithm Performance Models

Our first task is to determine how to navigate the tradeoffs in the Physical and Distillation layers identified in Section IV. Both of these tradeoffs involve an exchange between the time to create EPs and the infidelity of those EPs. However, they are not independent, as they operate on distinct layers using dependent key resources: the Physical layer can lower M2O pump power setting to decrease the infidelity of generated (raw) EPs at the cost of higher average generation time, while the Distillation layer then uses those raw EPs to create purified EPs, and again may decrease the infidelity of EPs at the cost of slower purified pair production by increasing the number of distillation rounds. To compare these two, we must find a common resource at which to evaluate the performance profile due to their combined effect, and then we must compare this to the demands of algorithms to guide a choice of performance. The unified model achieves this by determining the available internode gate performance produced by the network stack as a function of the operation of the Physical layers and then allowing us to evaluate benchmark algorithms executed on the MNQC using those internode gates.

Let us begin by determining the performance of the internode gates offered by the MNQC network stack. Using the unified model pipeline in Figure 7, we can link the models of the previous section in order to evaluate in-

ternode gate performance as a function of the Physical, Distillation, and Data layers in the MNQC stack. Figure 8 shows the available infidelity and average generation time of internode gates using raw EPs and distilled EPs. The black curves indicate the average execution time and infidelity of internode gates executed using EPs from the M2O conversion process or with successive rounds of entanglement distillation. At the upper left corner of each curve, operating M2O conversion with a high excitation probability  $P_e$  creates a higher EP generation rate and thus low execution time, at the cost of higher infidelity. As  $P_e$  is decreased, the infidelity decreases but the average execution time increases. This is precisely the trade-off at the Physical layer identified in Section IV and is reflected in the negative slope of the M2O curve.

Given a particular M2O excitation probability setting, which corresponds to a point along the black curve, the Distillation layer then navigates a similar tradeoff: relative to raw M2O EPs, distilled EPs will have a longer average production time but higher fidelity. At the Data layer, this translates to a longer average execution time but higher fidelity of internode gates. Each round decreases the infidelity, at the cost of increasing the average internode gate time. The interaction of the colored noise affecting the raw entangled pairs as a function of drive power with entanglement distillation leads to complex behavior of the resulting performance. For the fastest gates, no entanglement distillation should be used. For gates with lower infidelity, distillation should be used. The needs of algorithms will then dictate how the internode gate should be executed: given a targeted performance, the number of rounds as well as the excitation probability  $P_e$  shape the infidelity and link gate time achievable. If a compiler is able to select from a range of available internode gate times, then the MNQC stack must adjust the excitation probability of M2O generation dynamically to generate the highest fidelity gates for each internode gate time. For example, to achieve gates with lowest infidelity and an internode gate time below  $.01T_1$ , one should use the minimal  $P_e \approx .25$  and two rounds of distillation. On the other hand, to achieve the lowest infidelity possible at all, one should use  $P_e \approx .35$  and four rounds of distillation. Furthermore, depending on the desired execution time, the compiler may wish to select fewer rounds of distillation with a higher  $P_e$ , or vice versa. This is particularly important as the infidelity is as much as 2x worse when using the incorrect configuration.

Next we turn to see how this internode gate performance affects the performance of algorithms on an MNQC. Beginning with gate performance curves like that of Figure 8, simplified by including only the red bounding curve and removing the unfavorable region, we overlay on them the conditions for successful execution of a successful benchmark to create a ‘Gate-Algorithm Performance’ (GAP) plot. As before, we set  $T_1 = T_2 = 1\text{ms}$  and assume all local gates to take 100ns with a probability of depolarizing errors of .0001. The basis gates for these systems comprise the same basis gates as IBM-

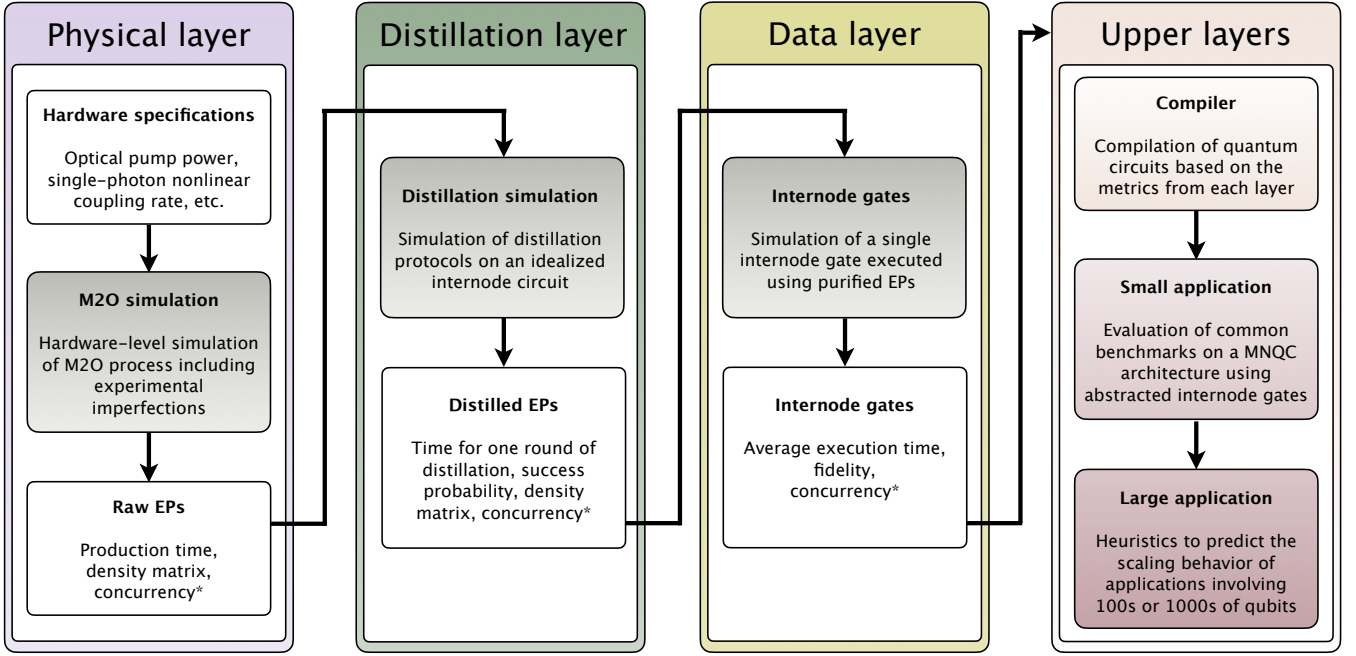


FIG. 7. Models of each layer of the MNQC model and the metrics passed between them. Each layer is simulated as detailed in Section IV, and the result is united to create an overall simulation of the MNQC. The concurrency metrics marked with an asterisk are set to unity here but can be used in a future generalized model that includes multiplexing of internode links for boosting the remote gate execution rate.

Quantum, and each algorithm is transpiled accordingly.

As a first example, we evaluate the effective Quantum Volume (QV) [164]. The QV is a measure of the size of the effective Hilbert space traversed by a quantum system before decoherence occurs. With a perfect internode link, the QV would be  $2^{10}$  (Fig 9); with no internode link it would be  $2^5$ . Hence this benchmark allows us to quantify the degree to which the multinode system outperforms any one of its nodes. To gauge the performance implications of performing distributed quantum computing, we perform a noisy simulation for each algorithm over this architecture, with the inter-node link having the respective gate time and fidelity attained from internode gate simulation.

The results of the QV benchmark are shown on a ‘Gate-Algorithm Performance’ (GAP plot) in Figure 10, which allows us to compare the available performance of gates produced by the MNQC network stack with the demands of algorithms we wish to execute. Times and fidelities that lead to successful completion of a QV circuit are denoted by shaded ‘success’ regions. Beginning from the unshaded region, lowering the infidelity and the gate average execution time allows for the successful execution of larger and larger QV circuits. Both parameters are key because while the infidelity of internode links directly causes noise, the long execution times allow errors to accumulate within the nodes.

On top of the shaded success regions, we overlay the available gate performances in a similar manner as in

Figure 8. The black line depicts gates executed using raw MZO generation, while the red lines denote internode gates using entanglement distillation. We can quickly see that the achievable performance is much slower and noisier than needed for QV circuits. Indeed, the rate and infidelity will require significant improvement for the MNQC to be able to achieve a QV that improves on the single node performance at all, and orders of magnitude improvement to achieve the maximum possible QV of  $2^{10}$ .

While QV gives a single number benchmark randomized over circuits built from all possible two-qubit gates, we also create a GAP plot for a benchmark battery [165, 166] to understand performance of the distillation model for specific algorithms. Our battery of tests is composed of: a Quantum Fourier Transform (QFT) benchmark, an ADDER benchmark, the Bernstein-Vazirani (BV) benchmark, and GHZ state distribution, in order of decreasing demands on the internode link. Again we see the need for faster and higher fidelity internode gates. However, using entanglement distillation, the GHZ and BV benchmarks can be achieved with high fidelity. Hence we see that even though entanglement distillation increases the internode gate time, its use is critical for enabling MNQCs to execute algorithms effectively.

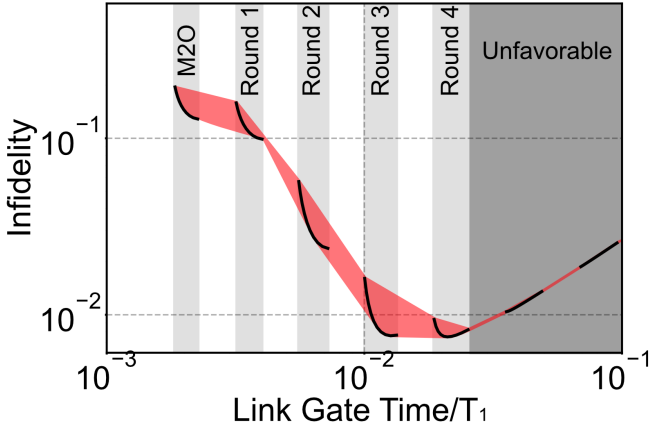


FIG. 8. Performance profiles of internode gates using (black line) only raw M2O entangled pairs or (red line) distilled pairs. Raw M2O pairs on the black line may be tuned for higher rate by increasing the excited state probability  $P_e$ . Successive rounds of distillation are indicated by the markers on the shaded regions. Using only M2O generation leads to the fastest internode gates at the cost of a high infidelity, while distillation reduces the infidelity at the cost of increasing the internode gate time. A limited number of rounds of distillation can be performed before internode errors degrade the qubits.

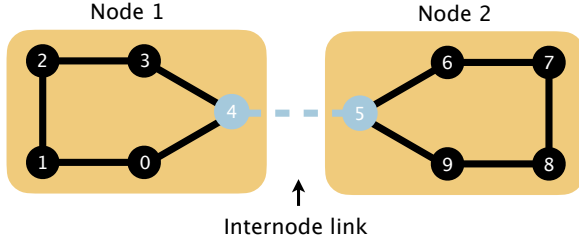


FIG. 9. Topology of a small MNQC that we simulate explicitly. The system consists of two five-qubit nodes with a single internode link, with entanglement generation, distillation, and remote gate execution abstracted into the link.

### C. Quantum Roofline Model

Because MNQCs are employed to create large quantum systems, we must be able to understand the scaling behavior of large systems in order to identify and navigate tradeoffs and performance bottlenecks. While the GAP models of the previous section gave us a manner to navigate tradeoffs in the MNQC network stack and determine performance requirements for small systems, they cannot scale to large systems as they require density matrix simulations.

In this section, we introduce a Quantum Roofline (Q-Roofline) model, based on the classical roofline model [79], which analyzes the scaling behavior of large systems. The Q-Roofline model allows us to determine whether quantum algorithms are bound by internode or local performance. It can then evaluate compiler performance by

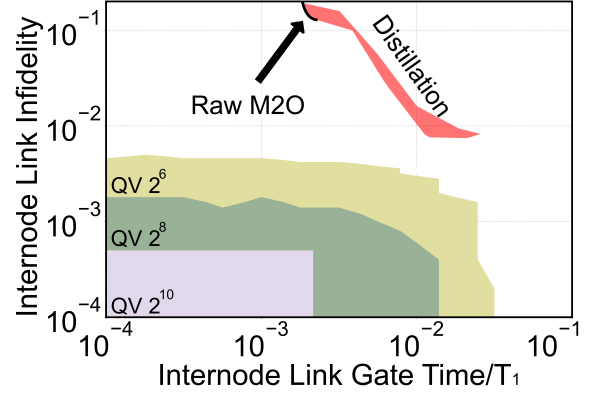


FIG. 10. Gate-algorithm performance plot of Quantum Volume evaluated on a 10-qubit system comprised of 2 ring QPUs connected with M2O and Entanglement Distillation simulation. Each distillation curve, denoted by the red region, has been truncated at the number of nested rounds at which its performance begins to degrade. The QV with no internode link is  $2^5$ . With presently available technology, including distillation, the QV of the MNQC does not reach  $2^6$ .

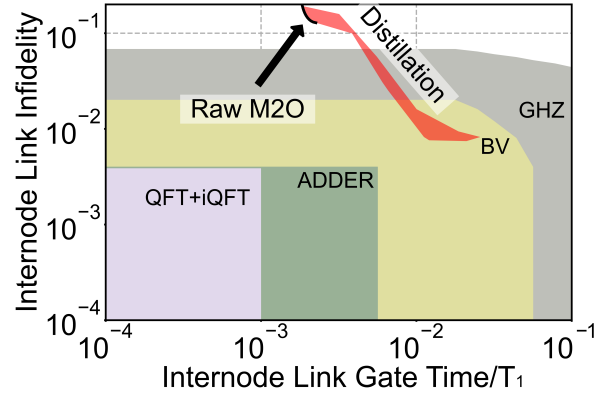


FIG. 11. Gate-algorithm performance plot of several benchmarks evaluated on a 10-qubit system comprised of 2 ring QPUs connected with M2O and Entanglement Distillation simulation. Shaded regions indicate performance of  $>90\%$  for the respective algorithm. Each distillation curve has been truncated at the number of nested rounds at which its performance begins to degrade.

determining whether the compiler has sufficiently balanced internode and local operations. The Q-Roofline model aims at modeling steady state behavior (e.g., averaging over the entire application) rather than instantaneous manner. However, when an application contains significantly distinct phases, one may draw particular Q-Roofline figures for each individual phase.

As a first example, we can use the Q-Roofline model to determine whether applications running on an MNQC are bottlenecked by internode or local performance. For a compiled circuit, we define the Computation-to-Communication Ratio (CCR) as the ratio of the number

of local gates of the algorithm versus remote internode gates over the entire circuit. On the other hand, given a quantum system, we define the machine CCR (MCCR) as the ratio of the rate of execution of local gates to the rate of execution of internode gates. Efficient compilation then seeks to match the balance of internode and local gates in the compiled circuit to that available to the machine, i.e. to match the CCR and MCCR, so as to maximize overall gate throughput while minimize circuit duration for the distributed circuit. We also define the gate density [165] as the occupancy of gates slots along the time evolution steps of a circuit (i.e., liveness defined in [166]), which provides an upper bound of performance when all remote gates become local. As an initial study on bound analysis, we assume the execution of computation and communication gates can be fully overlapped through the transpiler or runtime scheduler.

Figure 12 shows the Q-Roofline analysis of the application benchmarks from the previous section on the physical architecture in Figure 9. The vertical axis shows the rate of single-qubit gate execution. We take the time unit to be the average gate time. Thus, for this 10-qubit system, the computation performance upper-bound is 10 gates/time. We can draw an horizontal line to set the computation performance bound.

The horizontal axis of Figure 12 denotes the CCR of a circuit. Since there is only one inter-module link (Figure 9), given the duration of the remote gate is  $1.041E^{-6}$ s as shown in Figure 11, the internode gate duration is then 10.4 times that of a local gate (i.e., 100ns [167] as used in Section VB) and so the MCCR is 10.4. Using this (MCCR=10.4 and 10 gates/time) coordinate, we can locate a point  $\pi$  in Figure 12a. From that point, drawing a 45 degree line (following the definition of CCR and MCCR), we can obtain the communication performance bound for the targeted MNQC system.

Using these two bounds, we can understand whether internode or local performance bounds the application. The Roofline shape, showcasing the performance bounds, is purely dictated by the quantum hardware. The ridge point  $\pi$  defines the machine’s balance point [168]: if the compiled application’s post-transpilation CCR is less than  $\pi$ , it is communication bound in this machine; otherwise, it is computation bound. To see the exact bound, a vertical line can be drawn from the application’s CCR on the horizontal axis; the point it hits on the Roofline shape implies the performance bound.

In particular, let us evaluate the four benchmarks (i.e., BV, GHZ, ADDER and QFT). The GHZ benchmark shows the least demand of communication or the highest CCR, while QFT incorporates frequent entanglement operations through the inter-module link, showing the smallest CCR. The Adder and BV benchmarks display intermediate CCR. This is consistent with the difficulty of each benchmark to reach in Figure 11. In Figure 12a, all four benchmarks are communication bound given their CCRs in Table I and the settings of the system. However, none of them can hit the bounds due to their

| Algorithm | Qubits | Depth | 1q gate | 2q gate | Comm | CCR   | Density |
|-----------|--------|-------|---------|---------|------|-------|---------|
| GHZ       | 10     | 13    | 3       | 8       | 1    | 9.5   | 0.162   |
| BV        | 10     | 26    | 57      | 24      | 7    | 7.5   | 0.458   |
| QFT       | 10     | 633   | 323     | 439     | 164  | 3.662 | 0.242   |
| ADDER     | 10     | 219   | 101     | 177     | 55   | 4.136 | 0.258   |

TABLE I. Statistics of mapping four 10-qubit algorithm circuits to the small MNQC in Figure 9 using Qiskit (Version 0.33.0) transpiler. Depth refers to circuit depth post-transpilation. 1q gate and 2q gate refer to post-transpilation 1-qubit and 2-qubit gates. Comm refers to the number of internode link gates. Density refers to gate density.

poor gate density. Using QFT as an example, the CCR of QFT is nearly 3.7, but the gate density is merely 0.242, which means the low utilization of the local gates slots (due to application’s logic structure, transpiler behavior, and cost of intranode routing, etc.) limits its ability to even fully utilize the inter-module link, i.e., hit the communication bound. With a density of 0.242, in the best case, the computation performance is 2.42 gates/time, below the communication bound. The same conditions apply to the other three circuits. Therefore, in addition to the machine bound, one should also consider the circuit features such as gate density. Figure 12b shows a different scenario: let’s say we want to enhance the internode link fidelity from 0.9 to 0.99 through two rounds of distillation (see Figure 11). After the first round, the communication performance halves (MCCR=20.8) and we obtain the red slash by shifting right for a unit. Hence both QFT and ADDER are predicted to be communication bound despite their low gate density. Furthermore, through two rounds of distillation, the machine’s communication performance quarters (MCCR=31.2), and we obtain the green slash. Now, except for GHZ, the other three benchmarks QFT, ADDER, BV all become communication bound, with a delivery performance smaller than 2.42, 2.58 and 4.58 gates/time, respectively.

We can also see how the internode fidelity vs. execution time tradeoff that we have investigated affects the scaling performance of applications. From Figure 11, when the internode link gate time is  $1.041 \times 10^{-6}$ s, the link fidelity is about 0.805 with raw M2O. This results in an overall circuit execution fidelity of 0.9. With two rounds of distillation, the fidelity increases from 0.805 to 0.842 to 0.944 with the overhead of  $3 \times$  communication latency. This shifts the sloped line right by two units, as shown in Figure 12b. Note that each round of distillation doubles the communication latency, and both axes are in 2-log scale.

In particular, in the NISQ era, most fidelity enhancement techniques lead to certain performance degradation with overhead, as shown in Figure 13. Nevertheless, the Q-Roofline model shows how the compiler can play a key role in reaching the best scenario for an application circuit by matching the machine’s balance point. For example, when the application is communication bound, the compiler can increase the CCR to reach the balance

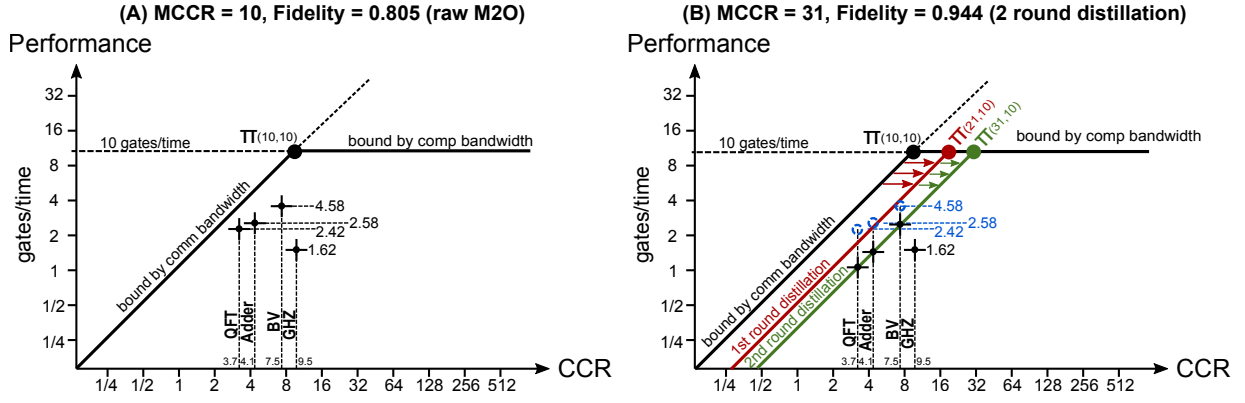


FIG. 12. Performance bound analysis for QFT, Adder, BV and GHZ on the 10-qubit MNQR system through Q-Roofline model.

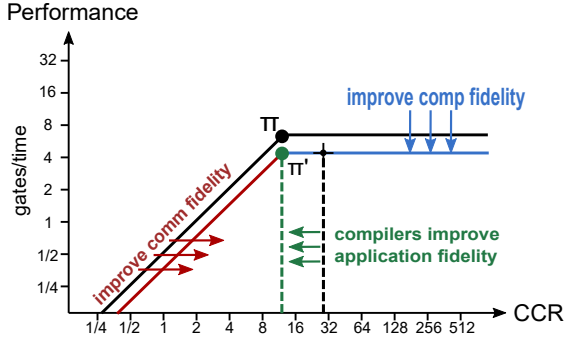


FIG. 13. The Q-Roofline model also shows how the tradeoffs in fidelity and internode gate execution time affect performance bottlenecks.

