# ARQUIN : Architectures for Multinode Superconducting Quantum Computers

JAMES ANG, Pacific Northwest National Laboratory, Richland, United States
GABRIELLA CARINI, Brookhaven National Laboratory, Upton, United States
YANZHU CHEN, Virginia Tech, Blacksburg, United States
ISAAC CHUANG, Massachusetts Institute of Technology, Cambridge, United States
MICHAEL DEMARCO, Brookhaven National Laboratory, Upton, United States and Massachusetts Institute of Technology, Cambridge, United States
SOPHIA ECONOMOU, Virginia Tech, Blacksburg, United States
ALEC EICKBUSCH, Yale University, New Haven, United States
ANDREI FARAON, Caltech, Pasadena, United States and IBM TJ Watson Research Center, Yorktown Heights, United States
KAI-MEI FU, University of Washington, Seattle, United States
STEVEN GIRVIN, Yale University, New Haven, United States
MICHAEL HATRIDGE, University of Pittsburgh, Pittsburgh, United States
ANDREW HOUCK, Princeton University, Princeton, United States
PAUL HILAIRE, Virginia Tech, Blacksburg, United States
KEVIN KRSULICH, IBM TJ Watson Research Center, Yorktown Heights, United States
ANG LI, Pacific Northwest National Laboratory, Richland, United States
CHENXU LIU, Virginia Tech, Blacksburg, United States and Pacific Northwest National Laboratory, Richland, United States
YUAN LIU, Massachusetts Institute of Technology, Cambridge, United States
MARGARET MARTONOSI, Princeton University, Princeton, United States
DAVID MCKAY, IBM TJ Watson Research Center, Yorktown Heights, United States
JIM MISEWICH, Brookhaven National Laboratory, Upton, United States
MARK RITTER, IBM TJ Watson Research Center, Yorktown Heights, United States




ACM 2643-6817/2024/7-ART
https://doi.org/10.1145/3674151

ROBERT SCHOELKOPF, Yale University, New Haven, United States
SAMUEL STEIN, Pacific Northwest National Laboratory, Richland, United States
SARA SUSSMAN, Princeton University, Princeton, United States
HONG TANG, Yale University, New Haven, United States
WEI TANG, Computer Science, Princeton University, Princeton, United States
TEAGUE TOMESH, Princeton University, Princeton, United States
NORM TUBMAN, NASA Ames Research Center, Moffett Field, United States
CHEN WANG, University of Massachusetts Amherst, Amherst, United States
NATHAN WIEBE, University of Toronto, Toronto, Canada and Pacific Northwest National Laboratory, Richland, United States
YONGXIN YAO, Ames Laboratory, Ames, United States and Kent State University, Kent, United States
DILLON YOST, NASA Ames Research Center, Moffett Field, United States
YIYU ZHOU, Yale University, New Haven, United States

Many proposals to scale quantum technology rely on modular or distributed designs wherein 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, internode gates in these systems may be two to three orders of magnitude noisier and slower than local operations. Surmounting the limitations of internode gates will require improvements in entanglement generation, use of entanglement distillation, and optimized software and compilers. Still, it remains unclear what performance is possible with current hardware and what performance algorithms require. In

Authors' Contact Information: James Ang, Pacific Northwest National Laboratory, Richland, Washington, United States; e-mail: ang@pnnl.gov; Gabriella Carini, Brookhaven National Laboratory, Upton, New York, United States; e-mail: carini@bnl.gov; Yanzhu Chen, Virginia Tech, Blacksburg, Virginia, United States; e-mail: yanzhuchen@vt.edu; Isaac Chuang, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States; e-mail: ichuang@mit.edu; Michael DeMarco, Brookhaven National Laboratory, Upton, New York, United States and Massachusetts Institute of Technology, Cambridge, Massachusetts, United States; e-mail: demarco@mit.edu; Sophia Economou, Virginia Tech, Blacksburg, Virginia, United States; e-mail: economou@vt.edu; Alec Eickbusch, Yale University, New Haven, Connecticut, United States; e-mail: alec.eickbusch@yale.edu; Andrei Faraon, Caltech, Pasadena, California, United States and IBM TJ Watson Research Center, Yorktown Heights, New York, United States; e-mail: faraon@caltech.edu; Kai-Mei Fu, University of Washington, Seattle, Washington, United States; e-mail: kaimeifu@uw.edu; Steven Girvin, Yale University, New Haven, Connecticut, United States; e-mail: steven.girvin@yale.edu; Michael Hatridge, University of Pittsburgh, Pittsburgh, Pennsylvania, United States; e-mail: hatridge@pitt.edu; Andrew Houck, Princeton University, Princeton, New Jersey, United States; e-mail: aahouck@princeton.edu; Paul Hilaire, Virginia Tech, Blacksburg, Virginia, United States; e-mail: paul.hilaire@quandela.com; Kevin Krsulich, IBM TJ Watson Research Center, Yorktown Heights, New York, United States; e-mail: kevin.krsulich@ibm.com; Ang Li, Pacific Northwest National Laboratory, Richland, Washington, United States; e-mail: ang.li@pnnl.gov; Chenxu Liu, Virginia Tech, Blacksburg, Virginia, United States and Pacific Northwest National Laboratory, Richland, Washington, United States; e-mail: Chenxu.liu@pnnl.gov; Yuan Liu, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States; e-mail: yuanliu@mit.edu; Margaret Martonosi, Princeton University, Princeton, New Jersey, United States; e-mail: mrm@cs.princeton.edu; David McKay, IBM TJ Watson Research Center, Yorktown Heights, New York, United States; e-mail: dcmckay@us.ibm.com; Jim Misewich, Brookhaven National Laboratory, Upton, New York, United States; e-mail: misewich@bnl.gov; Mark Ritter, IBM TJ Watson Research Center, Yorktown Heights, New York, United States; e-mail: mritter@us.ibm.com; Robert Schoelkopf, Yale University, New Haven, Connecticut, United States; e-mail: robert.schoelkopf@yale.edu; Samuel Stein, Pacific Northwest National Laboratory, Richland, Washington, United States; e-mail: samuel.stein@pnnl.gov; Sara Sussman, Princeton University, Princeton, New Jersey, United States; e-mail: sarafs@princeton.edu; Hong Tang, Yale University, New Haven, Connecticut, United States; e-mail: hong.tang@yale.edu; Wei Tang, Computer Science, Princeton University, Princeton, New Jersey, United States; e-mail: tangwei13579@gmail.com; teague tomesh, Princeton University, Princeton, New Jersey, United States; e-mail: ttomesh@princeton.edu; Norm Tubman, NASA Ames Research Center, Moffett Field, California, United States; e-mail: norman.m.tubman@nasa.gov; Chen Wang, University of Massachusetts Amherst, Amherst, Massachusetts, United States; e-mail: wangc@umass.edu; Nathan Wiebe, University of Toronto, Toronto, Ontario, Canada and Pacific Northwest National Laboratory, Richland, Washington, United States; e-mail: nathanwiebe@gmail.com; Yongxin Yao, Ames Laboratory, Ames, Iowa, United States and Kent State University, Kent, Ohio, United States; e-mail: ykent@iastate.edu; Dillon Yost, NASA Ames Research Center, Moffett Field, California, United States; e-mail: dyost@mit.edu; Yiyu Zhou, Yale University, New Haven, Connecticut, United States; e-mail: yiyu.zhou@yale.edu.

this paper, we employ a systems analysis approach to quantify overall MNQC performance in terms of hardware models of internode links, entanglement distillation, and local architecture. We show how to navigate tradeoffs in entanglement generation and distillation in the context of algorithm performance, lay out how compilers and software should balance between local and internode gates, and discuss when noisy quantum internode links have an advantage over purely classical links. We find that a factor of 10-100x better link performance is required and introduce a research roadmap for the co-design of hardware and software towards the realization of early MNQCs. While we focus on superconducting devices with optical interconnects, our approach is general across MNQC implementations

CCS Concepts: • **Hardware → Quantum computation**; • **Computing methodologies → Quantum mechanic simulation**; • **Computer systems organization → Distributed architectures**; **Quantum computing**.

Additional Key Words and Phrases: Quantum Computing, Quantum Computing Architecture, Multinode Quantum Computing, Distributed Quantum Computing, Transduction

## 1 Introduction

Modular, distributed, or multinode quantum computers (MNQCs) [8, 34, 91, 129, 162, 194, 195, 218, 288, 299, 309], wherein smaller devices or "nodes" are networked together [14] to make a unified multinode quantum computer, are considered a leading approach to building large scale quantum systems [116] without the difficulties of producing large monolithic devices [162]. Leading platforms include trapped-ion computers with multiple traps [38, 141, 214, 214], solid-state systems [28, 200, 219], atomic systems [192, 231, 296], and superconducting devices [40, 75, 76, 95, 107, 129, 145, 177, 197, 309].

In superconducting devices, a leading platform [217, 274] for quantum development, 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 dilution refrigerator required to cool the device [153]. Building links between devices in different refrigerators is thus a key capability [202]. Early-stage MNQCs with cryogenic links between refrigerators have been demonstrated [177], and when cryogenic links can be feasibly constructed they are a leading candidate for small systems [291]. 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 [313] and cable loss [16, 39, 45, 157, 169, 177, 291] become an issue. Rather than using cryogenic links, a system composed of devices housed in separate refrigerators with room-temperature microwave-to-optical (M2O) quantum internode links [56, 58, 118, 156, 160, 163, 189] between them is a more scalable proposal for building future MNQCs.

Critically, internode links in these systems are likely to be much noisier and slower than local gates and thus threaten the viability of MNQCs [15, 223]. These weak internode links hamper performance both by directly causing errors and by creating a computational bottleneck which allows decoherence [217, 274] 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 [56, 58, 118, 156, 160, 163]. In order to be viable, systems with quantum internode links must outperform not only monolithic quantum systems but also systems with only classical links between nodes [256, 257].

However, given the experimental challenges, it is unclear what MNQC system performance might be achieved with present hardware, what performance is required to execute future algorithms, and whether the gap between the two can be bridged. Evaluating MNQC performance presents a systems analysis challenge, combining hardware-level simulations of noise [15, 183, 237] with remote gate execution [151], distillation protocols [25, 81, 291], and overall algorithm demands. A further complication arises because MNQCs in fact balance three expensive resources: internode gates, local two-qubit gates [90, 285], and classical circuit cutting links [36, 191, 211, 216]. Compounding this, any analysis must navigate the exponential complexity of simulating both intranode and internode links combined with distillation, remote gates, and multinode algorithm execution [171]. Without a clear understanding of how all the components of an MNQC interact to affect system performance, future

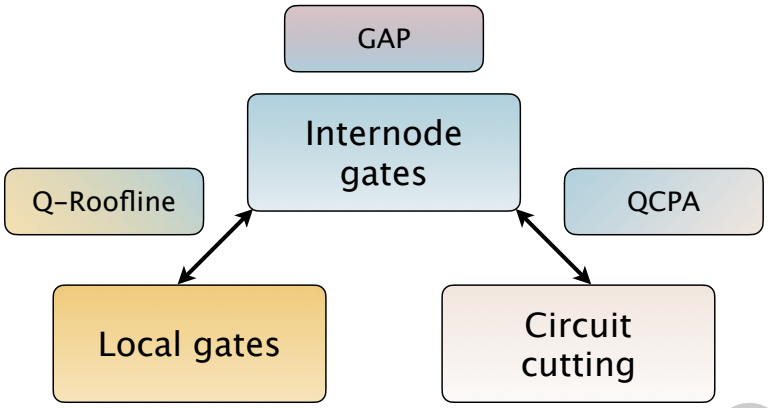



Fig. 1. Multinode Quantum Computers (MNQCs) must balance three key resources: internode gates, local computation, and circuit cutting gates. GAP represents Gate-Algorithm performance, indicating the latency and fidelity thresholds over a multinode quantum computer architecture on specific applications. Q-Roofline represents the Quantum Roofline model, proposed below under Section 5.3 - Quantum Roofline Model, and is a method for understanding the scaling behavior of larger systems, aimed at identifying tradeoffs and bottlenecks in scaling. QCPA represents the Quantum-Classical performance analysis, and describes the cost of error-mitigated internode links against the classical circuit cutting counterparts.

improvements in hardware and software may be incompatible and lead to reduced or even no improvements in MNQC performance [174].

Inspiration for this problem comes from classical computing, which has navigated similarly complex design constraints to build high-performance multi-node systems [13, 254, 267]. Early multicomputers including the ALEWIFE [4, 5] and BEOWULF [250] systems utilized existing hardware to lay the foundations of classical networks with distributed memory that evolved into interconnection architectures such as the Infiniband and Slingshot networks [72, 104, 130]. Physical hardware and software constraints were key to building modern interconnection architectures, from 'fat-tree' networks [168] that balance network bandwidth against the size of an architecture to adaptive routing [243] and complex network architectures [143] that maximize system performance while minimizing wiring overhead. From an architectural perspective, these tradeoffs are captured in 'Roofline' models [282] which quantifies the relative burden of local computation and memory communication. This approach of navigating tradeoffs by designing hardware and software jointly came to be called 'co-design' [11] and has played a significant role in the design of modern high-performance and exascale computing [20, 94].

More recently, 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 [57] and quantum internet [144], recent works have elaborated a vision for the development of a truly distributed quantum ecosystem [34, 64, 278], although hardware which is capable of delivering the requisite performance largely remains to be developed [238, 240]. Layered link protocols [67, 176, 215] focused on the preparation of nonlocal entanglement [6, 176] 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 [269] and dedicated compilers and frameworks [17, 41, 65, 283, 286]. With substantial progress envisioned in the realization of high-performance quantum interlinks [14], joint co-design of hardware and software will be key to enabling quantum networks [262].

In this paper, we present a systems analysis of MNQCs, determining what algorithm performance can be achieved with present M2O hardware, what hardware performance is required to enable advanced algorithms,

and quantitatively characterizing the impact of potential hardware and software research directions. We use a co-design layer architecture to characterize internode links and quantify the full range of available internode performance with present technology. Then, we integrate these results into three models: a 'Gate-Algorithm Performance' (GAP) model which determines achievable algorithm performance using internode links, a 'Quantum Roofline' (Q-Roofline) model which determines the relative costs of internode and local computation, and a 'Quantum-Classical Performance Analysis' (QCPA) which demonstrates the relative costs of error-mitigated internode links against classical "circuit cutting" links. We visualize these key resources and the corresponding higher level models in Figure 1. This quantitative systems analysis reveals a 100x performance improvement required to enable MNQCs, and allows the potential of various research directions to achieve this. 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 [64, 278] and cryogenic microwave links [35, 107, 291, 309].

The next section presents a review of superconducting transmon devices with M2O interlinks. In Section 3, we discuss the co-design architecture that allows the problem of internode links to be simplified by splitting the internode link into distinct layers. Section 4 then presents and analyzes models of each of the layers. Section 5 unifies these models into a full stack model, and presents the GAP, Q-Roofline, and QCPA analyses. In Section 6, 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 7 discusses potential applications of our methodology to other quantum platforms.



## 2 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, 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. 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. For simplicity, classical communication between nodes is taken to be fast, reliable, and well-synchronized to a single clock.

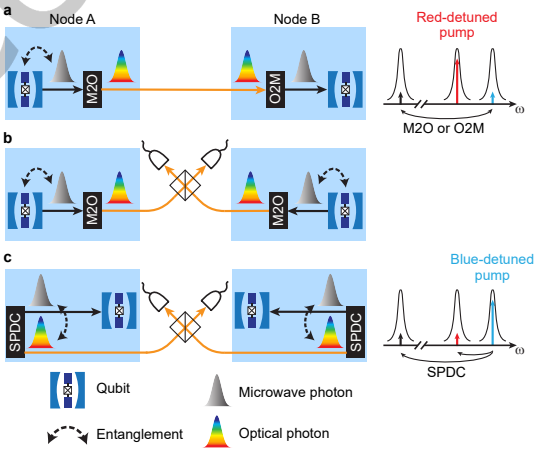



Fig. 2. Schemes for entanglement generation between remote nodes, including (a) pitch-and-catch, (b) heralded direct conversion, and (c) heralded SPDC.

Although various superconducting processor designs are being prototyped [19, 50], the most widely used architecture in both academic and industry settings is a 2D lattice of nearest-neighbor coupled transmons [3, 147] cooled to milli-Kelvin temperatures in dilution refrigerators. Transmons, strengthened by recent advancements in high-fidelity two-qubit gates [96, 252, 280] and enhanced qubit coherence nearing 0.5 ms [217, 274], are effective building blocks for quantum computing. Transmon processors with as many as 433

qubits are available [60] and steady progress is being made towards the goal of fault-tolerant quantum computing using error correction [3, 152, 251, 304].

Remote entanglement distribution experiments with transmon-qubit-based processors connected by cold microwave links have evolved substantially. A series of heralding-based probabilistic entanglement distribution experiments were performed between qubits on separate chips within a fridge [78, 197, 232]. 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 [177], and the resulting fidelities were largely limited by cable loss [16, 45, 157, 169, 177]. 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 [51, 308]. Heralded deterministic entanglement experiments have also been done [158] where sophisticated techniques were developed to mitigate cable loss [39]. Today's state-of-the-art deterministic entanglement distribution experiments are limited by cable loss [291, 307]. 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 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 [73, 165, 190], but these converters are not yet efficient and broadband enough for use in a remote entanglement distribution 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 [86, 99, 120, 124, 183, 239, 248, 290, 298], opto-magnonics [123, 301, 302, 311], electro-optomechanics [10, 12, 37, 73, 93, 117, 122, 134, 154, 190, 265], cold atoms [61, 264, 270] and rare-earth ions [22, 85, 88, 89, 207]. Reviews of recent experimental advances in M2O conversion can be found in Refs. [56, 58, 118, 156, 160, 163]. The conversion efficiency achieved in state-of-the-art experiments, however, remains far less than unity. Despite the relatively high on-chip conversion efficiency, total efficiency can be significantly lower due to the inevitable fiber-to-chip coupling loss and optical-pump-rejection filtering loss. In addition, because a high-power optical pump is needed to boost conversion efficiency, thermal 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 5 of Appendix B of the Supplemental Material.

In order to generate entanglement between separated refrigerators, one straightforward 'pitch-and-catch' method is to locally generate an 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 [42, 128, 188], 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 heralded direct conversion 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 $|01\rangle \pm |10\rangle$.

In addition, a remote entanglement generation scheme previously developed for atomic ensembles [21, 33, 42, 80, 179, 220] is another option for superconducting platforms [151] 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 which interfere in a beamsplitter to erase the which-path information, heralded by the single-photon detector between refrigerators. However, in the presence of high optical loss, a click from the optical detector might undesirably herald the $|11\rangle$ state, reducing fidelity. Therefore, in the following sections, we mainly focus on the heralded direct conversion scheme.

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. 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. 4). In comparison, transmon two-qubit gate infidelity has already been engineered down to less than 0.002 [280], 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 surmount the weak internode links.

## 3 Multinode Quantum Computing Architecture

Internode gates in multinode systems involve the complex interactions 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 [15, 183, 237] settings, entanglement distillation [25, 81, 291], and compiler optimization [24, 66, 90, 284, 285]. 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. Treating the asymmetric noise profile of the entanglement generation introduces required density matrix simulations, which are currently limited to $O(20)$ qubits [172, 208]. However, allotting just a few qubits for entanglement distillation, measurement ancillas, 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 and efficiently. We address this by abstracting the MNQC architecture into discrete layers, as visualized in Figure 3, enabling scalable simulation via phenomenological noise modelling of each components performance with respect to both latency and fidelity. We break down the MNQC stack and motivate this approach from the bottom up, from raw Entangled Pair generation over Microwave to Optical transduction, propagated up to the Application layer, where a virtual topology with relatively weaker edges couple processors.

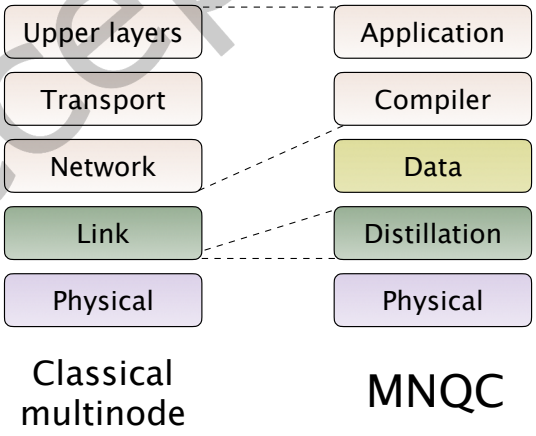



Fig. 3. Comparison of the layer stack for classical multinode architectures [130, 254] and our MNQC architecture.

Classical computing network architecture has long characterized similarly complex systems. 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 [72, 104, 130], 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 [254]). A key concept behind the operation of multinode systems is *transparency* [267]: 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.

Our goal in creating an architecture is to incorporate entanglement generation, distillation, and remote gate execution in a way which predicts the overall performance of the machine, makes clear how to navigate tradeoffs between these components, and renders the functioning of internode communication transparent to the upper layers of the stack. However, there are several key differences between quantum MNQCs and their classical counterparts (see Fig. 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 noise. 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 tenuous performance of links require much lower level access for the compiler, operating at the level of links between adjacent nodes in the network (Figure 4a), thus breaking true abstraction between the network and the computer.

An efficient method for the execution of remote gates is to use entangled pairs (EPs) produced from the M2O process, local operations, and classical internode communication to execute remote gates [110, 133]. 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 [74], 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. This is a considerable simplification for the upper layers, as the details of internode gate execution are now abstracted away and internode gates appear in the same manner as local gates, though slower and noisier.

Taken together, raw EPs, distilled EPs, and internode gates form a chain of key resources for internode gates, each produced from the previous, which are unique to multinode quantum systems. The key to creating the MNQC stack is to devote a 'layer' of the system to the production of each key resource. At the 'bottom' of the stack, M2O hardware produces raw EPs and is called the 'Physical layer' in analogy with the classical approach. Next, a 'Distillation layer' converts raw EPs into distilled EPs at a lower generation rate. Distilling EP pairs is a probabilistic protocols. Sequencing gates with probabilistic events requires repeat until success generation of entangled pairs. To capture the performance of this, we make use of mean execution times, and hence mean entangled pair generation time. Finally, a 'Data layer' manages the execution of internode gates, and exposes them as a resource to the Compiler and then Application layers, which sit atop the network stack. A comparison of this MNQC architecture with that from modern classical interconnection architectures is given in Fig. 3.

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

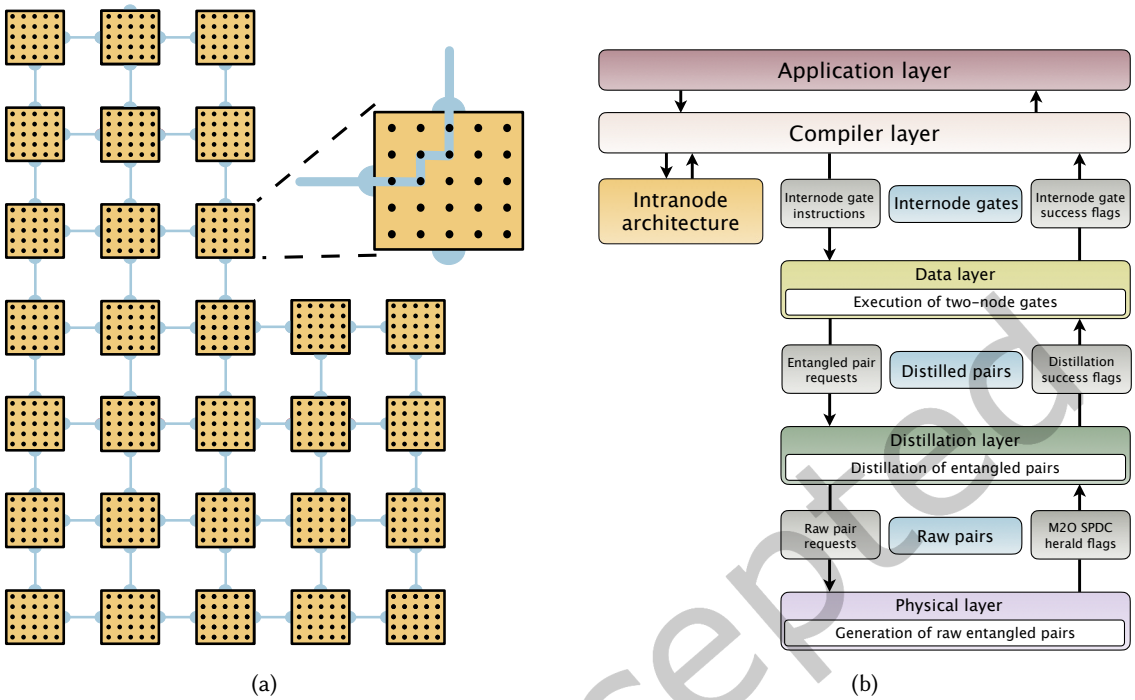



Fig. 4. *(color online). (a):* Schematic depiction of a multinode quantum system. Nodes (orange boxes) with many qubits (black dots) are connected with quantum links (blue lines) that allow internode gates. *(b):* Description of the layers, functions, and interfaces of the MNQC model. Our focus is on the MNQC network stack (the Physical, Distillation, and Data) layers and how their performance affects overall MNQC performance.



compiler. To retain flexibility, the MNQC stack may offer to the compiler several options for internode gates corresponding to distinct distillation settings, including even no distillation. However, the actual execution of those gates, including processing heralds from the M2O hardware, executing distillation and repeating until successful, and so on, should remain abstracted from the upper layers. On the other hand, overall responsibility for error correction will rest in the upper layers of the stack as it must span local and internode components. However, we note that the distillation layer provides a form of additional error correction on the link due to the deep connection between distillation and error correction [26].