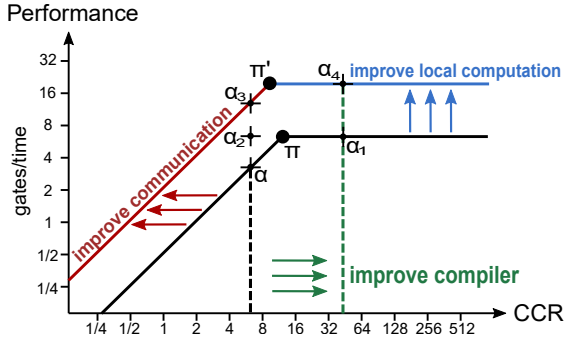


FIG. 14. Improvements to the local compute nodes and the communication operations between them increases the area beneath the hardware bounds in the Q-Roofline model. Within the software layer, optimizations performed by the compiler to minimize communication and maximize parallelism will move an application  $\alpha$  up and right.

point. On the other hand, when the application is computation bound, it can trade-off performance for fidelity (e.g., through distillation, error-mitigation, etc.) until again reaching the balance point.

Lastly, we may also use the Q-Roofline model to predict the effect of improvements to each layer on the scaling behavior of applications. Figure 14 illustrates how

technology advancement of local performance, internode operations (i.e. the MNQC network stack), and compilers would impact an application's performance scaling. As shown in the figure, (i) enhanced internode operations will shift the sloped communication bound of the Q-Roofline to the left, making it less likely that applications will be communication bound; (ii) improved quantum processors will lift the local computation bound up, leading to better system performance; (iii) better quantum compilers which minimize the number of communication operations between processors will contribute to larger CCRs, moving an application to the right along the  $x$ -axis and decreasing the chances of being communication bound. If an application is computation bound but has not saturated the device's local computation bandwidth, then a compiler which increases the parallelization of the program's instructions will increase the gate throughput and move the application upwards along the  $y$ -axis.

For example, through the performance scaling of local quantum devices and quantum interconnects, the machine's balance point  $\pi$  moves towards the upper-left to  $\pi'$ . Meanwhile, if an application is bound by communication at  $\alpha$ , (i) with only compiler improvement, the larger CCR renders the application from communication bound to computation bound, with a higher performance ( $\alpha_1$ ); (ii) with only communication improvement, the communication bound is lifted and performance improves to  $\alpha_2$ ; (iii) with both computation and communication improvement, the performance further improves to  $\alpha_3$ ; (iv) with all computation, communication and compiler improvement, the performance can arrive at  $\alpha_4$ . For quantum programs of a sufficiently large size, the compilation problem may become intractable and therefore the reported gate density and computation-to-communication ratio will be lower bounds on the true, optimal values.

Overall, the Q-Roofline model provides a way to conceptually and quantitatively balance the internode link performance with local performance. It allows us to identify and navigate bottlenecks by trading off internode and local computation intensity (i.e. adjusting the CCR) so that applications are balanced for the machines they are

executed on. For hardware designers, this information is useful for deciding whether it is most beneficial to increase compute or communication bandwidth or fidelity. On the software side, the location of an application with respect to both bandwidth bounds will inform the compiler whether it is better to focus on minimizing the number of remote operations to increase the CCR and unblock the application from the communication bound, or focus on maximizing parallelism.

#### D. Error Mitigation and Circuit Cutting

We have quantified the performance of MNQCs as a function of the internode gate time and fidelity, shown how to navigate the tradeoff between these two quantities, and examined the role that the Compiler and Application layers have in minimizing the use of the internode link. However, we have also seen the dramatic limitations of near-term MNQCs, whose performance only modestly exceeds that of a single node. Given this, we must determine whether the quantum link is worth building at all, or, more precisely, whether a quantum link can outperform a purely classical link.

For purely classical links, we could use classical circuit-knitting techniques [48, 49, 169] which execute circuits separately on the individual nodes many times to replicate a quantum link. On the quantum side, the use of multiple circuit executions allows us to consider error mitigation techniques. Here we compare the number of executions required for error mitigation to those required for circuit knitting in order to quantify the relative performance of quantum links and classical links. The key to achieving this is to combine the MNQC network simulations of internode gate execution time and fidelity from Section VB with models of error mitigation [170, 171] and circuit cutting [50–53].

For both error mitigation and circuit knitting, the number of circuits required scales exponentially with the number of circuit uses, i.e. as  $O(\gamma^k)$ , where  $k$  is the number of gates across the link and  $\gamma$  depends on which method we use and the underlying hardware performance. In the case of error mitigation, more executions are required to mitigate the loss in fidelity from the quantum link. In particular, for probabilistic error cancellation (PEC) [170, 171] the value of  $\gamma$  per gate is<sup>1</sup>

$$\gamma_{\text{PEC}}(d, F_p) = \left( \frac{d^2 F_p - 1}{d^2 - 1} \right)^{-4(d^2 - 1)/d^2}, \quad (1)$$

where  $d$  is the dimension of the gate ( $d = 4$  for a two-qubit gate) and  $F_p$  is the process fidelity. For an internode gate of fidelity  $F_{\text{LL}}$  and gate time  $T_{\text{LL}}$  the total

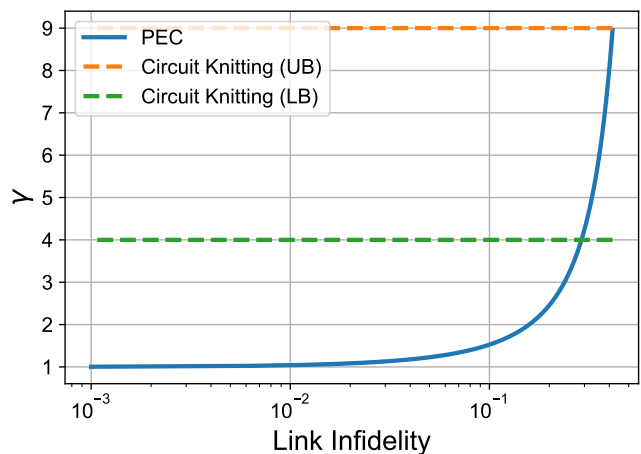


FIG. 15. Comparing the scaling of different methods to link circuit subsystems together where the number of circuits requires scales as  $O(\gamma^k)$  where  $k$  is the number of CX gates between the subsystems. For circuit knitting we give the upper-bound (UB) and lower-bound (LB), see Ref. [53]. Circuit knitting methods require no actual connection whereas PEC is a method to mitigate a lossy quantum link between the sections. A quantum link is almost always superior.

error due to the internode gate, including both the error of the operation and the (intranode) noise accumulated during the long internode gate execution time, is  $\gamma_{\text{PEC}}(4, F_{\text{LL}}) \gamma_{\text{PEC}}^{N_q}(2, e^{-T_{\text{LL}}/T_*})$ , where  $T_* = T_1 T_2 / (T_1 + T_2)$  is the effective fidelity lifetime of a qubit. In the case of circuit cutting or knitting [50–53], there is no quantum link, but one can emulate the  $2n$ -qubit system by running more circuits on the smaller devices and combining the results classically. In Ref. [53] it is shown that  $\gamma = 9$ , and that this can be reduced to  $\gamma = 4$  with local operations and classical communication. Since  $\gamma$  for circuit knitting is independent of the link fidelity, there is a crossover regime in which the circuit knitting procedures require less overhead.

The contrast between the procedures is summarized in Fig. 15. Despite the relatively poor performance of the internode link, it still develops a significant advantage over a purely classical link for link infidelity  $\lesssim .5$ . In the previous subsection, we found that the two-node infidelity is better than this in almost all cases when using M2O and entanglement distillation. Hence the quantum link is advantageous despite the noise and slow gate times. Moreover, this advantage is key when scaling the systems. For example, if we use  $k = 20$  internode gates during an algorithm, then the circuit cutting requires between  $10^{12}$  and  $10^{19}$  circuits while a quantum link with an infidelity of 10% requires only  $10^4$  circuits. A classical algorithm that scales as  $2^n$  needs about  $10^{18}$  steps. For example, as shown in the QCPA in Figure 16, the 10-qubit QFT circuit simulated for the benchmarks required 128 gates across the link, which for PEC at 2.5% infidelity of the link requires about  $10^6$  circuits to mit-

<sup>1</sup> This is  $\gamma^2$  from [171]

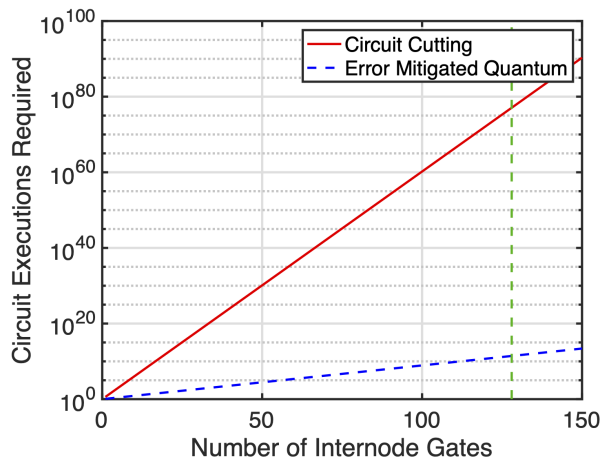


FIG. 16. Quantum-Classical Performance Analysis comparing the number of gates required for circuit cutting (red) and PEC (dotted blue) as a function of the number of internode gates. For the QFT benchmark, with 128 internode gates (green), circuit cutting is clearly infeasible.

igate while for circuit knitting would require a clearly infeasible  $10^{77}$  circuits.

While PEC is advantageous in many cases, it can be at a disadvantage if the internode gate is very long since then the  $\gamma_{\text{PEC}}$  increases due to the infidelity due to decoherence on all the other qubits. This effect is considerable if internode execution time  $T_{\text{link}}$  is on the order of  $T_*/N_q$ . Managing this effect will require balancing the number of qubits  $N_q$  in use, which sets  $\gamma$ , with the number of uses  $k$  of the internode link. Hence compilers that can maintain a high CCR are critical for maintaining the advantage of quantum links.

We have quantified the performance of each layer of the MNQC stack, showed how to navigate tradeoffs in the network layers, developed a model for evaluating Compiler and Application layer performance, and compared classical and circuit knitting approaches. In particular, we found that in order to navigate the tradeoff between the time and fidelity of internode operations at Physical and Distillation layers, the Physical layer should use initial states which depend on the number of rounds of distillation to be performed, which in turn require inputs from algorithm requirements. Furthermore, optimizing MNQC performance requires balancing local computation with internode computation which can be quantified using the Q-Roofline model. Moreover, we also saw the advantage of quantum links over classical circuit knitting approaches, which demonstrated that even with modest fidelity quantum links can reduce the number of circuits needed by many orders of magnitude over classical approaches.

## VI. TOWARDS A DISTRIBUTED QUANTUM COMPUTER: RESEARCH TARGETS

In the previous section, we saw that although quantum links outperform their classical counterparts, MNQCs will need considerable improvement to become viable models for scaling quantum computers. Developing MNQCs that can outperform any of their nodes and execute algorithms of practical importance will require improvements in each layer of the MNQC stack. In this section, we propose research directions that can deliver improved performance at each layer and illustrate how these improvements combine to improve MNQC performance in terms of the GAP, Q-Roofline, and QCPA models of the previous section.

### A. Physical Layer Improvements: M2O Conversion and Multiplexing

Improving internode gate performance is a key target for enabling performant MNQCs. The analysis of section V shows that MNQC performance is significantly bottlenecked by the low fidelity and generation rate of EPs, which lead to gate times and infidelities 10-1000x worse than what we expect from local gates. However, achieving improvements of this magnitude will require significant progress in current technology, or entirely new paradigms all together. Here we briefly describe three potential improvements: iterating on current M2O approaches, developing robust multiplexing, and, in the long term, developing a high-fidelity coupling between superconducting qubits and trapped ions. The speculative effects of these improvements are summarized in Table II.

Considerable progress may be made in the continued development of M2O devices. Metal reflectors [172, 173] and spot size converters [174, 175] have been experimentally demonstrated to minimize the insertion photon loss of grating couplers and edge couplers respectively, which can be applied to on-chip M2O converters to reduce fiber-to-chip coupling loss. Enhancement of the single-photon interaction rate is also critical, which requires further material and device optimization such as the minimization of mode volume [176, 177] and the improvement of optical and microwave resonator quality factors [128]. In addition, thermal added noise induced by optical pump heating needs to be well suppressed to reduce the conversion infidelity. Possible heat dissipation methods to be investigated include radiative cooling [178, 179], the use of superfluid helium for cooling [180], and the use of epitaxially grown superconducting materials [181, 182]. The bandwidth of the converters can be increased by operating the resonators in the overcoupled regime. Waveguide-based converters rather than resonator-based converters also present a potential route to broadband conversion. As a target for development, we present the performance of a hypothetical M2O converter (see Appendix B Ta-



|  | GAP | Q-Roofline | QCPA |
|--|-----|------------|------|
| <p><b>M2O Improvements:</b><br/>Iterated improvements to M2O devices and protocols can yield 10x higher rate and 10x lower infidelity. 5x or more</p>                |     |            |      |
| <p><b>M2O Multiplexing:</b><br/>Frequency and spatial multiplexing can increase effective EP generation rate 100x.</p>   |     |            |      |
| <p><b>Hi-fidelity Ion M2O:</b><br/>Ion-superconducting qubit coupling could increase rate 1000x using buffering and reduce infidelity 1000x using ion-ion links.</p> |     |            |      |

TABLE II. The effects of three classes of improvements to the physical layer, as demonstrated in the GAP, Q-Roofline, and QCPA analyses.

ble VI in Appendix B) as the purple curve in Fig. 17. Such a M2O converter can be used to achieve  $>1$  MHz production rate with an infidelity as low as  $\sim 0.05$ , which might be available in the future if the bandwidth, photon loss, and the single-photon nonlinear coupling rate of existing converters can be improved by one to two orders of magnitude. In the first row of Table II, we simulate these improvements using the method of the previous section; The lower time and infidelity of internode communication allow the ADDER benchmark to be executed on small architectures, while the balance between local and internode gates shifts towards allowing more internode gates and the gap between circuit cutting and quantum gates widens.

Besides experimental efforts, the development of protocols is another way to enhance the performance of current experiments. The fidelity of the direct conversion heralded scheme is primarily limited by the photon loss and thermal noise. The SPDC heralded scheme, however, is additionally limited by the possibility of multi-photon excitations in the resonator during the SPDC process [183]. Multi-photon excitations could potentially be suppressed through the use of an anharmonic resonator

[60]. In both schemes, the small probability that a photon is emitted simultaneously at both nodes, combined with the optical loss, will lead to a false heralding signal. One potential solution based on double-heralded detection has been proposed by Barrett and Kok [156] and experimentally realized with defects in crystals and trapped ions [184–187] and superconducting circuits [24]. While boosting the fidelity, this design requires two successful photon detections, and thus the success probability—as well as the entanglement generation rate—scales with the square of the photon detection probability. Alternative emerging protocols designed for M2O interfaces have also been proposed, such as the adaptive control protocol for reducing thermal noise [188], the active quantum feedback for deterministic entanglement generation [157], the continuous-variable quantum teleportation [189, 190] for high-fidelity state transfer, and time-bin [191] and frequency-bin [192] encoding for improved entanglement generation rate. One advantage of M2O links is that it may be possible to use the quantum control techniques available in circuit QED to use error correctable bosonic codes, several of which have recently exceeded the break-even point as quantum memories [193–195], for the com-

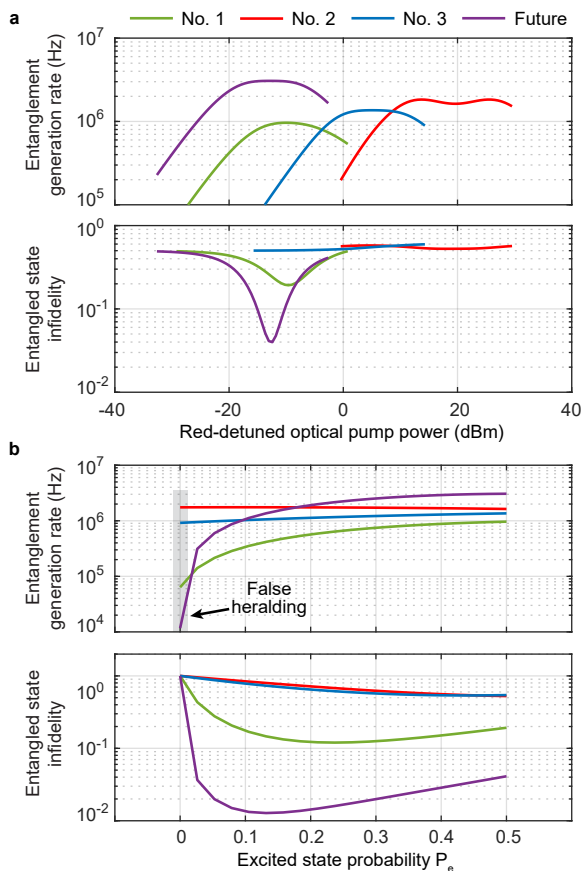


FIG. 17. The performance of a hypothetical M2O converter is shown as the purple curve. The data from Fig. 5 are also shown as a comparison.

munications.

Another key direction for improvement at the Physical layer is the use of multiplexing. As we saw in Section V, increasing the rate of pair generation is a key goal of the Physical layer. By operating multiple entanglement generation devices in parallel, we can increase the effective rate of entangled pair generation. Because decoherence accumulated while waiting for further EPs is a major source of internode noise, increasing the effective rate of EP generation reduces both the time and infidelity of internode communication, resulting in the dramatic effects shown in the second row of Table II.

Multiplexing M2O EP generation requires routing entangled photons generated in parallel channels into a superconducting node’s distillation module in real time. There are several methods for multiplexing flying qubits into a superconducting node. One multiplexing method that is promising for long-distance entanglement uses the “pitch-and-catch” framework, where the flying qubit is caught by a linear bus and swapped into the qubit coupled to that bus [24, 33, 35, 36]. Frequency-multiplexing the flying qubits would allow multiple flying qubits to be caught in parallel by the corresponding modes in the send/receive bus, and distributed into various cou-

pled qubits. One advantageous choice of bus-qubit coupler is the SNAIL (Superconducting Nonlinear Asymmetric Inductive eLement) [196] rather than currently used transmon couplers; the three-wave mixing interaction has reduced susceptibility to unwanted transitions/parametric processes compared to the four-wave mixing in transmon-based couplers. The SNAIL has been used to demonstrate successful all-to-all routing among 4 quantum modules [7]. The SNAIL can also be used as an alternative method for multiplexing flying qubits which is relevant for physically compact quantum computing within a single fridge. In this modality, a nonlinear SNAIL bus passively couples together all the qubits extending from it [7, 197].

Finally, hybrid technologies promise the greatest potential improvements, but also pose the most severe technical challenges [198]. In particular, a hybrid system using ions coupled to superconducting qubits [199–201] could allow for optical ion-ion links [202–204], between chips in separate dilution refrigerators. Significant technical challenges accompany hybrid ion-superconducting qubits [205]. However, techniques using molecular ions coupled to superconductors [206–210], while the use of ion chains for mode matching [211] can improve this coupling. As shown schematically in the third row of Table II, a superconducting-ion coupling would enable the rapid production of high-fidelity internode entangled pairs, likely limited chiefly by the rate and fidelity of the superconducting-ion coupling [201], and the use of ions as an extremely long-lived memory would also have significant implications for entanglement distillation and is discussed in the following section.

## B. Distillation Layer Improvements

In Section V we saw the key role entanglement distillation plays in enabling MNQC performance by improving the fidelity of EPs produced during the M2O process. However, entanglement purification performance is currently limited by the low yield of the purification protocols as well as by qubit decoherence during the purification. Potential improvements to this performance include careful co-design of protocols to adapt for the noise profile of M2O generation, the use of memory to prevent decoherence during the distillation process, and the use of long-term memories to allow buffering and effectively remove the fidelity bound. We tabulate these approaches and their speculative effects in Table III.

To improve the fidelity of these EPs, several entanglement purification protocols have been invented. In Appendix C we briefly introduce two purification protocols: the BBPSSW protocol [59] and the DEJMPS protocol [160]. In Section IV’s full stack simulation the DEJMPS protocol is used, as it provides more efficient purification compared to the BBPSSW protocol [160] and uses a small number of EPs to perform purification. We also notice that more advanced purification protocols,

|  | GAP | Q-Roofline | QCPA |
|--|-----|------------|------|
| <p><b>Distillation Protocol Co-Design:</b> Careful tailoring of distillation protocols to M2O noise may reduce infidelity 2x or more</p>         |     |            |      |
| <p><b>1ms Memory:</b> Protection from decoherence greatly improves distillation performance allowing new regimes of algorithms.</p>              |     |            |      |
| <p><b>10ms Memory:</b> Memory protects from decoherence and allows buffering of many EPs, allowing for high-fidelity, cheap internode gates.</p> |     |            |      |

TABLE III. The effects of three classes of improvements to the Distillation layer, as demonstrated in the GAP, Q-Roofline, and QCPA analyses.