Similar models for network architectures have been proposed in several pioneering works [67, 176, 209, 215]. 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 [67]. An excellent general overview of planned progress in quantum interlink technologies may be found in [14]; 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.

## 4 Models of the Superconducting MNQC Network Layers

The MNQC architecture from the previous section organizes entangled pair generation, entanglement distillation, and remote gate execution into layers. 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. In this paper, we specifically are tackling superconducting qubits, hence our focus on the challenges of coupling multi-fridge systems via a warm median such as optical links.

Beginning from the bottom of the stack, the Physical layer works to produce raw M2O EPs, which are quantified in terms of the heralded production rate and their density matrix. The Distillation layer is then 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 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 fidelity of each gate. Note that these metrics are simplified by characterizing key resources with the average times, rather than simulating the full probabilistic nature of the generation time. These metrics are denoted in Fig. 4b.

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 $|\phi_0\rangle = \sqrt{0.5}\,|g0\rangle + \sqrt{0.5}\,|e1\rangle$ at both nodes. The microwave photons are converted to optical photons via a M2O converter. The conversion efficiency are phenomenologically modeled by three beamsplitters. The first beamsplitter (representing the microwave resonator extraction efficiency) has a power transmission of $T_e = \gamma_{\mathrm{ext,e}}/\gamma_{\mathrm{tot,e}}$, where $\gamma_{\mathrm{ext,e}}$ is the external coupling rate of the microwave resonator, $\gamma_{\mathrm{int,e}}$ is the intrinsic decay rate of the microwave resonator, and $\gamma_{\mathrm{tot,e}} = \gamma_{\mathrm{ext,e}} + \gamma_{\mathrm{int,e}}$. Due to the pump-induced heating, the microwave resonator suffers from thermal added noise, which can be modeled as a thermal state $\rho_{\mathrm{th}}$

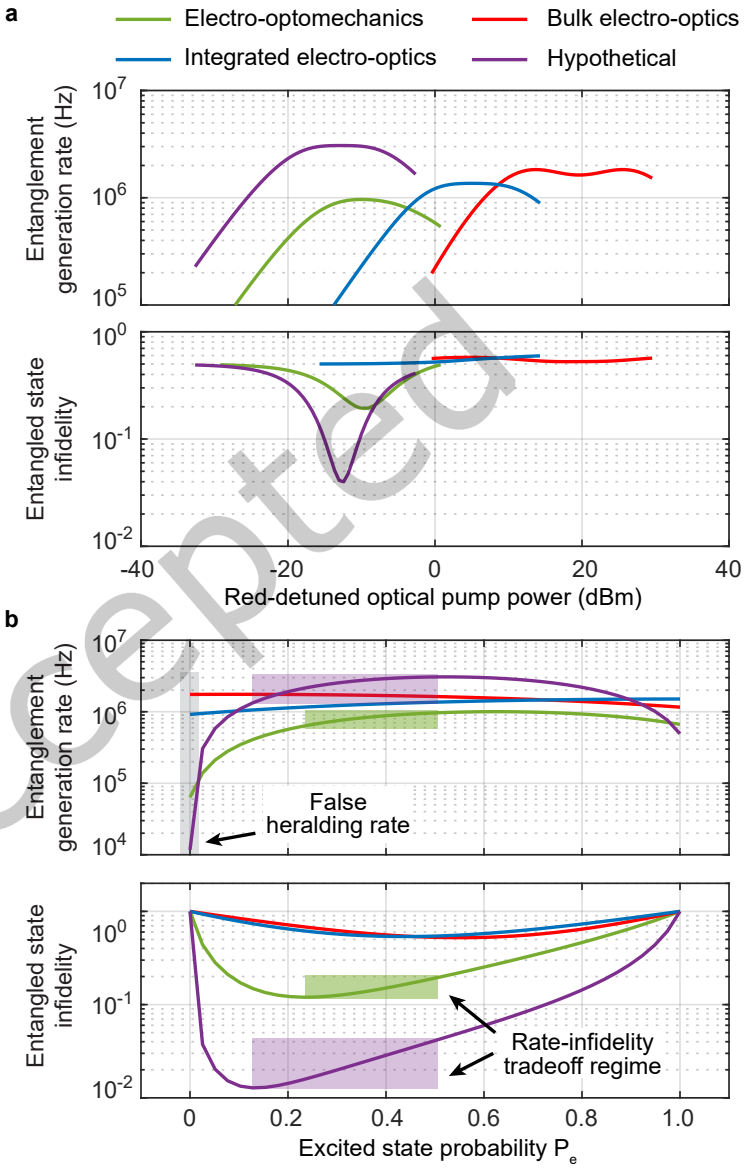



Fig. 5. The estimated Bell state generation rate and infidelity (log scales) for the scheme shown in Fig. 2(b). (a) the performance as a function of pump power with $P_e$ fixed at 0.5. (b) the performance as a function of $P_e$ with fixed pump power such that the cooperativity $C = 1$. The hypothetical curve applies to potential future devices as will be discussed in Section 6. Parameters used for simulations are shown in Appendix B Table 6. The rate at $P_e = 0$ is the false heralding rate triggered by thermal noise, and rate-infidelity tradeoff regimes (the green and purple shaded area) can be identified.

at another input port of the beamsplitter. The second beamsplitter (representing the intracavity conversion efficiency) has a power transmission of $T_{\text{in}} = 4C/(C+1)^2$, where $C = 4g^2/(\gamma_{\text{tot,e}}\gamma_{\text{tot,o}})$ is the cooperativity [86], $\gamma_{\text{tot,o}} = \gamma_{\text{ext,o}} + \gamma_{\text{int,o}}$ 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. The optical photons then interfere at a 50:50 beamsplitter.

We begin with an initial state $|\phi_0\rangle_A |\phi_0\rangle_B$ and numerically evolve the state with the Python QuTiP package [137] to obtain the density matrix $\rho_{\text{f}}$ after the 50:50 beamsplitter. The event that one detector measures 1 photon while the other detector measures 0 photon is considered a successful heralding, and the probability of a successful heralding can be calculated by tracing out the optical modes, i.e. $P_{\text{herald}} = \text{Tr}\,\langle 1,0|\,\rho_{\text{f}}\,|1,0\rangle$. Thus, the entanglement generation rate can be computed as $R = P_{\text{herald}}/t_{\text{tot}}$, where $t_{\text{tot}}$ is the total time period of one cycle. In the case of a successful heralding, the corresponding qubit state is $\rho_q = \langle 1,0|\,\rho_{\text{f}}\,|1,0\rangle\,/\text{Tr}\,\langle 1,0|\,\rho_{\text{f}}\,|1,0\rangle$, 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. More details on the simulation can be found in Appendix B.

The simulated entanglement infidelity and generation rate for current experimental platforms are shown in Fig. 5(a), using the parameter sets of three resonator-based M2O converters (see Appendix B Table 6) to perform the simulation. The intrinsic decay rate of both microwave and optical resonators is assumed to be five times lower than the present experiment values, which we expect to be achievable in the near term. We also include a hypothetical converter that may be available in the future, and the pathways for its experimental realization are discussed in Sec. 6.1. The highest M2O conversion efficiency is obtained at the pump power that reaches a unity cooperativity $C = 1$ [263], which consequently leads to the highest generation rate and lowest infidelity. For the electro-optomechanics converter (the green curve), the entangled qubit generation rate can approach 1 MHz with an infidelity near 0.2. However, for the bulk electro-optics (the red curve) and integrated electro-optics (the blue curve) converters, the infidelity remains above 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.

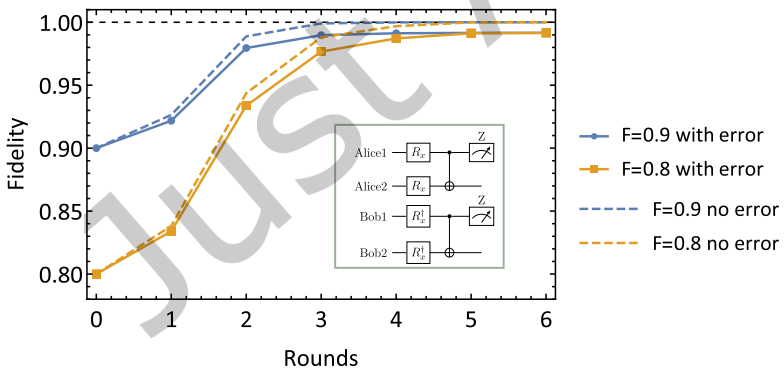



Fig. 6. The performance of entanglement distillation under the DEJMPS protocol [74] with and without errors from qubit decay, decoherence, and noisy local gates.

To enhance fidelity in the case of loss, the initial qubit-microwave photon state may not be fully entangled, i.e. $|\phi_0\rangle = \sqrt{1-P_e}\,|g0\rangle + \sqrt{P_e}\,|e1\rangle$ where $P_e$ is the probability of the excited qubit state and is experimentally tunable [197]. The pump power is fixed such that $C = 1$, and the results with $P_e$ tuned from 0 to 1 are shown in Fig. 5(b). The entanglement generation rate at $P_e = 0$ is thus the false heralding rate, which dominates for both electro-optics converters. The tuning of $P_e$ reveals a rate-infidelity tradeoff regime, which is highlighted as the shaded area in Fig. 5(b), 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. In Section 5, 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 also 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 of the Supplemental Material, we simulate the output of entanglement distillation using EPs produced from M2O conversion using the well-known DEJMPS [74] protocol, as a simple protocol that consumes 2 qubits and when successful, projects an EP pair into a higher fidelity bell state. In the case of failure, whereby a bell pair parity check fails, the EP pairs are scrapped. For simplicity, we set $T_1 = T_2 = 1$ms and assume all local gates to take 100ns with a probability of depolarizing errors of $\lambda = .0001$. The depolarizing channel is described by $E(\rho) = (1-\lambda)\rho + \lambda Tr[\rho]\frac{I}{2^n}$. Fig. 6 shows the fidelity of the Bell state shared between remote superconducting chips after $n$ rounds of recurrent entanglement purification performed using the DEJMPS protocol [74, 82]. The time for $n$ rounds of purification satisfies,

$$t_n = 2^{n-1}\tau + t_{n-1} + t_p. \tag{1}$$

where $t_{n,n-1}$ is the time till round $n$ $(n-1)$, $t_p$ is the measurement time, and $\tau$ is the single EP generation time. The state of the EPs after $n$ rounds of success purification is

$$\rho_{(n)} = \mathcal{E}(t_p)\left[\mathcal{P}\left[\mathcal{E}(t_{\text{idle},n-1})\left[\rho_{(n-1)}\right] \otimes \rho_{(n-1)}\right]\right], \tag{2}$$

where $\mathcal{E}(t)[\cdot]$ is a decay and decoherence error channel for idling time $t$, $\mathcal{P}[\cdot]$ is the channel corresponding to successful coincident measurements on two qubits for one round of purification, and $t_{\text{idle},n} = 2^n\tau$ (see Supplemental Material Appendix C for more information). 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 Fig. 6 shows the performance of entanglement distillation with (solid line) decoherence or (dotted line) an ideal memory that prevents decoherence. However, 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. The physical layer output is comprised of a density matrix output representing an entangled pair of qubits and a time to generate. The distillation layer simulates purifying the resulting EPs, returning an output density matrix and the total time required to generate EPs and purify them. For the data layer, taking only a CX gate is sufficient to provide computationally complete communication between nodes. Gate teleportation of the CX gate can be accomplished via the consumption of one (distilled) EP, two measurements, and two local CX gates [133]. Using the 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. Having generated a purified density matrix over some time, the purified EPs are piped into the data layer to simulate the performance of a single internode gate. This gate will have a fidelity and gate time attached to it, according to the entire stack below it, which characterises this internode link. These features are what motivate us to consider this low-level resource as a higher-level gate when viewed from the upper layers. This allows for seamless integration into modern day homogeneous gate set transpilers.

## 5 Full MNQC Analysis

We now have models of each layer in the MNQC network stack, from the Physical layer with M2O generation 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 to navigate tradeoffs across layers, or how to exchange internode gates with local computation and circuit cutting. In our design, we propagate the results of phenomenological error models through the stack, resulting in each layer contributing some error rate and latency which is propagated up to the highest level of the MNQC stack. In the case of injecting error correction, this would follow the same procedure whereby the layer below would propagate its error statistics into the error correction layer. This is a deliberate system noise model, as it provides feasible and scalable insights into near term approaches providing approximations of latency and fidelity thresholds.

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, a 'Gate-Algorithm Performance' (GAP) model 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. Next, the unified model output is fed into 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 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 circuit cutting links.

### 5.1 Unification of Layers into an Overall MNQC Model

The unified model aims to quantify overall MNQC performance as a function of the performance of each layer. Specifically, it takes hardware and software details of each layer as inputs, 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, the model returns 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, the unified model enables us to study the behavior of the Application and Compilation layers to determine the needs of algorithms, including simulation of algorithms running on small systems and tradeoff analysis for large systems.

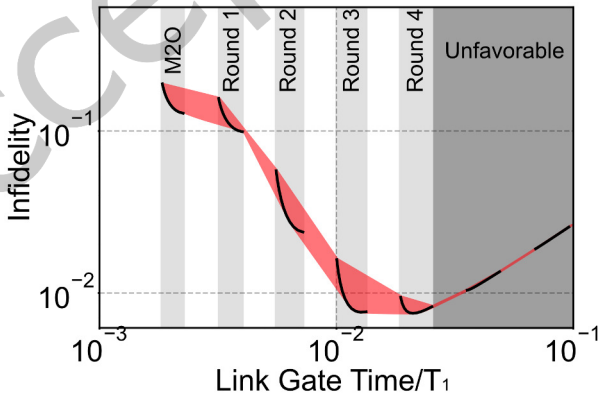



Fig. 7. Performance profiles of internode gates using (black line) only raw M2O entangled pairs or distilled pairs.

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. This comprises the outputs of the physical layer, which is passed to the Distillation layer. These outputs are fed into the next distillation layer to generate high-fidelity, purified EPs. Using the statistics recieved from the Physical layer,

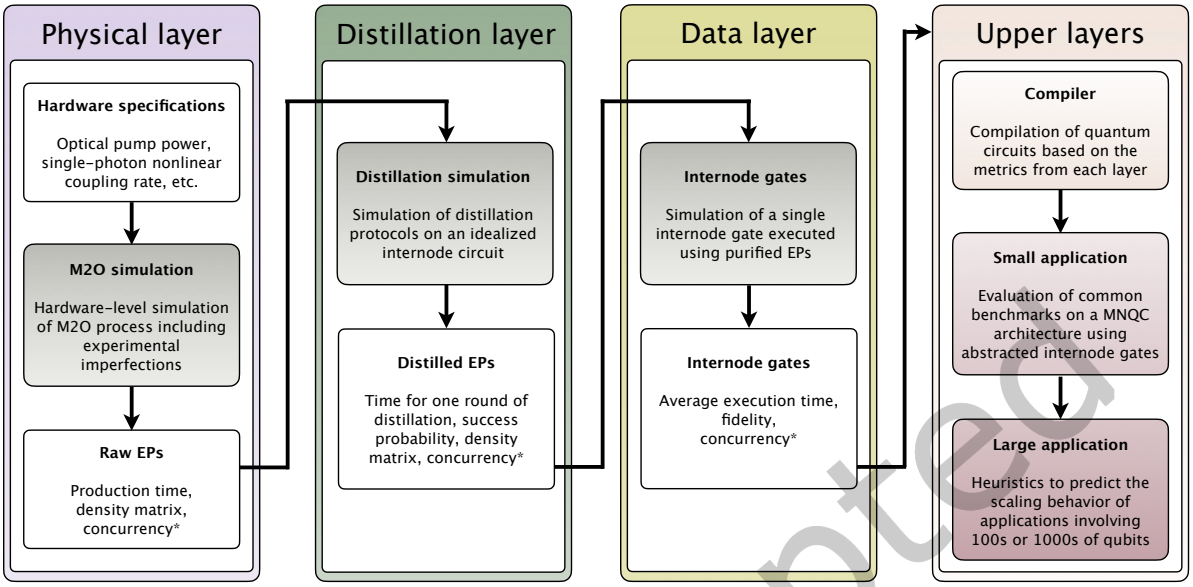



Fig. 8. Models of each layer of the MNQC model and the metrics passed between them. Each layer is simulated as detailed in Section 4, and the result is united to create an overal simulation of the MNQC. The concurrency metrics marked with an asterisk are set to unity here but can be used in a future generalized model that includes multiplexing of internode links. For future fault-tolerant error-corrected MNQC architectures, an additional layer is appended to the Upper Layer mediating the error correction subroutine, and each subsequent upper layer partition navigates the constraints of the error correction layer.

we simulate entanglement distillation, providing a suite of possible density matrices and respective generation times based on physical layer generation rate. 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. Within the data layer, a remote gate is simulated over 4 qubits, 2 of which representing a remote EPR pair, and 1 qubit on each node. A remote gate is simulated over these 4 qubits, attaining a final density matrix representing a noisy remote CX gate, with a total process time. Using this information, statistics representing the remote CX gate's latency and fidelity are provided to the compiler.

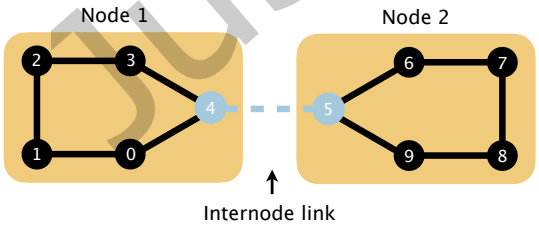



Fig. 9. Topology of a small MNQC that we simulate explicitly. The system consists of two five-qubit nodes with a single internode link.

To connect these results to the upper layers, we use the abstractions provided by the MNQC network stack, in particular the principle of transparency articulated in Section 3, 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.

The full simulation pipeline connects the M2O Physical layer simulation, Distillation, and Data layer simulations discussed in the previous section and unified

in the gate model to upper layers. Fig. 8 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 4, 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.

## 5.2 Gate-Algorithm Performance Models

Our first task is to quantify how the performance of internode links affects simple algorithms. This will allow us to determine how to navigate the tradeoffs in the Physical and Distillation layers identified in Section 4, both of which involve an exchange between the time to create EPs and the infidelity of those EPs. The unified model does this by determining the available internode gate performance produced by the network stack as a function of the operation of the Physical and Distillation layers and then allowing us to evaluate benchmark algorithms executed on the MNQC using internode gates.

We evaluate internode gate performance within the MNQC network stack using the unified model pipeline from Fig. 8. This links the outputs of the models across the Physical, Distillation, and Data layers, based on the experimental performance cited in the previous section and the Supplemental Material. Fig. 7 demonstrates the infidelity and average generation time of internode gates using raw EPs and distilled EPs. The black curves represent the tradeoff between execution time and infidelity, which is influenced by the excitation probability $P_e$ in the M2O conversion process. Higher $P_e$ values result in faster execution but increased infidelity, while lower values lead to reduced infidelity but longer execution times. This tradeoff, identified in Section 4, is reflected in the negative slope of the M2O curve.

At a specific M2O excitation probability setting, corresponding to a point on the black curve, the Distillation layer faces a tradeoff: distilled EPs offer higher fidelity but longer average production times compared to raw M2O EPs. This results in internode gates with higher fidelity but longer execution times at the Data layer. Each distillation round reduces infidelity at the expense of increased internode gate time. The interplay between colored noise, drive power, and entanglement distillation creates complex performance behavior. Fastest gates require no distillation, while lower-infidelity gates benefit from distillation rounds.

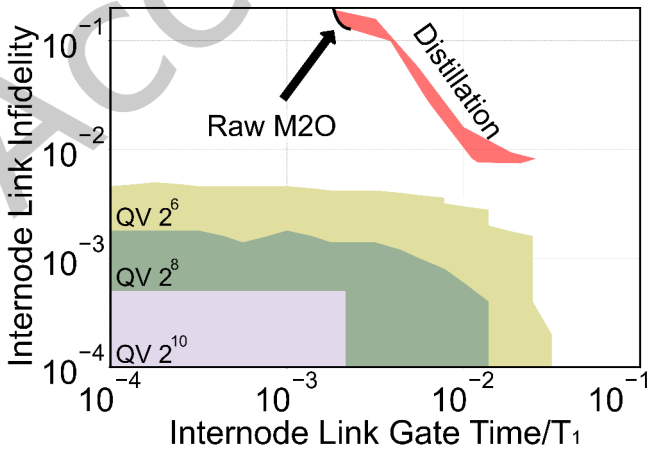



Fig. 10. Gate-algorithm performance plot of Quantum Volume. Each distillation curve, denoted by the red region, has been truncated at the number of nested rounds at which its performance begins to degrade.

Algorithm requirements then dictate internode gate execution: targeted performance, number of rounds, and excitation probability $P_e$ determine infidelity and link gate time. If a compiler selects from various internode gate times, the MNQC stack must dynamically adjust M2O generation's $P_e$ to produce the highest fidelity gates for each gate time. For instance, achieving the lowest infidelity and internode gate time below $.01T_1$ requires

minimal $P_e \approx .25$ and two distillation rounds. Conversely, for the lowest infidelity overall, use $P_e \approx .35$ and four distillation rounds.

Next we turn to see how this internode gate performance affects the performance of algorithms on an MNQC. At the outset, it is clear that error correction, a key objective of future MNQCs, is beyond the regime of consideration. Error correction consumes EPs at rate well in excess of a typical error correction cycle time; for example, executing a distance-$d_{sc}$ surface code $CX$ gate requires $d_{sc}$ gates [226] per error correction cycle time of $1\mu s$ [2, 53]. Given the performance profiles of internode gates, with a single gate of .99 fidelity taking $10\mu s$ (recall $T_1 = 1$ms), current hardware is insufficient to consider error corrected architectures.

Note, however, that that the distillation applied to a quantum link constitutes a form of low-level error correction [27]. Hence we consider algorithms on a simple toy system consisting of two five-qubit nodes with a single internode link with entanglement distillation. This will allow us to determine how hardware performance affects algorithm performance for a few simple benchmarks and guide the navigation of tradeoffs within the physical and distillation layers. As quantum links improve and full-blown error correction on multinode systems becomes feasible, the same analyses that we perform in the remainder of this section should be performed at the logical level.

Beginning with gate performance curves like that of Fig. 7, 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 $T_1 = T_2 = 1$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-Quantum, and each algorithm is transpiled accordingly.

As a first example, we evaluate the effective Quantum Volume (QV) [63]. 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 inter-node gate simulation.

The results of the QV benchmark are shown on a 'Gate-Algorithm Performance' (GAP plot) in Fig. 10, which compares the available performance of gates produced by the MNQC network stack with the demands of algorithms. 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. The available gate performances are overlaid on top of the shaded success regions in a similar manner as in Fig. 7. The black line depicts gates executed using raw M2O generation, while the red lines denote internode gates using entanglement distillation.

However, QV is a strenuous test, as it consists of circuits built from all possible two-qubit gates, and specific algorithms may fare better. Fig.11 shows GAP plot for a standardized battery of benchmarks [170, 261] 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. While the need for faster and higher fidelity internode gates is again apparent, the GHZ and BV benchmarks can be achieved with high fidelity using distillation. Hence we see that even though entanglement distillation increases the internode gate time, its use is critical for enabling MNQCs to execute algorithms effectively.

## 5.3 Quantum Roofline Model

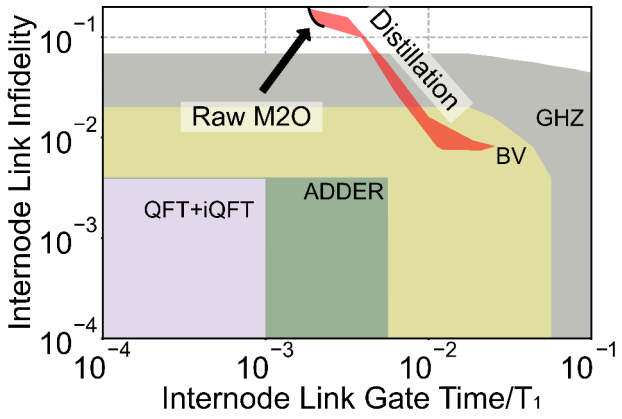



Fig. 11. Gate-algorithm performance plot of several benchmarks. Shaded regions indicate performance of >90% for the respective algorithm.

As shown in Fig. 10, the achievable performance is slower and noisier than needed for QV circuits, requiring an order-of-magnitude improvement in both rate and fidileity for the MNQC to be able to achieve a QV that improves on the single node performance at all. Approximately two orders of magnitude of improvement to both rate and fidelity are needed to achieve the maximum possible QV of $2^{10}$.

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 [282], 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 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 with $N_L$ local gates and $N_I$ internode gates, the Computation-to-Communication Ratio (CCR) is defined as:

$$CCR = \frac{N_L}{N_I} \tag{3}$$

On the other hand, a quantum system naturally executes these gates at different rates. Defining the time of execution of local gates as $T_L$ and that of internode gates as $T_I$, the Machine CCR (MCCR) is given by:

$$MCCR = \frac{T_I}{T_L} \tag{4}$$

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 set $CCR \simeq MCCR$, so as to maximize overall gate throughput while minimize circuit duration for the distributed circuit. Furthermore, the gate density [170] is defined as the occupancy of gates slots along the time evolution steps of a circuit (i.e., liveness defined in [261]), 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, with the time

| 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 1. Statistics of mapping four 10-qubit algorithm circuits to the small MNQC in Figure 9 using Qiskit (Version 0.33.0) transpiler. Comm refers to the number of internode link gates.