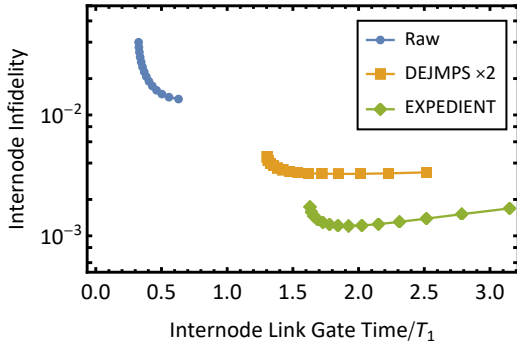


FIG. 18. Comparison of the EXPEDIENT purification protocol [212] with the two-round nested DEJMPS protocol [160].

e.g., double selection purification protocol [213], EXPEDIENT and STRINGENT purification protocols [212], may give better output EP fidelities after purification. In Fig. 18, we compare the performance of the two-round nested DEJMPS protocol with the EXPEDIENT protocol. The input EP state is from the M2O calculation. We notice the EXPEDIENT protocol uses 5 EPs in to-

tal to generate one EP with higher fidelity. Compared to the 2-round nested DEJMPS protocol, the EXPEDIENT protocol can give  $\sim 2$  times improvement. However, as it requires more EPs for each purification operation, the time for remote gate operation will be longer, and it will suffer more from the decoherence error if the EP generation is slow. However, even the DEJMPS protocol has a low purification yield. This is because for each purification, one of the two input EPs is destroyed. One direction for future work is to design new protocols for more efficient entanglement purification. Both the BBPSSW and DEJMPS protocols accept any raw EPs whose fidelity to the target state is greater than 0.5, without utilizing any other information about those states. One way to improve purification efficiency is to construct a precise error model for the raw EPs generated from the physical layer, and use that error information to design a more efficient purification protocol. This new protocol can either use hashing protocols with high finite yield [212, 214, 215] or require fewer rounds of nested purification to achieve high-fidelity EPs, so it could be used to implement more complex distributed algorithms. The improved performance of the EXPEDIENT protocol is shown in the first

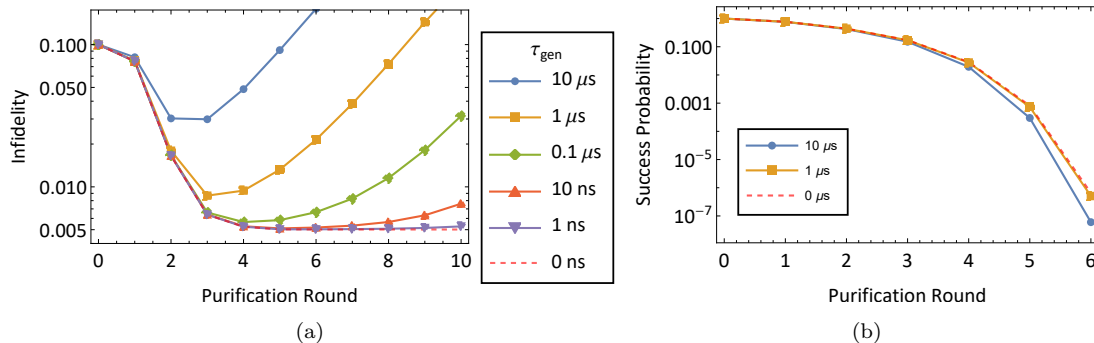


FIG. 19. The performance of the DEJMPS purification protocol [160] with sequential raw EP generation. In (a), we plot the infidelity of the output state from  $n$  rounds of purification using DEJMPS protocol. The generated raw EPs have fidelity 0.9. The superconducting qubits have  $T_1 = T_2 = 1 \text{ ms}$ . In (b), we plot the one-shot success probability as a function of nested purification rounds. The one-shot success probability decreases exponentially.

row of Table III.

In the full stack simulation (see Fig. 10 and 11), with the finite raw entangled pair generation rate, it is only practical to perform a few rounds of nested purification. In Fig. 19a, we calculate the fidelity of the output entangled pair after  $n$  rounds of purification. In this calculation, we especially show the effect of the finite rate of raw EP generation on the purification protocol. We observe a steady increase in output state infidelity due to the qubits relaxing and dephasing while waiting for more raw EPs to be generated. As discussed in Section VIA, the fastest raw EP generation rates are currently on the order of 1 MHz, so to make purification robust, effort must be made to reduce the raw EP generation time and increase qubit coherence times. In Fig. 19b we plot the single-shot success probability of  $n$  rounds of purification [see Eq. (C5) in Appendix C]. Due to the DEJMPS protocol’s low yield, even though the success probability of each single purification of two raw EPs can be close to unity, the overall single-shot success probability decreases exponentially as the number of nested purification rounds increases. This can also be seen from the fact that the number of terms in Eq. (C5) increases exponentially as the round increases. So the purification protocol’s low yield limits the practical benefit of doing many purification rounds in the entanglement distillation layer.

Furthermore, in our full-stack simulation, we assume that the local gates have depolarization error with probability 0.1%. However, in reality, the local gates between the compute qubits used to purify EPs may have larger errors. In Fig. 20, we consider the fidelity gain ( $\Delta F$ ) by performing a single round of entanglement purification to explore the effect of local imperfections. We consider the effect of CNOT gate error as well as qubit relaxation and dephasing during the purification protocol. We notice that even with CNOT gate error  $P_{\text{CNOT}} = 0.01$  [107], the efficiency of entanglement purification is noticeably affected compared to  $P_{\text{CNOT}} = 0.001$  case. To improve the performance of the purification layer and fully leverage the power of entanglement purification, local gate

error needs to be kept low.

One likely way to suppress relaxation and dephasing during purification is to use dedicated quantum memory elements. When the compute qubits are waiting for the next raw EP to arrive, their states can be swapped into quantum memory elements that have longer coherence times. This is particularly helpful for later rounds of distillation when the idle time on one of the two EPs from the previous round is substantial. In order to achieve this goal, the quantum memory elements need to have fast and high-fidelity SWAP gates with the compute qubits and they need to be stabilized against relaxation and de-

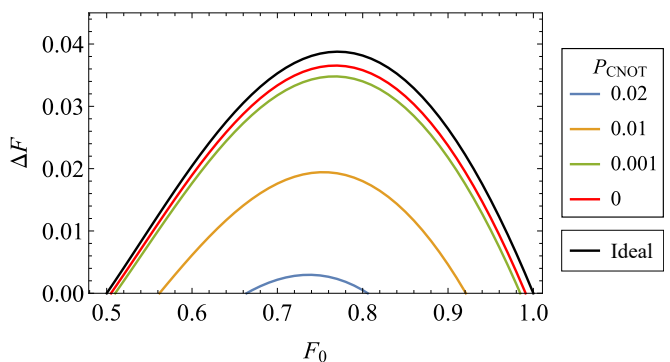


FIG. 20. The performance of the DEJMPS purification protocol with imperfections. We consider a single round of entanglement purification in the presence of imperfect gates between superconducting qubits and finite qubit coherence times. The CNOT gate error is modeled by depolarizing error channels with error probability  $P_{\text{CNOT}} = .0001$ . The fidelity of the imperfect EPs before purification is  $F_0$ , and after the purification we calculate the fidelity gain  $\Delta F = F_{\text{new}} - F_0$ . In addition, the superconducting qubits have finite relaxation and dephasing times for all but the case labeled “Ideal”. The qubit relaxation and dephasing times are  $T_1 = T_2 = 1 \text{ ms}$ . Each round of purification takes  $1 \mu\text{s}$ . The “Ideal” line is for the performance of the purification protocol with neither CNOT gate error nor decoherence error.

phasing, either by having naturally long coherence times or via active or autonomous quantum error correction [216]. The effect of a memory with a 1ms coherence time is shown in the second row of Table III, where we see that it dramatically reduces the achievable internode infidelities.

Transmon and fluxonium superconducting qubits have demonstrated high-fidelity two-qubit entangling gates [102, 106, 107, 217, 218], which make them good computing elements. Recent improvements in material processing and shielding/filtering have also boosted their coherence times towards 1 ms (as assumed in our simulations). However, 2D qubit coherence is often limited by dielectric loss from the substrate [219] and the interfaces [220], while 3D microwave modes can serve as even better memory elements [221] with potential lifetimes up to seconds [222]. Furthermore, 3D multimode cavities are a promising form of quantum memory element because a memory buffer with many storage modes can be created out of a single physical cavity, and high-fidelity SWAP gates in and out of the buffer can be performed by a single transmon [223, 224].

For these memory cavities to be effective in distillation protocols, an important area of improvement is the fidelity of cavity-qubit [223] or cavity-cavity SWAP [225] gates. These previous demonstrations rely on the shared nonlinearity of transmons to activate relatively slow four wave mixing processes. More recent experiments have shown that by using purpose-built parametric couplers one can perform much faster SWAP operations (100 ns or less) regardless of the nonlinearity of the swapping modes [226, 227], analogous to parametric two-qubit gates [228]. Implementation of these gates may allow storage of retrieval of EPs to and from quantum memory elements with infidelity at  $10^{-4}$  level.

An effective quantum memory exceeding 1ms, such as the hybrid superconducting-ion system discussed in the previous section, can have paradigm-shifting effects on both the fidelity and rate of internode communication because it can allow for the buffering [63] of entangled pairs. The effects of a 10ms quantum memory are schematically shown in the third row of Table III, where the buffering of memory reduces the time to execute internode gates during an algorithm to be comparable to that of local computation and thus results in a dramatic improvement in MNQC performance. In order to further increase the coherence time of the memory qubit, one could consider encoding the quantum information into a bosonic error correction code and implementing error correction [229–231]. Recently, active and autonomous stabilization of bosonic codes has been demonstrated close to or beyond the break-even point including the cat code [193, 216], the binomial code [232], and the GKP code [194, 233]. However, in these experiments the coherence of the error-corrected quantum memory is limited by that of the non-linear ancilla element used for stabilization. Eliminating this limitation and realizing a fault-tolerant bosonic memory well beyond the break-even point is an active

topic of research [234–236].

## C. Compiler and Application Improvements

While the lower levels of the MNQC stack determine the properties of the internode gates, it is the application and compiler layers that determine the use of internode gates and thus the performance of applications on a future MNQC. Much as in classical computing, developing compilers that can efficiently optimize around weaker internode links and applications that are adapted to multinode architectures will be critical for the success of MNQCs, and we must understand how improvements to these layer intersect with those of the rest of the stack. Determining the potential improvements for these layers involves considerable uncertainty as we do not have bounds for the performance achievable by compilers that have not been built (in the language of Section V, we do not have bounds on achievable CCRs), nor can we estimate the potential of algorithms yet to be discovered. Hence for these layers we take a schematic approach that still allows us to lay out a research agenda towards effective MNQCs.

At the compiler level, the perennial issues of qubit placement and routing must be overcome in addition to the complexities introduced by modular architectures containing heterogeneous qubit implementations and gate operations. Between the compiler and application layers, questions surrounding the software infrastructure responsible for workload management and resource sharing must be addressed. To overcome these issues we point to the similarities between distributed QC and classical HPC and discuss ways in which the strategies developed in the classical domain might be adapted to the quantum case. Finally, at the top of the stack we emphasize the need to profile and better understand distributed applications such that the information learned at the application level might help inform the co-design of the lower layers of the MNQC stack.

### 1. Compiler Improvements

Multinode systems pose a significant challenge for compilers due to both their scale and the complexities of balancing internode gates, local gates, and circuit cutting gates. As we have seen, internode operations are likely to remain more expensive and error-prone than local quantum gates, and therefore minimizing the communication overhead incurred during compilation will remain a primary concern.

Scale poses a challenge because assigning logical qubits to physical qubits, scheduling complex multi-qubit interactions, and routing physical qubits while respecting connectivity constraints become intractable as the number of qubits and gates in the program increase [237–239]. Current compilers are capable of translating large

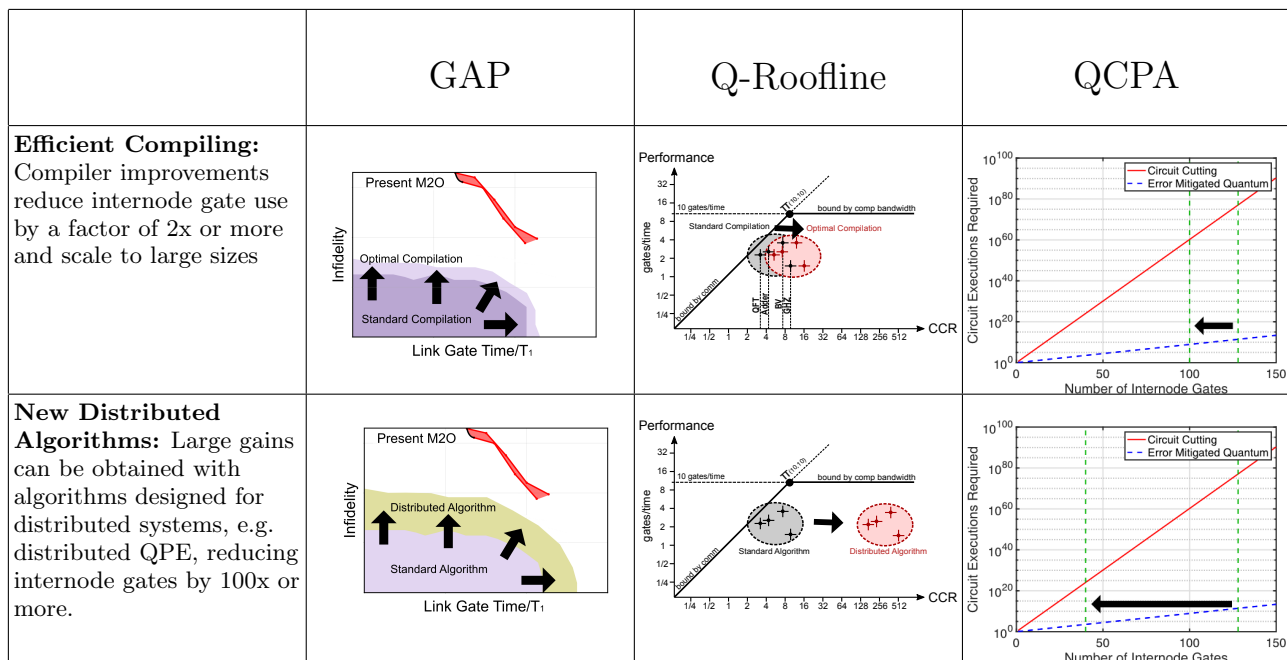


TABLE IV. Schematic depiction of the effects of improvements to the compiler and application layers, as demonstrated in the GAP, Q-Roofline, and QCPA analyses.

programs (containing more than  $10^6$  logical qubits and gates) into hardware-agnostic assembly programs [240], but are currently limited in their ability to map this to a hardware-compatible executable [241]. This problem is similar to the situation within classical high-performance computing where empirical studies have shown that the communication overhead quite often accounts for a larger portion of the program runtime than compute [242–244]. Quantum compilers may look to the field of classical HPC where load balancing has been extensively studied and efficient heuristic methods have been developed [245, 246]. Compiler-oriented partitioning, where circuit partitioning algorithms [64, 247, 248] are applied during compilation, can also be applied to optimize for minimal communications, maximum fidelity, and balanced workloads [54]. Once a program has been partitioned, distribution binds circuit partitions to module nodes and schedules inter-node communications. This is similar to qubit mapping, but at a coarser grain. The goal is to shorten the critical path (e.g., hide communication latency with local computation) and maximize program success rates while respecting communication dependencies constraints [54]. The architecture and metrics introduced in Figure 4 help to quantify the entanglement distillation process such that this information may be incorporated into a compiler to optimize distillation scheduling. Additional optimizations include buffer management [63], aggregation [55], and collective communication [63, 249].

Furthermore, good system performance may be achieved through efficient load balancing by boosting local occupancy and minimizing communication overheads. As discussed in Section V, for large quantum programs we

use the CCR of the compiled program as the performance metric for comparing among compilers, algorithms, and runtimes. Theoretically, the CCR is bounded by the number of local computations when no communication is ever needed. However, as shown in the Q-Roofline models of Figure 12, the connectivity constraints of the hardware may lead to an application becoming communication bound, and Figure 14 demonstrates how compiler optimizations may be used to mitigate this overhead. Developing techniques to incorporate gate fidelities into the Q-Roofline model to estimate program success rates of large scale quantum programs is a promising area of future research.

As a feature of user access, designing clusters of distributed quantum computers presents new and interesting challenges with regards to their software infrastructure. First, the appropriate level of abstraction for distributed quantum systems is an open question. Recent work has shown that quantum program success rates can be greatly improved by breaking layers of abstraction [250] and thus it will be necessary to balance quantum program success rates with user efficiency when designing distributed quantum systems. Secondly, while Section V is concerned with optimizing the compute throughput for a single application, multiple users submitting multiple dependent or independent job requests to a QC cluster presents set of challenges for efficient workload scheduling. In the classical paradigm, workload managers such as SLURM [251] are responsible for scheduling the available hardware resources to best meet the needs of the users. In a distributed quantum cluster, it appears that shared entanglement is likely to be

the most precious resource but the exact optimization objective itself and the specific management method still remain open questions.

Finally, MNQC systems lead to an interesting problem of software-hardware co-design because they may naturally support diverse heterogeneous architectures. Heterogeneity may manifest within the computation or communication within a distributed architecture. Individual nodes may consist of memory and compute regions implemented via different qubit modalities, and diverse technologies, implementing both quantum sensors and computers, may be used within a single quantum network. In this work we focused our analysis on combined quantum-classical communication channels which enable entanglement distribution [59] and teleportation [252]. However, other protocols may be used requiring only classical channels as in quantum circuit cutting [48, 49, 51] and entanglement forging [253], or solely quantum channels such as shuttling [254], direct state transfer [32] and cross-chip two-qubit gates [20]. Each protocol presents unique tradeoffs between fidelity, speed, and ease of implementation that any future compiler for a distributed system must consider.

## 2. Application and Algorithm Improvements

The design of a distributed quantum architecture will be heavily influenced by the workloads it is expected to encounter in practice. Profiling quantum programs to better understand the similarities and differences in their resource requirements is a critical area of future work. Prior work evaluating the performance of potential quantum architectures demonstrated that the match between hardware and application is important because quantum programs display different levels of computation versus communication [166, 255]. Our work in Section V and Figure 11 supports this view by demonstrating quantum applications' sensitivity to the parameters which characterize the quantum communication channels. Taking an example from classical computing, most applications can be assigned to one of a small number of application classes such as dense linear algebra, sparse linear algebra,  $N$ -body methods, and so on [68]. An important open question is understanding whether most quantum algorithms can similarly be grouped into a small number of general computational motifs.

In addition to profiling existing quantum applications, algorithm development – especially algorithms developed specifically for distributed systems – will play a critical role in the evolution of the field. Early investigations into distributed quantum applications include quantum telecomputation [256], distributed Shor's algorithm and arithmetic [252, 257, 258], distributed VQE (via classical networks [259] or quantum interconnects [260]), and distributed phase estimation [261].

In Section V, we noted that while many complex algorithms were unachievable using present technology, GHZ

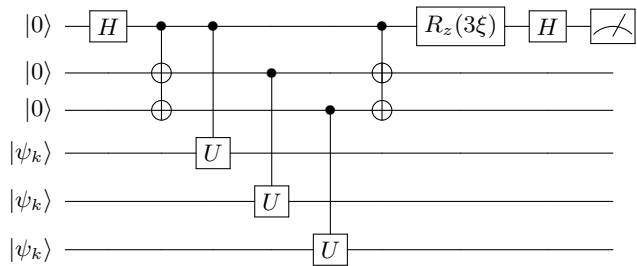


FIG. 21. Distributed Quantum Phase Estimation circuit on  $p = 3$  nodes. Although QPE is a high-depth algorithm, the highly communication-efficient structure of the distributed QPE circuit renders it a natural candidate for early MNQCs.

creation could be performed with high fidelity. In turn, this actually implies that Quantum Phase Estimation (QPE) is a good candidate for execution on early MNQCs. In fact, despite the fact that QPE is viewed as a high circuit depth algorithm, the multinode architecture can be used to increase the phase kickback coming from the controlled unitary operation and thus forms a good candidate for evaluation on an MNQC. Two strategies exist for such parallelism: the fully coherent approach of [262] which gives a reduction in the depth of phase estimation that is linear in the number of nodes and the approach that uses classical communication (found in the supplementary material of [261]). Both of these are reviewed in detail in Appendix A. In the case of an MNQC with quantum links, then we can use  $O(\frac{1}{\epsilon})$  nodes to perform phase estimation to accuracy  $\epsilon$  in  $O(1)$  time; in the case of purely classical links, then  $O(\frac{1}{\epsilon^2})$  nodes suffices to achieve the same bound.

In brief, the fully coherent version of distributed quantum phase estimation takes the form in Figure 21 [262]. It then follows from noting that the circuit returns the phase  $e^{i3\theta_k}$  from the phase kickback effect that in general this idea can be repeated  $p$  times to obtain  $p$  times the phase that would be seen with a single step of an iterative phase estimation procedure. However, the error in the internode link must be  $O(\frac{\epsilon^2}{\log(1/\epsilon)})$ , placing a significant demand on the performance of the MNQC stack. These properties thus make the QPE an intriguing early primitive for future early multinode machines using both classical and quantum links.

For quantum simulation, there are physical systems and model Hamiltonians that exhibit hybrid quantum-classical characters that can be naturally parallelized. One example of these model Hamiltonians is the quantum embedding descriptions of complex materials with multiple inequivalent impurity-bath subsystems [263–267] and quantum minimal entanglement typical thermal state sampling for finite-temperature simulations [268]. Additionally, in complex chemical systems such as metal-organic framework [269] and protein-ligand binding [270, 271], reaction centers typically contain transition metal

species that exhibit strong quantum effects while the rest of molecular backbone are largely classical and thus enables natural parallelization in simulation by utilizing locality of chemical processes.

Besides these rather straightforward quantum parallelizations, quantum algorithms and physical systems can also be tailored for calculations on the multinode quantum architecture with weak linkages. For example, the impurity-bath model in dynamical mean-field theory calculations can be optimized to minimize the direct interactions between the impurity and bath subsystems [272], and quantum transport systems of leads through nanocontacts naturally minimize the number of inter-node nonlocal gates [273, 274]. In quantum embedding calculations, the size of the fragment or cluster can be reduced and the level of theory for treating the bath may be performed at a lower mean field level which can minimize the number of entangled degrees of freedom with the fragment. Furthermore, simulations of the full electronic structure of periodic materials may lend themselves well to MNQC architectures. Electronic structure calculations at different in reciprocal momentum space can be parallelized with limited inter-node communication required [275]. Alternatively, real-space Wannier function representations to achieve compact encodings of electronic orbitals [276] may allow for parallelization of neighboring periodic cell images over separate nodes. Such approaches could facilitate electronic property calculations in the thermodynamic limit with smaller simulation cells, and thus a reduction in the qubit requirements per node.

These parallel schemes on electronic structure calculation can be directly applied to semi-classical *ab initio* molecular dynamics simulation such as the Born-Oppenheimer molecular dynamics [277]. A key step of the molecular dynamics simulation is to evaluate the electronic structure at different nuclei coordinates repeatedly, which naturally benefits from the distributed QPE protocol in Appendix A. Beyond semi-classical dynamics, it is possible to rephrase quantum dynamics as finding the ground state of a composite Hamiltonian [278], where parallelization protocols for embedding schemes as discussed above are promising to accelerate quantum dynamics. VQE approaches can also be designed to have structure that can take advantage of systems in which disjoint degrees of freedom are connected by small terms in a Hamiltonian (weak linkages) with only weak correlations between the subregions. Recent advances in classical simulations have been able to exploit this type of correlation structure with cluster algorithms [279]. A recent quantum algorithm has demonstrated that a VQE approach can be designed with the same advantages as a cluster algorithm [280]. Further research based on clustering algorithms may allow for new VQE approaches that are well suited for MNQC architectures.

## VII. OUTLOOK

We have quantified the potential performance of internode gates by building a layered architecture for internode link execution in MNQCs and developing a detailed quantitative model of each layer. By uniting these models, we were able to compare the available internode gate performance with the demands of algorithms in the GAP analysis, then reveal the relative costs of internode gates relative to local gates with the Q-Roofline model and relative to circuit cutting with the QCPA analysis. Our results paint a picture of the improvement in internode link performance needed to realize MNQCs links capable of competing with monolithic systems, and we laid out a research roadmap towards MNQCs, displaying potential improvements for each of the Physical, Distillation, and Application and Compiler layers in terms of the GAP, Q-Roofline, and QCPA models.

Going forward, these models provide benchmarks to quantify the impact of actual research developments as they are achieved. For future improvements in M2O technology improves, we can now directly predict the algorithms unlocked by improved fidelity and rate of M2O conversion. Similarly, as quantum memories are developed, we can determine how entanglement distillation will be improved, or how buffering of entangled pairs will allow new, more demanding computations to be completed successfully. For distributed compilers and algorithms, the Q-Roofline model provides a tangible metric for compiler performance, while the GAP and QCPA analyses demonstrate the impacts of improved efficiency and reduced reliance on internode communication. Uniting these analyses, we can now measure progress towards MNQCs that outperform monolithic systems.

More broadly, one of the most exciting future directions is extending this analysis to other platforms. While we have focused on superconducting systems with M2O interlinks as an MNQC, there is a wide array of platforms which have been envisioned as potential realizations of MNQCs. Borrowing from classical co-design [81, 82], our models are designed in a way which allows for different interconnection platforms to be analyzed in future work by changing only the Physical layer, while new distillation protocols can be used in the Distillation layer simulations and new local architecture can be changed in the Application and Compiler layer simulations. By interchanging these models, our approach can quantify the available performance across a range of systems from large scale quantum networks [85, 86] for distributed computing to smaller networks using cryogenic microwave links [7, 20, 30, 100, 281], which show considerable promise in the nearest term. Similarly, our approach can also characterize modular trapped ion systems [11–13] and neutral atom platforms [282, 283]. Hybrid systems consisting of several of these technologies [199–201, 284, 285] are a promising route for future networked systems and can also be treated within this framework, allowing this approach to treat the full range



of inter-operable distributed quantum systems. Just as our analysis revealed key tradeoffs and pointed the way for future technology development in superconducting M2O systems, a similar analysis of each of these other candidate platforms can quantify the performance currently available technology can offer as an MNQC, reveal key tradeoffs and interactions between components that may be make-or-break for multinode systems, and help develop research roadmaps that point the way towards the successful realization of MNQCs.

### VIII. KEY AUTHOR CONTRIBUTIONS

M. A. DeMarco led the project, developed the MNQC architecture, and guided simulation pipeline development. S. Stein built the simulation pipeline and performed simulations of entanglement distillation, remote gates, and benchmark execution. Y. Zhou performed simulations of the remote entanglement generation based on M2O converters and led writing on M2O simulation. C. Liu built models for the distillation layer and led the proposals for distillation improvements. T. Tomesh pro-

vided input on the MNQC architecture and led the proposals for compiler and application improvements. S. Sussman modeled coupling flying qubits into transmons and developed the proposals for the use of quantum memory for distillation. Y. Chen drafted the sections on protocols for entanglement generation. W. Tang drafted the sections on quantum compilers and multicomputing. P. Hilaire developed language on entanglement generation protocols and distributed architecture. D. McKay performed the error mitigation and circuit cutting analysis and edited the final draft. A. Li proposed the quantum roofline model, performed calculations for it, and guided the project. N. Wiebe wrote the analysis for distributed QPE. I. Chuang supervised the project and provided guidance in the architecture and writing.

### IX. ACKNOWLEDGEMENTS

This material is based upon work supported by the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers, Co-design Center for Quantum Advantage (C2QA) under contract number DE-SC0012704.

- 
- [1] S. Bravyi, O. Dial, J. M. Gambetta, D. Gil, and Z. Nazario, arXiv e-prints, arXiv:2209.06841 (2022), arXiv:2209.06841 [quant-ph].
- [2] IBM, “Ibm quantum roadmap,” Accessed: Sep. 20, 2022.
- [3] C. Monroe, R. Raussendorf, A. Ruthven, K. R. Brown, P. Maunz, L. M. Duan, and J. Kim, arXiv e-prints, arXiv:1208.0391 (2012), arXiv:1208.0391 [quant-ph].
- [4] K. Zhang, J. Thompson, X. Zhang, Y. Shen, Y. Lu, S. Zhang, J. Ma, V. Vedral, M. Gu, and K. Kim, Nature Communications **10**, 4692 (2019), arXiv:1907.12171 [quant-ph].
- [5] N. LaRacunte, K. N. Smith, P. Imany, K. L. Silverman, and F. T. Chong, (2022), arXiv:2201.08825 [quant-ph].
- [6] C. Monroe, R. Raussendorf, A. Ruthven, K. R. Brown, P. Maunz, L. M. Duan, and J. Kim, Phys. Rev. A **89**, 022317 (2014).
- [7] C. Zhou *et al.*, (2021), arXiv:2109.06848 [quant-ph].
- [8] M. Pompili, S. L. N. Hermans, S. Baier, H. K. C. Beukers, P. C. Humphreys, R. N. Schouten, R. F. L. Vermeulen, M. J. Tiggelman, L. dos Santos Martins, B. Dirkse, S. Wehner, and R. Hanson, Science **372**, 259 (2021), <https://www.science.org/doi/pdf/10.1126/science.abg1919>.
- [9] D. Awschalom, K. K. Berggren, H. Bernien, S. Bhave, L. D. Carr, P. Davids, S. E. Economou, D. Englund, A. Faraon, M. Fejer, S. Guha, M. V. Gustafsson, E. Hu, L. Jiang, J. Kim, B. Kozh, P. Kumar, P. G. Kwiat, M. Lončar, M. D. Lukin, D. A. Miller, C. Monroe, S. W. Nam, P. Narang, J. S. Orcutt, M. G. Raymer, A. H. Safavi-Naeini, M. Spiropulu, K. Srinivasan, S. Sun, J. Vučković, E. Waks, R. Walsworth, A. M. Weiner, and Z. Zhang, PRX Quantum **2**, 017002 (2021).
- [10] L. Gyongyosi and S. Imre, Sci. Rep. **11**, 5172 (2021).
- [11] C. D. Bruzewicz, J. Chiaverini, R. McConnell, and J. M. Sage, Applied Physics Reviews **6**, 021314 (2019), arXiv:1904.04178 [quant-ph].
- [12] J. M. Pino, J. M. Dreiling, C. Figgatt, J. P. Gaebler, S. A. Moses, M. S. Allman, C. H. Baldwin, M. Foss-Feig, D. Hayes, K. Mayer, C. Ryan-Anderson, and B. Neyenhuis, Nature (London) **592**, 209 (2021), arXiv:2003.01293 [quant-ph].
- [13] V. Kaushal, B. Lekitsch, A. Stahl, J. Hilder, D. Pijn, C. Schmiegelow, A. Bermudez, M. Müller, F. Schmidt-Kaler, and U. Poschinger, AVS Quantum Science **2**, 014101 (2020), arXiv:1912.04712 [quant-ph].
- [14] H. Bernien, B. Hensen, W. Pfaff, G. Koolstra, M. S. Blok, L. Robledo, T. H. Taminiau, M. Markham, D. J. Twitchen, L. Childress, and R. Hanson, Nature (London) **497**, 86 (2013), arXiv:1212.6136 [quant-ph].
- [15] K. Nemoto, M. Trupke, S. J. Devitt, B. Scharfenberger, K. Buczak, J. Schmiedmayer, and W. J. Munro, Scientific Reports **6**, 26284 (2016), arXiv:1412.5950 [quant-ph].
- [16] M. Pompili, S. L. N. Hermans, S. Baier, H. K. C. Beukers, P. C. Humphreys, R. N. Schouten, R. F. L. Vermeulen, M. J. Tiggelman, L. dos Santos Martins, B. Dirkse, S. Wehner, and R. Hanson, Science **372**, 259 (2021), arXiv:2102.04471 [quant-ph].
- [17] D. L. Moehring, P. Maunz, S. Olmschenk, K. C. Younge, D. N. Matsukevich, L. M. Duan, and C. Monroe, Nature (London) **449**, 68 (2007).
- [18] S. Ritter, C. Nölleke, C. Hahn, A. Reiserer, A. Neuzner, M. Uphoff, M. Mücke, E. Figueroa, J. Bochmann, and G. Rempe, Nature (London) **484**, 195 (2012), arXiv:1202.5955 [quant-ph].

- [19] C. B. Young, A. Safari, P. Huft, J. Zhang, E. Oh, R. Chinnarasu, and M. Saffman, *Applied Physics B: Lasers and Optics* **128**, 151 (2022), arXiv:2202.01634 [quant-ph].
- [20] A. Gold, J. P. Paquette, A. Stockklauser, M. J. Reagor, M. S. Alam, A. Bestwick, N. Didier, A. Nersisyan, F. Oruc, A. Razavi, B. Scharmman, E. A. Sete, B. Sur, D. Venturelli, C. J. Winkleblack, F. Wudarski, M. Harburn, and C. Rigetti, *npj Quantum Information* **7**, 142 (2021), arXiv:2102.13293 [quant-ph].
- [21] B. Foxen, J. Y. Mutus, E. Lucero, R. Graff, A. Megrant, Y. Chen, C. Quintana, B. Burkett, J. Kelly, E. Jeffrey, Y. Yang, A. Yu, K. Arya, R. Barends, Z. Chen, B. Chiaro, A. Dunsworth, A. Fowler, C. Gidney, M. Giustina, T. Huang, P. Klimov, M. Neeley, C. Neill, P. Roushan, D. Sank, A. Vainsencher, J. Wenner, T. C. White, and J. M. Martinis, *Quantum Science and Technology* **3**, 014005 (2018), arXiv:1708.04270 [quant-ph].
- [22] L. D. Burkhardt, J. D. Teoh, Y. Zhang, C. J. Axline, L. Frunzio, M. H. Devoret, L. Jiang, S. M. Girvin, and R. J. Schoelkopf, *PRX Quantum* **2**, 030321 (2021), arXiv:2004.06168 [quant-ph].
- [23] P. Magnard, S. Storz, P. Kurpiers, J. Schär, F. Marxer, J. Lütolf, T. Walter, J.-C. Besse, M. Gabureac, K. Reuer, *et al.*, *Phys. Rev. Lett.* **125**, 260502 (2020).
- [24] A. Narla, S. Shankar, M. Hatridge, Z. Leghtas, K. M. Sliwa, E. Zolys-Geller, S. O. Mundhada, W. Pfaff, L. Frunzio, R. J. Schoelkopf, and M. H. Devoret, *Physical Review X* **6**, 031036 (2016), arXiv:1603.03742 [quant-ph].
- [25] M. H. Devoret and R. J. Schoelkopf, *Science* **339**, 1169 (2013).
- [26] M. Kjaergaard, M. E. Schwartz, J. Braumüller, P. Krantz, J. I-Jan Wang, S. Gustavsson, and W. D. Oliver, arXiv e-prints, arXiv:1905.13641 (2019), arXiv:1905.13641 [quant-ph].
- [27] M. H. Devoret, A. Wallraff, and J. M. Martinis, arXiv e-prints, cond-mat/0411174 (2004), arXiv:cond-mat/0411174 [cond-mat.mes-hall].
- [28] S. Krinner, S. Storz, P. Kurpiers, P. Magnard, J. Heinsoo, R. Keller, J. Luetolf, C. Eichler, and A. Wallraff, arXiv e-prints, arXiv:1806.07862 (2018), arXiv:1806.07862 [quant-ph].
- [29] N. H. Nickerson, J. F. Fitzsimons, and S. C. Benjamin, *Phys. Rev. X* **4**, 041041 (2014).
- [30] H. Yan, Y. Zhong, H.-S. Chang, A. Bienfait, M.-H. Chou, C. R. Conner, E. Dumur, J. Grebel, R. G. Povey, and A. N. Cleland, *Phys. Rev. Lett.* **128**, 080504 (2022).
- [31] H. Zu, W. Dai, and A. T. A. M. de Waele, *Cryogenics* **121**, 103390 (2022).
- [32] C. J. Axline, L. D. Burkhardt, W. Pfaff, M. Zhang, K. Chou, P. Campagne-Ibarcq, P. Reinhold, L. Frunzio, S. M. Girvin, L. Jiang, M. H. Devoret, and R. J. Schoelkopf, *Nature Physics* **14**, 705 (2018).
- [33] P. Campagne-Ibarcq, E. Zolys-Geller, A. Narla, S. Shankar, P. Reinhold, L. Burkhardt, C. Axline, W. Pfaff, L. Frunzio, R. J. Schoelkopf, and M. H. Devoret, *Phys. Rev. Lett.* **120**, 200501 (2018).
- [34] P. Kurpiers, P. Magnard, T. Walter, B. Royer, M. Pechal, J. Heinsoo, Y. Salathé, A. Akin, S. Storz, J.-C. Besse, S. Gasparinetti, A. Blais, and A. Wallraff, *Nature* **558**, 264 (2018).
- [35] N. Leung, Y. Lu, S. Chakram, R. K. Naik, N. Earnest, R. Ma, K. Jacobs, A. N. Cleland, and D. I. Schuster, *npj Quantum Information* **5**, 18 (2019).
- [36] L. D. Burkhardt, J. D. Teoh, Y. Zhang, C. J. Axline, L. Frunzio, M. Devoret, L. Jiang, S. Girvin, and R. Schoelkopf, *PRX Quantum* **2**, 030321 (2021).
- [37] A. Place, L. Rodgers, P. Mundada, *et al.*, *Nat. Commun.* **12**, 1779 (2021).
- [38] C. Wang, X. Li, H. Xu, Z. Li, J. Wang, Z. Yang, Z. Mi, X. Liang, T. Su, C. Yang, G. Wang, W. Wang, Y. Li, M. Chen, C. Li, K. Linghu, J. Han, Y. Zhang, Y. Feng, Y. Song, T. Ma, J. Zhang, R. Wang, P. Zhao, W. Liu, G. Xue, Y. Jin, and H. Yu, *npj Quantum Information* **8**, 3 (2022).
- [39] X. Han, W. Fu, C.-L. Zou, L. Jiang, and H. X. Tang, *Optica* **8**, 1050 (2021).
- [40] G. Kurizki, P. Bertet, Y. Kubo, K. Mølmer, D. Petrosyan, P. Rabl, and J. Schmiedmayer, *Proc. Natl. Acad. Sci. USA* **112**, 3866 (2015).
- [41] N. J. Lambert, A. Rueda, F. Sedlmeir, and H. G. Schwefel, *Adv. Quantum Technol.* **3**, 1900077 (2020).
- [42] N. Lauk, N. Sinclair, S. Barzanjeh, J. P. Covey, M. Saffman, M. Spiropulu, and C. Simon, *Quantum Sci. Technol.* **5**, 020501 (2020).
- [43] Y. Chu and S. Gröblacher, *Appl. Phys. Lett.* **117**, 150503 (2020).
- [44] A. Clerk, K. Lehnert, P. Bertet, J. Petta, and Y. Nakamura, *Nat. Phys.* **16**, 257 (2020).
- [45] M. Mirhosseini, A. Sipahigil, M. Kalaei, and O. Painter, *Nature (London)* **588**, 599 (2020).
- [46] D. Awschalom, K. K. Berggren, H. Bernien, S. Bhave, L. D. Carr, P. Davids, S. E. Economou, D. Englund, A. Faraon, M. Fejer, S. Guha, M. V. Gustafsson, E. Hu, L. Jiang, J. Kim, B. Korzh, P. Kumar, P. G. Kwiat, M. Lončar, M. D. Lukin, D. A. B. Miller, C. Monroe, S. W. Nam, P. Narang, J. S. Orcutt, M. G. Raymer, A. H. Safavi-Naeini, M. Spiropulu, K. Srinivasan, S. Sun, J. Vučković, E. Waks, R. Walsworth, A. M. Weiner, and Z. Zhang, arXiv e-prints, arXiv:1912.06642 (2019), arXiv:1912.06642 [quant-ph].
- [47] C. Qiao, Y. Zhao, G. Zhao, and H. Xu, in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications Workshops* (2022).
- [48] W. Tang, T. Tomesh, M. Suchara, J. Larson, and M. Martonosi, in *Proceedings of the 26th ACM International conference on architectural support for programming languages and operating systems* (2021) pp. 473–486.
- [49] W. Tang and M. Martonosi, arXiv preprint arXiv:2207.00933 (2022).
- [50] S. Bravyi, G. Smith, and J. A. Smolin, *Phys. Rev. X* **6**, 021043 (2016).
- [51] T. Peng, A. W. Harrow, M. Ozols, and X. Wu, *Physical Review Letters* **125**, 150504 (2020).
- [52] K. Mitarai and K. Fujii, *New Journal of Physics* **23**, 023021 (2020).
- [53] C. Piveteau and D. Sutter, arXiv preprint 2205.00016 (2022).
- [54] D. Ferrari, A. S. Cacciapuoti, M. Amoretti, and M. Calceffi, arXiv preprint arXiv:2012.09680 (2020).
- [55] A. Wu, H. Zhang, G. Li, A. Shabani, Y. Xie, and Y. Ding, arXiv preprint arXiv:2207.11674 (2022).
- [56] A. Rueda, F. Sedlmeir, M. C. Collodo, U. Vogl, B. Stiller, G. Schunk, D. V. Strekalov, C. Marquardt, J. M. Fink, O. Painter, *et al.*, *Optica* **3**, 597 (2016).