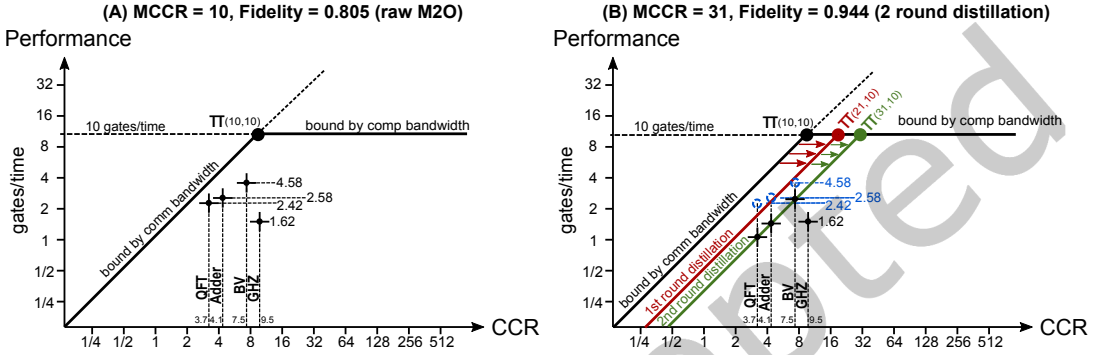



Fig. 12. Performance bound analysis for QFT, Adder, BV and GHZ on the 10-qubit MNQR system through Q-Roofline model.

unit taken to be the average gate time. Thus, for this 10-qubit system, the computation performance upper-bound is 10 gates/time. From this point, a horizontal line is drawn 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 [279] as used in Section 5.2) and so the MCCR is 10.4. This coordinate (MCCR=10.4 and 10 gates/time), determines the balance point $\pi$ in Figure 12a. From that point, drawing a 45-degree line (following the definition of CCR and MCCR), defines 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 [173]: 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 consider the Q-Roofline model evaluated for 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 1 and the settings of the system.

However, none of them can hit the bounds due to their 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: suppose that we seek to enhance the inter-node 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 and so the communication bound shifts right by one unit (shown in red). 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), leading two the green communication bound. 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.

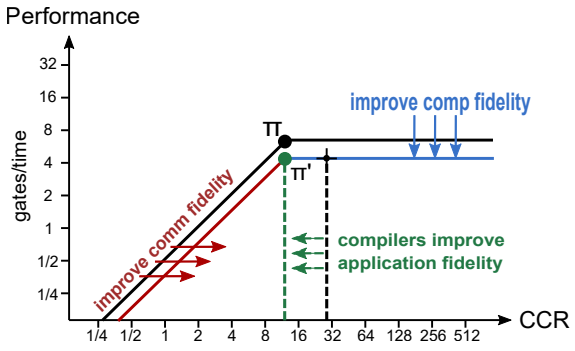



Fig. 13. The Q-Roofline model also shows how the tradeoffs in fidelity and internode gate execution time affect performance bottlenecks.

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× 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 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 $\alpha$ (Fig 14), (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.

### 5.4 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, under what conditions a quantum link can outperform a purely classical "circuit-cutting" link?

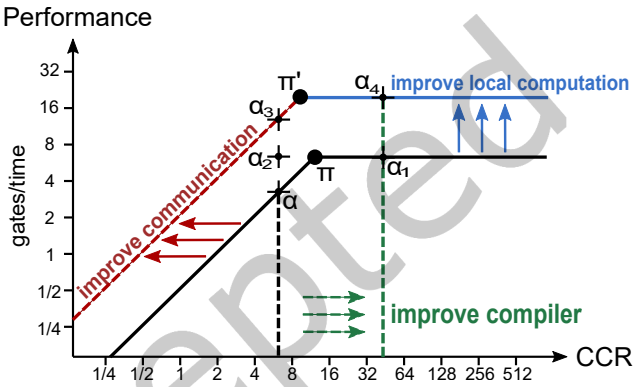



Fig. 14. Improvements to the local compute nodes and the communication operations between them increases the area beneath the hardware bounds in the Q-Roofline model.

We consider classical circuit-knitting techniques [255–257] which execute circuits separately on individual nodes in an MNQC 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 5.2 with models of error mitigation [258, 266] and circuit cutting [36, 191, 211, 216].

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 is used 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) the value of $\gamma$, defined in [266], per gate has been shown [258, 266] to be $\gamma_{\mathrm{PEC}}(d, F_{\mathrm{p}}) = \left(\frac{d^2 F_{\mathrm{p}} - 1}{d^2 - 1}\right)^{-4(d^2-1)/d^2}$, where $d$ is the dimension of the gate ($d = 4$ for a two-qubit gate) and $F_p$ is the process fidelity. For circuit knitting, this is a fixed value.

For an internode gate of fidelity $F_{\mathrm{LL}}$ and gate time $T_{\mathrm{LL}}$ the total 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 $\gamma_{\mathrm{PEC}}(4, F_{\mathrm{LL}})\gamma_{\mathrm{PEC}}^{N_q}(2, e^{-T_{\mathrm{LL}}/T_*})$, 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 [36, 191, 211, 216], there is no quantum link, but one can emulate the larger two-node system by running more circuits on the smaller devices and combining the results classically.

In Ref. [216] 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.

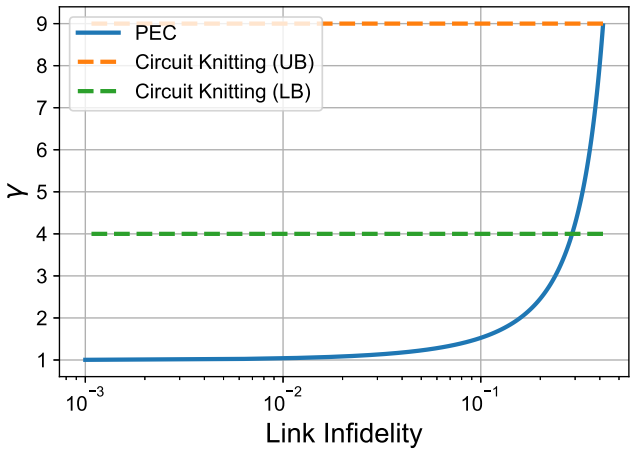



Fig. 15. Comparing the scaling of different methods to link circuit subsystems together where the number of circuits requires scales as $O(\gamma^k)$ where $k$ is the number of CX gates between the subsystems. A quantum link, even if lossy, is almost always superior.

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 for an algorithm with $k = 20$ internode gates, 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 Fig. 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 mitigate 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.

## 6 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, illustrate how these improvements combine to improve MNQC performance, and when able give estimates of the potential performance improvements in terms of the GAP, Q-Roofline, and QCPA models of the previous section. Section 6.1 outlines the potential system improvements from the physical layer, 6.2 improvements in distillation, and 6.3 in compiler improvements, and 6.4 provides an outlook on the path towards error correction.

### 6.1 Physical Layer Improvements: M2O Conversion and Multiplexing

Improving internode gate performance is a key target for enabling performant MNQCs. The analysis of section 5 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.

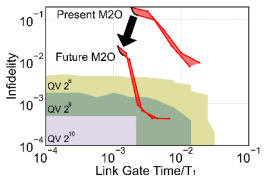
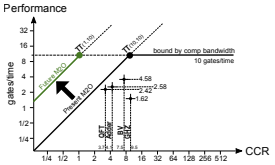
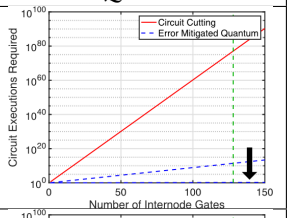
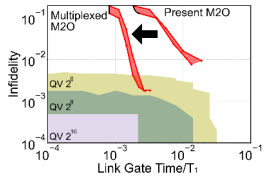
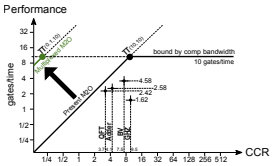
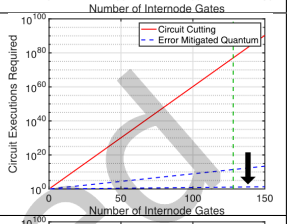
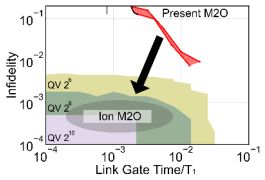
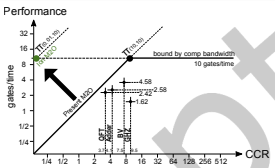
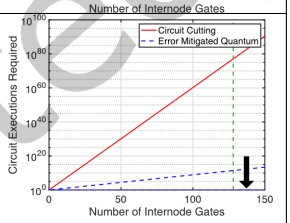
| | GAP | Q-Roofline | QCPA |
|---|---|---|---|
| **M2O Improvements:** Iterated improvements to M2O devices and protocols can yield 10x higher rate and 10x lower infidelity. |  |  |  |
| **M2O Multiplexing:** Frequency and spatial multiplexing can increase effective EP generation rate 100x. |  |  |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

Table 2. The effects of three classes of improvements to the physical layer, as demonstrated in the GAP, Q-Roofline, and QCPA analyses.



Here we discuss three ways of improving M2O performance: iterative progress on current technology to increase the rate and fidelity of M2O conversion; multiplexing existing M2O schemes to increase the rate; and finally entirely different methods of pursuing M2O conversion. We outline these research directions and provide quantitative estimates, through simulation or reference, of what overall MNQC performance they enable through the GAP, Q-Roofline, and QCPA analyses in Table 2.

Considerable progress may be made in the continued development of M2O devices. Metal reflectors [139, 149] and spot size converters [18, 199] 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 [55, 125, 175] and the use of materials with stronger nonlinearity [245]. The improvement of optical and microwave resonator [167] quality factors is also crucial for boosting the resonator extraction efficiency and intracavity field enhancement. The optical quality factor can potentially be improved by etching recipe optimization [312], the use of high-reflectivity mirrors [135], the careful waveguide design for scattering loss suppression [221], and the low-roughness material polishing [239]. The microwave quality factor, however, is majorly limited by the optical pump heating effect. 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 [276, 289], the use of superfluid helium for cooling [164], and the use of epitaxially grown superconducting materials [54, 292]. 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. Given these efforts, we propose a hypothetical M2O converter (see Appendix B Table II) as the purple curve in Fig. 5 that we wish to be available in the future. Such a M2O converter can be used to achieve >1 MHz production rate with an infidelity as low as ~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 2, 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.

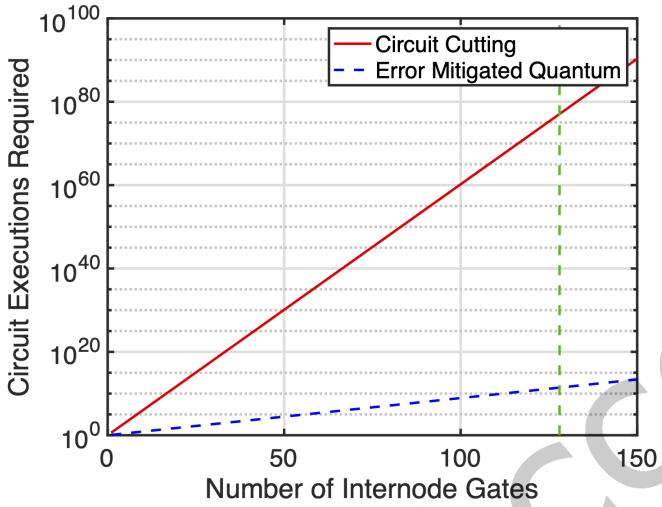



Fig. 16. Quantum-Classical Performance Analysis comparing the number of gates required for circuit cutting (red) and PEC (dotted blue) as a function of the number of internode gates.

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 heralded SPDC scheme, however, is additionally limited by the possibility of multi-photon excitations in the resonator during the SPDC process [113]. Multi-photon excitations could potentially be suppressed through the use of an anharmonic resonator [151]. 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 [21] and experimentally realized with defects in crystals and trapped ions [29, 47, 121, 213] and superconducting circuits [197]. 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 [300], the active quantum feedback for deterministic entanglement generation [179], the continuous-variable quantum teleportation [236, 287] for high-fidelity state transfer, and time-bin [306] and frequency-bin [305] 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 [126, 205, 247], for the communications.

Another key direction for improvement at the Physical layer is the use of multiplexing. As we saw in Section 5, 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 2. Particularly key here is enabling the use of error correction, which will

require bell pair generation well much faster than the typical cycle time of $1\mu s$ [2, 53], likely through the use of multiple multiplexed connections between nodes.

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 [39, 45, 169, 197]. 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 coupled qubits. One advantageous choice of bus-qubit coupler is the SNAIL (Superconducting Nonlinear Asymmetric Inductive eLement) [97] 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 [309]. 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 [184, 309].

Finally, hybrid technologies promise the greatest potential improvements, but also pose the most severe technical challenges [272]. In particular, a hybrid system using ions coupled to superconducting qubits [68, 71, 142] could allow for optical ion-ion links [181, 185, 204], between chips in separate dilution refrigerators. Significant technical challenges accompany hybrid ion-superconducting qubits [268]. However, techniques using molecular ions coupled to superconductors [7, 224, 225, 242, 275], while the use of ion chains for mode matching [159] can improve this coupling. As shown schematically in the third row of Table 2, a superconducting-ion coupling would effectively enable the rapid production of high-fidelity internode entangled pairs by using the long lifetime of ions to buffer entanglement, likely limited chiefly by the rate and fidelity of the superconducting-ion coupling [142].

## 6.2 Distillation Layer Improvements

In Section 5 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.

Here we consider three potential directions for improvements to the distillation layer. First, distillation protocols may be co-designed to adapt for the specific noise profile of the M2O generation considered in the physical layer. Secondly, we can use quantum memories to mitigate qubit decoherence while the system waits for additional Bell pairs. Finally, extremely long-lived quantum memories, with lifetimes much longer than the time required to generate all EPs required for a computation, allow for buffered execution and effectively remove the fidelity bound. We tabulate these

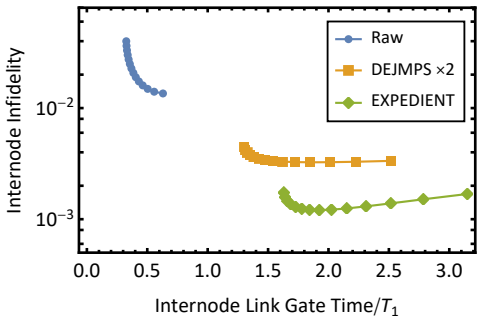



Fig. 17. Comparison of the EXPEDIENT purification protocol [203] with the two-round nested DEJMPS protocol [74]. Each point indicates a potential infidelity and latency correspdonding to the distillation protocol. Expediant attains lower infidelities, at the cost of increased bell pair consumption and latency when compared to the DEJMPS protocol.

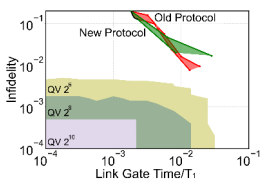
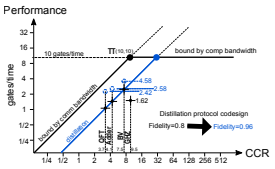
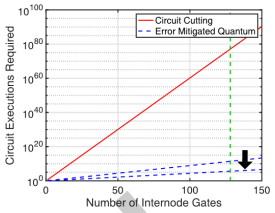
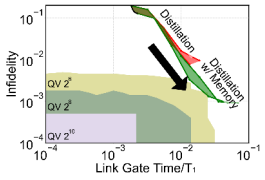
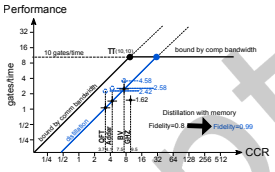
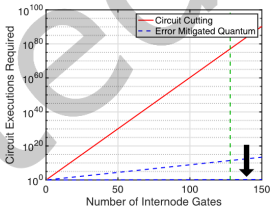
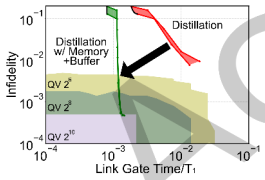
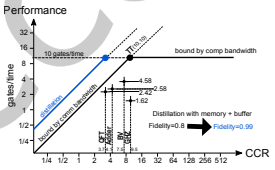
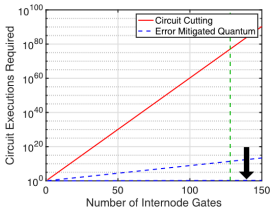
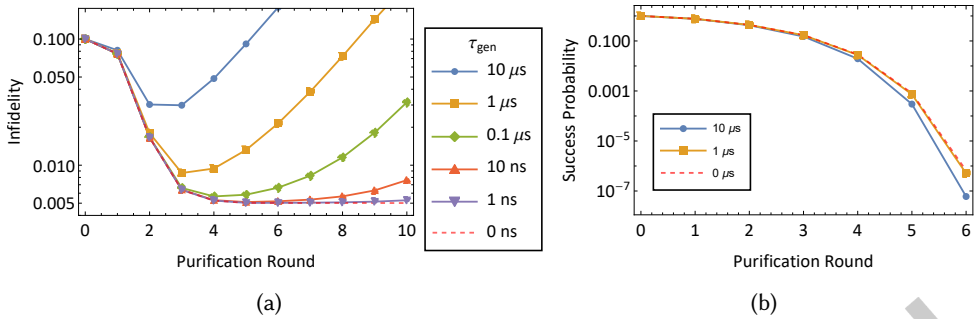
| | GAP | Q-Roofline | QCPA |
|---|---|---|---|
| **Distillation Protocol Co-Design:** Careful tailoring of distillation protocols to M2O noise may reduce infidelity 2x or more |  |  |  |
| **1ms Memory:** Protection from decoherence greatly improves distillation performance allowing new regimes of algorithms.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

Table 3. The effects of three classes of improvements to the Distillation layer, as demonstrated in the GAP, Q-Roofline, and QCPA analyses.



approaches and their effects on MNQC performance in Table 3.

To improve the fidelity of distilled EPs, several entanglement purification protocols can be used. In Appendix C of the Supplemental Material, we briefly introduce two purification protocols: the BBPSSW protocol [25] and the DEJMPS protocol [74]. In Section 4's full stack simulation the DEJMPS protocol is used, as it provides more efficient purification compared to the BBPSSW protocol [74] and uses a small number of EPs to perform purification. We also notice that more advanced purification protocols, e.g., double selection purification protocol [100], EXPEDIENT and STRINGENT purification protocols [203], may give better output EP fidelities after purification. In Fig. 17, 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 total to generate




(a) (b)

Fig. 18. The performance of the DEJMPS purification protocol [74] 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. In (b), we plot the one-shot success probability as a function of nested purification rounds.

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 [83, 109, 203] or require fewer rounds of nested purification to achieve high-fidelity EPs, so it could be used to implement more complex distributed algorithms. Distillation protocols optimized for the noise channel of the raw EP generation may be particularly helpful [131, 150] The improved performance of the EXPEDIENT protocol is shown in the first row of Table 3 (see Appendix C for more details).

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. 18a, 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 6.1, 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. 18b we plot the single-shot success probability of $n$ rounds of purification [see Eq. (C5) in Appendix C of the Supplemental Material]. 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

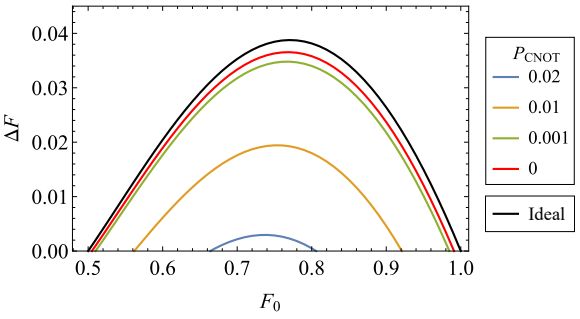



Fig. 19. The performance of the DEJMPS purification protocol with imperfections. We consider a single round of entanglement distillation in the presence of imperfect gates between superconducting qubits and finite qubit coherence times.

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 0.001. However, in reality, the local gates between the compute qubits used to purify EPs may have larger errors. In Fig. 19, 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 $P_{\text{CNOT}} = 0.01$ [280], 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 dephasing, either by having naturally long coherence times or via active or autonomous quantum error correction [105]. The effect of a memory with a 1ms coherence time is shown in the second row of Table 3, where we see that it dramatically reduces the achievable internode infidelities.

Transmon and fluxonium superconducting qubits have demonstrated high-fidelity two-qubit entangling gates [19, 79, 92, 252, 280], 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 [227] and the interfaces [273], while 3D microwave modes can serve as even better memory elements [229] with potential lifetimes up to seconds [233]. 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 [48, 49].

For these memory cavities to be effective in distillation protocols, an important area of improvement is the fidelity of cavity-qubit [49] or cavity-cavity SWAP [103] gates. These previous demonstrations rely on the shared nonlinearity of transmons to activate relatively slow four wave mixing processes. Recent experiments have shown that purpose-built parametric couplers can perform much faster SWAP operations (100 ns or less) regardless of the nonlinearity of the swapping modes [70, 114], analogous to parametric two-qubit gates [228]. Implementation of these gates may allow SWAPs of EPs to and from quantum memory elements with infidelity at the $10^{-4}$ level.

In the long term, 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 [284] of entangled pairs. The effects of a 10ms quantum memory are schematically shown in the third row of Table 3, 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 [43, 138, 259]. Recently, active and autonomous stabilization of bosonic codes has been demonstrated close to or beyond the break-even point including the cat code [105, 205], the binomial code [127], and the GKP code [44, 247]. However, in these experiments the coherence of the error-corrected quantum

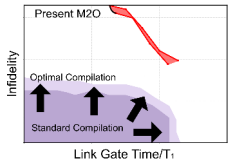
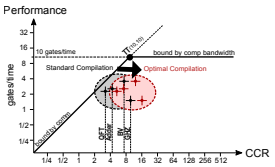
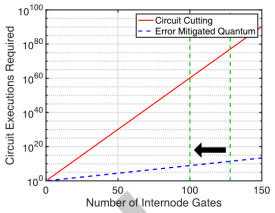
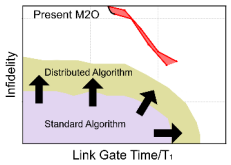
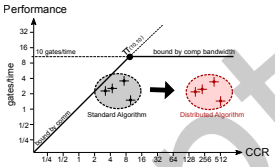
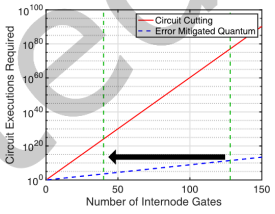
| | GAP | Q-Roofline | QCPA |
|---|---|---|---|
| **Efficient Compiling:** Compiler improvements reduce internode gate use by a factor of 2x or more and scale to large sizes |  |  |  |
| **Distributed Algorithms:** Large gains obtained with algorithms designed for distributed systems, reducing internode gates by 100x or more. |  |  |  |

Table 4. Schematic depiction of the effects of improvements to the compiler and application layers, as demonstrated in the GAP, Q-Roofline, and QCPA analyses.



memory is limited by that of the nonlinear 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 [111, 222, 234].

## 6.3 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 5, 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 and show how to quantify progress 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.

*6.3.1 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 program 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 [32, 62, 246]. Current compilers are capable of translating large programs (containing more than $10^6$ program qubits and gates) into hardware-*agnostic* assembly programs [132], but are currently limited in their ability to map this to a hardware-compatible executable [253]. 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 [101, 155, 206]. Quantum compilers may look to the field of classical HPC where load balancing has been extensively studied and efficient heuristic methods have been developed [140, 210]. Compiler-oriented partitioning, where circuit partitioning algorithms [9, 66, 69] are applied during compilation, can also be applied to optimize for minimal communications, maximum fidelity, and balanced workloads [90]. Once a program has been partitioned, distribution binds circuit partitions to module nodes and schedules inter-node communications with the goal of shortening the critical path (e.g., hide communication latency with local computation) and maximize program success rates while respecting communication dependencies constraints [90]. The architecture and metrics introduced in Fig. 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 [284], aggregation [285], and collective communication [119, 284].

Furthermore, good system performance may be achieved through efficient load balancing by boosting local occupancy and minimizing communication overheads. As discussed in Section 5, 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 Fig. 12, the connectivity constraints of the hardware may lead to an application becoming communication bound, and Fig. 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 [244] and thus it will be necessary to balance quantum program success rates with user efficiency when designing distributed quantum systems. Secondly, while Section 5 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` [295] 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 [25] and teleportation [186]. However, other protocols may be used requiring only classical channels as in quantum circuit cutting [211, 256, 257] and entanglement forging [84], or solely quantum channels such as shuttling [235], direct state transfer [16] and cross-chip two-qubit gates [107]. Each protocol presents unique tradeoffs between fidelity, speed, and ease of implementation that any future compiler for a distributed system must consider.

*6.3.2 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 [260, 261]. Our work in Section 5 and Fig. 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 [13]. 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 [112], distributed Shor's algorithm and arithmetic [186, 187, 294], distributed VQE (via classical networks [249] or quantum interconnects [77]), and distributed phase estimation [230].

In Section 5, we noted that while many complex algorithms were unachievable using present technology, GHZ creation could be performed with high fidelity. In turn, this actually implies that Quantum Phase Estimation (QPE) is among the best candidates for execution on early MNQCs. 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 [146] which gives a reduction in the depth of phase estimation that is linear in the number of nodes and an approach that uses classical communication [230]. Both of these are reviewed in detail in Appendix A of the Supplemental Material. 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 Fig. 20 [146]. 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.