- [57] T. P. McKenna, J. D. Witmer, R. N. Patel, W. Jiang, R. Van Laer, P. Arrangoiz-Arriola, E. A. Wollack, J. F. Herrmann, and A. H. Safavi-Naeini, *Optica* **7**, 1737 (2020).
- [58] W. Dür and H.-J. Briegel, *Phys. Rev. Lett.* **90**, 067901 (2003).
- [59] C. H. Bennett, G. Brassard, S. Popescu, B. Schumacher, J. A. Smolin, and W. K. Wootters, *Phys. Rev. Lett.* **76**, 722 (1996).
- [60] S. Krastanov, H. Raniwala, J. Holzgrafe, K. Jacobs, M. Lončar, M. J. Reagor, and D. R. Englund, *Phys. Rev. Lett.* **127**, 040503 (2021).
- [61] A. Li, O. Subasi, X. Yang, and S. Krishnamoorthy, in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis* (IEEE, 2020) pp. 1–15.
- [62] T. E. O’Brien, B. Tarasinski, and L. DiCarlo, *npj Quantum Information* **3**, 1 (2017).
- [63] A. Wu, Y. Ding, and A. Li, arXiv preprint arXiv:2208.06724 (2022).
- [64] D. Dadkhah, M. Zomorodi, S. E. Hosseini, P. Plawiak, and X. Zhou, *IEEE Access* **10**, 70329 (2022).
- [65] R. Beals, S. Brierley, O. Gray, A. W. Harrow, S. Kutin, N. Linden, D. Shepherd, and M. Stather, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **469**, 20120686 (2013).
- [66] G. Li, A. Wu, Y. Shi, A. Javadi-Abhari, Y. Ding, and Y. Xie, in *Proceedings of the Eight Annual ACM International Conference on Nanoscale Computing and Communication*, NANOCOM ’21 (Association for Computing Machinery, New York, NY, USA, 2021).
- [67] A. Tanenbaum and M. van Steen, *Distributed Systems: Principles and Paradigms* (Pearson Prentice Hall, 2007).
- [68] K. Asanovic, R. Bodik, B. C. Catanzaro, J. J. Gebis, P. Husbands, K. Keutzer, D. A. Patterson, W. L. Plishker, J. Shalf, S. W. Williams, *et al.*, (2006).
- [69] M. van Steen and A. S. Tanenbaum, *Computing* **98**, 967 (2016).
- [70] A. Agarwal, R. Bianchini, D. Chaiken, K. Johnson, D. Kranz, J. Kubiatowicz, B.-H. Lim, K. Mackenzie, and D. Yeung, in *Proceedings 22nd Annual International Symposium on Computer Architecture* (1995) pp. 2–13.
- [71] A. Agarwal, R. Bianchini, D. Chaiken, F. Chong, K. Johnson, D. Kranz, J. Kubiatowicz, B.-H. Lim, K. Mackenzie, and D. Yeung, *Proceedings of the IEEE* **87**, 430 (1999).
- [72] T. Sterling, E. Lusk, and W. Gropp, *Beowulf Cluster Computing with Linux*, 2nd ed. (MIT Press, Cambridge, MA, USA, 2003).
- [73] D. De Sensi, S. Di Girolamo, K. H. McMahon, D. Roweth, and T. Hoefler, arXiv e-prints , arXiv:2008.08886 (2020), arXiv:2008.08886 [cs.DC].
- [74] *Infiniband Architecture Specification Volume 1*, Infiniband Trade Association, 1st ed.
- [75] A. Gara, M. A. Blumrich, D. Chen, G.-T. Chiu, P. Coates, M. E. Giampapa, R. A. Haring, P. Heidelberger, D. Hoenicke, G. V. Kopsay, *et al.*, *IBM Journal of research and development* **49**, 195 (2005).
- [76] C. E. Leiserson, *IEEE Transactions on Computers* **C-34**, 892 (1985).
- [77] S. L. Scott and et al., “The cray t3e network: Adaptive routing in a high performance 3d torus,” (1996).
- [78] J. Kim, W. J. Dally, S. Scott, and D. Abts, in *2008 International Symposium on Computer Architecture* (2008) pp. 77–88.
- [79] S. Williams, A. Waterman, and D. Patterson, *Communications of the ACM* **52**, 65 (2009).
- [80] J. A. Ang, in *Proceedings of the 2015 International Symposium on Memory Systems*, MEMSYS ’15 (Association for Computing Machinery, New York, NY, USA, 2015) p. 51–52.
- [81] R. Barrett, S. Borkar, S. Dosanjh, S. Hammond, M. Heroux, X. Hu, J. Luitjens, S. Parker, J. Shalf, and L. Tang, *Advances in Parallel Computing* **24**, 141 (2013).
- [82] I. Foster and W. Gentzsch, *High Performance Computing: From Grids and Clouds to Exascale*, Advances in parallel computing (IOS Press, 2011).
- [83] J. I. Cirac, P. Zoller, H. J. Kimble, and H. Mabuchi, *Phys. Rev. Lett.* **78**, 3221 (1997), arXiv:quant-ph/9611017 [quant-ph].
- [84] H. J. Kimble, *Nature (London)* **453**, 1023 (2008), arXiv:0806.4195 [quant-ph].
- [85] D. Cuomo, M. Caleffi, and A. S. Cacciapuoti, arXiv e-prints , arXiv:2002.11808 (2020), arXiv:2002.11808 [quant-ph].
- [86] S. Wehner, D. Elkouss, and R. Hanson, *Science* **362**, 9288 (2018).
- [87] N. Sangouard, C. Simon, H. de Riedmatten, and N. Gisin, *Rev. Mod. Phys.* **83**, 33 (2011).
- [88] Q. Ruihong and M. Ying, *Journal of Physics: Conference Series* **1237**, 052032 (2019).
- [89] S. W. Loke, (2022), arXiv:2208.10127 [cs.DC].
- [90] A. Dahlberg, M. Skrzypczyk, T. Coopmans, L. Wubben, F. Rozpedek, M. Pompili, A. Stolk, P. Pawelczak, R. Knegjens, J. de Oliveira Filho, R. Hanson, and S. Wehner, arXiv e-prints , arXiv:1903.09778 (2019), arXiv:1903.09778 [quant-ph].
- [91] A. Pirker and W. Dür, *New Journal of Physics* **21**, 033003 (2019), arXiv:1810.03556 [quant-ph].
- [92] H. Ai, Y.-Y. Fang, C.-R. Feng, Z. Peng, and Z.-L. Xiang, *Phys. Rev. Applied* **17**, 054021 (2022).
- [93] M. Victora, S. Krastanov, A. Sanchez de la Cerda, S. Willis, and P. Narang, arXiv e-prints , arXiv:2011.11644 (2020), arXiv:2011.11644 [quant-ph].
- [94] D. Cuomo, M. Caleffi, K. Krsulich, F. Tramonto, G. Agliardi, E. Prati, and A. S. Cacciapuoti, (2021), arXiv:2112.14139 [quant-ph].
- [95] A. Wu, Y. Ding, and A. Li, (2022), arXiv:2208.06724 [quant-ph].
- [96] S. F. Bush, W. A. Challener, and G. Mantelet, *AVS Quantum Science* **3**, 030501 (2021).
- [97] A. Wu, H. Zhang, G. Li, A. Shabani, Y. Xie, and Y. Ding, (2022), arXiv:2207.11674 [quant-ph].
- [98] J. M. Baker, C. Duckering, A. Hoover, and F. T. Chong, arXiv e-prints , arXiv:2005.12259 (2020), arXiv:2005.12259 [quant-ph].
- [99] T. Tomesh and M. Martonosi, *IEEE Micro* **41**, 33 (2021).
- [100] S. Bravyi, D. Gosset, and Y. Liu, *Phys. Rev. Lett.* **128**, 220503 (2022).
- [101] C. Chamberland, K. Noh, P. Arrangoiz-Arriola, E. T. Campbell, C. T. Hann, J. Iverson, H. Putterman, T. C. Bohdanowicz, S. T. Flammia, A. Keller, G. Refael, J. Preskill, L. Jiang, A. H. Safavi-Naeini, O. Painter, and F. G. Brandão, *PRX Quantum* **3**, 010329 (2022).

- [102] F. Bao, H. Deng, D. Ding, R. Gao, X. Gao, C. Huang, X. Jiang, H.-S. Ku, Z. Li, X. Ma, X. Ni, J. Qin, Z. Song, H. Sun, C. Tang, T. Wang, F. Wu, T. Xia, W. Yu, F. Zhang, G. Zhang, X. Zhang, J. Zhou, X. Zhu, Y. Shi, J. Chen, H.-H. Zhao, and C. Deng, *Phys. Rev. Lett.* **129**, 010502 (2022).
- [103] J. Koch, T. M. Yu, J. Gambetta, A. A. Houck, D. I. Schuster, J. Majer, A. Blais, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf, *Phys. Rev. A* **76**, 042319 (2007).
- [104] R. Acharya, I. Aleiner, R. Allen, T. I. Andersen, M. Ansmann, F. Arute, K. Arya, A. Asfaw, J. Atalaya, R. Babbush, D. Bacon, J. C. Bardin, J. Basso, A. Bengtsson, S. Boixo, G. Bortoli, A. Bourassa, J. Bovaird, L. Brill, M. Broughton, B. B. Buckley, D. A. Buell, T. Burger, B. Burkett, N. Bushnell, Y. Chen, Z. Chen, B. Chiaro, J. Cogran, R. Collins, P. Conner, W. Courtney, A. L. Crook, B. Curtin, D. M. Debroy, A. D. T. Barba, S. Demura, A. Dunsworth, D. Eppens, C. Erickson, L. Faoro, E. Farhi, R. Fatemi, L. F. Burgos, E. Forati, A. G. Fowler, B. Foxen, W. Giang, C. Gidney, D. Gilboa, M. Giustina, A. G. Dau, J. A. Gross, S. Habegger, M. C. Hamilton, M. P. Harrigan, S. D. Harrington, O. Higgott, J. Hilton, M. Hoffmann, S. Hong, T. Huang, A. Huff, W. J. Huggins, L. B. Ioffe, S. V. Isakov, J. Iveland, E. Jeffrey, Z. Jiang, C. Jones, P. Juhas, D. Kafri, K. Kechedzhi, J. Kelly, T. Khattar, M. Khezri, M. Kieferová, S. Kim, A. Kitaev, P. V. Klimov, A. R. Klots, A. N. Korotkov, F. Kostritsa, J. M. Kreikebaum, D. Landhuis, P. Laptev, K.-M. Lau, L. Laws, J. Lee, K. Lee, B. J. Lester, A. Lill, W. Liu, A. Locharla, E. Lucero, F. D. Malone, J. Marshall, O. Martin, J. R. McClean, T. McCourt, M. McEwen, A. Megrant, B. M. Costa, X. Mi, K. C. Miao, M. Mohseni, S. Montazeri, A. Morvan, E. Mount, W. Mruczkiewicz, O. Naaman, M. Neeley, C. Neill, A. Nersisyan, H. Neven, M. Newman, J. H. Ng, A. Nguyen, M. Nguyen, M. Y. Niu, T. E. O'Brien, A. Opremcak, J. Platt, A. Petukhov, R. Potter, L. P. Pryadko, C. Quintana, P. Roushan, N. C. Rubin, N. Saei, D. Sank, K. Sankaragomathi, K. J. Satzinger, H. F. Schurkus, C. Schuster, M. J. Shearn, A. Shorter, V. Shvarts, J. Skrzynny, V. Smelyanskiy, W. C. Smith, G. Sterling, D. Strain, M. Szalay, A. Torres, G. Vidal, B. Villalonga, C. V. Heidweiller, T. White, C. Xing, Z. J. Yao, P. Yeh, J. Yoo, G. Young, A. Zalcman, Y. Zhang, and N. Zhu, arXiv preprint arXiv:2207.06431 (2022).
- [105] B. Foxen, C. Neill, A. Dunsworth, P. Roushan, B. Chiaro, A. Megrant, J. Kelly, Z. Chen, K. Satzinger, R. Barends, F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, S. Boixo, D. Buell, B. Burkett, Y. Chen, R. Collins, E. Farhi, A. Fowler, C. Gidney, M. Giustina, R. Graff, M. Harrigan, T. Huang, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, P. Klimov, A. Korotkov, F. Kostritsa, D. Landhuis, E. Lucero, J. McClean, M. McEwen, X. Mi, M. Mohseni, J. Y. Mutus, O. Naaman, M. Neeley, M. Niu, A. Petukhov, C. Quintana, N. Rubin, D. Sank, V. Smelyanskiy, A. Vainsencher, T. C. White, Z. Yao, P. Yeh, A. Zalcman, H. Neven, and J. M. Martinis (Google AI Quantum), *Phys. Rev. Lett.* **125**, 120504 (2020).
- [106] Y. Sung, L. Ding, J. Braumüller, A. Vepsäläinen, B. Kannan, M. Kjaergaard, A. Greene, G. O. Samach, C. McNally, D. Kim, A. Melville, B. M. Niedzielski, M. E. Schwartz, J. L. Yoder, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *Phys. Rev. X* **11**, 021058 (2021).
- [107] K. X. Wei, E. Magesan, I. Lauer, S. Srinivasan, D. F. Bogorin, S. Carnevale, G. A. Keefe, Y. Kim, D. Klaus, W. Landers, N. Sundaresan, C. Wang, E. J. Zhang, M. Steffen, O. E. Dial, D. C. McKay, and A. Kandala, *Phys. Rev. Lett.* **129**, 060501 (2022).
- [108] H. Collins and C. Nay, "Ibm unveils 400 qubit-plus quantum processor and next-generation ibm quantum system two," .
- [109] N. Sundaresan, T. J. Yoder, Y. Kim, M. Li, E. H. Chen, G. Harper, T. Thorbeck, A. W. Cross, A. D. Córcoles, and M. Takita, arXiv preprint arXiv:2203.07205 (2022).
- [110] Y. Zhao, Y. Ye, H.-L. Huang, Y. Zhang, D. Wu, H. Guan, Q. Zhu, Z. Wei, T. He, S. Cao, F. Chen, T.-H. Chung, H. Deng, D. Fan, M. Gong, C. Guo, S. Guo, L. Han, N. Li, S. Li, Y. Li, F. Liang, J. Lin, H. Qian, H. Rong, H. Su, L. Sun, S. Wang, Y. Wu, Y. Xu, C. Ying, J. Yu, C. Zha, K. Zhang, Y.-H. Huo, C.-Y. Lu, C.-Z. Peng, X. Zhu, and J.-W. Pan, *Phys. Rev. Lett.* **129**, 030501 (2022).
- [111] S. Krinner, N. Lacroix, A. Remm, A. Di Paolo, E. Genois, C. Leroux, C. Hellings, S. Lazar, F. Swiadek, J. Herrmann, G. J. Norris, C. K. Andersen, M. Müller, A. Blais, C. Eichler, and A. Wallraff, *Nature* **605**, 669 (2022).
- [112] N. Roch, M. E. Schwartz, F. Motzoi, C. Macklin, R. Vijay, A. W. Eddins, A. N. Korotkov, K. B. Whaley, M. Sarovar, and I. Siddiqi, *Phys. Rev. Lett.* **112**, 170501 (2014).
- [113] C. Dickel, J. J. Wesdorp, N. K. Langford, S. Peiter, R. Sagastizabal, A. Bruno, B. Criger, F. Motzoi, and L. DiCarlo, *Phys. Rev. B* **97**, 064508 (2018).
- [114] Y. P. Zhong, H.-S. Chang, K. J. Satzinger, M.-H. Chou, A. Bienfait, C. R. Conner, É. Dumur, J. Grebel, G. A. Peairs, R. G. Povey, D. I. Schuster, and A. N. Cleland, *Nature Physics* **15**, 741 (2019).
- [115] H.-S. Chang, Y. P. Zhong, A. Bienfait, M.-H. Chou, C. R. Conner, E. Dumur, J. Grebel, G. A. Peairs, R. G. Povey, K. J. Satzinger, and A. N. Cleland, *Phys. Rev. Lett.* **124**, 240502 (2020).
- [116] P. Kurpiers, M. Pechal, B. Royer, P. Magnard, T. Walter, J. Heinsoo, Y. Salathé, A. Akin, S. Storz, J.-C. Besse, S. Gasparinetti, A. Blais, and A. Wallraff, *Phys. Rev. Applied* **12**, 044067 (2019).
- [117] 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, and A. N. Cleland, *Nature* **590**, 571 (2021).
- [118] M. Mirhosseini, A. Sipahigil, M. Kalaei, and O. Painter, *Nature* **588**, 599 (2020).
- [119] R. Delaney, M. Urmey, S. Mittal, B. Brubaker, J. Kindem, P. Burns, C. Regal, and K. Lehnert, *Nature* **606**, 489 (2022).
- [120] F. Lecocq, F. Quinlan, K. Cicak, J. Aumentado, S. Didams, and J. Teufel, *Nature* **591**, 575 (2021).
- [121] J. Holzgrafe, N. Sinclair, D. Zhu, A. Shams-Ansari, M. Colangelo, Y. Hu, M. Zhang, K. K. Berggren, and M. Lončar, *Optica* **7**, 1714 (2020).
- [122] Y. Xu, A. A. Sayem, L. Fan, C.-L. Zou, S. Wang, R. Cheng, W. Fu, L. Yang, M. Xu, and H. X. Tang,

- Nat. Commun. **12**, 4453 (2021).
- [123] L. Fan, C.-L. Zou, R. Cheng, X. Guo, X. Han, Z. Gong, S. Wang, and H. X. Tang, *Sci. Adv.* **4**, eaar4994 (2018).
- [124] W. Fu, M. Xu, X. Liu, C.-L. Zou, C. Zhong, X. Han, M. Shen, Y. Xu, R. Cheng, S. Wang, *et al.*, *Phys. Rev. A* **103**, 053504 (2021).
- [125] A. Youssefi, I. Shomroni, Y. J. Joshi, N. R. Bernier, A. Lukashchuk, P. Urich, L. Qiu, and T. J. Kippenberg, *Nat. Electron.* **4**, 326 (2021).
- [126] M. Soltani, M. Zhang, C. Ryan, G. J. Ribeill, C. Wang, and M. Loncar, *Phys. Rev. A* **96**, 043808 (2017).
- [127] W. Hease, A. Rueda, R. Sahu, M. Wulf, G. Arnold, H. G. Schwefel, and J. M. Fink, *PRX Quantum* **1**, 020315 (2020).
- [128] R. Sahu, W. Hease, A. Rueda, G. Arnold, L. Qiu, and J. M. Fink, *Nat. Commun.* **13**, 1276 (2022).
- [129] X. Zhang, N. Zhu, C.-L. Zou, and H. X. Tang, *Phys. Rev. Lett.* **117**, 123605 (2016).
- [130] R. Hisatomi, A. Osada, Y. Tabuchi, T. Ishikawa, A. Noguchi, R. Yamazaki, K. Usami, and Y. Nakamura, *Phys. Rev. B* **93**, 174427 (2016).
- [131] N. Zhu, X. Zhang, X. Han, C.-L. Zou, C. Zhong, C.-H. Wang, L. Jiang, and H. X. Tang, *Optica* **7**, 1291 (2020).
- [132] X. Zhang, C.-L. Zou, L. Jiang, and H. X. Tang, *Phys. Rev. Lett.* **113**, 156401 (2014).
- [133] R. W. Andrews, R. W. Peterson, T. P. Purdy, K. Cicak, R. W. Simmonds, C. A. Regal, and K. W. Lehnert, *Nat. Phys.* **10**, 321 (2014).
- [134] A. P. Higginbotham, P. Burns, M. Urmey, R. Peterson, N. Kampel, B. Brubaker, G. Smith, K. Lehnert, and C. Regal, *Nat. Phys.* **14**, 1038 (2018).
- [135] G. Arnold, M. Wulf, S. Barzanjeh, E. Redchenko, A. Rueda, W. J. Hease, F. Hassani, and J. M. Fink, *Nat. Commun.* **11**, 4460 (2020).
- [136] B. M. Brubaker, J. M. Kindem, M. D. Urmey, S. Mittal, R. D. Delaney, P. S. Burns, M. R. Vissers, K. W. Lehnert, and C. A. Regal, *Phys. Rev. X* **12**, 021062 (2022).
- [137] A. Kumar, A. Suleymanzade, M. Stone, L. Taneja, A. Anferov, D. I. Schuster, and J. Simon, *arXiv:2207.10121* (2022).
- [138] A. Vainsencher, K. Satzinger, G. Peairs, and A. Cleland, *Appl. Phys. Lett.* **109**, 033107 (2016).
- [139] W. Jiang, C. J. Sarabalis, Y. D. Dahmani, R. N. Patel, F. M. Mayor, T. P. McKenna, R. Van Laer, and A. H. Safavi-Naeini, *Nat. Commun.* **11**, 1166 (2020).
- [140] M. Forsch, R. Stockill, A. Wallucks, I. Marinković, C. Gärtner, R. A. Norte, F. van Otten, A. Fiore, K. Srinivasan, and S. Gröblacher, *Nat. Phys.* **16**, 69 (2020).
- [141] X. Han, W. Fu, C. Zhong, C.-L. Zou, Y. Xu, A. A. Sayem, M. Xu, S. Wang, R. Cheng, L. Jiang, *et al.*, *Nat. Commun.* **11**, 3237 (2020).
- [142] H.-T. Tu, K.-Y. Liao, Z.-X. Zhang, X.-H. Liu, S.-Y. Zheng, S.-Z. Yang, X.-D. Zhang, H. Yan, and S.-L. Zhu, *Nat. Photon.* **16**, 291 (2022).
- [143] T. Vogt, C. Gross, J. Han, S. B. Pal, M. Lam, M. Kiffner, and W. Li, *Phys. Rev. A* **99**, 023832 (2019).
- [144] J. P. Covey, A. Sipahigil, and M. Saffman, *Phys. Rev. A* **100**, 012307 (2019).
- [145] C. O'Brien, N. Lauk, S. Blum, G. Morigi, and M. Fleischhauer, *Phys. Rev. Lett.* **113**, 063603 (2014).
- [146] X. Fernandez-Gonzalvo, Y.-H. Chen, C. Yin, S. Rogge, and J. J. Longdell, *Phys. Rev. A* **92**, 062313 (2015).
- [147] X. Fernandez-Gonzalvo, S. P. Horvath, Y.-H. Chen, and J. J. Longdell, *Phys. Rev. A* **100**, 033807 (2019).
- [148] J. G. Bartholomew, J. Rochman, T. Xie, J. M. Kindem, A. Ruskuc, I. Craiciu, M. Lei, and A. Faraon, *Nat. Commun.* **11**, 3266 (2020).
- [149] J. R. Everts, M. C. Berrington, R. L. Ahlefeldt, and J. J. Longdell, *Phys. Rev. A* **99**, 063830 (2019).
- [150] C. Cabrillo, J. I. Cirac, P. García-Fernández, and P. Zoller, *Phys. Rev. A* **59**, 1025 (1999).
- [151] J. Minář, H. De Riedmatten, C. Simon, H. Zbinden, and N. Gisin, *Phys. Rev. A* **77**, 052325 (2008).
- [152] P. C. Humphreys, N. Kalb, J. P. Morits, R. N. Schouten, R. F. Vermeulen, D. J. Twitchen, M. Markham, and R. Hanson, *Nature* **558**, 268 (2018).
- [153] S. Bose, P. L. Knight, M. B. Plenio, and V. Vedral, *Phys. Rev. Lett.* **83**, 5158 (1999).
- [154] L.-M. Duan, M. D. Lukin, J. I. Cirac, and P. Zoller, *Nature* **414**, 413 (2001).
- [155] I. E. Protsenko, G. Reymond, N. Schlosser, and P. Grangier, *Phys. Rev. A* **66**, 062306 (2002).
- [156] S. D. Barrett and P. Kok, *Phys. Rev. A* **71**, 060310 (2005).
- [157] L. Martin, F. Motzoi, H. Li, M. Sarovar, and K. B. Whaley, *Physical Review A* **92**, 062321 (2015).
- [158] D. Gottesman and I. L. Chuang, *Nature (London)* **402**, 390 (1999), *arXiv:quant-ph/9908010* [quant-ph].
- [159] L. Jiang, J. M. Taylor, A. S. Sørensen, and M. D. Lukin, *Phys. Rev. A* **76**, 062323 (2007).
- [160] D. Deutsch, A. Ekert, R. Jozsa, C. Macchiavello, S. Popescu, and A. Sanpera, *Phys. Rev. Lett.* **77**, 2818 (1996).
- [161] M. Pant, H. Krovi, D. Towsley, L. Tassioulas, L. Jiang, P. Basu, D. Englund, and S. Guha, *arXiv e-prints*, *arXiv:1708.07142* (2017), *arXiv:1708.07142* [quant-ph].
- [162] M. Tsang, *Phys. Rev. A* **84**, 043845 (2011).
- [163] W. Dür and H. J. Briegel, *Reports on Progress in Physics* **70**, 1381 (2007), *arXiv:0705.4165* [quant-ph].
- [164] A. W. Cross, L. S. Bishop, S. Sheldon, P. D. Nation, and J. M. Gambetta, *Phys. Rev. A* **100**, 032328 (2019), *arXiv:1811.12926* [quant-ph].
- [165] A. Li, S. Stein, S. Krishnamoorthy, and J. Ang, *ACM Transactions on Quantum Computing* (2022), 10.1145/3550488, just Accepted.
- [166] T. Tomesh, P. Gokhale, V. Omole, G. S. Ravi, K. N. Smith, J. Vizslai, X.-C. Wu, N. Hardavellas, M. R. Martonosi, and F. T. Chong, in *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)* (IEEE, 2022) pp. 587–603.
- [167] K. Wei, E. Magesan, I. Lauer, S. Srinivasan, D. Bogorin, S. Carnevale, G. Keefe, Y. Kim, D. Klaus, W. Landers, *et al.*, *Physical Review Letters* **129**, 060501 (2022).
- [168] A. Li, Y. Tay, A. Kumar, and H. Corporaal, in *Proceedings of the 24th international symposium on high-performance parallel and distributed computing* (2015) pp. 101–106.
- [169] W. Tang and M. Martonosi, *arXiv preprint arXiv:2205.05836* (2022).
- [170] K. Temme, S. Bravyi, and J. M. Gambetta, *Phys. Rev. Lett.* **119**, 180509 (2017).
- [171] E. van den Berg, Z. K. Mineev, A. Kandala, and K. Temme, *arXiv preprint 2201.09866* (2022).
- [172] S. Kang, R. Zhang, Z. Hao, D. Jia, F. Gao, F. Bo, G. Zhang, and J. Xu, *Opt. Lett.* **45**, 6651 (2020).