We now turn to algorithms beyond distributed QPE. 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 [23, 148, 161, 281, 293] and quantum minimal entanglement typical thermal state sampling for finite-temperature simulations [196]. Additionally, in complex chemical systems

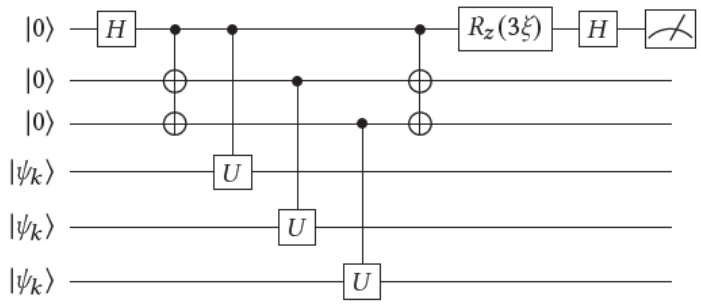



Fig. 20. Distributed Quantum Phase Estimation circuit on $p = 3$ nodes. Although QPE is a high-depth algorithm, the highly communication-efficient structure of the distributed QPE circuit renders it a natural candidate for early MNQCs.



such as metal-organic framework [310] and protein-ligand binding [98, 277], 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 [271], and quantum transport systems of leads through nanocontacts naturally minimize the number of inter-node nonlocal gates [108, 115]. 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 [106]. Alternatively, real-space Wannier function representations to achieve compact encodings of electronic orbitals [59] 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 [180]. 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 of the Supplemental Material. Beyond semi-classical dynamics, it is possible to rephrase quantum dynamics as finding the ground state of a composite Hamiltonian [182], 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 [1]. A recent quantum algorithm has demonstrated that a VQE approach can be designed with the same advantages as a cluster algorithm [303]. Further research based on clustering algorithms may allow for new VQE approaches that are well suited for MNQC architectures.

### 6.4 Towards Error Correction for MNQCs

In Section 5.2, we found that current M2O hardware performance was insufficient to allow full error correction. However, error correction remains a goal for MNQCs in the far future, and here we lay out what hardware progress should be targeted for error correction, and how this changes our proposed architecture and analyses.

The route towards error correction has been illuminated through recent work [226] which indicates, for the particular case of surface code, that the threshold at a boundary can be as high as 10% if enough pairs are produced to enable an internode gate. The key is then producing sufficent entanglement to allow for lattice surgery operations, i.e. achieving O(10) internode gates with infidelity .1 or less per cycle time of O(1 $\mu s$), for an average internode gate time O(100(ns)).

Referring to the GAP analysis, we see that current technology suffers from generation rates too slow and infidelity too high to allow error correction. However, the results of this section show that this objective is potentially achievable through iterated M2O improvements combined with multiplexing, which can reduce gate time to 1-2 μs and further decrease that time by a factor of 10x or more. New transduction schemes, such as superconducting-ion coupling that permits buffering, may also contribute to this goal.

To accommodate error correction, our proposed architecture would need to be modified, potentially by including an Error Correction layer. This layer must coordinate local and internode operations across the lowest level of the upper layers, facilitating lattice surgery operations in surface code, for example. Additionally, it must negotiate with the data layer (which interacts with the distillation layer) to execute internode operations at the required speed and fidelity to maintain a threshold below the limit. In particular, the role of the network stack, which we focus on in this paper, would remain largely unchanged, providing internode gates at given times and fidelities to the upper layers as requested.

Moreover, the method of analysis we present in this paper can be adapted to accommodate error correction. For instance, the gap clock toy model transforms into a system of 10 logical qubits, and the quantum roofline plot's bandwidth and computation bounds are interpreted in terms of logical gates rather than physical gates, where the tradeoffs between bandwith and computation may be even more relevant for fault-tolerant operation. Similarly, the trade-offs between classical and quantum gates become increasingly important due to the overhead of logical quantum internode gates. Circuit cutting will continue to play a role in algorithm execution, complementing error-corrected quantum internode gates. Future work will explore a qualitative analysis of error-corrected fault-tolerant systems to facilitate ongoing hardware development.

## 7 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 [20, 94], 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 [64, 278] for distributed computing to smaller networks using cryogenic microwave links [35, 107, 178, 291, 309], which show considerable promise in the nearest term. Similarly, our approach can also characterize modular trapped ion systems [38, 141, 214, 214] and neutral atom platforms [30, 297]. Hybrid systems consisting of several of these technologies [68, 71, 142, 201, 241] 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.

## 8 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, benchmark execution, and led the publishing effort. 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 provided 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. H. Tang guided the development of the physical layer simulations and roadmap. C. Wang, R. Schoelkopf, M. Hatridge, A.A. Houck, S. M. Girvin, and A. Eickbusch guided the development of hardware models, protocols, and quantum memory. S. Economou guided the simulations of entanglement distillation and the distillation roadmap. K. M. Fu and A. Faraon contributed to novel entanglement generation protocols. J. Ang, K. Krsulich, M. Ritter, M. Martonosi contributed quantum compiler metrics and classical architecture comparisons. G. Carini and J. Misewich contibuted to the roadmap towards distributed and interoperable QC. Y. Liu, Y.X. Yao, D. Yost, N. Tubman contributed potential distributed quantum algorithms. 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.

## 9 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.

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231.

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

External Interest Disclosure: RJS is a founder and equity holder in Quantum Circuits, Inc. (QCI). MH and SMG are consultants and equity holders in QCI.

## 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 [146] 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 [230]). 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.

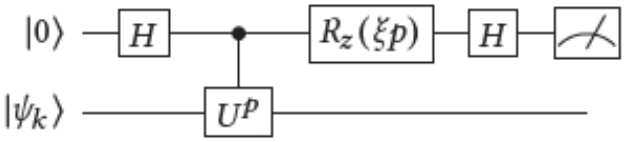
The central challenge of phase estimation is that the optimal scaling is given by the Heisenberg limit $|\hat{\theta}_k - \theta_k| \leq \frac{\pi}{T}$ for any quantum algorithm that estimates $\theta_k$ using $T$ applications of the unitary $U$. For applications in chemistry, these errors need to be on the order of $10^{-4}$ or smaller [230], 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

$$O\left(\frac{\log(T)}{\delta\epsilon_\theta}\right) \tag{5}$$

Next we need to invoke the parallelized phase estimation procedure of [230]. 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




(6)

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 [230] can be employed). By choosing $p$ adaptively using the heuristic in [230] 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

$$\text{Depth}_{U,\text{Cl}} = O\left(\frac{\log(1/\epsilon)}{\delta\epsilon_\theta}\right). \tag{7}$$

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 in the main text. 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 (6) 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

$$\mathrm{Depth}_{U,\mathrm{Qm}} = O\left(\frac{\log(1/\epsilon)}{\delta\epsilon_\theta}\right); \tag{8}$$

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

$$\|\Lambda^{\circ m} - \tilde{\Lambda}^{\circ m}\|_\diamond \leq m\Delta. \tag{9}$$

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

$$\|\mathrm{Tr}(\Lambda^{\circ m}(\rho)Q) - \mathrm{Tr}(\tilde{\Lambda}^{\circ m}(\rho)Q)\| \leq m\|Q\|\Delta. \tag{10}$$

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

$$\Delta = \frac{\epsilon}{m\pi} = O\left(\frac{\epsilon}{T\log(T)}\right) \tag{11}$$

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

$$\Delta = \frac{\epsilon}{m\pi} = O\left(\frac{\epsilon^2}{\log(1/\epsilon)}\right) \tag{12}$$

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 $O(T\text{polylog}(T\log(T)/\epsilon)) = \widetilde{O}(T\text{polylog}(1/\epsilon))$ 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 [87]. 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

$$\text{Depth}_{U,Dist} = O\left(\frac{\log(1/\epsilon)}{\delta\epsilon} + \frac{1}{\epsilon T} + \gamma T\text{polylog}(T/\epsilon)\right) \tag{13}$$

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

$$\text{Depth}^{\text{opt}}_{U,Dist} = \tilde{O}\left(\frac{\log(1/\epsilon)}{\delta\epsilon} + \sqrt{\frac{\gamma}{\epsilon}}\right). \tag{14}$$

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 [46] 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 [166] 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

$$\text{Depth}_{U,Dist} = O\left(\frac{\alpha\log(1/\epsilon)}{\delta E_{\text{gap}}} + \frac{\alpha^4}{\epsilon^4 T} + \gamma T\text{polylog}(T/\epsilon)\right) \tag{15}$$

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.

## B M2O Conversion Simulation

The simulation model used for entangled generation is shown in Fig. 21(a) [197]. 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 [197]. 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. 21(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 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) [99]. 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 [86], $\gamma_{\text{tot,o}} = \gamma_{\text{ext,o}} + \gamma_{\text{int,o}}$ 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

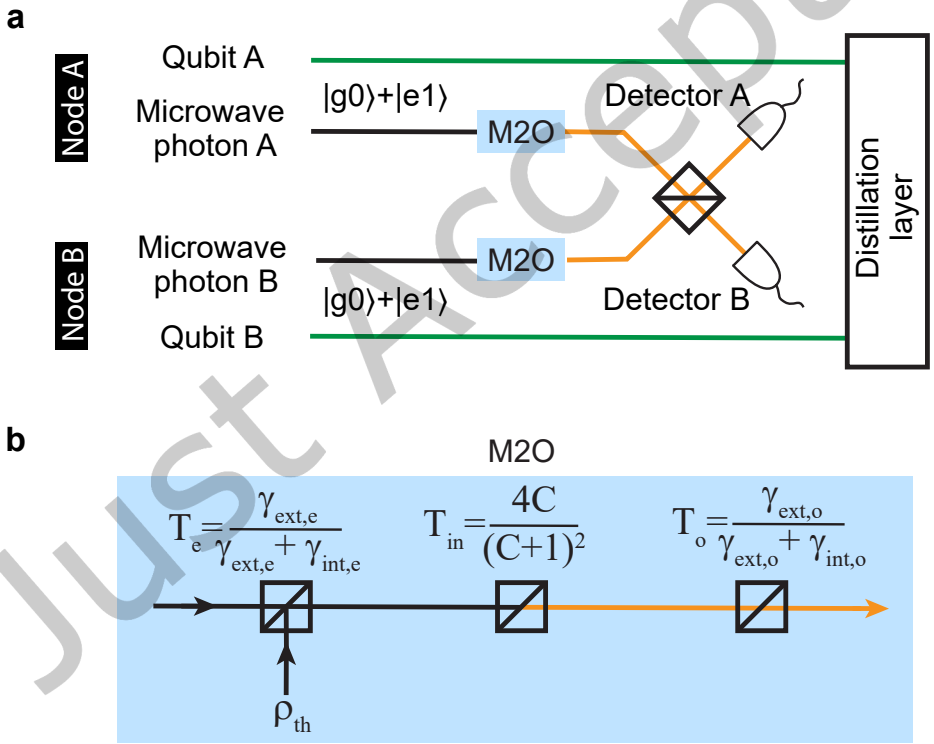



Fig. 21. (a), the diagram of the scheme used for remote entangled qubit generation. (b), the M2O converter is modeled by a series of beamsplitters. The first beamsplitter represents the microwave resonator extraction efficiency, the second beamsplitter represents the intracavity M2O conversion efficiency, 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.

| No. | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Platform | Electro-optomechanics | Bulk electro-optics | Integrated electro-optics | Optomagnonics | Rare-earth ions | Cold atoms |
| Single-photon coupling rate (Hz) | $\frac{g_{\rm om}}{2\pi} = 60$<br>$\frac{g_{\rm em}}{2\pi} = 1.6$ | $\frac{g_{\rm eo}}{2\pi} = 37$ | $\frac{g_{\rm eo}}{2\pi} = 750$ | $\frac{g_{\rm mago}}{2\pi} = 17.2$ | - | - |
| Cavity decay rate (Hz) | $\frac{\gamma_{\rm ext,o}}{2\pi} = 2.1\times10^6$<br>$\frac{\gamma_{\rm int,o}}{2\pi} = 5.6\times10^5$<br>$\frac{\gamma_{\rm ext,e}}{2\pi} = 1.4\times10^6$<br>$\frac{\gamma_{\rm int,e}}{2\pi} = 1.3\times10^6$ | $\frac{\gamma_{\rm ext,o}}{2\pi} = 1.5\times10^7$<br>$\frac{\gamma_{\rm int,o}}{2\pi} = 1.1\times10^7$<br>$\frac{\gamma_{\rm ext,e}}{2\pi} = 5.6\times10^6$<br>$\frac{\gamma_{\rm int,e}}{2\pi} = 8.1\times10^6$ | $\frac{\gamma_{\rm ext,o}}{2\pi} = 3.3\times10^7$<br>$\frac{\gamma_{\rm int,o}}{2\pi} = 1.4\times10^8$<br>$\frac{\gamma_{\rm ext,e}}{2\pi} = 3.2\times10^6$<br>$\frac{\gamma_{\rm int,e}}{2\pi} = 5.8\times10^6$ | $\frac{\gamma_{\rm ext,o}}{2\pi} = 4.8\times10^7$<br>$\frac{\gamma_{\rm int,o}}{2\pi} = 1.5\times10^9$<br>$\frac{\gamma_{\rm ext,e}}{2\pi} = 1.7\times10^8$<br>$\frac{\gamma_{\rm int,e}}{2\pi} = 8.5\times10^7$ | - | - |
| Cooperativity | $C_{\rm om} = 4.5\times10^4$<br>$C_{\rm em} = 1.0\times10^4$ | $C_{\rm eo} = 0.92$ | $C_{\rm eo} = 0.04$ | $C_{\rm mago} = 4.1\times10^{-7}$<br>$C_{\rm mage} = 0.8$ | - | - |
| Efficiency | $\eta_{\rm tot} = 0.19$<br>$\eta_{\rm in} = 0.59$ | $\eta_{\rm tot} = 0.14$<br>$\eta_{\rm in} = 0.99$ | $\eta_{\rm tot} = 0.01$<br>$\eta_{\rm in} = 0.15$ | $\eta_{\rm tot} = 1.1\times10^{-8}$<br>$\eta_{\rm in} = 5.2\times10^{-7}$ | $\eta_{\rm tot} = 1.26\times10^{-5}$ | $\eta_{\rm tot} = 0.82$ |
| Bandwidth (Hz) | $6.1\times10^3$ | - | - | $1.6\times10^7$ | - | $1\times10^6$ |
| Added noise $n_{\rm add}$ | 1.4 | 0.41 | - | - | - | 0.8 |
| Environment temperature (K) | 0.04 | 0.01 | 1.9 | 300 | 4.6 | 300 |
| Reference | [73] | [239] | [290] | [311] | [22] | [264] |

Table 5. Summary of M2O converter performances on different experimental platforms. The definitions of parameters are discussed in Appendix B.

| Platform | EOM | Bulk EO | Integrated EO | Future |
|---|---|---|---|---|
| $\frac{g_0}{2\pi}$ (Hz) | 60 | 37 | 750 | 1000 |
| $\frac{\gamma_{\rm ext,o}}{2\pi}$ (Hz) | $2.1\times10^6$ | $1.5\times10^7$ | $3.3\times10^7$ | $10^7$ |
| $\frac{\gamma_{\rm int,o}}{2\pi}$ (Hz) | $1.1\times10^5$ | $2.2\times10^6$ | $2.8\times10^7$ | $2\times10^5$ |
| $\frac{\gamma_{\rm ext,e}}{2\pi}$ (Hz) | $1.4\times10^6$ | $5.6\times10^6$ | $3.2\times10^6$ | $10^7$ |
| $\frac{\gamma_{\rm int,e}}{2\pi}$ (Hz) | $2.6\times10^5$ | $1.6\times10^6$ | $1.2\times10^6$ | $2\times10^5$ |

Table 6. Parameter sets used for M2O entanglement generation simulation based on the electro-optomechanics (EOM) and electro-optics (EO) platform. The optical and microwave resonator intrinsic decay rate are made 5 times lower than the original values in Table 5. The last column presents a hypothetical parameter set which we wish to be available in the future.



of $T_o = \gamma_{\rm ext,o}/(\gamma_{\rm ext,o} + \gamma_{\rm int,o})$ which represents the optical resonator extraction efficiency. We assume an optical detector dark count rate of 50 Hz [198].

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 [137] to obtain the density matrix $\rho_{\rm f}$ 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 $B \approx \gamma_{\rm tot,e}$ [86], 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 $t_{\rm tot} = t_1 + t_2 + t_3$. 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 $P_{\rm herald} = {\rm Tr}\,\langle 1,0|\,\rho_{\rm f}\,|1,0\rangle$. Thus, the entanglement generation rate can be computed as $R = P_{\rm herald}/t_{\rm tot}$. In the case of a successful heralding, the corresponding qubit state is $\rho_q = \langle 1,0|\,\rho_{\rm f}\,|1,0\rangle\,/{\rm Tr}\,\langle 1,0|\,\rho_{\rm f}\,|1,0\rangle$, 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 6. The first three parameter sets come from Table 5, 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 [102, 136, 312] and hence we still refer to them as 'current M2O' in the manuscript, despite several optimistic assumptions made above. In addition, although the No. 1 converter in Table 5 is based on electro-optomechanical effects which require different formulas to calculate its conversion efficiency and bandwidth [12], 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 in the main text. 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 of the main text. 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.

## 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 the fidelity of the Bell state to unity [25]. 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

$$F_{\text{new}} = \frac{F^2 + (1-F)^2/9}{F^2 + 2F(1-F)/3 + 5(1-F)^2/9}. \tag{16}$$

Following this work, Deutsch *et al* proposed a similar method (DEJMPS), which improves the efficiency of the purification protocol [74]. 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 [74].

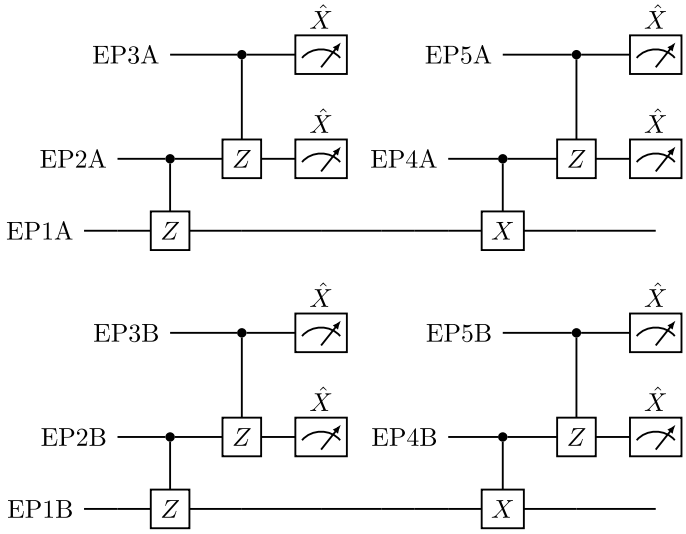



Fig. 22. The EXPEDIENT purification protocol gate sequence [203]. EP1A and EP1B are the two qubits corresponding to the same EP.

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 [83]. 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. [74], 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,

$$\rho_{\text{new}} = \mathcal{P}[\rho_{\text{old},1} \otimes \rho_{\text{old},2}], \tag{17}$$

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. [74].

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

$$t_n = 2^{n-1}\tau + t_{n-1} + t_p. \tag{18}$$

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

$$\rho_{(n)} = \mathcal{E}(t_p)\left[\mathcal{P}\left[\mathcal{E}(t_{\text{idle},n-1})\left[\rho_{(n-1)}\right]\otimes\rho_{(n-1)}\right]\right] \tag{19}$$

The probabilistic nature of entanglement purification is from the measurement on one pair of Bell states. As pointed out in Ref. [74], 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

$$P_n = \prod_{j=1}^{n} p_j^{2^{j-1}}. \tag{20}$$

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.

For completion, in Fig. 22, we show the EXPEDIENT purification protocol gate sequence. It consumes five EPs (EP1 to EP5) to get an improved EP. Similar to the Deutsch protocol, we consider possible decay and decoherence errors while the EPs are idling in the quantum registers and the error on the entangling gates between the qubits. Note that there is competition between the EP generation and the two-qubit gate operations. For example, when the EP2 is prepared, and EP1 and EP2 are applying the control-$Z$ gate, the EP3 is attempting to be generated. If the gate time is short compared to the EP generation time $\tau$, after the control-$Z$ gate, EP1 and EP2 are idling. Otherwise, the EP3 needs to wait for the control-$Z$ gate, which will suffer decay and decoherence errors instead. For example, after taking the measurement of EP2 and EP3, when $\tau_g < \tau$, the state of EP1 is

$$\begin{aligned}\rho_{\text{new}} = \mathcal{P}_{XX}\{U_{\text{CZ},32}\mathcal{E}(\tau-\tau_g)\left[U_{\text{CZ},21}\mathcal{E}(\tau)[\rho_0]\otimes\rho_0\right.\\ \left.\times U_{\text{CZ},21}^{\dagger}\right]\otimes\rho_0 U_{\text{CZ},32}^{\dagger}\Big\},\end{aligned} \tag{21}$$

where $\mathcal{P}_{XX}$ is the quantum channel for a success $\hat{X}\hat{X}$ coincident measurements on EP2 and EP3, $U_{\text{CZ}}$ are the unitary for the pair of CZ gates, $\rho_0$ is the state of a raw EP. While if $\tau g > \tau$, the state of EP1 is

$$\rho_{\text{new}} = \mathcal{P}_{XX}\{U_{\text{CZ},32}U_{\text{CZ},21}\mathcal{E}(\tau)[\rho_0]\otimes\rho_0\times U_{\text{CZ},21}^{\dagger}\otimes\mathcal{E}(\tau-\tau_g)[\rho_0]U_{\text{CZ},32}^{\dagger}\Big\}.$$

A similar analysis is applied when EP4 and EP5 are used for purification.

## 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 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 of the main text, and optionally distilled by the protocol underlined therein. 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 remone 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.

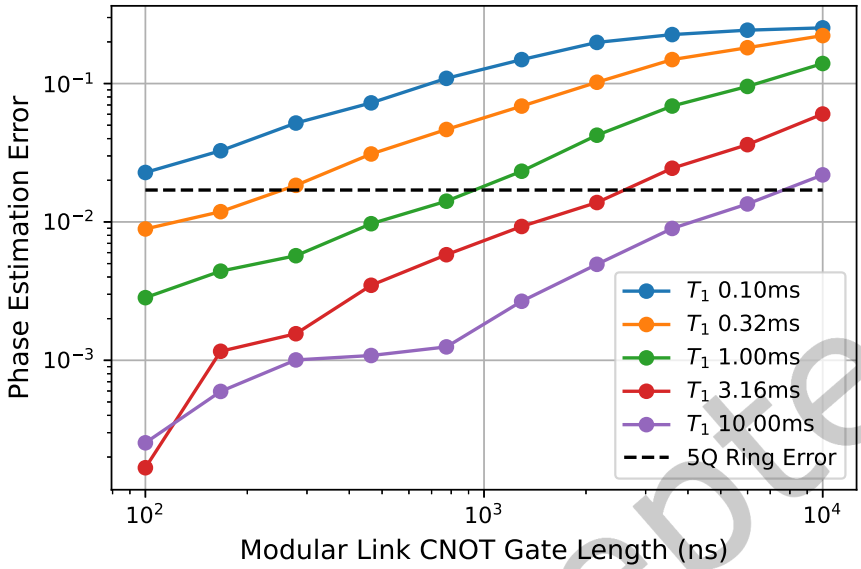



Fig. 23. QPE Error for a 10-qubit algorithm as a function of the modular gate length for different values of $T_1$.

## 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 of the main text 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.

## 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 $T_* = T_1T_2/(T_1 + T_2)$ is the effective fidelity lifetime of a local qubit.

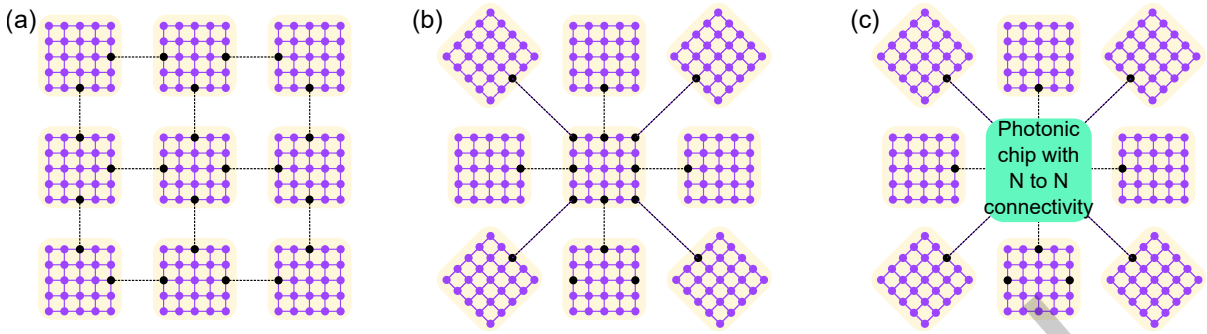



Fig. 24. (a) Array-like, (b) star-like architectures of a distributed quantum computer. (c) Star-like architecture with an N-to-N photonic chip enabling EP generation between any communication qubits from other superconducting chips.

Requiring a given overall logarithmic fidelity of $\log F$ then requires:

$$\log F = \text{const.} \sim log(e^{\xi_I} + e^{\xi_T}) \tag{22}$$

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.

## 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 appart 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 [31, 52, 212]. Using a central photonic node has been for example envisioned for trapped ions [193]. A universal photonic chip is an $N \times N$ linear interferometer based on phase-controled 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)).

## References


[1] Vibin Abraham and Nicholas J. Mayhall. 2020. Selected Configuration Interaction in a Basis of Cluster State Tensor Products. *Journal of Chemical Theory and Computation* 16, 10 (2020), 6098–6113. https://doi.org/10.1021/acs.jctc.0c00141 arXiv:https://doi.org/10.1021/acs.jctc.0c00141 PMID: 32846094.

[2] Rajeev Acharya et al. 2022. Suppressing quantum errors by scaling a surface code logical qubit. (7 2022). arXiv:2207.06431 [quant-ph]

[3] Rajeev Acharya, Igor Aleiner, Richard Allen, Trond I. Andersen, Markus Ansmann, Frank Arute, Kunal Arya, Abraham Asfaw, Juan Atalaya, Ryan Babbush, Dave Bacon, Joseph C. Bardin, Joao Basso, Andreas Bengtsson, Sergio Boixo, Gina Bortoli, Alexandre Bourassa, Jenna Bovaird, Leon Brill, Michael Broughton, Bob B. Buckley, David A. Buell, Tim Burger, Brian Burkett, Nicholas Bushnell, Yu Chen, Zijun Chen, Ben Chiaro, Josh Cogan, Roberto Collins, Paul Conner, William Courtney, Alexander L. Crook, Ben Curtin, Dripto M. Debroy, Alexander Del Toro Barba, Sean Demura, Andrew Dunsworth, Daniel Eppens, Catherine Erickson, Lara Faoro, Edward Farhi, Reza Fatemi, Leslie Flores Burgos, Ebrahim Forati, Austin G. Fowler, Brooks Foxen, William Giang, Craig Gidney, Dar Gilboa, Marissa Giustina, Alejandro Grajales Dau, Jonathan A. Gross, Steve Habegger, Michael C. Hamilton, Matthew P. Harrigan, Sean D. Harrington, Oscar Higgott, Jeremy Hilton, Markus Hoffmann, Sabrina Hong, Trent Huang, Ashley Huff, William J. Huggins, Lev B. Ioffe, Sergei V. Isakov, Justin Iveland, Evan Jeffrey, Zhang Jiang, Cody Jones, Pavol Juhas, Dvir Kafri, Kostyantyn Kechedzhi, Julian Kelly, Tanuj Khattar, Mostafa Khezri, Mária Kieferová, Seon Kim, Alexei Kitaev, Paul V. Klimov, Andrey R. Klots, Alexander N. Korotkov, Fedor Kostritsa, John Mark Kreikebaum, David Landhuis, Pavel Laptev, Kim-Ming Lau, Lily Laws, Joonho Lee, Kenny Lee, Brian J. Lester, Alexander Lill, Wayne Liu, Aditya Locharla, Erik Lucero, Fionn D. Malone, Jeffrey Marshall, Orion Martin, Jarrod R. McClean, Trevor Mccourt, Matt McEwen, Anthony Megrant, Bernardo Meurer Costa, Xiao Mi, Kevin C. Miao, Masoud Mohseni, Shirin Montazeri, Alexis Morvan, Emily Mount, Wojciech Mruczkiewicz, Ofer Naaman, Matthew Neeley, Charles Neill, Ani Nersisyan, Hartmut Neven, Michael Newman, Jiun How Ng, Anthony Nguyen, Murray Nguyen, Murphy Yuezhen Niu, Thomas E. O'Brien, Alex Opremcak, John Platt, Andre Petukhov, Rebecca Potter, Leonid P. Pryadko, Chris Quintana, Pedram Roushan, Nicholas C. Rubin, Negar Saei, Daniel Sank, Kannan Sankaragomathi, Kevin J. Satzinger, Henry F. Schurkus, Christopher Schuster, Michael J. Shearn, Aaron Shorter, Vladimir Shvarts, Jindra Skruzny, Vadim Smelyanskiy, W. Clarke Smith, George Sterling, Doug Strain, Marco Szalay, Alfredo Torres, Guifre Vidal, Benjamin Villalonga, Catherine Vollgraff Heidweiller, Theodore White, Cheng Xing, Z. Jamie Yao, Ping Yeh, Juhwan Yoo, Grayson Young, Adam Zalcman, Yaxing Zhang, and Ningfeng Zhu. 2022. Suppressing quantum errors by scaling a surface code logical qubit. *arXiv preprint arXiv:2207.06431* (2022).

[4] A. Agarwal, R. Bianchini, D. Chaiken, F.T. Chong, K.L. Johnson, D. Kranz, J.D. Kubiatowicz, Beng-Hong Lim, K. Mackenzie, and D. Yeung. 1999. The MIT Alewife Machine. *Proc. IEEE* 87, 3 (1999), 430–444. https://doi.org/10.1109/5.747864

[5] A. Agarwal, R. Bianchini, D. Chaiken, K.L. Johnson, D. Kranz, J. Kubiatowicz, Beng-Hong Lim, K. Mackenzie, and D. Yeung. 1995. The MIT Alewife machine: architecture and performance. In *Proceedings 22nd Annual International Symposium on Computer Architecture*. 2–13. https://doi.org/10.1109/ISCA.1995.524544

[6] Hao Ai, Ying-Yü Fang, Cheng-Rui Feng, Zhihui Peng, and Ze-Liang Xiang. 2022. Multinode State Transfer and Nonlocal State Preparation via a Unidirectional Quantum Network. *Phys. Rev. Applied* 17 (May 2022), 054021. Issue 5. https://doi.org/10.1103/PhysRevApplied.17.054021

[7] A. Andre, D. DeMille, J. M. Doyle, M. D. Lukin, S. E. Maxwell, P. Rabl, R. Schoelkopf, and P. Zoller. 2006. Polar molecules near superconducting resonators: a coherent, all-electrical, molecule-mesoscopic interface. *arXiv e-prints*, Article quant-ph/0605201 (May 2006), quant-ph/0605201 pages. arXiv:quant-ph/0605201 [quant-ph]

[8] Pablo Andres-Martinez, Tim Forrer, Daniel Mills, Jun-Yi Wu, Luciana Henaut, Kentaro Yamamoto, Mio Murao, and Ross Duncan. 2023. Distributing circuits over heterogeneous, modular quantum computing network architectures. *arXiv preprint arXiv:2305.14148* (2023).

[9] Pablo Andres-Martinez and Chris Heunen. 2019. Automated distribution of quantum circuits via hypergraph partitioning. *Physical Review A* 100, 3 (2019), 032308.

[10] Reed W Andrews, Robert W Peterson, Tom P Purdy, Katarina Cicak, Raymond W Simmonds, Cindy A Regal, and Konrad W Lehnert. 2014. Bidirectional and efficient conversion between microwave and optical light. *Nat. Phys.* 10, 4 (2014), 321–326.

[11] James A. Ang. 2015. High Performance Computing Co-Design Strategies. In *Proceedings of the 2015 International Symposium on Memory Systems* (Washington DC, DC, USA) *(MEMSYS '15)*. Association for Computing Machinery, New York, NY, USA, 51–52. https://doi.org/10.1145/2818950.2818959

[12] G Arnold, Matthias Wulf, Shabir Barzanjeh, ES Redchenko, A Rueda, William J Hease, Farid Hassani, and Johannes M Fink. 2020. Converting microwave and telecom photons with a silicon photonic nanomechanical interface. *Nat. Commun.* 11 (2020), 4460.

[13] Krste Asanovic, Ras Bodik, Bryan Christopher Catanzaro, Joseph James Gebis, Parry Husbands, Kurt Keutzer, David A Patterson, William Lester Plishker, John Shalf, Samuel Webb Williams, et al. 2006. The landscape of parallel computing research: A view from berkeley. (2006).

[14] David Awschalom, Karl K. Berggren, Hannes Bernien, Sunil Bhave, Lincoln D. Carr, Paul Davids, Sophia E. Economou, Dirk Englund, Andrei Faraon, Martin Fejer, Saikat Guha, Martin V. Gustafsson, Evelyn Hu, Liang Jiang, Jungsang Kim, Boris Korzh, Prem Kumar, Paul G. Kwiat, Marko Lončar, Mikhail D. Lukin, David A.B. Miller, Christopher Monroe, Sae Woo Nam, Prineha Narang, Jason S.

Orcutt, Michael G. Raymer, Amir H. Safavi-Naeini, Maria Spiropulu, Kartik Srinivasan, Shuo Sun, Jelena Vučković, Edo Waks, Ronald Walsworth, Andrew M. Weiner, and Zheshen Zhang. 2021. Development of Quantum Interconnects (QuICs) for Next-Generation Information Technologies. *PRX Quantum* 2 (Feb 2021), 017002. Issue 1. https://doi.org/10.1103/PRXQuantum.2.017002

[15] David Awschalom, Karl K. Berggren, Hannes Bernien, Sunil Bhave, Lincoln D. Carr, Paul Davids, Sophia E. Economou, Dirk Englund, Andrei Faraon, Marty Fejer, Saikat Guha, Martin V. Gustafsson, Evelyn Hu, Liang Jiang, Jungsang Kim, Boris Korzh, Prem Kumar, Paul G. Kwiat, Marko Lončar, Mikhail D. Lukin, David A. B. Miller, Christopher Monroe, Sae Woo Nam, Prineha Narang, Jason S. Orcutt, Michael G. Raymer, Amir H. Safavi-Naeini, Maria Spiropulu, Kartik Srinivasan, Shuo Sun, Jelena Vučković, Edo Waks, Ronald Walsworth, Andrew M. Weiner, and Zheshen Zhang. 2019. Development of Quantum InterConnects for Next-Generation Information Technologies. *arXiv e-prints*, Article arXiv:1912.06642 (Dec. 2019), arXiv:1912.06642 pages. arXiv:1912.06642 [quant-ph]

[16] Christopher J Axline, Luke D Burkhart, Wolfgang Pfaff, Mengzhen Zhang, Kevin Chou, Philippe Campagne-Ibarcq, Philip Reinhold, Luigi Frunzio, S M Girvin, Liang Jiang, M H Devoret, and R J Schoelkopf. 2018. On-demand quantum state transfer and entanglement between remote microwave cavity memories. *Nature Physics* 14, 7 (2018), 705–710. https://doi.org/10.1038/s41567-018-0115-y

[17] Jonathan M. Baker, Casey Duckering, Alexander Hoover, and Frederic T. Chong. 2020. Time-Sliced Quantum Circuit Partitioning for Modular Architectures. *arXiv e-prints*, Article arXiv:2005.12259 (May 2020), arXiv:2005.12259 pages. arXiv:2005.12259 [quant-ph]

[18] B Ben Bakir, A Vazquez de Gyves, R Orobtchouk, P Lyan, C Porzier, A Roman, and J-M Fedeli. 2010. Low-loss (<1 dB) and polarization-insensitive edge fiber couplers fabricated on 200-mm silicon-on-insulator wafers. *IEEE Photon. Technol. Lett.* 22, 11 (2010), 739–741.

[19] Feng Bao, Hao Deng, Dawei Ding, Ran Gao, Xun Gao, Cupjin Huang, Xun Jiang, Hsiang-Sheng Ku, Zhisheng Li, Xizheng Ma, Xiaotong Ni, Jin Qin, Zhijun Song, Hantao Sun, Chengchun Tang, Tenghui Wang, Feng Wu, Tian Xia, Wenlong Yu, Fang Zhang, Gengyan Zhang, Xiaohang Zhang, Jingwei Zhou, Xing Zhu, Yaoyun Shi, Jianxin Chen, Hui-Hai Zhao, and Chunqing Deng. 2022. Fluxonium: An Alternative Qubit Platform for High-Fidelity Operations. *Phys. Rev. Lett.* 129 (Jun 2022), 010502. Issue 1. https://doi.org/10.1103/PhysRevLett.129.010502

[20] R.F. Barrett, S. Borkar, S.S. Dosanjh, S.D. Hammond, Michael Heroux, X.S. Hu, Justin Luitjens, S.G. Parker, John Shalf, and L. Tang. 2013. On the role of co-design in high performance computing. *Advances in Parallel Computing* 24 (01 2013), 141–155. https://doi.org/10.3233/978-1-61499-324-7-141

[21] Sean D. Barrett and Pieter Kok. 2005. Efficient high-fidelity quantum computation using matter qubits and linear optics. *Phys. Rev. A* 71 (Jun 2005), 060310. Issue 6. https://doi.org/10.1103/PhysRevA.71.060310

[22] John G Bartholomew, Jake Rochman, Tian Xie, Jonathan M Kindem, Andrei Ruskuc, Ioana Craiciu, Mi Lei, and Andrei Faraon. 2020. On-chip coherent microwave-to-optical transduction mediated by ytterbium in $YVO_4$. *Nat. Commun.* 11 (2020), 3266.

[23] Bela Bauer, Dave Wecker, Andrew J. Millis, Matthew B. Hastings, and Matthias Troyer. 2016. Hybrid Quantum-Classical Approach to Correlated Materials. *Phys. Rev. X* 6 (Sep 2016), 031045. Issue 3. https://doi.org/10.1103/PhysRevX.6.031045

[24] Robert Beals, Stephen Brierley, Oliver Gray, Aram W Harrow, Samuel Kutin, Noah Linden, Dan Shepherd, and Mark Stather. 2013. Efficient distributed quantum computing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 469, 2153 (2013), 20120686.

[25] Charles H. Bennett, Gilles Brassard, Sandu Popescu, Benjamin Schumacher, John A. Smolin, and William K. Wootters. 1996. Purification of Noisy Entanglement and Faithful Teleportation via Noisy Channels. *Phys. Rev. Lett.* 76 (Jan 1996), 722–725. Issue 5. https://doi.org/10.1103/PhysRevLett.76.722

[26] Charles H. Bennett, David P. DiVincenzo, John A. Smolin, and William K. Wootters. 1996. Mixed state entanglement and quantum error correction. *Phys. Rev. A* 54 (1996), 3824–3851. https://doi.org/10.1103/PhysRevA.54.3824 arXiv:quant-ph/9604024

[27] Charles H. Bennett, David P. Divincenzo, John A. Smolin, and William K. Wootters. 1996. Mixed-state entanglement and quantum error correction. *Physical Review A*. 54, 5 (Nov. 1996), 3824–3851. https://doi.org/10.1103/PhysRevA.54.3824 arXiv:quant-ph/9604024 [quant-ph]

[28] 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. 2013. Heralded entanglement between solid-state qubits separated by three metres. *Nature*. 497, 7447 (May 2013), 86–90. https://doi.org/10.1038/nature12016 arXiv:1212.6136 [quant-ph]

[29] 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. 2013. Heralded entanglement between solid-state qubits separated by three metres. *Nature* 497, 7447 (01 May 2013), 86–90. https://doi.org/10.1038/nature12016

[30] Dolev Bluvstein et al. 2022. A quantum processor based on coherent transport of entangled atom arrays. *Nature* 604, 7906 (2022), 451–456. https://doi.org/10.1038/s41586-022-04592-6 arXiv:2112.03923 [quant-ph]

[31] Wim Bogaerts, Daniel Pérez, José Capmany, David AB Miller, Joyce Poon, Dirk Englund, Francesco Morichetti, and Andrea Melloni. 2020. Programmable photonic circuits. *Nature* 586, 7828 (2020), 207–216.

[32] Édouard Bonnet, Tillmann Miltzow, and Paweł Rzażewski. 2018. Complexity of token swapping and its variants. *Algorithmica* 80, 9 (2018), 2656–2682.

[33] S. Bose, P. L. Knight, M. B. Plenio, and V. Vedral. 1999. Proposal for Teleportation of an Atomic State via Cavity Decay. *Phys. Rev. Lett.* 83 (Dec 1999), 5158–5161. Issue 24. https://doi.org/10.1103/PhysRevLett.83.5158

[34] Sergey Bravyi, Oliver Dial, Jay M. Gambetta, Dario Gil, and Zaira Nazario. 2022. The Future of Quantum Computing with Superconducting Qubits. *arXiv e-prints*, Article arXiv:2209.06841 (Sept. 2022), arXiv:2209.06841 pages. arXiv:2209.06841 [quant-ph]

[35] Sergey Bravyi, David Gosset, and Yinchen Liu. 2022. How to Simulate Quantum Measurement without Computing Marginals. *Phys. Rev. Lett.* 128 (Jun 2022), 220503. Issue 22. https://doi.org/10.1103/PhysRevLett.128.220503

[36] Sergey Bravyi, Graeme Smith, and John A. Smolin. 2016. Trading Classical and Quantum Computational Resources. *Phys. Rev. X* 6 (Jun 2016), 021043. Issue 2. https://doi.org/10.1103/PhysRevX.6.021043

[37] 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. 2022. Optomechanical Ground-State Cooling in a Continuous and Efficient Electro-Optic Transducer. *Phys. Rev. X* 12 (Jun 2022), 021062. Issue 2. https://doi.org/10.1103/PhysRevX.12.021062

[38] Colin D. Bruzewicz, John Chiaverini, Robert McConnell, and Jeremy M. Sage. 2019. Trapped-ion quantum computing: Progress and challenges. *Applied Physics Reviews* 6, 2, Article 021314 (June 2019), 021314 pages. https://doi.org/10.1063/1.5088164 arXiv:1904.04178 [quant-ph]

[39] Luke D. Burkhart, James D. Teoh, Yaxing Zhang, Christopher J. Axline, Luigi Frunzio, M.H. Devoret, Liang Jiang, S.M. Girvin, and R.J. Schoelkopf. 2021. Error-Detected State Transfer and Entanglement in a Superconducting Quantum Network. *PRX Quantum* 2 (Aug 2021), 030321. Issue 3. https://doi.org/10.1103/PRXQuantum.2.030321

[40] Luke D. Burkhart, James D. Teoh, Yaxing Zhang, Christopher J. Axline, Luigi Frunzio, M. H. Devoret, Liang Jiang, S. M. Girvin, and R. J. Schoelkopf. 2021. Error-Detected State Transfer and Entanglement in a Superconducting Quantum Network. *PRX Quantum* 2, 3, Article 030321 (Aug. 2021), 030321 pages. https://doi.org/10.1103/PRXQuantum.2.030321 arXiv:2004.06168 [quant-ph]

[41] Stephen F. Bush, William A. Challener, and Guillaume Mantelet. 2021. A perspective on industrial quantum networks. *AVS Quantum Science* 3, 3, Article 030501 (Sept. 2021), 030501 pages. https://doi.org/10.1116/5.0051881

[42] C. Cabrillo, J. I. Cirac, P. García-Fernández, and P. Zoller. 1999. Creation of entangled states of distant atoms by interference. *Phys. Rev. A* 59 (Feb 1999), 1025–1033. Issue 2. https://doi.org/10.1103/PhysRevA.59.1025

[43] Weizhou Cai, Yuwei Ma, Weiting Wang, Chang-Ling Zou, and Luyan Sun. 2021. Bosonic quantum error correction codes in superconducting quantum circuits. *Fundamental Research* 1, 1 (2021), 50–67. https://doi.org/10.1016/j.fmre.2020.12.006

[44] 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. 2020. Quantum error correction of a qubit encoded in grid states of an oscillator. *Nature* 584 (2020). Issue 7821. https://doi.org/10.1038/s41586-020-2603-3

[45] P. Campagne-Ibarcq, E. Zalys-Geller, A. Narla, S. Shankar, P. Reinhold, L. Burkhart, C. Axline, W. Pfaff, L. Frunzio, R. J. Schoelkopf, and M. H. Devoret. 2018. Deterministic Remote Entanglement of Superconducting Circuits through Microwave Two-Photon Transitions. *Phys. Rev. Lett.* 120 (May 2018), 200501. Issue 20. https://doi.org/10.1103/PhysRevLett.120.200501

[46] Earl Campbell. 2019. Random compiler for fast Hamiltonian simulation. *Physical review letters* 123, 7 (2019), 070503.

[47] B. Casabone, A. Stute, K. Friebe, B. Brandstätter, K. Schüppert, R. Blatt, and T. E. Northup. 2013. Heralded Entanglement of Two Ions in an Optical Cavity. *Phys. Rev. Lett.* 111 (Sep 2013), 100505. Issue 10. https://doi.org/10.1103/PhysRevLett.111.100505

[48] Srivatsan Chakram, Kevin He, Akash V Dixit, Andrew E Oriani, Ravi K Naik, Nelson Leung, Hyeokshin Kwon, Wen-Long Ma, Liang Jiang, and David I Schuster. 2022. Multimode photon blockade. *Nature Physics* 18, 8 (2022), 879–884. https://doi.org/10.1038/s41567-022-01630-y

[49] Srivatsan Chakram, Andrew E Oriani, Ravi K Naik, Akash V Dixit, Kevin He, Ankur Agrawal, Hyeokshin Kwon, and David I Schuster. 2021. Seamless high-Q microwave cavities for multimode circuit quantum electrodynamics. *Phys. Rev. Lett.* 127, 10 (2021), 107701.

[50] Christopher Chamberland, Kyungjoo Noh, Patricio Arrangoiz-Arriola, Earl T. Campbell, Connor T. Hann, Joseph Iverson, Harald Putterman, Thomas C. Bohdanowicz, Steven T. Flammia, Andrew Keller, Gil Refael, John Preskill, Liang Jiang, Amir H. Safavi-Naeini, Oskar Painter, and Fernando G.S.L. Brandão. 2022. Building a Fault-Tolerant Quantum Computer Using Concatenated Cat Codes. *PRX Quantum* 3 (Feb 2022), 010329. Issue 1. https://doi.org/10.1103/PRXQuantum.3.010329

[51] H.-S. Chang, Y. P. Zhong, A. Bienfait, M.-H. Chou, C. R. Conner, É. Dumur, J. Grebel, G. A. Peairs, R. G. Povey, K. J. Satzinger, and A. N. Cleland. 2020. Remote Entanglement via Adiabatic Passage Using a Tunably Dissipative Quantum Communication System. *Phys. Rev. Lett.* 124 (Jun 2020), 240502. Issue 24. https://doi.org/10.1103/PhysRevLett.124.240502

[52] Xia Chen, Milan M Milosevic, Stevan Stanković, Scott Reynolds, Thalia Dominguez Bucio, Ke Li, David J Thomson, Frederic Gardes, and Graham T Reed. 2018. The emergence of silicon photonics as a flexible technology platform. *Proc. IEEE* 106, 12 (2018), 2101–2116.

[53] Zijun Chen et al. 2021. Exponential suppression of bit or phase flip errors with repetitive error correction. (2 2021). https://doi.org/10.1038/s41586-021-03588-y arXiv:2102.06132 [quant-ph]

[54] Risheng Cheng, John Wright, Huili G Xing, Debdeep Jena, and Hong X Tang. 2020. Epitaxial niobium nitride superconducting nanowire single-photon detectors. *Appl. Phys. Lett.* 117, 13 (2020), 132601.

[55] Hyeongrak Choi, Mikkel Heuck, and Dirk Englund. 2017. Self-similar nanocavity design with ultrasmall mode volume for single-photon nonlinearities. *Phys. Rev. Lett.* 118, 22 (2017), 223605.

[56] Yiwen Chu and Simon Gröblacher. 2020. A perspective on hybrid quantum opto-and electromechanical systems. *Appl. Phys. Lett.* 117, 15 (2020), 150503.

[57] J. I. Cirac, P. Zoller, H. J. Kimble, and H. Mabuchi. 1997. Quantum State Transfer and Entanglement Distribution among Distant Nodes in a Quantum Network. *Physical Review Letters*. 78, 16 (April 1997), 3221–3224. https://doi.org/10.1103/PhysRevLett.78.3221 arXiv:quant-ph/9611017 [quant-ph]

[58] AA Clerk, KW Lehnert, P Bertet, JR Petta, and Y Nakamura. 2020. Hybrid quantum systems with circuit quantum electrodynamics. *Nat. Phys.* 16, 3 (2020), 257–267.

[59] Laura Clinton, Toby Cubitt, Brian Flynn, Filippo Maria Gambetta, Joel Klassen, Ashley Montanaro, Stephen Piddock, Raul A Santos, and Evan Sheridan. 2022. Towards near-term quantum simulation of materials. *arXiv preprint arXiv:2205.15256* (2022).

[60] Hugh Collins and Chris Nay. [n. d.]. IBM unveils 400 qubit-plus quantum processor and next-generation IBM Quantum System Two. https://newsroom.ibm.com/2022-11-09-IBM-Unveils-400-Qubit-Plus-Quantum-Processor-and-Next-Generation-IBM-Quantum-System-Two

[61] Jacob P Covey, Alp Sipahigil, and Mark Saffman. 2019. Microwave-to-optical conversion via four-wave mixing in a cold ytterbium ensemble. *Phys. Rev. A* 100, 1 (2019), 012307.

[62] Alexander Cowtan, Silas Dilkes, Ross Duncan, Alexandre Krajenbrink, Will Simmons, and Seyon Sivarajah. 2019. On the qubit routing problem. *arXiv preprint arXiv:1902.08091* (2019).

[63] Andrew W. Cross, Lev S. Bishop, Sarah Sheldon, Paul D. Nation, and Jay M. Gambetta. 2019. Validating quantum computers using randomized model circuits. *Physical Review A*. 100, 3, Article 032328 (Sept. 2019), 032328 pages. https://doi.org/10.1103/PhysRevA.100.032328 arXiv:1811.12926 [quant-ph]

[64] Daniele Cuomo, Marcello Caleffi, and Angela Sara Cacciapuoti. 2020. Towards a Distributed Quantum Computing Ecosystem. *arXiv e-prints*, Article arXiv:2002.11808 (Feb. 2020), arXiv:2002.11808 pages. arXiv:2002.11808 [quant-ph]

[65] Daniele Cuomo, Marcello Caleffi, Kevin Krsulich, Filippo Tramonto, Gabriele Agliardi, Enrico Prati, and Angela Sara Cacciapuoti. 2021. Optimized compiler for Distributed Quantum Computing. (12 2021). arXiv:2112.14139 [quant-ph]

[66] Davood Dadkhah, Mariam Zomorodi, Seyed Ebrahim Hosseini, Pawel Plawiak, and Xujuan Zhou. 2022. Reordering and partitioning of distributed quantum circuits. *IEEE Access* 10 (2022), 70329–70341.