- [173] I. Krasnokutska, R. J. Chapman, J.-L. J. Tambasco, and A. Peruzzo, *Opt. Express* **27**, 17681 (2019).
- [174] J. Nauriyal, M. Song, R. Yu, and J. Cardenas, *Optica* **6**, 549 (2019).
- [175] B. B. Bakir, A. V. de Gyves, R. Orobtchouk, P. Lyan, C. Porzier, A. Roman, and J.-M. Fedeli, *IEEE Photon. Technol. Lett.* **22**, 739 (2010).
- [176] M. Li, J. Ling, Y. He, U. A. Javid, S. Xue, and Q. Lin, *Nat. Commun.* **11**, 4123 (2020).
- [177] S. Hönl, Y. Popoff, D. Caimi, A. Beccari, T. J. Kippenberg, and P. Seidler, *Nat. Commun.* **13**, 2065 (2022).
- [178] M. Xu, X. Han, C.-L. Zou, W. Fu, Y. Xu, C. Zhong, L. Jiang, and H. X. Tang, *Phys. Rev. Lett.* **124**, 033602 (2020).
- [179] Z. Wang, M. Xu, X. Han, W. Fu, S. Puri, S. Girvin, H. X. Tang, S. Shankar, and M. Devoret, *Phys. Rev. Lett.* **126**, 180501 (2021).
- [180] P. Lebrun and L. Tavian, arXiv:1501.07156 (2015).
- [181] R. Yan, G. Khalsa, S. Vishwanath, Y. Han, J. Wright, S. Rouvimov, D. S. Katzer, N. Nepal, B. P. Downey, D. A. Muller, *et al.*, *Nature* **555**, 183 (2018).
- [182] R. Cheng, J. Wright, H. G. Xing, D. Jena, and H. X. Tang, *Appl. Phys. Lett.* **117**, 132601 (2020).
- [183] S. Guha, H. Krovi, C. A. Fuchs, Z. Dutton, J. A. Slater, C. Simon, and W. Tittel, *Phys. Rev. A* **92**, 022357 (2015).
- [184] H. Bernien, B. Hensen, W. Pfaff, G. Koolstra, M. S. Blok, L. Robledo, T. H. Taminiau, M. Markham, D. J. Twitchen, L. Childress, and R. Hanson, *Nature* **497**, 86 (2013).
- [185] W. Pfaff, B. J. Hensen, H. Bernien, S. B. van Dam, M. S. Blok, T. H. Taminiau, M. J. Tiggelman, R. N. Schouten, M. Markham, D. J. Twitchen, and R. Hanson, *Science* **345**, 532 (2014), <https://www.science.org/doi/pdf/10.1126/science.1253512>.
- [186] B. Hensen, H. Bernien, A. E. Dréau, A. Reiserer, N. Kalb, M. S. Blok, J. Ruitenbergh, R. F. L. Vermeulen, R. N. Schouten, C. Abellán, W. Amaya, V. Pruneri, M. W. Mitchell, M. Markham, D. J. Twitchen, D. Elkouss, S. Wehner, T. H. Taminiau, and R. Hanson, *Nature* **526**, 682 (2015).
- [187] B. Casabone, A. Stute, K. Friebe, B. Brandstätter, K. Schüppert, R. Blatt, and T. E. Northup, *Phys. Rev. Lett.* **111**, 100505 (2013).
- [188] M. Zhang, C.-L. Zou, and L. Jiang, *Phys. Rev. Lett.* **120**, 020502 (2018).
- [189] J. Wu, C. Cui, L. Fan, and Q. Zhuang, *Phys. Rev. Appl.* **16**, 064044 (2021).
- [190] A. Rueda, W. Hease, S. Barzanjeh, and J. M. Fink, *npj Quantum Inf.* **5**, 1 (2019).
- [191] C. Zhong, Z. Wang, C. Zou, M. Zhang, X. Han, W. Fu, M. Xu, S. Shankar, M. H. Devoret, H. X. Tang, *et al.*, *Phys. Rev. Lett.* **124**, 010511 (2020).
- [192] C. Zhong, X. Han, H. X. Tang, and L. Jiang, *Phys. Rev. A* **101**, 032345 (2020).
- [193] N. Ofek, A. Petrenko, R. Heeres, P. Reinhold, Z. Leghtas, B. Vlastakis, Y. Liu, L. Frunzio, S. M. Girvin, L. Jiang, M. Mirrahimi, M. H. Devoret, and R. J. Schoelkopf, *Nature* **536** (2016), 10.1038/nature18949.
- [194] V. V. Sivak, A. Eickbusch, B. Royer, S. Singh, I. Tsioutsios, S. Ganjam, A. Miano, B. L. Brock, A. Z. Ding, L. Frunzio, S. M. Girvin, R. J. Schoelkopf, and M. H. Devoret, arXiv e-prints, arXiv:2211.09116 (2022), arXiv:2211.09116 [quant-ph].
- [195] L. Hu, Y. Ma, W. Cai, X. Mu, Y. Xu, W. Wang, Y. Wu, H. Wang, Y. P. Song, C. L. Zou, S. M. Girvin, L. M. Duan, and L. Sun, *Nature Physics* **15**, 503 (2019), arXiv:1805.09072 [quant-ph].
- [196] N. E. Frattini, U. Vool, S. Shankar, A. Narla, K. M. Sliwa, and M. H. Devoret, *Applied Physics Letters* **110** (2017), 10.1063/1.4984142.
- [197] E. McKinney, M. Xia, C. Zhou, P. Lu, M. Hatridge, and A. K. Jones, arXiv preprint arXiv:2205.04387 (2022).
- [198] M. Wallquist, K. Hammerer, P. Rabl, M. Lukin, and P. Zoller, *Physica Scripta Volume T* **137**, 014001 (2009), arXiv:0911.3835 [quant-ph].
- [199] N. Daniilidis and H. Häffner, *Annual Review of Condensed Matter Physics* **4**, 83 (2013), <https://doi.org/10.1146/annurev-conmatphys-030212-184253>.
- [200] D. De Motte, A. R. Grounds, M. Reháč, A. Rodríguez Blanco, B. Lekitsch, G. S. Giri, P. Neillinger, G. Oelsner, E. Il'ichev, M. Grajcar, and W. K. Hensinger, *Quantum Information Processing* **15**, 5385 (2016), arXiv:1510.07298 [quant-ph].
- [201] D. Kielpinski, D. Kafri, M. J. Woolley, G. J. Milburn, and J. M. Taylor, *Phys. Rev. Lett.* **108**, 130504 (2012), arXiv:1111.5999 [quant-ph].
- [202] R. J. Niffenegger, J. Stuart, C. Sorace-Agaskar, D. Kharras, S. Bramhavar, C. D. Bruzewicz, W. Loh, R. T. Maxson, R. McConnell, D. Reens, G. N. West, J. M. Sage, and J. Chiaverini, *Nature (London)* **586**, 538 (2020), arXiv:2001.05052 [quant-ph].
- [203] K. K. Mehta, C. Zhang, M. Malinowski, T.-L. Nguyen, M. Stadler, and J. P. Home, *Nature (London)* **586**, 533 (2020), arXiv:2002.02258 [quant-ph].
- [204] P. Maunz, S. Olmschenk, D. Hayes, D. N. Matsukevich, L. M. Duan, and C. Monroe, *Phys. Rev. Lett.* **102**, 250502 (2009), arXiv:0902.2136 [quant-ph].
- [205] J. Verdú, H. Zoubi, C. Koller, J. Majer, H. Ritsch, and J. Schmiedmayer, *Phys. Rev. Lett.* **103**, 043603 (2009).
- [206] D. I. Schuster, L. S. Bishop, I. L. Chuang, D. DeMille, and R. J. Schoelkopf, *Phys. Rev. A* **83**, 012311 (2011).
- [207] A. Andre, D. DeMille, J. M. Doyle, M. D. Lukin, S. E. Maxwell, P. Rabl, R. Schoelkopf, and P. Zoller, arXiv e-prints, quant-ph/0605201 (2006), arXiv:quant-ph/0605201 [quant-ph].
- [208] S. X. Wang, Y. Ge, J. Labaziewicz, E. Dauler, K. Berggren, and I. L. Chuang, *Applied Physics Letters* **97**, 244102 (2010), arXiv:1010.6108 [quant-ph].
- [209] P. Rabl, D. DeMille, J. M. Doyle, M. D. Lukin, R. J. Schoelkopf, and P. Zoller, *Phys. Rev. Lett.* **97**, 033003 (2006).
- [210] P. Rabl and P. Zoller, *Phys. Rev. A* **76**, 042308 (2007).
- [211] L. Lamata, D. R. Leibbrandt, I. L. Chuang, J. I. Cirac, M. D. Lukin, V. Vuletić, and S. F. Yelin, *Phys. Rev. Lett.* **107**, 030501 (2011).
- [212] N. H. Nickerson, Y. Li, and S. C. Benjamin, *Nat. Commun.* **4**, 1756 (2013).
- [213] K. Fujii and K. Yamamoto, *Phys. Rev. A* **80**, 042308 (2009).
- [214] W. Dür and H. J. Briegel, *Reports on Progress in Physics* **70**, 1381 (2007).
- [215] S. Krastanov, V. V. Albert, and L. Jiang, *Quantum* **3**, 123 (2019).
- [216] J. M. Gertler, B. Baker, J. Li, S. Shirol, J. Koch, and C. Wang, *Nature* **590**, 243 (2021).

- [217] Q. Ficheux, L. B. Nguyen, A. Somoroff, H. Xiong, K. N. Nesterov, M. G. Vavilov, and V. E. Manucharyan, *Phys. Rev. X* **11**, 021026 (2021).
- [218] E. Dogan, D. Rosenstock, L. L. Guevel, H. Xiong, R. A. Mencia, A. Somoroff, K. N. Nesterov, M. G. Vavilov, V. E. Manucharyan, and C. Wang, arXiv preprint arXiv:2204.11829 (2022).
- [219] A. P. Read, B. J. Chapman, C. U. Lei, J. C. Curtis, S. Ganjam, L. Krayzman, L. Frunzio, and R. J. Schoelkopf, arXiv preprint arXiv:2206.14334 (2022).
- [220] C. Wang, C. Axline, Y. Y. Gao, T. Brecht, Y. Chu, L. Frunzio, M. H. Devoret, and R. J. Schoelkopf, *Applied Physics Letters* **107**, 162601 (2015).
- [221] M. Reagor, W. Pfaff, C. Axline, R. W. Heeres, N. Ofek, K. Sliwa, E. Holland, C. Wang, J. Blumoff, K. Chou, M. J. Hatridge, L. Frunzio, M. H. Devoret, L. Jiang, and R. J. Schoelkopf, *Phys. Rev. B* **94**, 014506 (2016).
- [222] A. Romanenko, R. Pilipenko, S. Zorzetti, D. Frolov, M. Awida, S. Belomestnykh, S. Posen, and A. Grassellino, *Phys. Rev. Applied* **13**, 034032 (2020).
- [223] S. Chakram, A. E. Oriani, R. K. Naik, A. V. Dixit, K. He, A. Agrawal, H. Kwon, and D. I. Schuster, *Phys. Rev. Lett.* **127**, 107701 (2021).
- [224] S. Chakram, K. He, A. V. Dixit, A. E. Oriani, R. K. Naik, N. Leung, H. Kwon, W.-L. Ma, L. Jiang, and D. I. Schuster, *Nature Physics* **18**, 879 (2022).
- [225] Y. Y. Gao, B. J. Lester, Y. Zhang, C. Wang, S. Rosenblum, L. Frunzio, L. Jiang, S. M. Girvin, and R. J. Schoelkopf, *Physical Review X* **8**, 021073 (2018).
- [226] S. de Graaf, B. J. Chapman, J. C. Curtis, Y. Zhang, N. E. Frattini, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf, *Bulletin of the American Physical Society* (2022).
- [227] C. Guinn, S. Sussman, P. S. Mundada, A. Vrajitoarea, C. Leroux, A. Place, C. L. Calonne, A. Di Paolo, A. Petrescu, A. Blais, and A. A. Houck, *Bulletin of the American Physical Society* (2022).
- [228] M. Reagor, C. B. Osborn, N. Tezak, A. Staley, G. Prawiroatmodjo, M. Scheer, N. Alidoust, E. A. Sete, N. Didier, M. P. d. Silva, E. Acala, J. Angeles, A. Bestwick, M. Block, B. Bloom, A. Bradley, C. Bui, S. Caldwell, L. Capelluto, R. Chilcott, J. Cordova, G. Crossman, M. Curtis, S. Deshpande, T. E. Bouayadi, D. Gershovich, S. Hong, A. Hudson, P. Karalekas, K. Kuang, M. Lenihan, R. Manenti, T. Manning, J. Marshall, Y. Mohan, W. O'Brien, J. Otterbach, A. Papageorge, J.-P. Paquette, M. Pelstring, A. Polloreno, V. Rawat, C. A. Ryan, R. Renzas, N. Rubin, D. Russel, M. Rust, D. Scarabelli, M. Selvanayagam, R. Sinclair, R. Smith, M. Suska, T.-W. To, M. Vahidpour, N. Vodrahalli, T. Whyland, K. Yadav, W. Zeng, and C. T. Rigetti, *Science Advances* **4**, eaao3603 (2018).
- [229] B. M. Terhal, J. Conrad, and C. Vuillot, "Towards scalable bosonic quantum error correction," (2020).
- [230] W. Cai, Y. Ma, W. Wang, C.-L. Zou, and L. Sun, *Fundamental Research* **1**, 50 (2021).
- [231] A. Joshi, K. Noh, and Y. Y. Gao, "Quantum information processing with bosonic qubits in circuit qed," (2021).
- [232] L. Hu, Y. Ma, W. Cai, X. Mu, Y. Xu, W. Wang, Y. Wu, H. Wang, Y. P. Song, C. L. Zou, S. M. Girvin, L. M. Duan, and L. Sun, *Nature Physics* **15** (2019), 10.1038/s41567-018-0414-3.
- [233] P. Campagne-Ibarcq, A. Eickbusch, S. Touzard, E. Zalys-Geller, N. E. Frattini, V. V. Sivak, P. Reinhold, S. Puri, S. Shankar, R. J. Schoelkopf, L. Frunzio, M. Mirrahimi, and M. H. Devoret, *Nature* **584** (2020), 10.1038/s41586-020-2603-3.
- [234] S. Rosenblum, P. Reinhold, M. Mirrahimi, L. Jiang, L. Frunzio, and R. J. Schoelkopf, *Science* **361** (2018), 10.1126/science.aat3996.
- [235] S. Puri, A. Grimm, P. Campagne-Ibarcq, A. Eickbusch, K. Noh, G. Roberts, L. Jiang, M. Mirrahimi, M. H. Devoret, and S. M. Girvin, *Phys. Rev. X* **9**, 041009 (2019).
- [236] A. L. Grimsmo and S. Puri, *PRX Quantum* **2**, 020101 (2021).
- [237] M. Y. Siraichi, V. F. d. Santos, C. Collange, and F. M. Q. Pereira, in *Proceedings of the 2018 International Symposium on Code Generation and Optimization* (2018) pp. 113–125.
- [238] A. Cowtan, S. Dilkes, R. Duncan, A. Krajenbrink, W. Simmons, and S. Sivarajah, arXiv preprint arXiv:1902.08091 (2019).
- [239] É. Bonnet, T. Miltzow, and P. Rzażewski, *Algorithmica* **80**, 2656 (2018).
- [240] A. Javadi-Abhari, *Towards a scalable software stack for resource estimation and optimization in general-purpose quantum computers*, Ph.D. thesis, Princeton University (2017).
- [241] B. Tan and J. Cong, *IEEE Transactions on Computers* **70**, 1363 (2020).
- [242] S. Kumar, Y. Sabharwal, R. Garg, and P. Heidelberger, in *2008 37th International Conference on Parallel Processing (IEEE, 2008)* pp. 320–329.
- [243] T. Fujiwara, P. Malakar, K. Reda, V. Vishwanath, M. E. Papka, and K.-L. Ma, in *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)* (IEEE, 2017) pp. 59–70.
- [244] M. Orenes-Vera, A. Manocha, J. Balkind, F. Gao, J. L. Aragón, D. Wentzlaff, and M. Martonosi, in *ISCA* (2022) pp. 817–830.
- [245] F. Pellegrini, *Combinatorial Scientific Computing*, 373 (2012).
- [246] G. Karypis and V. Kumar, *SIAM Journal on scientific Computing* **20**, 359 (1998).
- [247] P. Andres-Martinez and C. Heunen, *Physical Review A* **100**, 032308 (2019).
- [248] Z. Davarzani, M. Zomorodi-Moghadam, M. Houshmand, and M. Nouri-baygi, *Quantum Information Processing* **19**, 1 (2020).
- [249] T. Häner, D. S. Steiger, T. Hoefler, and M. Troyer, in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (2021) pp. 1–13.
- [250] Y. Shi, P. Gokhale, P. Murali, J. M. Baker, C. Duckering, Y. Ding, N. C. Brown, C. Chamberland, A. Javadi-Abhari, A. W. Cross, *et al.*, *Proceedings of the IEEE* **108**, 1353 (2020).
- [251] A. B. Yoo, M. A. Jette, and M. Grondona, in *Workshop on job scheduling strategies for parallel processing* (Springer, 2003) pp. 44–60.
- [252] R. V. Meter, W. Munro, K. Nemoto, and K. M. Itoh, *ACM Journal on Emerging Technologies in Computing Systems (JETC)* **3**, 1 (2008).

- [253] A. Eddins, M. Motta, T. P. Gujarati, S. Bravyi, A. Mezzacapo, C. Hadfield, and S. Sheldon, *PRX Quantum* **3**, 010309 (2022).
- [254] M. A. Rowe, A. Ben-Kish, B. Demarco, D. Leibfried, V. Meyer, J. Beall, J. Britton, J. Hughes, W. M. Itano, B. Jelenković, C. Langer, T. Rosenband, and D. J. Wineland, *Quantum Info. Comput.* **2**, 257–271 (2002).
- [255] D. D. Thaker, T. S. Metodi, A. W. Cross, I. L. Chuang, and F. T. Chong, in *33rd International Symposium on Computer Architecture (ISCA'06)* (IEEE, 2006) pp. 378–390.
- [256] L. K. Grover, arXiv preprint quant-ph/9704012 (1997).
- [257] A. Yimsiriwattana and S. J. Lomonaco Jr, in *Quantum Information and Computation II*, Vol. 5436 (SPIE, 2004) pp. 360–372.
- [258] R. D. V. Meter III, arXiv preprint quant-ph/0607065 (2006).
- [259] S. Stein, N. Wiebe, Y. Ding, P. Bo, K. Kowalski, N. Baker, J. Ang, and A. Li, in *Proceedings of the 49th Annual International Symposium on Computer Architecture* (2022) pp. 59–71.
- [260] S. DiAdamo, M. Ghibaudi, and J. Cruise, arXiv preprint arXiv:2101.02504 (2021).
- [261] M. Reiher, N. Wiebe, K. M. Svore, D. Wecker, and M. Troyer, *Proceedings of the national academy of sciences* **114**, 7555 (2017).
- [262] E. Knill, G. Ortiz, and R. D. Somma, *Physical Review A* **75**, 012328 (2007).
- [263] G. Kotliar, S. Y. Savrasov, K. Haule, V. S. Oudovenko, O. Parcollet, and C. A. Marianetti, *Rev. Mod. Phys.* **78**, 865 (2006).
- [264] N. Lanatà, Y. Yao, C.-Z. Wang, K.-M. Ho, and G. Kotliar, *Phys. Rev. X* **5**, 011008 (2015).
- [265] B. Bauer, D. Wecker, A. J. Millis, M. B. Hastings, and M. Troyer, *Phys. Rev. X* **6**, 031045 (2016).
- [266] Y. Yao, F. Zhang, C.-Z. Wang, K.-M. Ho, and P. P. Orth, *Phys. Rev. Research* **3**, 013184 (2021).
- [267] M. Welborn, T. Tsuchimochi, and T. Van Voorhis, *The Journal of Chemical Physics* **145**, 074102 (2016).
- [268] M. Motta, C. Sun, A. T. K. Tan, M. J. O'Rourke, E. Ye, A. J. Minnich, F. G. S. L. Brandão, and G. K.-L. Chan, *Nature Physics* **16**, 205 (2020), arXiv:1901.07653 [quant-ph].
- [269] H.-C. Zhou, J. R. Long, and O. M. Yaghi, *Chemical reviews* **112**, 673 (2012).
- [270] H. Fu, H. Chen, M. Blazhynska, E. Goulard Coderc de Lacam, F. Szczepaniak, A. Pavlova, X. Shao, J. C. Gumbart, F. Dehez, B. Roux, *et al.*, *Nature Protocols* **17**, 1114 (2022).
- [271] A. Warshel, *Angewandte Chemie International Edition* **53**, 10020 (2014).
- [272] E. P. E. K. A. L. D. Vollhardt, ed., *DMFT: From Infinite Dimensions to Real Materials: Lecture Notes of the Autumn School on Correlated Electrons* (Institute for Advanced Simulation and German Research School for Simulation Sciences, 2018).
- [273] D. Goldhaber-Gordon, H. Shtrikman, D. Mahalu, D. Abusch-Magder, U. Meirav, and M. A. Kastner, *Nature (London)* **391**, 156 (1998), arXiv:cond-mat/9707311 [cond-mat.str-el].
- [274] S. Gustavsson, R. Leturcq, M. Studer, I. Shorubalko, T. Ihn, K. Ensslin, D. C. Driscoll, and A. C. Gossard, *Surface Science Reports* **64**, 191 (2009), arXiv:0905.4675 [cond-mat.mes-hall].
- [275] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, *et al.*, *Journal of physics: Condensed matter* **21**, 395502 (2009).
- [276] L. Clinton, T. Cubitt, B. Flynn, F. M. Gambetta, J. Klassen, A. Montanaro, S. Piddock, R. A. Santos, and E. Sheridan, arXiv preprint arXiv:2205.15256 (2022).
- [277] D. Marx and J. Hutter, *Ab initio molecular dynamics: basic theory and advanced methods* (Cambridge University Press, 2009).
- [278] J. R. McClean, J. A. Parkhill, and A. Aspuru-Guzik, *Proceedings of the National Academy of Sciences* **110**, E3901 (2013).
- [279] V. Abraham and N. J. Mayhall, *Journal of Chemical Theory and Computation* **16**, 6098 (2020), pMID: 32846094, <https://doi.org/10.1021/acs.jctc.0c00141>.
- [280] Y. Zhang, L. Cincio, C. F. A. Negre, P. Czarnik, P. Coles, P. M. Anisimov, S. M. Mniszewski, S. Tretiak, and P. A. Dub, arXiv e-prints, arXiv:2106.07619 (2021).
- [281] P. Magnard, S. Storz, P. Kurpiers, J. Schär, F. Marxer, J. Lütolf, T. Walter, J. C. Besse, M. Gabureac, K. Reuer, A. Akin, B. Royer, A. Blais, and A. Wallraff, *Phys. Rev. Lett.* **125**, 260502 (2020), arXiv:2008.01642 [quant-ph].
- [282] C. B. Young, A. Safari, P. Huft, J. Zhang, E. Oh, R. Chinnarasu, and M. Saffman, *Appl. Phys. B* **128**, 151 (2022), arXiv:2202.01634 [quant-ph].
- [283] D. Bluvstein *et al.*, *Nature* **604**, 451 (2022), arXiv:2112.03923 [quant-ph].
- [284] C. Schuck, X. Guo, L. Fan, X. Ma, M. Poot, and H. X. Tang, *Nature Communications* **7**, 10352 (2016), arXiv:1511.07081 [quant-ph].
- [285] T. Neuman, M. Eichenfield, M. E. Trusheim, L. Hackett, P. Narang, and D. Englund, *npj Quantum Information* **7**, 121 (2021).
- [286] K. Fang, X. Wang, M. Tomamichel, and R. Duan, *IEEE Transactions on Information Theory* **65**, 6454 (2019).
- [287] E. Campbell, *Physical review letters* **123**, 070503 (2019).
- [288] J. Lee, D. W. Berry, C. Gidney, W. J. Huggins, J. R. McClean, N. Wiebe, and R. Babbush, *PRX Quantum* **2**, 030305 (2021).
- [289] C. M. Natarajan, M. G. Tanner, and R. H. Hadfield, *Supercond. Sci. Technol.* **25**, 063001 (2012).
- [290] J. R. Johansson, P. D. Nation, and F. Nori, *Comput. Phys. Commun.* **183**, 1760 (2012).
- [291] R. Gao, N. Yao, J. Guan, L. Deng, J. Lin, M. Wang, L. Qiao, W. Fang, and Y. Cheng, *Chin. Opt. Lett.* **20**, 011902 (2022).
- [292] R. Zhuang, J. He, Y. Qi, and Y. Li, *Adv. Mater.*, 2208113 (2022).
- [293] W. Jin, Q.-F. Yang, L. Chang, B. Shen, H. Wang, M. A. Leal, L. Wu, M. Gao, A. Feshali, M. Paniccia, *et al.*, *Nat. Photon.* **15**, 346 (2021).
- [294] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, *Nature communications* **5**, 1 (2014).
- [295] X. Chen, M. M. Milosevic, S. Stanković, S. Reynolds, T. D. Bucio, K. Li, D. J. Thomson, F. Gardes, and G. T. Reed, *Proceedings of the IEEE* **106**, 2101 (2018).
- [296] W. Bogaerts, D. Pérez, J. Capmany, D. A. Miller, J. Poon, D. Englund, F. Morichetti, and A. Melloni,