[67] Axel Dahlberg, Matthew Skrzypczyk, Tim Coopmans, Leon Wubben, Filip Rozpedek, Matteo Pompili, Arian Stolk, Przemysław Pawełczak, Robert Knegjens, Julio de Oliveira Filho, Ronald Hanson, and Stephanie Wehner. 2019. A Link Layer Protocol for Quantum Networks. *arXiv e-prints*, Article arXiv:1903.09778 (March 2019), arXiv:1903.09778 pages. arXiv:1903.09778 [quant-ph]

[68] Nikos Daniilidis and Hartmut Häffner. 2013. Quantum Interfaces Between Atomic and Solid-State Systems. *Annual Review of Condensed Matter Physics* 4, 1 (2013), 83–112. https://doi.org/10.1146/annurev-conmatphys-030212-184253 arXiv:https://doi.org/10.1146/annurev-conmatphys-030212-184253

[69] Zohreh Davarzani, Mariam Zomorodi-Moghadam, Mahboobeh Houshmand, and Mostafa Nouri-baygi. 2020. A dynamic programming approach for distributing quantum circuits by bipartite graphs. *Quantum Information Processing* 19, 10 (2020), 1–18.

[70] Stijn de Graaf, Benjamin J Chapman, Jacob C Curtis, Yaxing Zhang, Nicholas E Frattini, Michel H Devoret, Steven M Girvin, and Robert J Schoelkopf. 2022. Fast parametrically driven entangling gates in superconducting circuits using a tunable coupler. *Bulletin of the American Physical Society* (2022).

[71] D. De Motte, A. R. Grounds, M. Rehák, A. Rodriguez Blanco, B. Lekitsch, G. S. Giri, P. Neilinger, G. Oelsner, E. Il'ichev, M. Grajcar, and W. K. Hensinger. 2016. Experimental system design for the integration of trapped-ion and superconducting qubit systems. *Quantum Information Processing* 15, 12 (Dec. 2016), 5385–5414. https://doi.org/10.1007/s11128-016-1368-y arXiv:1510.07298 [quant-ph]

[72] Daniele De Sensi, Salvatore Di Girolamo, Kim H. McMahon, Duncan Roweth, and Torsten Hoefler. 2020. An In-Depth Analysis of the Slingshot Interconnect. *arXiv e-prints*, Article arXiv:2008.08886 (Aug. 2020), arXiv:2008.08886 pages. arXiv:2008.08886 [cs.DC]

[73] RD Delaney, MD Urmey, S Mittal, BM Brubaker, JM Kindem, PS Burns, CA Regal, and KW Lehnert. 2022. Superconducting-qubit readout via low-backaction electro-optic transduction. *Nature* 606, 7914 (2022), 489–493.

[74] David Deutsch, Artur Ekert, Richard Jozsa, Chiara Macchiavello, Sandu Popescu, and Anna Sanpera. 1996. Quantum Privacy Amplification and the Security of Quantum Cryptography over Noisy Channels. *Phys. Rev. Lett.* 77 (Sep 1996), 2818–2821. Issue 13. https://doi.org/10.1103/PhysRevLett.77.2818

[75] M. H. Devoret and R. J. Schoelkopf. 2013. Superconducting Circuits for Quantum Information: An Outlook. *Science* 339, 6124 (March 2013), 1169–1174. https://doi.org/10.1126/science.1231930

[76] M. H. Devoret, A. Wallraff, and J. M. Martinis. 2004. Superconducting Qubits: A Short Review. *arXiv e-prints*, Article cond-mat/0411174 (Nov. 2004), cond-mat/0411174 pages. arXiv:cond-mat/0411174 [cond-mat.mes-hall]

[77] Stephen DiAdamo, Marco Ghibaudi, and James Cruise. 2021. Distributed quantum computing and network control for accelerated VQE. *arXiv preprint arXiv:2101.02504* (2021).

[78] C. Dickel, J. J. Wesdorp, N. K. Langford, S. Peiter, R. Sagastizabal, A. Bruno, B. Criger, F. Motzoi, and L. DiCarlo. 2018. Chip-to-chip entanglement of transmon qubits using engineered measurement fields. *Phys. Rev. B* 97 (Feb 2018), 064508. Issue 6. https://doi.org/10.1103/PhysRevB.97.064508

[79] Ebru Dogan, Dario Rosenstock, Loïck Le Guevel, Haonan Xiong, Raymond A. Mencia, Aaron Somoroff, Konstantin N. Nesterov, Maxim G. Vavilov, Vladimir E. Manucharyan, and Chen Wang. 2022. Demonstration of the Two-Fluxonium Cross-Resonance Gate.

*arXiv preprint arXiv:2204.11829* (2022).

[80] L.-M. Duan, M. D. Lukin, J. I. Cirac, and P. Zoller. 2001. Long-distance quantum communication with atomic ensembles and linear optics. *Nature* 414, 6862 (22 Nov 2001), 413–418. https://doi.org/10.1038/35106500

[81] W. Dür and H.-J. Briegel. 2003. Entanglement Purification for Quantum Computation. *Phys. Rev. Lett.* 90 (Feb 2003), 067901. Issue 6. https://doi.org/10.1103/PhysRevLett.90.067901

[82] W. Dür and H. J. Briegel. 2007. Entanglement purification and quantum error correction. *Reports on Progress in Physics* 70, 8 (Aug. 2007), 1381–1424. https://doi.org/10.1088/0034-4885/70/8/R03 arXiv:0705.4165 [quant-ph]

[83] W Dür and H J Briegel. 2007. Entanglement purification and quantum error correction. *Reports on Progress in Physics* 70, 8 (jul 2007), 1381–1424. https://doi.org/10.1088/0034-4885/70/8/r03

[84] Andrew Eddins, Mario Motta, Tanvi P Gujarati, Sergey Bravyi, Antonio Mezzacapo, Charles Hadfield, and Sarah Sheldon. 2022. Doubling the size of quantum simulators by entanglement forging. *PRX Quantum* 3, 1 (2022), 010309.

[85] Jonathan R Everts, Matthew C Berrington, Rose L Ahlefeldt, and Jevon J Longdell. 2019. Microwave to optical photon conversion via fully concentrated rare-earth-ion crystals. *Phys. Rev. A* 99, 6 (2019), 063830.

[86] Linran Fan, Chang-Ling Zou, Risheng Cheng, Xiang Guo, Xu Han, Zheng Gong, Sihao Wang, and Hong X Tang. 2018. Superconducting cavity electro-optics: a platform for coherent photon conversion between superconducting and photonic circuits. *Sci. Adv.* 4, 8 (2018), eaar4994.

[87] Kun Fang, Xin Wang, Marco Tomamichel, and Runyao Duan. 2019. Non-asymptotic entanglement distillation. *IEEE Transactions on Information Theory* 65, 10 (2019), 6454–6465.

[88] Xavier Fernandez-Gonzalvo, Yu-Hui Chen, Chunming Yin, Sven Rogge, and Jevon J Longdell. 2015. Coherent frequency up-conversion of microwaves to the optical telecommunications band in an Er:YSO crystal. *Phys. Rev. A* 92, 6 (2015), 062313.

[89] Xavier Fernandez-Gonzalvo, Sebastian P Horvath, Yu-Hui Chen, and Jevon J Longdell. 2019. Cavity-enhanced Raman heterodyne spectroscopy in $Er^{3+}$:$Y_2SiO_5$ for microwave to optical signal conversion. *Phys. Rev. A* 100, 3 (2019), 033807.

[90] Davide Ferrari, Angela Sara Cacciapuoti, Michele Amoretti, and Marcello Caleffi. 2020. Compiler design for distributed quantum computing. *arXiv preprint arXiv:2012.09680* (2020).

[91] Davide Ferrari, Stefano Carretta, and Michele Amoretti. 2023. A modular quantum compilation framework for distributed quantum computing. *IEEE Transactions on Quantum Engineering* (2023).

[92] Quentin Ficheux, Long B. Nguyen, Aaron Somoroff, Haonan Xiong, Konstantin N. Nesterov, Maxim G. Vavilov, and Vladimir E. Manucharyan. 2021. Fast Logic with Slow Qubits: Microwave-Activated Controlled-Z Gate on Low-Frequency Fluxoniums. *Phys. Rev. X* 11 (May 2021), 021026. Issue 2. https://doi.org/10.1103/PhysRevX.11.021026

[93] Moritz Forsch, Robert Stockill, Andreas Wallucks, Igor Marinković, Claus Gärtner, Richard A Norte, Frank van Otten, Andrea Fiore, Kartik Srinivasan, and Simon Gröblacher. 2020. Microwave-to-optics conversion using a mechanical oscillator in its quantum ground state. *Nat. Phys.* 16, 1 (2020), 69–74.

[94] I. Foster and W. Gentzsch. 2011. *High Performance Computing: From Grids and Clouds to Exascale.* IOS Press. https://books.google.com/books?id=5Uzge1cBsHoC

[95] B. Foxen, J. Y. Mutus, E. Lucero, R. Graff, A. Megrant, Yu Chen, C. Quintana, B. Burkett, J. Kelly, E. Jeffrey, Yan Yang, Anthony Yu, K. Arya, R. Barends, Zijun 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 John M. Martinis. 2018. Qubit compatible superconducting interconnects. *Quantum Science and Technology* 3, 1 (Jan. 2018), 014005. https://doi.org/10.1088/2058-9565/aa94fc arXiv:1708.04270 [quant-ph]

[96] B. Foxen, C. Neill, A. Dunsworth, P. Roushan, B. Chiaro, A. Megrant, J. Kelly, Zijun Chen, K. Satzinger, R. Barends, F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, S. Boixo, D. Buell, B. Burkett, Yu 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. 2020. Demonstrating a Continuous Set of Two-Qubit Gates for Near-Term Quantum Algorithms. *Phys. Rev. Lett.* 125 (Sep 2020), 120504. Issue 12. https://doi.org/10.1103/PhysRevLett.125.120504

[97] N. E. Frattini, U. Vool, S. Shankar, A. Narla, K. M. Sliwa, and M. H. Devoret. 2017. 3-wave mixing Josephson dipole element. *Applied Physics Letters* 110 (2017). Issue 22. https://doi.org/10.1063/1.4984142

[98] Haohao Fu, Haochuan Chen, Marharyta Blazhynska, Emma Goulard Coderc de Lacam, Florence Szczepaniak, Anna Pavlova, Xueguang Shao, James C Gumbart, François Dehez, Benoît Roux, et al. 2022. Accurate determination of protein: ligand standard binding free energies from molecular dynamics simulations. *Nature Protocols* 17, 4 (2022), 1114–1141.

[99] Wei Fu, Mingrui Xu, Xianwen Liu, Chang-Ling Zou, Changchun Zhong, Xu Han, Mohan Shen, Yuntao Xu, Risheng Cheng, Sihao Wang, et al. 2021. Cavity electro-optic circuit for microwave-to-optical conversion in the quantum ground state. *Phys. Rev. A* 103, 5 (2021), 053504.

[100] Keisuke Fujii and Katsuji Yamamoto. 2009. Entanglement purification with double selection. *Phys. Rev. A* 80 (Oct 2009), 042308. Issue 4. https://doi.org/10.1103/PhysRevA.80.042308

[101] Takanori Fujiwara, Preeti Malakar, Khairi Reda, Venkatram Vishwanath, Michael E Papka, and Kwan-Liu Ma. 2017. A visual analytics system for optimizing communications in massively parallel applications. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 59–70.

[102] Renhong Gao, Ni Yao, Jianglin Guan, Li Deng, Jintian Lin, Min Wang, Lingling Qiao, Wei Fang, and Ya Cheng. 2022. Lithium niobate microring with ultra-high Q factor above $10^8$. *Chin. Opt. Lett.* 20, 1 (2022), 011902.

[103] Yvonne Y. Gao, Brian J. Lester, Yaxing Zhang, Chen Wang, Serge Rosenblum, Luigi Frunzio, Liang Jiang, S. M. Girvin, and Robert J. Schoelkopf. 2018. Programmable Interference between Two Microwave Quantum Memories. *Physical Review X* 8, 2 (June 2018), 021073. https://doi.org/10.1103/PhysRevX.8.021073

[104] Alan Gara, Matthias A Blumrich, Dong Chen, GL-T Chiu, Paul Coteus, Mark E Giampapa, Ruud A Haring, Philip Heidelberger, Dirk Hoenicke, Gerard V Kopcsay, et al. 2005. Overview of the Blue Gene/L system architecture. *IBM Journal of research and development* 49, 2.3 (2005), 195–212.

[105] Jeffrey M Gertler, Brian Baker, Juliang Li, Shruti Shirol, Jens Koch, and Chen Wang. 2021. Protecting a bosonic qubit with autonomous quantum error correction. *Nature* 590 (2021), 243–248. Issue 7845. https://doi.org/10.1038/s41586-021-03257-0

[106] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, et al. 2009. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of physics: Condensed matter* 21, 39 (2009), 395502.

[107] Alysson Gold, J. P. Paquette, Anna Stockklauser, Matthew J. Reagor, M. Sohaib Alam, Andrew Bestwick, Nicolas Didier, Ani Nersisyan, Feyza Oruc, Armin Razavi, Ben Scharmann, Eyob A. Sete, Biswajit Sur, Davide Venturelli, Cody James Winkleblack, Filip Wudarski, Mike Harburn, and Chad Rigetti. 2021. Entanglement across separate silicon dies in a modular superconducting qubit device. *npj Quantum Information* 7, Article 142 (Jan. 2021), 142 pages. https://doi.org/10.1038/s41534-021-00484-1 arXiv:2102.13293 [quant-ph]

[108] D. Goldhaber-Gordon, Hadas Shtrikman, D. Mahalu, David Abusch-Magder, U. Meirav, and M. A. Kastner. 1998. Kondo effect in a single-electron transistor. *Nature.* 391, 6663 (Jan. 1998), 156–159. https://doi.org/10.1038/34373 arXiv:cond-mat/9707311 [cond-mat.str-el]

[109] Kenneth Goodenough, Sébastian de Bone, Vaishnavi L. Addala, Stefan Krastanov, Sarah Jansen, Dion Gijswijt, and David Elkouss. 2023. Near-term $n$ to $k$ distillation protocols using graph codes. *arXiv e-prints*, Article arXiv:2303.11465 (March 2023), arXiv:2303.11465 pages. https://doi.org/10.48550/arXiv.2303.11465 arXiv:2303.11465 [quant-ph]

[110] Daniel Gottesman and Isaac L. Chuang. 1999. Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations. *Nature.* 402, 6760 (Nov. 1999), 390–393. https://doi.org/10.1038/46503 arXiv:quant-ph/9908010 [quant-ph]

[111] Arne L. Grimsmo and Shruti Puri. 2021. Quantum Error Correction with the Gottesman-Kitaev-Preskill Code. *PRX Quantum* 2 (Jun 2021), 020101. Issue 2. https://doi.org/10.1103/PRXQuantum.2.020101

[112] Lov K Grover. 1997. Quantum telecomputation. *arXiv preprint quant-ph/9704012* (1997).

[113] Saikat Guha, Hari Krovi, Christopher A. Fuchs, Zachary Dutton, Joshua A. Slater, Christoph Simon, and Wolfgang Tittel. 2015. Rate-loss analysis of an efficient quantum repeater architecture. *Phys. Rev. A* 92 (Aug 2015), 022357. Issue 2. https://doi.org/10.1103/PhysRevA.92.022357

[114] Charles Guinn, Sara Sussman, Pranav S Mundada, Andrei Vrajitoarea, Catherine Leroux, Alexander Place, Camille Le Calonnec, Agustin Di Paolo, Alexandru Petrescu, Alexandre Blais, and Andrew A Houck. 2022. Fast parametrically driven entangling gates in superconducting circuits using a tunable coupler. *Bulletin of the American Physical Society* (2022).

[115] S. Gustavsson, R. Leturcq, M. Studer, I. Shorubalko, T. Ihn, K. Ensslin, D. C. Driscoll, and A. C. Gossard. 2009. Electron counting in quantum dots. *Surface Science Reports* 64, 6 (June 2009), 191–232. https://doi.org/10.1016/j.surfrep.2009.02.001 arXiv:0905.4675 [cond-mat.mes-hall]

[116] Laszlo Gyongyosi and Sandor Imre. 2021. Scalable distributed gate-model quantum computers. *Sci. Rep.* 11, 1 (2021), 5172. https://doi.org/10.1038/s41598-020-76728-5

[117] Xu Han, Wei Fu, Changchun Zhong, Chang-Ling Zou, Yuntao Xu, Ayed Al Sayem, Mingrui Xu, Sihao Wang, Risheng Cheng, Liang Jiang, et al. 2020. Cavity piezo-mechanics for superconducting-nanophotonic quantum interface. *Nat. Commun.* 11 (2020), 3237.

[118] Xu Han, Wei Fu, Chang-Ling Zou, Liang Jiang, and Hong X Tang. 2021. Microwave-optical quantum frequency conversion. *Optica* 8, 8 (2021), 1050–1064.

[119] Thomas Häner, Damian S Steiger, Torsten Hoefler, and Matthias Troyer. 2021. Distributed quantum computing with qmpi. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–13.

[120] William Hease, Alfredo Rueda, Rishabh Sahu, Matthias Wulf, Georg Arnold, Harald GL Schwefel, and Johannes M Fink. 2020. Bidirectional electro-optic wavelength conversion in the quantum ground state. *PRX Quantum* 1, 2 (2020), 020315.

[121] B. Hensen, H. Bernien, A. E. Dréau, A. Reiserer, N. Kalb, M. S. Blok, J. Ruitenberg, 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. 2015. Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature* 526, 7575 (01 Oct 2015), 682–686. https://doi.org/10.1038/nature15759

[122] Andrew P Higginbotham, PS Burns, MD Urmey, RW Peterson, NS Kampel, BM Brubaker, G Smith, KW Lehnert, and CA Regal. 2018. Harnessing electro-optic correlations in an efficient mechanical converter. *Nat. Phys.* 14, 10 (2018), 1038–1042.
[123] Ryusuke Hisatomi, Alto Osada, Yutaka Tabuchi, Toyofumi Ishikawa, Atsushi Noguchi, Rekishu Yamazaki, Koji Usami, and Yasunobu Nakamura. 2016. Bidirectional conversion between microwave and light via ferromagnetic magnons. *Phys. Rev. B* 93, 17 (2016), 174427.
[124] Jeffrey Holzgrafe, Neil Sinclair, Di Zhu, Amirhassan Shams-Ansari, Marco Colangelo, Yaowen Hu, Mian Zhang, Karl K Berggren, and Marko Lončar. 2020. Cavity electro-optics in thin-film lithium niobate for efficient microwave-to-optical transduction. *Optica* 7, 12 (2020), 1714–1720.
[125] Simon Hönl, Youri Popoff, Daniele Caimi, Alberto Beccari, Tobias J Kippenberg, and Paul Seidler. 2022. Microwave-to-optical conversion with a gallium phosphide photonic crystal cavity. *Nat. Commun.* 13 (2022), 2065.
[126] 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. 2019. Quantum error correction and universal gate set operation on a binomial bosonic logical qubit. *Nature Physics* 15, 5 (Feb. 2019), 503–508. https://doi.org/10.1038/s41567-018-0414-3 arXiv:1805.09072 [quant-ph]
[127] 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. 2019. Quantum error correction and universal gate set operation on a binomial bosonic logical qubit. *Nature Physics* 15 (2019). Issue 5. https://doi.org/10.1038/s41567-018-0414-3
[128] Peter C Humphreys, Norbert Kalb, Jaco PJ Morits, Raymond N Schouten, Raymond FL Vermeulen, Daniel J Twitchen, Matthew Markham, and Ronald Hanson. 2018. Deterministic delivery of remote entanglement on a quantum network. *Nature* 558, 7709 (2018), 268–273.
[129] IBM. [n. d.]. IBM Quantum Roadmap. https://research.ibm.com/blog/ibm-quantum-roadmap-2025 Accessed: Sep. 20, 2022.
[130] Infiniband Trade Association [n. d.]. *Infiniband Architecture Specification Volume 1* (1.2.1 ed.). Infiniband Trade Association.
[131] Sarah Jansen, Kenneth Goodenough, Sébastian de Bone, Dion Gijswijt, and David Elkouss. 2022. Enumerating all bilocal Clifford distillation protocols through symmetry reduction. *Quantum* 6 (May 2022), 715. https://doi.org/10.22331/q-2022-05-19-715 arXiv:2103.03669 [quant-ph]
[132] Ali Javadi-Abhari. 2017. *Towards a scalable software stack for resource estimation and optimization in general-purpose quantum computers.* Ph. D. Dissertation. Princeton University.
[133] Liang Jiang, Jacob M. Taylor, Anders S. Sørensen, and Mikhail D. Lukin. 2007. Distributed quantum computation based on small quantum registers. *Phys. Rev. A* 76 (Dec 2007), 062323. Issue 6. https://doi.org/10.1103/PhysRevA.76.062323
[134] Wentao Jiang, Christopher J Sarabalis, Yanni D Dahmani, Rishi N Patel, Felix M Mayor, Timothy P McKenna, Raphaël Van Laer, and Amir H Safavi-Naeini. 2020. Efficient bidirectional piezo-optomechanical transduction between microwave and optical frequency. *Nat. Commun.* 11 (2020), 1166.
[135] Naijun Jin, Charles A McLemore, David Mason, James P Hendrie, Yizhi Luo, Megan L Kelleher, Prashanta Kharel, Franklyn Quinlan, Scott A Diddams, and Peter T Rakich. 2022. Micro-fabricated mirrors with finesse exceeding one million. *Optica* 9, 9 (2022), 965–970.
[136] Warren Jin, Qi-Fan Yang, Lin Chang, Boqiang Shen, Heming Wang, Mark A Leal, Lue Wu, Maodong Gao, Avi Feshali, Mario Paniccia, et al. 2021. Hertz-linewidth semiconductor lasers using CMOS-ready ultra-high-Q microresonators. *Nat. Photon.* 15, 5 (2021), 346–353.
[137] J Robert Johansson, Paul D Nation, and Franco Nori. 2012. QuTiP: An open-source Python framework for the dynamics of open quantum systems. *Comput. Phys. Commun.* 183, 8 (2012), 1760–1772.
[138] Atharv Joshi, Kyungjoo Noh, and Yvonne Y. Gao. 2021. Quantum information processing with bosonic qubits in circuit QED. Issue 3. https://doi.org/10.1088/2058-9565/abe989
[139] Shuting Kang, Ru Zhang, Zhenzhong Hao, Di Jia, Feng Gao, Fang Bo, Guoquan Zhang, and Jingjun Xu. 2020. High-efficiency chirped grating couplers on lithium niobate on insulator. *Opt. Lett.* 45, 24 (2020), 6651–6654.
[140] George Karypis and Vipin Kumar. 1998. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing* 20, 1 (1998), 359–392.
[141] V. Kaushal, B. Lekitsch, A. Stahl, J. Hilder, D. Pijn, C. Schmiegelow, A. Bermudez, M. Müller, F. Schmidt-Kaler, and U. Poschinger. 2020. Shuttling-based trapped-ion quantum information processing. *AVS Quantum Science* 2, 1, Article 014101 (Feb. 2020), 014101 pages. https://doi.org/10.1116/1.5126186 arXiv:1912.04712 [quant-ph]
[142] D. Kielpinski, D. Kafri, M. J. Woolley, G. J. Milburn, and J. M. Taylor. 2012. Quantum Interface between an Electrical Circuit and a Single Atom. *Physical Review Letters.* 108, 13, Article 130504 (March 2012), 130504 pages. https://doi.org/10.1103/PhysRevLett.108.130504 arXiv:1111.5999 [quant-ph]
[143] John Kim, Wiliam J. Dally, Steve Scott, and Dennis Abts. 2008. Technology-Driven, Highly-Scalable Dragonfly Topology. In *2008 International Symposium on Computer Architecture.* 77–88. https://doi.org/10.1109/ISCA.2008.19
[144] H. J. Kimble. 2008. The quantum internet. *Nature.* 453, 7198 (June 2008), 1023–1030. https://doi.org/10.1038/nature07127 arXiv:0806.4195 [quant-ph]
[145] Morten Kjaergaard, Mollie E. Schwartz, Jochen Braumüller, Philip Krantz, Joel I-Jan Wang, Simon Gustavsson, and William D. Oliver. 2019. Superconducting Qubits: Current State of Play. *arXiv e-prints*, Article arXiv:1905.13641 (May 2019), arXiv:1905.13641 pages. arXiv:1905.13641 [quant-ph]

[146] Emanuel Knill, Gerardo Ortiz, and Rolando D Somma. 2007. Optimal quantum measurements of expectation values of observables. *Physical Review A* 75, 1 (2007), 012328.
[147] Jens Koch, Terri M. Yu, Jay Gambetta, A. A. Houck, D. I. Schuster, J. Majer, Alexandre Blais, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf. 2007. Charge-insensitive qubit design derived from the Cooper pair box. *Phys. Rev. A* 76 (Oct 2007), 042319. Issue 4. https://doi.org/10.1103/PhysRevA.76.042319
[148] G. Kotliar, S. Y. Savrasov, K. Haule, V. S. Oudovenko, O. Parcollet, and C. A. Marianetti. 2006. Electronic structure calculations with dynamical mean-field theory. *Rev. Mod. Phys.* 78 (Aug 2006), 865–951. Issue 3. https://doi.org/10.1103/RevModPhys.78.865
[149] Inna Krasnokutska, Robert J Chapman, Jean-Luc J Tambasco, and Alberto Peruzzo. 2019. High coupling efficiency grating couplers on lithium niobate on insulator. *Opt. Express* 27, 13 (2019), 17681–17685.
[150] Stefan Krastanov, Victor V. Albert, and Liang Jiang. 2019. Optimized Entanglement Purification. *Quantum* 3 (Feb. 2019), 123. https://doi.org/10.22331/q-2019-02-18-123
[151] Stefan Krastanov, Hamza Raniwala, Jeffrey Holzgrafe, Kurt Jacobs, Marko Lončar, Matthew J Reagor, and Dirk R Englund. 2021. Optically heralded entanglement of superconducting systems in quantum networks. *Phys. Rev. Lett.* 127, 4 (2021), 040503.
[152] Sebastian Krinner, Nathan Lacroix, Ants Remm, Agustin Di Paolo, Elie Genois, Catherine Leroux, Christoph Hellings, Stefania Lazar, Francois Swiadek, Johannes Herrmann, Graham J Norris, Christian Kraglund Andersen, Markus Müller, Alexandre Blais, Christopher Eichler, and Andreas Wallraff. 2022. Realizing repeated quantum error correction in a distance-three surface code. *Nature* 605, 7911 (2022), 669–674. https://doi.org/10.1038/s41586-022-04566-8
[153] Sebastian Krinner, Simon Storz, Philipp Kurpiers, Paul Magnard, Johannes Heinsoo, Raphael Keller, Janis Luetolf, Christopher Eichler, and Andreas Wallraff. 2018. Engineering cryogenic setups for 100-qubit scale superconducting circuit systems. *arXiv e-prints*, Article arXiv:1806.07862 (June 2018), arXiv:1806.07862 pages. arXiv:1806.07862 [quant-ph]
[154] Aishwarya Kumar, Aziza Suleymanzade, Mark Stone, Lavanya Taneja, Alexander Anferov, David I Schuster, and Jonathan Simon. 2022. Quantum-limited millimeter wave to optical transduction. *arXiv:2207.10121* (2022).
[155] Sameer Kumar, Yogish Sabharwal, Rahul Garg, and Philip Heidelberger. 2008. Optimization of all-to-all communication on the blue gene/l supercomputer. In *2008 37th International Conference on Parallel Processing*. IEEE, 320–329.
[156] Gershon Kurizki, Patrice Bertet, Yuimaru Kubo, Klaus Mølmer, David Petrosyan, Peter Rabl, and Jörg Schmiedmayer. 2015. Quantum technologies with hybrid systems. *Proc. Natl. Acad. Sci. USA* 112, 13 (2015), 3866–3873.
[157] 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. 2018. Deterministic quantum state transfer and remote entanglement using microwave photons. *Nature* 558, 7709 (2018), 264–267. https://doi.org/10.1038/s41586-018-0195-y
[158] 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. 2019. Quantum Communication with Time-Bin Encoded Microwave Photons. *Phys. Rev. Applied* 12 (Oct 2019), 044067. Issue 4. https://doi.org/10.1103/PhysRevApplied.12.044067
[159] L. Lamata, D. R. Leibrandt, I. L. Chuang, J. I. Cirac, M. D. Lukin, V. Vuletić, and S. F. Yelin. 2011. Ion Crystal Transducer for Strong Coupling between Single Ions and Single Photons. *Phys. Rev. Lett.* 107 (Jul 2011), 030501. Issue 3. https://doi.org/10.1103/PhysRevLett.107.030501
[160] Nicholas J Lambert, Alfredo Rueda, Florian Sedlmeir, and Harald GL Schwefel. 2020. Coherent conversion between microwave and optical photons—an overview of physical implementations. *Adv. Quantum Technol.* 3, 1 (2020), 1900077.
[161] Nicola Lanatà, Yongxin Yao, Cai-Zhuang Wang, Kai-Ming Ho, and Gabriel Kotliar. 2015. Phase Diagram and Electronic Structure of Praseodymium and Plutonium. *Phys. Rev. X* 5 (Jan 2015), 011008. Issue 1. https://doi.org/10.1103/PhysRevX.5.011008
[162] Nicholas LaRacuente, Kaitlin N. Smith, Poolad Imany, Kevin L. Silverman, and Frederic T. Chong. 2022. Short-Range Microwave Networks to Scale Superconducting Quantum Computation. (1 2022). arXiv:2201.08825 [quant-ph]
[163] Nikolai Lauk, Neil Sinclair, Shabir Barzanjeh, Jacob P Covey, Mark Saffman, Maria Spiropulu, and Christoph Simon. 2020. Perspectives on quantum transduction. *Quantum Sci. Technol.* 5, 2 (2020), 020501.
[164] Philippe Lebrun and L Tavian. 2015. Cooling with superfluid helium. *arXiv:1501.07156* (2015).
[165] Florent Lecocq, Franklyn Quinlan, Katarina Cicak, Jose Aumentado, SA Diddams, and JD Teufel. 2021. Control and readout of a superconducting qubit using a photonic link. *Nature* 591, 7851 (2021), 575–579.
[166] Joonho Lee, Dominic W Berry, Craig Gidney, William J Huggins, Jarrod R McClean, Nathan Wiebe, and Ryan Babbush. 2021. Even more efficient quantum computations of chemistry through tensor hypercontraction. *PRX Quantum* 2, 3 (2021), 030305.
[167] Moonjoo Lee. 2019. Ultrahigh-quality-factor superconducting microwave resonator on diamond for quantum information processing. *Jap. J. Appl. Phys.* 58, 10 (2019), 100914.
[168] Charles E. Leiserson. 1985. Fat-trees: Universal networks for hardware-efficient supercomputing. *IEEE Trans. Comput.* C-34, 10 (1985), 892–901. https://doi.org/10.1109/TC.1985.6312192
[169] N. Leung, Y. Lu, S. Chakram, R. K. Naik, N. Earnest, R. Ma, K. Jacobs, A. N. Cleland, and D. I. Schuster. 2019. Deterministic bidirectional communication and remote entanglement generation between superconducting qubits. *npj Quantum Information* 5, 1 (15 Feb 2019), 18. https://doi.org/10.1038/s41534-019-0128-0

[170] Ang Li, Samuel Stein, Sriram Krishnamoorthy, and James Ang. 2022. QASMBench: A Low-Level Quantum Benchmark Suite for NISQ Evaluation and Simulation. *ACM Transactions on Quantum Computing* (jul 2022). https://doi.org/10.1145/3550488 Just Accepted.

[171] Ang Li, Omer Subasi, Xiu Yang, and Sriram Krishnamoorthy. 2020. Density Matrix Quantum Circuit Simulation via the BSP Machine on Modern GPU Clusters. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–15. https://doi.org/10.1109/SC41405.2020.00017

[172] Ang Li, Omer Subasi, Xiu Yang, and Sriram Krishnamoorthy. 2020. Density matrix quantum circuit simulation via the BSP machine on modern GPU clusters. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 1–15.

[173] Ang Li, YC Tay, Akash Kumar, and Henk Corporaal. 2015. Transit: A visual analytical model for multithreaded machines. In *Proceedings of the 24th international symposium on high-performance parallel and distributed computing*. 101–106.

[174] Gushu Li, Anbang Wu, Yunong Shi, Ali Javadi-Abhari, Yufei Ding, and Yuan Xie. 2021. On the Co-Design of Quantum Software and Hardware. In *Proceedings of the Eight Annual ACM International Conference on Nanoscale Computing and Communication* (Virtual Event, Italy) *(NANOCOM '21)*. Association for Computing Machinery, New York, NY, USA, Article 15, 7 pages. https://doi.org/10.1145/3477206.3477464

[175] Mingxiao Li, Jingwei Ling, Yang He, Usman A Javid, Shixin Xue, and Qiang Lin. 2020. Lithium niobate photonic-crystal electro-optic modulator. *Nat. Commun.* 11 (2020), 4123.

[176] Seng W. Loke. 2022. From Distributed Quantum Computing to Quantum Internet Computing: an Overview. (8 2022). arXiv:2208.10127 [cs.DC]

[177] Paul Magnard, Simon Storz, Philipp Kurpiers, Josua Schär, Fabian Marxer, Janis Lütolf, T Walter, J-C Besse, Mihai Gabureac, Kevin Reuer, et al. 2020. Microwave quantum link between superconducting circuits housed in spatially separated cryogenic systems. *Phys. Rev. Lett.* 125, 26 (2020), 260502.

[178] 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. 2020. Microwave Quantum Link between Superconducting Circuits Housed in Spatially Separated Cryogenic Systems. *Physical Review Letters*. 125, 26, Article 260502 (Dec. 2020), 260502 pages. https://doi.org/10.1103/PhysRevLett.125.260502 arXiv:2008.01642 [quant-ph]

[179] Leigh Martin, Felix Motzoi, Hanhan Li, Mohan Sarovar, and K Birgitta Whaley. 2015. Deterministic generation of remote entanglement with active quantum feedback. *Physical Review A* 92, 6 (2015), 062321.

[180] Dominik Marx and Jürg Hutter. 2009. *Ab initio molecular dynamics: basic theory and advanced methods*. Cambridge University Press.

[181] P. Maunz, S. Olmschenk, D. Hayes, D. N. Matsukevich, L. M. Duan, and C. Monroe. 2009. Heralded Quantum Gate between Remote Quantum Memories. *Physical Review Letters*. 102, 25, Article 250502 (June 2009), 250502 pages. https://doi.org/10.1103/PhysRevLett.102.250502 arXiv:0902.2136 [quant-ph]

[182] Jarrod R McClean, John A Parkhill, and Alán Aspuru-Guzik. 2013. Feynman's clock, a new variational principle, and parallel-in-time quantum dynamics. *Proceedings of the National Academy of Sciences* 110, 41 (2013), E3901–E3909.

[183] Timothy P McKenna, Jeremy D Witmer, Rishi N Patel, Wentao Jiang, Raphaël Van Laer, Patricio Arrangoiz-Arriola, E Alex Wollack, Jason F Herrmann, and Amir H Safavi-Naeini. 2020. Cryogenic microwave-to-optical conversion using a triply resonant lithium-niobate-on-sapphire transducer. *Optica* 7, 12 (2020), 1737–1745.

[184] Evan McKinney, Mingkang Xia, Chao Zhou, Pinlei Lu, Michael Hatridge, and Alex K. Jones. 2022. Co-Designed Architectures for Modular Superconducting Quantum Computers. *arXiv preprint arXiv:2205.04387* (2022).

[185] Karan K. Mehta, Chi Zhang, Maciej Malinowski, Thanh-Long Nguyen, Martin Stadler, and Jonathan P. Home. 2020. Integrated optical multi-ion quantum logic. *Nature*. 586, 7830 (Oct. 2020), 533–537. https://doi.org/10.1038/s41586-020-2823-6 arXiv:2002.02258 [quant-ph]

[186] Rodney Van Meter, WJ Munro, Kae Nemoto, and Kohei M Itoh. 2008. Arithmetic on a distributed-memory quantum multicomputer. *ACM Journal on Emerging Technologies in Computing Systems (JETC)* 3, 4 (2008), 1–23.

[187] Rodney Doyle Van Meter III. 2006. Architecture of a quantum multicomputer optimized for shor's factoring algorithm. *arXiv preprint quant-ph/0607065* (2006).

[188] Jiří Minář, Hugues De Riedmatten, Christoph Simon, Hugo Zbinden, and Nicolas Gisin. 2008. Phase-noise measurements in long-fiber interferometers for quantum-repeater applications. *Phys. Rev. A* 77, 5 (2008), 052325.

[189] Mohammad Mirhosseini, Alp Sipahigil, Mahmoud Kalaee, and Oskar Painter. 2020. Superconducting qubit to optical photon transduction. *Nature*. 588, 7839 (Jan. 2020), 599–603. https://doi.org/10.1038/s41586-020-3038-6

[190] Mohammad Mirhosseini, Alp Sipahigil, Mahmoud Kalaee, and Oskar Painter. 2020. Superconducting qubit to optical photon transduction. *Nature* 588, 7839 (2020), 599–603.

[191] Kosuke Mitarai and Keisuke Fujii. 2020. Constructing a virtual two-qubit gate by sampling single-qubit operations. *New Journal of Physics* 23 (2020), 023021. https://doi.org/10.1088/1367-2630/abd7bc

[192] D. L. Moehring, P. Maunz, S. Olmschenk, K. C. Younge, D. N. Matsukevich, L. M. Duan, and C. Monroe. 2007. Entanglement of single-atom quantum bits at a distance. *Nature*. 449, 7158 (Sept. 2007), 68–71. https://doi.org/10.1038/nature06118

[193] C Monroe, R Raussendorf, A Ruthven, KR Brown, P Maunz, L-M Duan, and J Kim. 2014. Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects. *Physical Review A* 89, 2 (2014), 022317.

[194] C. Monroe, R. Raussendorf, A. Ruthven, K. R. Brown, P. Maunz, L. M. Duan, and J. Kim. 2012. Large Scale Modular Quantum Computer Architecture with Atomic Memory and Photonic Interconnects. *arXiv e-prints*, Article arXiv:1208.0391 (Aug. 2012), arXiv:1208.0391 pages. arXiv:1208.0391 [quant-ph]

[195] C. Monroe, R. Raussendorf, A. Ruthven, K. R. Brown, P. Maunz, L. M. Duan, and J. Kim. 2014. Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects. *Physical Review A*. 89, 2, Article 022317 (Feb. 2014), 022317 pages. https://doi.org/10.1103/PhysRevA.89.022317

[196] Mario Motta, Chong Sun, Adrian T. K. Tan, Matthew J. O'Rourke, Erika Ye, Austin J. Minnich, Fernando G. S. L. Brandão, and Garnet Kin-Lic Chan. 2020. Determining eigenstates and thermal states on a quantum computer using quantum imaginary time evolution. *Nature Physics* 16, 2 (Jan. 2020), 205–210. https://doi.org/10.1038/s41567-019-0704-4 arXiv:1901.07653 [quant-ph]

[197] A. Narla, S. Shankar, M. Hatridge, Z. Leghtas, K. M. Sliwa, E. Zalys-Geller, S. O. Mundhada, W. Pfaff, L. Frunzio, R. J. Schoelkopf, and M. H. Devoret. 2016. Robust Concurrent Remote Entanglement Between Two Superconducting Qubits. *Physical Review X* 6, 3, Article 031036 (July 2016), 031036 pages. https://doi.org/10.1103/PhysRevX.6.031036 arXiv:1603.03742 [quant-ph]

[198] Chandra M Natarajan, Michael G Tanner, and Robert H Hadfield. 2012. Superconducting nanowire single-photon detectors: physics and applications. *Supercond. Sci. Technol.* 25, 6 (2012), 063001.

[199] Juniyali Nauriyal, Meiting Song, Raymond Yu, and Jaime Cardenas. 2019. Fiber-to-chip fusion splicing for low-loss photonic packaging. *Optica* 6, 5 (2019), 549–552.

[200] Kae Nemoto, Michael Trupke, Simon J. Devitt, Burkhard Scharfenberger, Kathrin Buczak, Jörg Schmiedmayer, and William J. Munro. 2016. Photonic Quantum Networks formed from $NV^-$ centers. *Scientific Reports* 6, Article 26284 (May 2016), 26284 pages. https://doi.org/10.1038/srep26284 arXiv:1412.5950 [quant-ph]

[201] Tomáš Neuman, Matt Eichenfield, Matthew E. Trusheim, Lisa Hackett, Prineha Narang, and Dirk Englund. 2021. A phononic interface between a superconducting quantum processor and quantum networked spin memories. *npj Quantum Information* 7, Article 121 (Jan. 2021), 121 pages. https://doi.org/10.1038/s41534-021-00457-4

[202] Naomi H. Nickerson, Joseph F. Fitzsimons, and Simon C. Benjamin. 2014. Freely Scalable Quantum Technologies Using Cells of 5-to-50 Qubits with Very Lossy and Noisy Photonic Links. *Phys. Rev. X* 4 (Dec 2014), 041041. Issue 4. https://doi.org/10.1103/PhysRevX.4.041041

[203] Naomi H. Nickerson, Ying Li, and Simon C. Benjamin. 2013. Topological quantum computing with a very noisy network and local error rates approaching one percent. *Nat. Commun.* 4, 1 (23 Apr 2013), 1756. https://doi.org/10.1038/ncomms2773

[204] R. J. Niffenegger, J. Stuart, C. Sorace-Agaskar, D. Kharas, S. Bramhavar, C. D. Bruzewicz, W. Loh, R. T. Maxson, R. McConnell, D. Reens, G. N. West, J. M. Sage, and J. Chiaverini. 2020. Integrated multi-wavelength control of an ion qubit. *Nature*. 586, 7830 (Oct. 2020), 538–542. https://doi.org/10.1038/s41586-020-2811-x arXiv:2001.05052 [quant-ph]

[205] Nissim Ofek, Andrei Petrenko, Reinier Heeres, Philip Reinhold, Zaki Leghtas, Brian Vlastakis, Yehan Liu, Luigi Frunzio, S. M. Girvin, L. Jiang, Mazyar Mirrahimi, M. H. Devoret, and R. J. Schoelkopf. 2016. Extending the lifetime of a quantum bit with error correction in superconducting circuits. *Nature* 536 (2016). Issue 7617. https://doi.org/10.1038/nature18949

[206] Marcelo Orenes-Vera, Aninda Manocha, Jonathan Balkind, Fei Gao, Juan L Aragón, David Wentzlaff, and Margaret Martonosi. 2022. Tiny but mighty: designing and realizing scalable latency tolerance for manycore SoCs.. In *ISCA*. 817–830.

[207] Christopher O'Brien, Nikolai Lauk, Susanne Blum, Giovanna Morigi, and Michael Fleischhauer. 2014. Interfacing superconducting qubits and telecom photons via a rare-earth-doped crystal. *Phys. Rev. Lett.* 113, 6 (2014), 063603.

[208] Thomas E O'Brien, B Tarasinski, and Leo DiCarlo. 2017. Density-matrix simulation of small surface codes under current and projected experimental noise. *npj Quantum Information* 3, 1 (2017), 1–8.

[209] Mihir Pant, Hari Krovi, Don Towsley, Leandros Tassiulas, Liang Jiang, Prithwish Basu, Dirk Englund, and Saikat Guha. 2017. Routing entanglement in the quantum internet. *arXiv e-prints*, Article arXiv:1708.07142 (Aug. 2017), arXiv:1708.07142 pages. arXiv:1708.07142 [quant-ph]

[210] François Pellegrini. 2012. Scotch and PT-scotch graph partitioning software: an overview. *Combinatorial Scientific Computing* (2012), 373–406.

[211] Tianyi Peng, Aram W Harrow, Maris Ozols, and Xiaodi Wu. 2020. Simulating large quantum circuits on a small quantum computer. *Physical Review Letters* 125, 15 (2020), 150504.

[212] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J Love, Alán Aspuru-Guzik, and Jeremy L O'brien. 2014. A variational eigenvalue solver on a photonic quantum processor. *Nature communications* 5, 1 (2014), 1–7.

[213] 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. 2014. Unconditional quantum teleportation between distant solid-state quantum bits. *Science* 345, 6196 (2014), 532–535. https://doi.org/10.1126/science.1253512 arXiv:https://www.science.org/doi/pdf/10.1126/science.1253512

[214] 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. 2021. Demonstration of the trapped-ion quantum CCD computer architecture. *Nature*. 592, 7853 (Jan. 2021), 209–213. https://doi.org/10.1038/s41586-021-03318-4 arXiv:2003.01293 [quant-ph]

[215] A. Pirker and W. Dür. 2019. A quantum network stack and protocols for reliable entanglement-based networks. *New Journal of Physics* 21, 3, Article 033003 (March 2019), 033003 pages. https://doi.org/10.1088/1367-2630/ab05f7 arXiv:1810.03556 [quant-ph]

[216] Christophe Piveteau and David Sutter. 2022. Circuit knitting with classical communication. *arXiv preprint 2205.00016* (2022).
[217] A.P.M. Place, L.V.H. Rodgers, P. Mundada, et al. 2021. New material platform for superconducting transmon qubits with coherence times exceeding 0.3 milliseconds. *Nat. Commun.* 12 (2021), 1779.
[218] 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. 2021. Realization of a multinode quantum network of remote solid-state qubits. *Science* 372, 6539 (2021), 259–264. https://doi.org/10.1126/science.abg1919 arXiv:https://www.science.org/doi/pdf/10.1126/science.abg1919
[219] 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. 2021. Realization of a multinode quantum network of remote solid-state qubits. *Science* 372, 6539 (April 2021), 259–264. https://doi.org/10.1126/science.abg1919 arXiv:2102.04471 [quant-ph]
[220] I. E. Protsenko, G. Reymond, N. Schlosser, and P. Grangier. 2002. Conditional quantum logic using two atomic qubits. *Phys. Rev. A* 66 (Dec 2002), 062306. Issue 6. https://doi.org/10.1103/PhysRevA.66.062306
[221] Matthew W Puckett, Kaikai Liu, Nitesh Chauhan, Qiancheng Zhao, Naijun Jin, Haotian Cheng, Jianfeng Wu, Ryan O Behunin, Peter T Rakich, Karl D Nelson, et al. 2021. 422 Million intrinsic quality factor planar integrated all-waveguide resonator with sub-MHz linewidth. *Nat. Commun.* 12, 1 (2021), 934.
[222] Shruti Puri, Alexander Grimm, Philippe Campagne-Ibarcq, Alec Eickbusch, Kyungjoo Noh, Gabrielle Roberts, Liang Jiang, Mazyar Mirrahimi, Michel H. Devoret, and S. M. Girvin. 2019. Stabilized Cat in a Driven Nonlinear Cavity: A Fault-Tolerant Error Syndrome Detector. *Phys. Rev. X* 9 (Oct 2019), 041009. Issue 4. https://doi.org/10.1103/PhysRevX.9.041009
[223] Chunming Qiao, Yangming Zhao, Gongming Zhao, and Hongli Xu. 2022. Quantum Data Networking for Distributed Quantum Computing: Opportunities and Challenges. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications Workshops*. https://doi.org/10.1109/INFOCOMWKSHPS54753.2022.9798138
[224] P. Rabl, D. DeMille, J. M. Doyle, M. D. Lukin, R. J. Schoelkopf, and P. Zoller. 2006. Hybrid Quantum Processors: Molecular Ensembles as Quantum Memory for Solid State Circuits. *Phys. Rev. Lett.* 97 (Jul 2006), 033003. Issue 3. https://doi.org/10.1103/PhysRevLett.97.033003
[225] P. Rabl and P. Zoller. 2007. Molecular dipolar crystals as high-fidelity quantum memory for hybrid quantum computing. *Phys. Rev. A* 76 (Oct 2007), 042308. Issue 4. https://doi.org/10.1103/PhysRevA.76.042308
[226] Joshua Ramette, Josiah Sinclair, Nikolas P. Breuckmann, and Vladan Vuletić. 2023. Fault-Tolerant Connection of Error-Corrected Qubits with Noisy Links. *arXiv e-prints*, Article arXiv:2302.01296 (Feb. 2023), arXiv:2302.01296 pages. https://doi.org/10.48550/arXiv.2302.01296 arXiv:2302.01296 [quant-ph]
[227] Alexander P. Read, Benjamin J. Chapman, Chan U Lei, Jacob C. Curtis, Suhas Ganjam, Lev Krayzman, Luigi Frunzio, and Robert J. Schoelkopf. 2022. Precision measurement of the microwave dielectric loss of sapphire in the quantum regime with parts-per-billion sensitivity. *arXiv preprint arXiv:2206.14334* (2022).
[228] Matthew Reagor, Christopher B. Osborn, Nikolas Tezak, Alexa Staley, Guenevere Prawiroatmodjo, Michael Scheer, Nasser Alidoust, Eyob A. Sete, Nicolas Didier, Marcus P. da Silva, Ezer Acala, Joel Angeles, Andrew Bestwick, Maxwell Block, Benjamin Bloom, Adam Bradley, Catvu Bui, Shane Caldwell, Lauren Capelluto, Rick Chilcott, Jeff Cordova, Genya Crossman, Michael Curtis, Saniya Deshpande, Tristan El Bouayadi, Daniel Girshovich, Sabrina Hong, Alex Hudson, Peter Karalekas, Kat Kuang, Michael Lenihan, Riccardo Manenti, Thomas Manning, Jayss Marshall, Yuvraj Mohan, William O'Brien, Johannes Otterbach, Alexander Papageorge, Jean-Philip Paquette, Michael Pelstring, Anthony Polloreno, Vijay Rawat, Colm A. Ryan, Russ Renzas, Nick Rubin, Damon Russel, Michael Rust, Diego Scarabelli, Michael Selvanayagam, Rodney Sinclair, Robert Smith, Mark Suska, Ting-Wai To, Mehrnoosh Vahidpour, Nagesh Vodrahalli, Tyler Whyland, Kamal Yadav, William Zeng, and Chad T. Rigetti. 2018. Demonstration of universal parametric entangling gates on a multi-qubit lattice. *Science Advances* 4, 2 (Feb. 2018), eaao3603. https://doi.org/10.1126/sciadv.aao3603
[229] Matthew Reagor, Wolfgang Pfaff, Christopher Axline, Reinier W. Heeres, Nissim Ofek, Katrina Sliwa, Eric Holland, Chen Wang, Jacob Blumoff, Kevin Chou, Michael J. Hatridge, Luigi Frunzio, Michel H. Devoret, Liang Jiang, and Robert J. Schoelkopf. 2016. Quantum memory with millisecond coherence in circuit QED. *Phys. Rev. B* 94 (Jul 2016), 014506. Issue 1. https://doi.org/10.1103/PhysRevB.94.014506
[230] Markus Reiher, Nathan Wiebe, Krysta M Svore, Dave Wecker, and Matthias Troyer. 2017. Elucidating reaction mechanisms on quantum computers. *Proceedings of the national academy of sciences* 114, 29 (2017), 7555–7560.
[231] Stephan Ritter, Christian Nölleke, Carolin Hahn, Andreas Reiserer, Andreas Neuzner, Manuel Uphoff, Martin Mücke, Eden Figueroa, Joerg Bochmann, and Gerhard Rempe. 2012. An elementary quantum network of single atoms in optical cavities. *Nature.* 484, 7393 (April 2012), 195–200. https://doi.org/10.1038/nature11023 arXiv:1202.5955 [quant-ph]
[232] 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. 2014. Observation of Measurement-Induced Entanglement and Quantum Trajectories of Remote Superconducting Qubits. *Phys. Rev. Lett.* 112 (Apr 2014), 170501. Issue 17. https://doi.org/10.1103/PhysRevLett.112.170501
[233] A. Romanenko, R. Pilipenko, S. Zorzetti, D. Frolov, M. Awida, S. Belomestnykh, S. Posen, and A. Grassellino. 2020. Three-Dimensional Superconducting Resonators at $T < 20$ mK with Photon Lifetimes up to $\tau = 2$ s. *Phys. Rev. Applied* 13 (Mar 2020), 034032. Issue 3. https://doi.org/10.1103/PhysRevApplied.13.034032

[234] S. Rosenblum, P. Reinhold, M. Mirrahimi, Liang Jiang, L. Frunzio, and R. J. Schoelkopf. 2018. Fault-tolerant detection of a quantum error. *Science* 361 (2018). Issue 6399. https://doi.org/10.1126/science.aat3996

[235] M. A. Rowej, 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. 2002. Transport of Quantum States and Separation of Ions in a Dual RF Ion Trap. *Quantum Info. Comput.* 2, 4 (jun 2002), 257–271.