Nature **586**, 207 (2020).

- [297] C. Monroe, R. Raussendorf, A. Ruthven, K. Brown, P. Maunz, L.-M. Duan, and J. Kim, Physical Review A **89**, 022317 (2014).

## Appendix A: Distributed QPE

Quantum phase estimation (QPE) is perhaps surprisingly one of the easiest subroutines in quantum computing to distribute in MNQC systems. This is in spite of the fact that QPE is often viewed as a high circuit depth algorithm. In this section we will discuss two approaches for distributing quantum phase estimation. Two strategies exist for such parallelism: the fully coherent approach of [262] which gives a reduction in the depth of phase estimation that is linear in the number of nodes and the approach that uses classical communication (found in the supplementary material of [261]). Our aim in this section is to review these approaches and place bounds on the channel fidelity needed to see an advantage from the former approach.

The task of phase estimation is to provide an estimate of an eigenphase of a unitary operation  $U$ . Specifically, assume that for unitary  $U \in \mathbb{C}^{2^n \times 2^n}$  that  $|\psi_n\rangle$  are eigenvectors such that  $U|\psi_n\rangle = e^{i\theta_n}|\psi_n\rangle$  for real valued  $\theta_n$ . The aim of the phase estimation problem is to find, for any  $\epsilon > 0$  and probability of success at least  $2/3$ , an estimate  $\hat{\theta}_k$  such that there exists  $\theta_k$  that obeys  $|\hat{\theta}_k - \theta_k| \leq \epsilon$ . In practice, the phase estimation problem is usually more specific and a particular eigenphase is desired. In which, case the user must provide a quantum state that has high-overlap with the target eigenstate for this protocol to succeed with high probability.

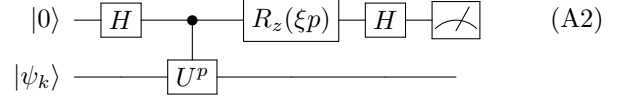
The central challenge of phase estimation is that the optimal scaling is given by the Heisenberg limit  $|\hat{\theta}_k - \theta_k| \leq \frac{\pi}{T}$  for any quantum algorithm that estimates  $\theta_k$  using  $T$  applications of the unitary  $U$ . For applications in chemistry, these errors need to be on the order of  $10^{-4}$  or smaller [261], necessitating a large number of applications of the underlying unitary. Our aim is to distribute these executions of the unitary over the network in such a way so that the phase estimation can be performed in low depth. Specifically, if we will see that if we have an MNQC then in the most extreme case we can use  $T$  nodes to perform phase estimation in  $O(1)$  time and in the case where classical interconnects are used then  $T^2$  nodes suffices to achieve the same bound.

We will begin with the simplest case wherein each node can only communicate classically with each of the other nodes. Let us assume that in each case a quantum state  $|\phi_k\rangle$  can be prepared such that  $|\langle \phi_k | \psi_k \rangle|^2 = 1 - \delta$  for target state  $|\psi_k\rangle$ . Further, let us assume that for all  $j \neq k$ ,  $|\theta_j - \theta_k| \geq \epsilon_\theta$ . We begin our protocol by applying phase estimation to prepare each of the states on the  $T^2$  nodes. This requires  $O(\log(1/\gamma)/\epsilon_\theta)$  number of applications of the underlying  $U$  on each node to ensure the correct eigenvalue with probability of failure at most  $1 - \gamma$  using conventional phase estimation. The number of trials needed per node before a successful state preparation is observed is geometrically distributed. If  $\gamma \geq 2/3$  then the probability distribution function shows that the number of trials needed before the probability of observing no successful preparations is  $O(1/T^2)$  is  $O(\log(T^2)/\delta)$ .

From the union bound, the probability of any of the runs requiring more than this is  $O(1)$  and thus the total depth (as quantified by the number of unitary circuits applied to prepare the  $O(T^2)$  independent eigenstates is in

$$O\left(\frac{\log(T)}{\delta\epsilon_\theta}\right) \quad (\text{A1})$$

Next we need to invoke the parallelized phase estimation procedure of [261]. Each experiment in this algorithm involves communicating to each of the nodes and requesting it to perform  $U^p$  for some value of  $p$  and measuring the result in the iterative phase estimation circuit, which takes the following form



The likelihood of measuring zero for this circuit is  $\cos^2(p(\theta - \xi)/2)$ . The circuit is repeated  $T^2$  times, one for each node, and the results are communicated back to the classical head node. From the central limit theorem, the distribution on the number of zeros observed out of the  $T^2$  measurements is approximately a Gaussian with mean  $T^2 \cos^2(p(\theta - \xi)/2)$ . As the likelihood is approximately Gaussian, the method of conjugate priors can be used to efficiently update an initially Gaussian prior distribution on  $\theta$  to find a posterior distribution in polynomial time (alternatively if one does not want to use the central limit theorem the Monte-Carlo methods of [261] can be employed). By choosing  $p$  adaptively using the heuristic in [261] they show that  $O(\log(T))$  suffices to achieve error in the estimate  $O(1/T)$ . Further each such experiment requires evolution time at most  $O(T/\sqrt{C})$  where  $C$  is the number of nodes devoted to the phase estimation. In cases where  $C = O(T^2)$ , such as our setting, this implies that the depth of each phase estimation job (as quantified by the number of sequential operations of  $U$ ) is  $O(1)$ .

The above tasks are repeated  $O(\log(T))$  times by each node and therefore the depth of the phase estimation algorithm once the state  $|\psi_k\rangle$  is prepared on each node. The depth of the classical communication version of the QPE algorithm appropriate for a MNQC setting with  $T^2 = O(1/\epsilon^2)$  is dominated by the state preparation step, which can be performed using  $T^2$  workers in depth

$$\text{Depth}_{U,C1} = O\left(\frac{\log(1/\epsilon)}{\delta\epsilon_\theta}\right). \quad (\text{A3})$$

Note that through the use of fixed point amplitude amplification rather than statistical sampling, the depth can further be reduced by to  $O\left(\frac{\log(1/\epsilon)\text{polylog}(1/\epsilon_\theta)}{\sqrt{\delta\epsilon_\theta}}\right)$ ; however, the use of this technique will require additional ancillae and comparison logic to implement the required reflections about the estimated energy returned by a coherent (as opposed to iterative) phase estimation procedure

and thus we focus our attention on the non-amplified case.

As no quantum communication is needed for this algorithm there are no further errors if we assume that we are working in a model wherein all intra-node operations are error free but inter-node operations have intrinsic error associated with them. This also makes this application a good baseline comparison to judge the quantum version of phase estimation.

The fully coherent version of distributed quantum phase estimation takes the form in Figure 21 [262]. It then follows from noting that the circuit returns the phase  $e^{i3\theta_k}$  from the phase kickback effect that in general this idea can be repeated  $p$  times to obtain  $p$  times the phase that would be seen with a single step of an iterative phase estimation procedure. Specifically, in both cases the probability of measuring zero is  $\cos^2(p(\theta_k - \xi)/2)$  per experiment.

The protocol for implementing this circuit on a quantum MNQC works as follows.

1. In parallel prepare a state  $|\phi_k\rangle$  on each of  $T$  nodes on the quantum computer.
2. Use the above phase estimation procedure and prior knowledge of  $\theta_k$  to ensure that each state is  $|\psi_k\rangle$  with probability  $1 - O(1/T)$ .
3. For each invocation of the circuit of (A2) in the implementation of an iterative phase estimation algorithm (such as Robust Phase Estimation) replace the circuit with the following procedure:
  - (a) Prepare a  $T$  qubit GHZ state on the head node.
  - (b) Send one qubit of the GHZ state to each of the  $T$  worker nodes.
  - (c) For each worker, apply controlled  $U$  using the share of the GHZ state to their state  $|\psi_k\rangle$ .
  - (d) Return all qubits to head node.
  - (e) Apply single qubit rotation (if required by iterative phase estimation protocol), invert GHZ preparation and measure qubit 0 and return result as outcome of measurement for the step of ITPE.

The above protocol works because the likelihood as argued above is precisely the same in the distributed algorithm as it would be in the ordinary algorithm for phase estimation. As the core element of an iterative phase estimation procedure is the inference of the most likely eigenphase given a set of experimental data, the inference procedure will take precisely the same form since the likelihood function is the same. Thus the protocol allows us to trivially parallelize any iterative phase estimation procedure over the  $T$  workers.

Iterative phase estimation procedures such as Robust Phase Estimation require  $O(\log(T))$  rounds if we desire an error of  $O(1/T)$ . Each such round can be executed

in constant depth (as measured by the number of layers of controlled  $U$  gates executed). It further requires  $2T$  applications of a communication channel from the head node to the workers. For simplicity, let us assume that the interaction graph is star graph wherein the root is the head node so that all workers can directly communicate with the head node. In settings where a more restricted topology is present, the communication will need to be chained between the workers involved to distribute the GHZ state. Regardless, the total number of bits that need to be sent by the protocol is in  $O(T \log(T))$  and the overall depth as mentioned is logarithmic. Thus, assuming that the cost of any entanglement distillation is negligible, the overall depth of the algorithm is also

$$\text{Depth}_{U, \text{Qm}} = O\left(\frac{\log(1/\epsilon)}{\delta\epsilon_\theta}\right); \quad (\text{A4})$$

however, the number of workers needed to achieve this limit is quadratically smaller than the case where only classical communication is permitted.

Next let us assume that the channel that describes communication between the head node and the workers is within diamond distance  $\Delta$  from the ideal channel. That is to say if  $\Lambda$  is the ideal channel that swaps a qubit between the two nodes and  $\tilde{\Lambda}$  is the actual quantum channel then  $\|\Lambda - \tilde{\Lambda}\|_\diamond \leq \Delta$ . Here the diamond norm is the supremum of the induced trace norm between the inputs and the outputs of the channel when all possible input states (including states that are entangled with qubits that are not put through the channel) are considered. An important property of the diamond norm is that it is sub-additive meaning that for any positive integer  $m$  the composition of  $m$  channels obeys

$$\|\Lambda^{om} - \tilde{\Lambda}^{om}\|_\diamond \leq m\Delta. \quad (\text{A5})$$

Thus by the von Neumann trace inequality, for any observable  $Q$  and input state  $\rho$

$$\|\text{Tr}(\Lambda^{om}(\rho)Q) - \text{Tr}(\tilde{\Lambda}^{om}(\rho)Q)\| \leq m\|Q\|\Delta. \quad (\text{A6})$$

Thus as the observable for phase estimation has norm at most  $\pi$  it follows that the maximum error that is observable from the invocation of the channel in this fashion is  $m\pi\Delta$ . This implies that if we wish the error in the estimated phase to be at most  $\epsilon$  from communication between the head node and the workers then it suffices to take

$$\Delta = \frac{\epsilon}{m\pi} = O\left(\frac{\epsilon}{T \log(T)}\right) \quad (\text{A7})$$

Setting  $T = O(1/\epsilon)$  as well suffices to remove the  $O(1/\epsilon)$  overhead from phase estimation from the circuit depth while guaranteeing that we hit a fixed accuracy target

$$\Delta = \frac{\epsilon}{m\pi} = O\left(\frac{\epsilon^2}{\log(1/\epsilon)}\right) \quad (\text{A8})$$

This suggests that the error in the quantum communication channel must be exceptionally small in order guarantee (without further assumptions) that the overall error in the phase estimation protocol is small. Further, such applications are likely to be impractical without entanglement distillation or possibly virtual distillation.

Given that  $\epsilon$  is sufficiently low, entanglement distillation can be used to implement this channel. In order to distill states with this level of error we need  $O(T \text{polylog}(T \log(T)/\epsilon)) = \tilde{O}(T \text{polylog}(1/\epsilon))$  noisy uses of a channel connecting the head node with the workers in order to distill high enough fidelity states to teleport within the desired accuracy [286]. Thus if we assume that the depth of required to communicate between the nodes is  $\gamma \geq 0$  times the depth required to implement  $U$  (where  $\gamma$  will often but not always be less than 1) the total depth of the algorithm using  $T$  workers is

$$\text{Depth}_{U,Dist} = O\left(\frac{\log(1/\epsilon)}{\delta\epsilon} + \frac{1}{\epsilon T} + \gamma T \text{polylog}(T/\epsilon)\right) \quad (\text{A9})$$

This shows that as the number of workers increases, a favorable tradeoff in the depth of the circuit can be achieved. Specifically, such an optimal tradeoff is obtained when  $T \approx \Theta(\sqrt{1/(\epsilon\gamma)})$ . Given this choice, the optimized depth reads

$$\text{Depth}_{U,Dist}^{\text{opt}} = \tilde{O}\left(\frac{\log(1/\epsilon)}{\delta\epsilon} + \sqrt{\frac{\gamma}{\epsilon}}\right). \quad (\text{A10})$$

Thus if  $\gamma$  is viewed as a constant, then this approach can attain quadratically better depth scaling than with the error tolerance than the naïve phase estimation algorithm permits. However, this is not necessarily better than the case where no quantum interconnects are used if  $\gamma$  is not sufficiently small.

In order to understand the gulf between this let us assume that the phase estimation step used to validate the eigenstate has circuit depth  $\alpha \log(1/\epsilon)$ , where in the case of Hamiltonian simulation  $\alpha$  would be the sums of the absolute values of the coefficients and corresponds to a simulation method such as qubitization being used. Next, let us assume that each of the workers uses a low-order method such as Qdrift [287] to perform the simulation. In this case, we would take the Qdrift approximation to  $e^{-iHt}$  for a sufficiently short value of  $t$  and perform phase estimation on the result to precision  $\epsilon t$ .

The work of [288] shows that  $O(\alpha^4/\epsilon^4)$  exponentials need to be simulated to perform phase estimation to within error  $\epsilon$  using Qdrift. If we assume that we can parallelize  $T$  of them over our workers, the combined cost of phase estimation becomes

$$\text{Depth}_{U,Dist} = O\left(\frac{\alpha \log(1/\epsilon)}{\delta E_{\text{gap}}} + \frac{\alpha^4}{\epsilon^4 T} + \gamma T \text{polylog}(T/\epsilon)\right) \quad (\text{A11})$$

In the limit of negligible  $\gamma$ , this protocol can achieve depth  $\alpha \log(1/\epsilon)/\delta E_{\text{gap}}$  by choosing  $T = \alpha^3(\delta E_{\text{gap}})/\epsilon^4$ .

This shows that in a regime where parallelism is cheap that a simulation experiment can be carried out whose depth scales only with that required to verify that each worker possesses a copy of the groundstate. Note that by replacing QDrift with another simulation algorithm, such as Trotter formulas or Qubitization, we cannot get the same depth optimal result because we will be limited by the circuit depth needed to implement those protocols. A single segment of QDrift can be executed in constant depth and thus is the only known algorithm that can meet the above scaling.

## Appendix B: M2O Conversion Simulation

The simulation model used for entangled generation is shown in Fig. 22(a) [24]. At both nodes, the qubit and a microwave photon are prepared in an entangled state  $|\phi_0\rangle = \sqrt{1-P_e}|g0\rangle + \sqrt{P_e}|e1\rangle$ , where  $0 \leq P_e \leq 0.5$  is the probability of excited qubit-microwave photon state and is experimentally tunable [24]. The microwave photons are up-converted to optical photons and then interfere in a beamsplitter. The M2O converters are phenomenologically modeled as a series of beamsplitters as shown in Fig. 22(b). The first beamsplitter has a power transmission of  $T_e = \gamma_{\text{ext},e}/\gamma_{\text{tot},e}$ , where  $T_e$  is the extraction efficiency of the microwave resonator,  $\gamma_{\text{ext},e}$  is the external coupling rate of the microwave resonator,  $\gamma_{\text{int},e}$  is the intrinsic decay rate of the microwave resonator, and  $\gamma_{\text{tot},e} = \gamma_{\text{ext},e} + \gamma_{\text{int},e}$  is the total decay rate of the microwave resonator. Due to the pump-induced heating, the microwave resonator suffers from thermal added noise, and we model it as a thermal state  $\rho_{\text{th}}(n_{\text{add}})$  at another input port of the beamsplitter, and its mean photon

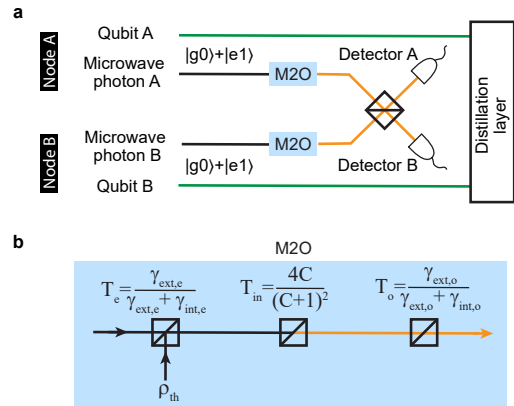


FIG. 22. (a), the diagram of the scheme used for remote entangled qubit generation. (b), the M2O converter is modeled by a series of beamsplitters. The first beamsplitter represents the microwave resonator extraction efficiency, the second beamsplitter represents the intracavity M2O conversion efficiency, and the third beamsplitter represents the optical resonator extraction efficiency. Thermal added noise is modeled by a thermal state  $\rho_{\text{th}}$  at the first beamsplitter.

number is  $n_{\text{add}}$ . In our simulation, we assume  $n_{\text{add}}$  depends linearly on the optical pump power  $P$  as  $n_{\text{add}} = k_{\text{add}}P$ , and we assume  $n_{\text{add}} = 1$  photon when  $P = 1$  mW (i.e.,  $k_{\text{add}} = 1$  photon/1 mW) [124]. The second beamsplitter has a power transmission of  $T_{\text{in}} = 4C/(C+1)^2$ , where  $T_{\text{in}}$  represents the intracavity M2O conversion efficiency for electro-optics converters,  $C = 4g^2/(\gamma_{\text{tot,e}}\gamma_{\text{tot,o}})$  is the cooperativity [123],  $\gamma_{\text{tot,o}} = \gamma_{\text{ext,o}} + \gamma_{\text{int,o}}$  is the total decay rate of the optical resonator,  $\gamma_{\text{ext,o}}$  is the external coupling rate of the optical resonator,  $\gamma_{\text{int,o}}$  is the internal decay rate of the optical resonator,  $g = g_0\sqrt{n_p}$  is the nonlinear coupling rate,  $g_0$  is the single-photon nonlinear coupling rate,  $n_p = 4\gamma_{\text{ext,o}}P/[\hbar\omega(\gamma_{\text{ext,o}} + \gamma_{\text{int,o}})^2]$  is the intracavity pump photon number, and  $\omega$  is the pump photon frequency. The last beamsplitter has a power transmission of  $T_o = \gamma_{\text{ext,o}}/(\gamma_{\text{ext,o}} + \gamma_{\text{int,o}})$  which represents the optical resonator extraction efficiency. We assume an optical detector dark count rate of 50 Hz [289].