[236] Alfredo Rueda, William Hease, Shabir Barzanjeh, and Johannes M Fink. 2019. Electro-optic entanglement source for microwave to telecom quantum state transfer. *npj Quantum Inf.* 5, 1 (2019), 1–11.

[237] Alfredo Rueda, Florian Sedlmeir, Michele C Collodo, Ulrich Vogl, Birgit Stiller, Gerhard Schunk, Dmitry V Strekalov, Christoph Marquardt, Johannes M Fink, Oskar Painter, et al. 2016. Efficient microwave to optical photon conversion: an electro-optical realization. *Optica* 3, 6 (2016), 597–604.

[238] Qiao Ruihong and Meng Ying. 2019. Research Progress Of Quantum Repeaters. *Journal of Physics: Conference Series* 1237, 5 (jun 2019), 052032. https://doi.org/10.1088/1742-6596/1237/5/052032

[239] Rishabh Sahu, William Hease, Alfredo Rueda, Georg Arnold, Liu Qiu, and Johannes M Fink. 2022. Quantum-enabled operation of a microwave-optical interface. *Nat. Commun.* 13 (2022), 1276.

[240] Nicolas Sangouard, Christoph Simon, Hugues de Riedmatten, and Nicolas Gisin. 2011. Quantum repeaters based on atomic ensembles and linear optics. *Rev. Mod. Phys.* 83 (Mar 2011), 33–80. Issue 1. https://doi.org/10.1103/RevModPhys.83.33

[241] C. Schuck, X. Guo, L. Fan, X. Ma, M. Poot, and H. X. Tang. 2016. Quantum interference in heterogeneous superconducting-photonic circuits on a silicon chip. *Nature Communications* 7, Article 10352 (Jan. 2016), 10352 pages. https://doi.org/10.1038/ncomms10352 arXiv:1511.07081 [quant-ph]

[242] D. I. Schuster, Lev S. Bishop, I. L. Chuang, D. DeMille, and R. J. Schoelkopf. 2011. Cavity QED in a molecular ion trap. *Phys. Rev. A* 83 (Jan 2011), 012311. Issue 1. https://doi.org/10.1103/PhysRevA.83.012311

[243] Steven L. Scott and et al. 1996. The Cray T3E Network: Adaptive Routing in a High Performance 3D Torus.

[244] Yunong Shi, Pranav Gokhale, Prakash Murali, Jonathan M Baker, Casey Duckering, Yongshan Ding, Natalie C Brown, Christopher Chamberland, Ali Javadi-Abhari, Andrew W Cross, et al. 2020. Resource-efficient quantum computing by breaking abstractions. *Proc. IEEE* 108, 8 (2020), 1353–1370.

[245] Georgios Sinatkas, Thomas Christopoulos, Odysseas Tsilipakos, and Emmanouil E Kriezis. 2021. Electro-optic modulation in integrated photonics. *J. Appl. Phys.* 130, 1 (2021), 010901.

[246] Marcos Yukio Siraichi, Vinícius Fernandes dos Santos, Caroline Collange, and Fernando Magno Quintão Pereira. 2018. Qubit allocation. In *Proceedings of the 2018 International Symposium on Code Generation and Optimization*. 113–125.

[247] 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. 2022. Real-time quantum error correction beyond break-even. *arXiv e-prints*, Article arXiv:2211.09116 (Nov. 2022), arXiv:2211.09116 pages. arXiv:2211.09116 [quant-ph]

[248] Mohammad Soltani, Mian Zhang, Colm Ryan, Guilhem J Ribeill, Cheng Wang, and Marko Loncar. 2017. Efficient quantum microwave-to-optical conversion using electro-optic nanophotonic coupled resonators. *Phys. Rev. A* 96, 4 (2017), 043808.

[249] Samuel Stein, Nathan Wiebe, Yufei Ding, Peng Bo, Karol Kowalski, Nathan Baker, James Ang, and Ang Li. 2022. EQC: ensembled quantum computing for variational quantum algorithms. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*. 59–71.

[250] Thomas Sterling, Ewing Lusk, and William Gropp. 2003. *Beowulf Cluster Computing with Linux* (2 ed.). MIT Press, Cambridge, MA, USA.

[251] Neereja Sundaresan, Theodore J. Yoder, Youngseok Kim, Muyuan Li, Edward H. Chen, Grace Harper, Ted Thorbeck, Andrew W. Cross, Antonio D. Córcoles, and Maika Takita. 2022. Matching and maximum likelihood decoding of a multi-round subsystem quantum error correction experiment. *arXiv preprint arXiv:2203.07205* (2022).

[252] Youngkyu Sung, Leon Ding, Jochen Braumüller, Antti Vepsäläinen, Bharath Kannan, Morten Kjaergaard, Ami Greene, Gabriel O. Samach, Chris McNally, David Kim, Alexander Melville, Bethany M. Niedzielski, Mollie E. Schwartz, Jonilyn L. Yoder, Terry P. Orlando, Simon Gustavsson, and William D. Oliver. 2021. Realization of High-Fidelity CZ and $ZZ$-Free iSWAP Gates with a Tunable Coupler. *Phys. Rev. X* 11 (Jun 2021), 021058. Issue 2. https://doi.org/10.1103/PhysRevX.11.021058

[253] Bochen Tan and Jason Cong. 2020. Optimality study of existing quantum computing layout synthesis tools. *IEEE Trans. Comput.* 70, 9 (2020), 1363–1373.

[254] A.S. Tanenbaum and M. van Steen. 2007. *Distributed Systems: Principles and Paradigms*. Pearson Prentice Hall. https://books.google.com/books?id=DL8ZAQAAIAAJ

[255] Wei Tang and Margaret Martonosi. 2022. Cutting Quantum Circuits to Run on Quantum and Classical Platforms. *arXiv preprint arXiv:2205.05836* (2022).

[256] Wei Tang and Margaret Martonosi. 2022. ScaleQC: A Scalable Framework for Hybrid Computation on Quantum and Classical Processors. *arXiv preprint arXiv:2207.00933* (2022).

[257] Wei Tang, Teague Tomesh, Martin Suchara, Jeffrey Larson, and Margaret Martonosi. 2021. Cutqc: using small quantum computers for large quantum circuit evaluations. In *Proceedings of the 26th ACM International conference on architectural support for programming languages and operating systems*. 473–486.

[258] Kristan Temme, Sergey Bravyi, and Jay M. Gambetta. 2017. Error Mitigation for Short-Depth Quantum Circuits. *Phys. Rev. Lett.* 119 (Nov 2017), 180509. Issue 18. https://doi.org/10.1103/PhysRevLett.119.180509

[259] B. M. Terhal, J. Conrad, and C. Vuillot. 2020. Towards scalable bosonic quantum error correction. Issue 4. https://doi.org/10.1088/2058-9565/ab98a5

[260] Darshan D Thaker, Tzvetan S Metodi, Andrew W Cross, Isaac L Chuang, and Frederic T Chong. 2006. Quantum memory hierarchies: Efficient designs to match available parallelism in quantum computing. In *33rd International Symposium on Computer Architecture (ISCA'06)*. IEEE, 378–390.

[261] Teague Tomesh, Pranav Gokhale, Victory Omole, Gokul Subramanian Ravi, Kaitlin N Smith, Joshua Viszlai, Xin-Chuan Wu, Nikos Hardavellas, Margaret R Martonosi, and Frederic T Chong. 2022. Supermarq: A scalable quantum benchmark suite. In *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 587–603.

[262] T. Tomesh and M. Martonosi. 2021. Quantum Codesign. *IEEE Micro* 41, 05 (sep 2021), 33–40. https://doi.org/10.1109/MM.2021.3094461

[263] Mankei Tsang. 2011. Cavity quantum electro-optics. II. Input-output relations between traveling optical and microwave fields. *Phys. Rev. A* 84, 4 (2011), 043845.

[264] Hai-Tao Tu, Kai-Yu Liao, Zuan-Xian Zhang, Xiao-Hong Liu, Shun-Yuan Zheng, Shu-Zhe Yang, Xin-Ding Zhang, Hui Yan, and Shi-Liang Zhu. 2022. High-efficiency coherent microwave-to-optics conversion via off-resonant scattering. *Nat. Photon.* 16, 4 (2022), 291–296.

[265] Amit Vainsencher, KJ Satzinger, GA Peairs, and AN Cleland. 2016. Bi-directional conversion between microwave and optical frequencies in a piezoelectric optomechanical device. *Appl. Phys. Lett.* 109, 3 (2016), 033107.

[266] Ewout van den Berg, Zlatko K. Minev, Abhinav Kandala, and Kristan Temme. 2022. Probabilistic error cancellation with sparse Pauli-Lindblad models on noisy quantum processors. *arXiv preprint 2201.09866* (2022).

[267] Maarten van Steen and Andrew S Tanenbaum. 2016. A brief introduction to distributed systems. *Computing* 98, 10 (2016), 967–1009.

[268] J. Verdú, H. Zoubi, Ch. Koller, J. Majer, H. Ritsch, and J. Schmiedmayer. 2009. Strong Magnetic Coupling of an Ultracold Gas to a Superconducting Waveguide Cavity. *Phys. Rev. Lett.* 103 (Jul 2009), 043603. Issue 4. https://doi.org/10.1103/PhysRevLett.103.043603

[269] Michelle Victora, Stefan Krastanov, Alexander Sanchez de la Cerda, Steven Willis, and Prineha Narang. 2020. Purification and Entanglement Routing on Quantum Networks. *arXiv e-prints*, Article arXiv:2011.11644 (Nov. 2020), arXiv:2011.11644 pages. arXiv:2011.11644 [quant-ph]

[270] Thibault Vogt, Christian Gross, Jingshan Han, Sambit B Pal, Mark Lam, Martin Kiffner, and Wenhui Li. 2019. Efficient microwave-to-optical conversion using Rydberg atoms. *Phys. Rev. A* 99, 2 (2019), 023832.

[271] E. Pavarini E. Koch A. Lichtenstein D. Vollhardt (Ed.). 2018. *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.

[272] M. Wallquist, K. Hammerer, P. Rabl, M. Lukin, and P. Zoller. 2009. Hybrid quantum devices and quantum engineering. *Physica Scripta Volume T* 137, Article 014001 (Dec. 2009), 014001 pages. https://doi.org/10.1088/0031-8949/2009/T137/014001 arXiv:0911.3835 [quant-ph]

[273] C. Wang, C. Axline, Y. Y. Gao, T. Brecht, Y. Chu, L. Frunzio, M. H. Devoret, and R. J. Schoelkopf. 2015. Surface participation and dielectric loss in superconducting qubits. *Applied Physics Letters* 107, 16 (Oct. 2015), 162601. https://doi.org/10.1063/1.4934486

[274] Chenlu Wang, Xuegang Li, Huikai Xu, Zhiyuan Li, Junhua Wang, Zhen Yang, Zhenyu Mi, Xuehui Liang, Tang Su, Chuhong Yang, Guangyue Wang, Wenyan Wang, Yongchao Li, Mo Chen, Chengyao Li, Kehuan Linghu, Jiaxiu Han, Yingshan Zhang, Yulong Feng, Yu Song, Teng Ma, Jingning Zhang, Ruixia Wang, Peng Zhao, Weiyang Liu, Guangming Xue, Yirong Jin, and Haifeng Yu. 2022. Towards practical quantum computers: transmon qubit with a lifetime approaching 0.5 milliseconds. *npj Quantum Information* 8, 1 (2022), 3. https://doi.org/10.1038/s41534-021-00510-2

[275] Shannon X. Wang, Yufei Ge, Jaroslaw Labaziewicz, Eric Dauler, Karl Berggren, and Isaac L. Chuang. 2010. Superconducting microfabricated ion traps. *Applied Physics Letters* 97, 24, Article 244102 (Dec. 2010), 244102 pages. https://doi.org/10.1063/1.3526733 arXiv:1010.6108 [quant-ph]

[276] Zhixin Wang, Mingrui Xu, Xu Han, Wei Fu, Shruti Puri, SM Girvin, Hong X Tang, S Shankar, and MH Devoret. 2021. Quantum microwave radiometry with a superconducting qubit. *Phys. Rev. Lett.* 126, 18 (2021), 180501.

[277] Arieh Warshel. 2014. Multiscale modeling of biological functions: from enzymes to molecular machines (Nobel Lecture). *Angewandte Chemie International Edition* 53, 38 (2014), 10020–10031.

[278] Stephanie Wehner, David Elkouss, and Ronald Hanson. 2018. Quantum internet: A vision for the road ahead. *Science* 362, 6412 (Oct. 2018), 9288. https://doi.org/10.1126/science.aam9288

[279] KX Wei, E Magesan, I Lauer, S Srinivasan, DF Bogorin, S Carnevale, GA Keefe, Y Kim, D Klaus, W Landers, et al. 2022. Hamiltonian Engineering with Multicolor Drives for Fast Entangling Gates and Quantum Crosstalk Cancellation. *Physical Review Letters* 129, 6 (2022), 060501.

[280] 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. 2022. Hamiltonian Engineering with Multicolor Drives for

Fast Entangling Gates and Quantum Crosstalk Cancellation. *Phys. Rev. Lett.* 129 (Aug 2022), 060501. Issue 6. https://doi.org/10.1103/PhysRevLett.129.060501

[281] Matthew Welborn, Takashi Tsuchimochi, and Troy Van Voorhis. 2016. Bootstrap embedding: An internally consistent fragment-based method. *The Journal of Chemical Physics* 145, 7 (2016), 074102.

[282] Samuel Williams, Andrew Waterman, and David Patterson. 2009. Roofline: an insightful visual performance model for multicore architectures. *Commun. ACM* 52, 4 (2009), 65–76.

[283] Anbang Wu, Yufei Ding, and Ang Li. 2022. CollComm: Enabling Efficient Collective Quantum Communication Based on EPR buffering. (8 2022). arXiv:2208.06724 [quant-ph]

[284] Anbang Wu, Yufei Ding, and Ang Li. 2022. CollComm: Enabling Efficient Collective Quantum Communication Based on EPR buffering. *arXiv preprint arXiv:2208.06724* (2022).

[285] Anbang Wu, Hezi Zhang, Gushu Li, Alireza Shabani, Yuan Xie, and Yufei Ding. 2022. AutoComm: A Framework for Enabling Efficient Communication in Distributed Quantum Programs. *arXiv preprint arXiv:2207.11674* (2022).

[286] Anbang Wu, Hezi Zhang, Gushu Li, Alireza Shabani, Yuan Xie, and Yufei Ding. 2022. AutoComm: A Framework for Enabling Efficient Communication in Distributed Quantum Programs. (7 2022). arXiv:2207.11674 [quant-ph]

[287] Jing Wu, Chaohan Cui, Linran Fan, and Quntao Zhuang. 2021. Deterministic microwave-optical transduction based on quantum teleportation. *Phys. Rev. Appl.* 16, 6 (2021), 064044.

[288] Jun-Yi Wu, Kosuke Matsui, Tim Forrer, Akihito Soeda, Pablo Andrés-Martínez, Daniel Mills, Luciana Henaut, and Mio Murao. 2023. Entanglement-efficient bipartite-distributed quantum computing. *Quantum* 7 (2023), 1196.

[289] Mingrui Xu, Xu Han, Chang-Ling Zou, Wei Fu, Yuntao Xu, Changchun Zhong, Liang Jiang, and Hong X Tang. 2020. Radiative cooling of a superconducting resonator. *Phys. Rev. Lett.* 124, 3 (2020), 033602.

[290] Yuntao Xu, Ayed Al Sayem, Linran Fan, Chang-Ling Zou, Sihao Wang, Risheng Cheng, Wei Fu, Likai Yang, Mingrui Xu, and Hong X Tang. 2021. Bidirectional interconversion of microwave and light with thin-film lithium niobate. *Nat. Commun.* 12 (2021), 4453.

[291] Haoxiong Yan, Youpeng Zhong, Hung-Shen Chang, Audrey Bienfait, Ming-Han Chou, Christopher R. Conner, Étienne Dumur, Joel Grebel, Rhys G. Povey, and Andrew N. Cleland. 2022. Entanglement Purification and Protection in a Superconducting Quantum Network. *Phys. Rev. Lett.* 128 (Feb 2022), 080504. Issue 8. https://doi.org/10.1103/PhysRevLett.128.080504

[292] Rusen Yan, Guru Khalsa, Suresh Vishwanath, Yimo Han, John Wright, Sergei Rouvimov, D Scott Katzer, Neeraj Nepal, Brian P Downey, David A Muller, et al. 2018. GaN/NbN epitaxial semiconductor/superconductor heterostructures. *Nature* 555, 7695 (2018), 183–189.

[293] Yongxin Yao, Feng Zhang, Cai-Zhuang Wang, Kai-Ming Ho, and Peter P. Orth. 2021. Gutzwiller hybrid quantum-classical computing approach for correlated materials. *Phys. Rev. Research* 3 (Feb 2021), 013184. Issue 1. https://doi.org/10.1103/PhysRevResearch.3.013184

[294] Anocha Yimsiriwattana and Samuel J Lomonaco Jr. 2004. Distributed quantum computing: A distributed Shor algorithm. In *Quantum Information and Computation II*, Vol. 5436. SPIE, 360–372.

[295] Andy B Yoo, Morris A Jette, and Mark Grondona. 2003. Slurm: Simple linux utility for resource management. In *Workshop on job scheduling strategies for parallel processing*. Springer, 44–60.

[296] C. B. Young, A. Safari, P. Huft, J. Zhang, E. Oh, R. Chinnarasu, and M. Saffman. 2022. An architecture for quantum networking of neutral atom processors. *Applied Physics B: Lasers and Optics* 128, 8, Article 151 (Aug. 2022), 151 pages. https://doi.org/10.1007/s00340-022-07865-0 arXiv:2202.01634 [quant-ph]

[297] C. B. Young, A. Safari, P. Huft, J. Zhang, E. Oh, R. Chinnarasu, and M. Saffman. 2022. An architecture for quantum networking of neutral atom processors. *Appl. Phys. B* 128, 8 (2022), 151. https://doi.org/10.1007/s00340-022-07865-0 arXiv:2202.01634 [quant-ph]

[298] Amir Youssefi, Itay Shomroni, Yash J Joshi, Nathan R Bernier, Anton Lukashchuk, Philipp Uhrich, Liu Qiu, and Tobias J Kippenberg. 2021. A cryogenic electro-optic interconnect for superconducting devices. *Nat. Electron.* 4, 5 (2021), 326–332.

[299] Kuan Zhang, Jayne Thompson, Xiang Zhang, Yangchao Shen, Yao Lu, Shuaining Zhang, Jiajun Ma, Vlatko Vedral, Mile Gu, and Kihwan Kim. 2019. Modular quantum computation in a trapped ion system. *Nature Communications* 10, Article 4692 (Oct. 2019), 4692 pages. https://doi.org/10.1038/s41467-019-12643-2 arXiv:1907.12171 [quant-ph]

[300] Mengzhen Zhang, Chang-Ling Zou, and Liang Jiang. 2018. Quantum transduction with adaptive control. *Phys. Rev. Lett.* 120, 2 (2018), 020502.

[301] Xufeng Zhang, Na Zhu, Chang-Ling Zou, and Hong X Tang. 2016. Optomagnonic whispering gallery microresonators. *Phys. Rev. Lett.* 117, 12 (2016), 123605.

[302] Xufeng Zhang, Chang-Ling Zou, Liang Jiang, and Hong X Tang. 2014. Strongly coupled magnons and cavity microwave photons. *Phys. Rev. Lett.* 113, 15 (2014), 156401.

[303] Yu Zhang, Lukasz Cincio, Christian F. A. Negre, Piotr Czarnik, Patrick Coles, Petr M. Anisimov, Susan M. Mniszewski, Sergei Tretiak, and Pavel A. Dub. 2021. Variational Quantum Eigensolver with Reduced Circuit Complexity. *arXiv e-prints*, Article arXiv:2106.07619 (June 2021), arXiv:2106.07619 pages.

[304] Youwei Zhao, Yangsen Ye, He-Liang Huang, Yiming Zhang, Dachao Wu, Huijie Guan, Qingling Zhu, Zuolin Wei, Tan He, Sirui Cao, Fusheng Chen, Tung-Hsun Chung, Hui Deng, Daojin Fan, Ming Gong, Cheng Guo, Shaojun Guo, Lianchen Han, Na Li, Shaowei Li, Yuan Li, Futian Liang, Jin Lin, Haoran Qian, Hao Rong, Hong Su, Lihua Sun, Shiyu Wang, Yulin Wu, Yu Xu, Chong Ying, Jiale Yu, Chen Zha, Kaili

Zhang, Yong-Heng Huo, Chao-Yang Lu, Cheng-Zhi Peng, Xiaobo Zhu, and Jian-Wei Pan. 2022. Realization of an Error-Correcting Surface Code with Superconducting Qubits. *Phys. Rev. Lett.* 129 (Jul 2022), 030501. Issue 3. https://doi.org/10.1103/PhysRevLett.129.030501

[305] Changchun Zhong, Xu Han, Hong X Tang, and Liang Jiang. 2020. Entanglement of microwave-optical modes in a strongly coupled electro-optomechanical system. *Phys. Rev. A* 101, 3 (2020), 032345.

[306] Changchun Zhong, Zhixin Wang, Changling Zou, Mengzhen Zhang, Xu Han, Wei Fu, Mingrui Xu, S Shankar, Michel H Devoret, Hong X Tang, et al. 2020. Proposal for heralded generation and detection of entangled microwave–optical-photon pairs. *Phys. Rev. Lett.* 124, 1 (2020), 010511.

[307] Youpeng Zhong, Hung-Shen Chang, Audrey Bienfait, Étienne Dumur, Ming-Han Chou, Christopher R. Conner, Joel Grebel, Rhys G. Povey, Haoxiong Yan, David I. Schuster, and Andrew N. Cleland. 2021. Deterministic multi-qubit entanglement in a quantum network. *Nature* 590, 7847 (01 Feb 2021), 571–575. https://doi.org/10.1038/s41586-021-03288-7

[308] 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. 2019. Violating Bell's inequality with remotely connected superconducting qubits. *Nature Physics* 15, 8 (01 Aug 2019), 741–744. https://doi.org/10.1038/s41567-019-0507-7

[309] Chao Zhou et al. 2021. A modular quantum computer based on a quantum state router. (9 2021). arXiv:2109.06848 [quant-ph]

[310] Hong-Cai Zhou, Jeffrey R Long, and Omar M Yaghi. 2012. Introduction to metal–organic frameworks. *Chemical reviews* 112, 2 (2012), 673–674.

[311] Na Zhu, Xufeng Zhang, Xu Han, Chang-Ling Zou, Changchun Zhong, Chiao-Hsuan Wang, Liang Jiang, and Hong X Tang. 2020. Waveguide cavity optomagnonics for microwave-to-optics conversion. *Optica* 7, 10 (2020), 1291–1297.

[312] Rongjin Zhuang, Jinze He, Yifan Qi, and Yang Li. 2022. High-Q Thin Film Lithium Niobate Microrings Fabricated with Wet Etching. *Adv. Mater.* (2022), 2208113.

[313] H. Zu, W. Dai, and A. T. A. M. de Waele. 2022. Development of dilution refrigerators-A review. *Cryogenics* 121, Article 103390 (Jan. 2022), 103390 pages. https://doi.org/10.1016/j.cryogenics.2021.103390

Received 30 August 2023; revised 4 April 2024; accepted 24 May 2024