In our simulation, we begin with an initial state  $|\phi_0\rangle_A|\phi_0\rangle_B$  and numerically evolve the state with the Python QuTiP package [290] to obtain the density matrix  $\rho_f$  after the interfering beamsplitter. We assume that it takes  $t_1 = 50$  ns to prepare the initial states by local gate operations. We also assume the microwave photon and optical photon transmission loss is zero. We note that high photon transmission loss can decrease the entanglement generation rate and the fidelity and thus fails the next Distillation layer. Although zero transmission loss is experimentally unavailable yet, we still make this assumption for the purpose of illustrating the workflow of our stack model, and the consequent rate and fidelity can be understood as on-chip metrics. The nonzero transmission loss can be easily included into the model by incorporating the transmission loss to the optical/microwave cavity extraction efficiency. The converter bandwidth can be approximated as  $B \approx \gamma_{\text{tot,e}}$  [123], and we thus assume a pump pulse duration of  $t_2 = 1/B$  and a resonator reset time  $t_3 = 1/B$ . Hence, the total time duration for one period is  $t_{\text{tot}} = t_1 + t_2 + t_3$ . The event that detector A measures 1 photon while detector B measures 0 photon is considered a successful heralding, and the probability of a successful heralding can be calculated as  $P_{\text{herald}} = \text{Tr}\langle 1,0|\rho_f|1,0\rangle$ . Thus, the entanglement generation rate can be computed as  $R = P_{\text{herald}}/t_{\text{tot}}$ . In the case of a successful heralding, the corresponding qubit state is  $\rho_q = \langle 1,0|\rho_f|1,0\rangle/\text{Tr}\langle 1,0|\rho_f|1,0\rangle$ , and the entanglement fidelity is  $F = \langle \Psi^+|\rho_q|\Psi^+\rangle$ , where  $|\Psi^+\rangle = (|ge\rangle + |eg\rangle)/\sqrt{2}$  is the target qubit Bell state.

The parameter sets used for simulation are shown in Table VI. The first three parameter sets come from Table V, but both the microwave and optical intrinsic decay rate are 5 times smaller than the original values to allow a higher generation rate and lower infidelity and thus enable the next Distillation layer. We assume these parameter sets are experimentally available relatively soon given the recent progress in low-loss nonlinear optical material fabrication [291–293] and hence we still refer to them as ‘current M2O’ in the manuscript, despite sev-

eral optimistic assumptions made above. In addition, although the No. 1 converter in Table V is based on electro-optomechanical effects which require different formulas to calculate its conversion efficiency and bandwidth [135], we treat it as an electro-optic converter for simplicity, because this work aims at presenting a simulation model rather than a comprehensive analysis on various types of converters. We also present a hypothetical parameter set that we wish to be available in the future. The entangled pair generation rate, entanglement fidelity, and the density matrices are used as the input of the next distillation layer.

The result of simulation is shown in Fig 5. We first set  $P_e = 0.5$  and sweep the pump power as shown in Fig 5(a). For the No. 1 parameter set, one can see that the highest generation rate and the lowest infidelity are obtained at the pump power corresponding to  $C = 1$ , where the conversion efficiency is maximized. The entangled qubit state generation rate can reach 1 MHz with an infidelity near 0.2. However, for No. 2 and No. 3 parameter sets, the infidelity remains above 0.5, because a high pump power is needed for  $C = 1$ , and the generation rate is dominated by the false heralding triggered by the thermal added noise. The false heralding rate can be observed in Fig 5(b), where we fix the pump power such that the cooperativity  $C = 1$  while sweeping  $0 \leq P_e \leq 0.5$ . The entanglement generation rate at  $P_e = 0$  is thus the false heralding rate, which dominates for No. 2 and No. 3 parameter sets. The tuning of  $P_e$  reveals a rate-infidelity tradeoff regime, which is highlighted as the green shaded area, where the rate increases but the infidelity also increases with an increasing  $P_e$ . In this regime, a larger  $P_e$  allows more optical photons to be generated, but it also increases the error of having two nodes in the excited states simultaneously. The simulation results including the ‘future’ parameter set are shown in Fig. 17. It can be observed that a large bandwidth, low loss, and low thermal noise are key to a high generation rate and low infidelity to enable the MNQC.

### Appendix C: Entanglement Distillation Simulation

In the entanglement distillation layer, we use raw EPs generated from the physical layer and perform entanglement distillation on them to generate higher fidelity EPs, at the cost of a slower generation time. Specifically, we take the heralding raw entangled state generation rate and the density matrix as inputs, perform the entanglement distillation, and report the distillation results to the Data layer. The output information to the Data layer includes the success distilled state density matrix, the distillation time and the success probability to the specified number of rounds of nested distillation.

To improve the quality of the remote entanglement is one of the key problems in the community of quantum communication. Historically, Bennett *et al.* proposed a protocol, to purify the imperfect Bell state and improve

| No.                              | 1  | 2  | 3  | 4  | 5                                  | 6                   |
|----------------------------------|--|--|--|--|------------------------------------|---------------------|
| Platform                         | Electro-optomechanics  | Electro-optics   | Electro-optics   | Optomagnonics  | Rare-earth ions                    | Cold atoms          |
| Single-photon coupling rate (Hz) | $\frac{g_{om}}{2\pi} = 60$<br>$\frac{g_{em}}{2\pi} = 1.6$  | $\frac{g_{eo}}{2\pi} = 37$   | $\frac{g_{eo}}{2\pi} = 750$  | $\frac{g_{mago}}{2\pi} = 17.2$   | -                                  | -                   |
| Cavity decay rate (Hz)           | $\frac{\gamma_{ext,o}}{2\pi} = 2.1 \times 10^6$<br>$\frac{\gamma_{int,o}}{2\pi} = 5.6 \times 10^5$<br>$\frac{\gamma_{ext,e}}{2\pi} = 1.4 \times 10^6$<br>$\frac{\gamma_{int,e}}{2\pi} = 1.3 \times 10^6$ | $\frac{\gamma_{ext,o}}{2\pi} = 1.5 \times 10^7$<br>$\frac{\gamma_{int,o}}{2\pi} = 1.1 \times 10^7$<br>$\frac{\gamma_{ext,e}}{2\pi} = 5.6 \times 10^6$<br>$\frac{\gamma_{int,e}}{2\pi} = 8.1 \times 10^6$ | $\frac{\gamma_{ext,o}}{2\pi} = 3.3 \times 10^7$<br>$\frac{\gamma_{int,o}}{2\pi} = 1.4 \times 10^8$<br>$\frac{\gamma_{ext,e}}{2\pi} = 3.2 \times 10^6$<br>$\frac{\gamma_{int,e}}{2\pi} = 5.8 \times 10^6$ | $\frac{\gamma_{ext,o}}{2\pi} = 4.8 \times 10^7$<br>$\frac{\gamma_{int,o}}{2\pi} = 1.5 \times 10^9$<br>$\frac{\gamma_{ext,e}}{2\pi} = 1.7 \times 10^8$<br>$\frac{\gamma_{int,e}}{2\pi} = 8.5 \times 10^7$ | -                                  | -                   |
| Cooperativity                    | $C_{om} = 4.5 \times 10^4$<br>$C_{em} = 1.0 \times 10^4$   | $C_{eo} = 0.92$  | $C_{eo} = 0.04$  | $C_{mago} = 4.1 \times 10^{-7}$<br>$C_{mage} = 0.8$  | -                                  | -                   |
| Efficiency                       | $\eta_{tot} = 0.19$<br>$\eta_{in} = 0.59$  | $\eta_{tot} = 0.14$<br>$\eta_{in} = 0.99$  | $\eta_{tot} = 0.01$<br>$\eta_{in} = 0.15$  | $\eta_{tot} = 1.1 \times 10^{-8}$<br>$\eta_{in} = 5.2 \times 10^{-7}$  | $\eta_{tot} = 1.26 \times 10^{-5}$ | $\eta_{tot} = 0.82$ |
| Bandwidth (Hz)                   | $6.1 \times 10^3$  | -  | -  | $1.6 \times 10^7$  | -                                  | $1 \times 10^6$     |
| Added noise $n_{add}$            | 1.4  | 0.41   | -  | -  | -                                  | 0.8                 |
| Environment temperature (K)      | 0.04   | 0.01   | 1.9  | 300  | 4.6                                | 300                 |
| Reference                        | [119]  | [128]  | [122]  | [131]  | [148]                              | [142]               |

TABLE V. Summary of M2O converter performances on different experimental platforms. The definitions of parameters are discussed in Appendix B.

| No.                                | 1                 | 2                 | 3                 | Future          |
|------------------------------------|-------------------|-------------------|-------------------|-----------------|
| $\frac{g_0}{2\pi}$ (Hz)            | 60                | 37                | 750               | $10^3$          |
| $\frac{\gamma_{ext,o}}{2\pi}$ (Hz) | $2.1 \times 10^6$ | $1.5 \times 10^7$ | $3.3 \times 10^7$ | $10^7$          |
| $\frac{\gamma_{int,o}}{2\pi}$ (Hz) | $1.1 \times 10^5$ | $2.2 \times 10^6$ | $2.8 \times 10^7$ | $2 \times 10^5$ |
| $\frac{\gamma_{ext,e}}{2\pi}$ (Hz) | $1.4 \times 10^6$ | $5.6 \times 10^6$ | $3.2 \times 10^6$ | $10^7$          |
| $\frac{\gamma_{int,e}}{2\pi}$ (Hz) | $2.6 \times 10^5$ | $1.6 \times 10^6$ | $1.2 \times 10^6$ | $2 \times 10^5$ |

TABLE VI. Parameter sets used for M2O entanglement generation simulation. The No. 1, No. 2 and No. 3 parameter sets come from Table V. The optical and microwave resonator intrinsic decay rate are made 5 times lower than the original values. The last column presents a hypothetical parameter set which we wish to be available in the future.

the fidelity of the Bell state to unity [59]. In this protocol, each round of purification protocol will consume a pair of imperfect Bell states to generate an imperfect Bell state with higher fidelity and entanglement with less than unit fidelity. Suppose the remote superconducting qubits are in a Bell state (spin singlet) with imperfection and the state fidelity is  $F$ , after one round of entanglement purification, the fidelity is improved to

$$F_{new} = \frac{F^2 + (1 - F)^2/9}{F^2 + 2F(1 - F)/3 + 5(1 - F)^2/9}. \quad (C1)$$

Following this work, Deutsch *et al* proposed a similar method (DEJMPS), which improves the efficiency of the purification protocol [160]. This protocol avoids random bilateral single-qubit rotations to depolarize the imperfect state but uses the local operation to change into the Bell-diagonal basis. The outcome fidelity depends on the overlap to the other three Bell basis states [160].

With the above two purification protocols, one way

to generate Bell states of remote superconducting qubits close to unit fidelity is to use the recurrence purification scheme [214]. In this scheme, in order to perform  $n$  rounds of entanglement purification, we need to prepare  $2^n$  EPs of imperfect Bell states of superconducting qubits. In each round, the states from the last step undergo the pairwise entanglement purification to get states with higher entanglement.

In our purification layer, the entanglement purification is performed based on DEJMPS protocol in Ref. [160], while the effects of experimental imperfections are also considered. The experimental imperfection can come from two sources, (1) the error on the local two-qubit gates between superconducting qubits, and (2) the decay and decoherence error on the qubits while the qubits are idling. Specifically, in (2), we consider the idling from either waiting for the qubits are being measured during the purification process or waiting for the generation of required raw EPs.

In the purification simulation, we take the superconducting qubits to have lifetime  $T_1$  and coherence time  $T_2$ . From the simulation of the physical layer, we extract the density matrix ( $\rho_0$ ) of the raw Bell pair with the generation rate ( $r$ ). We assume that even with multiplexing, the raw Bell pair generation can still be considered sequential. Therefore, the average generation time of each pair is  $\tau = 1/r$ . For the error source (1), we assume unit fidelity local operations, as throughout our simulation stack we assume that local operations are perfect. We consider an instantaneous purification operation described by the quantum channel,

$$\rho_{new} = \mathcal{P}[\rho_{old,1} \otimes \rho_{old,2}], \quad (C2)$$

where two EPs of entangled states with density matrices

$\rho_{\text{old},1}$  and  $\rho_{\text{old},2}$  are purified and get a single pair of qubits in the state  $\rho_{\text{new}}$ . Again, the actual implementation of the entangled purification is based on Ref. [160].

For the error source (2), we need to estimate the idling time for superconducting qubits. Providing the two EPs of Bell states are ready for purification, we assume the local gate operations between the superconducting qubits and the measurements take  $t_p$  time. This is modeled by a decay and decoherence error channel, noted as  $\mathcal{E}(t_p)[\rho]$ , applied to the Bell state after the purification process. To estimate the total time for  $n$  nested rounds of entanglement purification, we assume the time for  $(n-1)$  rounds of purification takes  $t_{n-1}$  time. In the  $n$ -th round, the first pair of Bell states is generated from the  $(n-1)$ -th round, which takes  $t_{n-1}$  time. The second bell state used in the  $n$ -th round starts from  $t_{\text{idle},n-1} = 2^{n-1}\tau$ , while it also takes another  $t_{n-1}$  to generate the second Bell state for the  $n$ -th round. Therefore, the first Bell pair needs to wait for another  $t_{\text{idle},n-1}$  time. This is also modeled by a decay and decoherence error channel applied to the first Bell states used for the  $n$ -th round of purification. Further, we can construct the following recurrence relation for the  $n$  rounds of purification

$$t_n = 2^{n-1}\tau + t_{n-1} + t_p. \quad (\text{C3})$$

The state of the EPs after  $n$  rounds of success purification is

$$\rho_{(n)} = \mathcal{E}(t_p) [\mathcal{P} [\mathcal{E}(t_{\text{idle},n-1}) [\rho_{(n-1)}] \otimes \rho_{(n-1)}]] \quad (\text{C4})$$

The probabilistic nature of entanglement purification is from the measurement on one pair of Bell states. As pointed out in Ref. [160], if the measurement outcomes on two qubits in a single pair of input states coincide, the purification is considered as a success. The probability of having coincident measurement outcomes is the success probability for each round of purification, noted as  $p_j$  for the  $j$ -th round. The overall success probability of  $n$  rounds of purification can be calculated as

$$P_n = \prod_{j=1}^n p_j^{2^{j-1}}. \quad (\text{C5})$$

After the distillation calculation finishes, the required time  $t_n$ , the success state density matrix  $\rho_{(n)}$ , and the success probability  $P_n$  of  $n$  rounds of purification is passed to the Data layer for Internode gate simulation.

#### Appendix D: Internode Gate Simulation

Having generated a distilled EP between modules, the next step in performing multinode quantum computing is inter-node operations. As a CX gate is computationally complete communication between nodes, here we focus on the case of only internode CX gates. Gate teleportation of the CX gate can be accomplished via the consumption of one EP, two measurements, and two local

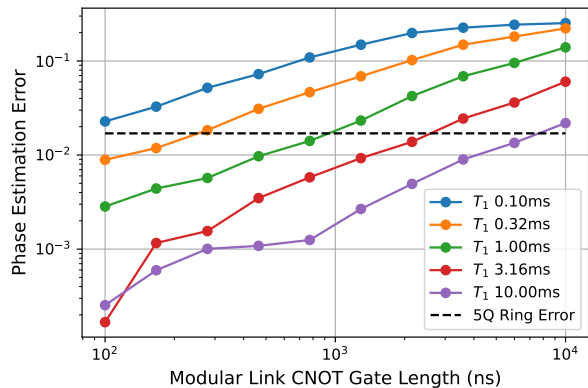


FIG. 23. QPE Error for a 10-qubit algorithm as a function of the modular gate length for different values of  $T_1$ .

CX gates. Simulation of the inter-node gate requires the use of a gate teleportation protocol, combined with the EP generated via M2O simulation under Section IV, and optionally distilled by the protocol underlined in Section IV. Simulation of the internode CX gate comprises beginning with an EP, represented by the density matrix output from entanglement distillation. We model local operations involved in the execution of the remote CX gate as having a local gate time of 100ns, suffering depolarizing errors with a probability of .0001, and taking  $T_1 = T_2 = 1\mu$  s. Having performed the protocol, the density matrix is captured at the protocol output, and reduced to represent the two data qubits. This density matrix is compared with the ideal simulation of a CX gate between two qubits, and used to report an overall teleported gate fidelity, and over all inter-module CX gate time.

#### Appendix E: Quantum Phase Estimation Benchmark

Another approach to the question of when to add a quantum link is by looking at algorithms where the answer improves in precision as the size of the system increases. For example, in quantum phase estimation (QPE) the number of ancilla  $n_A$  sets the precision of the phase estimate as  $1/2^{n_A}$ . For this problem, our phase unitary is a Z rotation with phase  $\phi = 0.658203$ . We setup the problem on the grid defined by (a) of Fig. 9 where the phase is applied to Q4. If there is no modular link then we solve on the  $n = 5$  qubit ring with  $n_A = 4$ . Assuming the gates on this ring are zero length and perfect, the minimum relative error is 1.7%. If we add the inter-module link to the problem to add 5 more qubits ( $n_A = 9$ ) then we can improve the relative error to zero. However, the error will increase if the fidelity of the link is less than one. We assume the link has a finite time to operate during which the link qubits, and all other qubits, will incur an error. The results are shown in

Fig. 23. Again, this gives a estimate on the order of gate errors where adding a link will lead to improvement in the problem space. Although the QPE problem has more optimal solutions on small systems, such as iterative phase estimation, it is an example that is easily extended into other problem spaces. For example, when using VQE to estimate molecular energies using more qubits allows for more molecular orbitals, which may lead to improved accuracy.

### Appendix F: Success Region Shapes

We can understand the roughly rectangular shape of the success regions in the GAPPs as follows. The axes of these plots are in terms of the logarithmic internode infidelity,  $\xi_I = \log I_{\text{local}}$ , and the logarithmic average execution time,  $\xi_T = \log(T_{\text{link}})$ . The infidelity due to local errors during the internode gate is  $\log I_{\text{local}} \sim N_q(\xi_T - \log T_*)$  where  $N_q$  is the number of local qubits and  $T_* = T_1 T_2 / (T_1 + T_2)$  is the effective fidelity lifetime of a local qubit. Requiring a given overall logarithmic fidelity of  $\log F$  then requires:

$$\log F = \text{const.} \sim \log(e^{\xi_I} + e^{\xi_T}) \quad (\text{F1})$$

which can be seen to lead to a roughly rectangular shape, as  $\log(e^x + e^y) = \text{const.}$  leads to a roughly rectangular curve.

### Appendix G: MNQC Networks and Layout

In the long run, having low-loss telecom communications between multiple QPUs could add even more flexibility on the overall architecture of the MNQC. Finding the QPU connectivity layout that best facilitates the implementation of quantum algorithms is an important question to be addressed. Should the QPUs be arranged

in an array-like structure (see Fig. 24(a))? This solution might be more appealing for implementation since the number of communication qubits per QPU remains constant. However, the distance between different QPUs might be an issue for compilation since a CX gate between two QPUs situated far apart might require a large number of inter-fridge CX gates. Should a better option be a star-like structure (see Fig. 24(b)), using a central node which is specialized for communication between the other nodes? This solution might reduce the average distance between QPUs and we discuss its potential in the following.

In the previous discussions, we have envisioned multiple QPUs communicating thanks to communication qubits, M2O converters, and fixed fiber links between different QPUs. In that setting however, we may not exploit to its full benefits the flexibility and adaptivity that allow photonic communications. Over the years, the integrated photonics community has developed reconfigurable silicon photonic hardware, allowing to manipulate photons more efficiently [294–296]. Using a central photonic node has been for example envisioned for trapped ions [297]. A universal photonic chip is an  $N \times N$  linear interferometer based on phase-controlled Mach-Zehnder interferometer and phase-shifters which allows to realize arbitrary unitary transformation on the input ports. By connecting the communication qubits to the input ports of such a chip and the output ports to single-photon detectors, we can centralize the heralded entanglement generation protocols between communication qubits through that interferometer. Moreover, contrary to the fixed structure where communication qubits are paired, such photonic chips should enable  $N$ -to- $N$  connectivity: each communication qubit can be connected to any other one. Therefore, using a central photonic node could allow to distribute easily EPs between any QPUs and thus drastically increase the overall modular quantum computer architecture (see Fig. 24(c)).



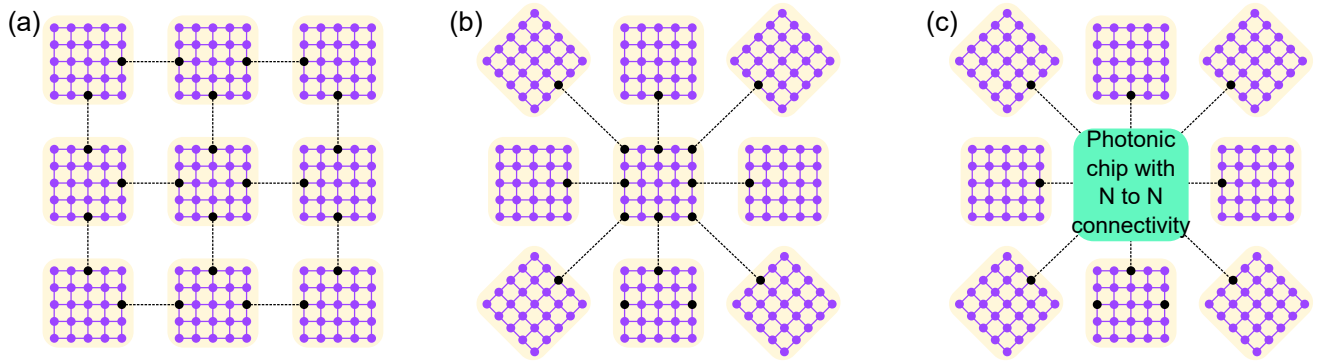


FIG. 24. (a) Array-like, (b) star-like architectures of a distributed quantum computer. (c) Star-like architecture with an N-to-N photonic chip enabling EP generation between any communication qubits from other superconducting chips.