This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MM.2021.3092706, IEEE Micro

Special Issue in Quantum Computing
Editor: Lizy K. John, ljohn@ece.utexas.edu

On Double Full-Stack Communications-Enabled Architectures for Multi-Core Quantum Computers

S. Rodrigo, S. Abadal, E. Alarcón
NanoNetworking Center in Catalonia, Universitat Politècnica de Catalunya

[1]M. Bandic, [1]J. van Someren, [1,2]C. G. Almudever
[1]QuTech, Delft University of Technology
[2]Computer Engineering Department, Technical University of Valencia

***Abstract*—Despite its tremendous potential, it is still unclear how quantum computing will scale to satisfy the requirements of its most powerful applications. Among other issues, there are hard limits to the number of qubits that can be integrated into a single chip. Multi-core architectures are a firm candidate for unlocking the scalability of quantum processors. Nonetheless, the vulnerability and complexity of quantum communications make this a challenging approach. A comprehensive design should imply consolidating the communications stack in the quantum computer architecture. In this paper, we explain how this vision, by entangling communications and computation in the core of the design, may help to solve the open challenges. We also summarize the first results of our application of structured design methodologies backing this vision. With our work, we hope to contribute with design guidelines that may help unleash the potential of quantum computing.**

THE DISCOVERY of quantum mechanics is leading yet another revolution in science. By leveraging properties such as superposition and entanglement, quantum computing promises unprecedented processing power and unconditional security, changing forever crucial areas such as cryptography, biochemistry, big data analysis, or artificial intelligence [1]. However, quantum state decoherence (i.e. loss of quantum information due to unwanted interactions with the environment) and complexity of qubit control, together with many other engineering challenges [2] pose significant obstacles in the development of quantum computing, compromising quantum computers' scalability. Despite sustained and remarkable advances in the quality and the number of qubits

integrated into a single chip, there is still much room for improvement to have practical quantum computers that may demonstrate the full potential that quantum computing has in store [3].

Some previous works have proposed multi-chip architectures for unlocking these scalability issues [4]. Putting together currently available small-sized computing nodes or *cores* as opposed to packing more qubits into monolithic quantum chips alleviates indeed the requirements for control circuits and improves qubit isolation. However, multi-core quantum computers come with their own set of challenges. Particularly, interconnecting quantum chips is far from being a simple task. Quantum data cannot be copied and time is crucial, as quantum decoherence steadily corrupts qubits. Because of this, we postulate that to lay firm foundations for multi- chip quantum computer architectures, a deeply entangled design between computation and communications is essential. Furthermore, a thorough analysis of these systems is needed to study whether they effectively enable scalability of quantum computers, and determine the resource overheads and computational costs of such architectures.

A gap exists between work on single-core quantum chips and that on large-scale distributed quantum computing and the quantum Internet[5]. This article aims at proposing our vision of the potential of a communications-aware multi-core quantum computer architecture filling this gap. We present a double full-stack layered vision combining communications with single- chip quantum computer designs, summarize the first results we have obtained using design space exploration in our efforts devoted to characterizing the costs and needs of this approach, and outline the main challenges specific to these architectures.

CONNECTING QUANTUM CORES

Multi-core quantum architectures may recall the revolution of classical multi-core classical computing. However, while multi-core classical processors enabled the potential of parallelism and solved existing energy and thermal issues, multi-core (or multi-chip) quantum computing comes as a solution to correlated errors and control issues, which limit its potential even for small computations. Therefore, interconnecting quantum nodes, which is a challenge itself, is not a matter of upgrading quantum computers but unlocking their prospects of success.

Qubits are operated and read out using quantum gates which, as opposed to what is done in classical computing, are applied *in-place*.Moreover, interactions between qubits (e.g. performing a two-qubit gate) can only occur when they are physically adjacent. Therefore, if we need two qubits to interact, we have to place them close together so that we can apply the needed operation between them. Inside a quantum computing node, these moves are performed usually by means of swapping gates, i.e. quantum states are exchanged among qubits by applying a chain of SWAPs (see, e.g. Section VI of [2]).

However, transferring a quantum state among quantum chips is a complex task: it cannot be done using classical communications, and, due to the no-cloning theorem (i.e. an arbitrary unknown quantum state cannot be copied), qubit retransmissions are impossible. Even more importantly, communication latencies have to be as low as possible, to minimize the effect of the constant degradation on the qubit to be transmitted due to quantum decoherence.

Aiming at overcoming these obstacles, different quantum interconnects techniques that enable quantum state transfer are employed. The two most important are ion shuttling and quantum teleportation. Although they are still in a nascent stage, both have been demonstrated experimentally at different scales [6], [7], [5]. Ion shuttling is a technique to physically move qubits using electromagnetic fields in order to place together the ones that need to inter-operate. This technique is used in a specific implementation of qubits, ion traps. Using multiplexed architectures such as the quantum charge-coupled device (QCCD), some experiments have shown coherent shuttling of ion qubits through 2D junctions over millimetre distances in microsecond timescales. However, with today's technology, its latency and complexity do not scale well for more than ~ 100 qubits, and optical interconnects are needed to scale to larger platforms [4]. For a deeper look into the state of the art of this technology, see e.g. [6]).

Qubit teleportation, on the other hand, uses a pair of entangled photons and a classical channel to transfer quantum information without having

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MM.2021.3092706, IEEE Micro

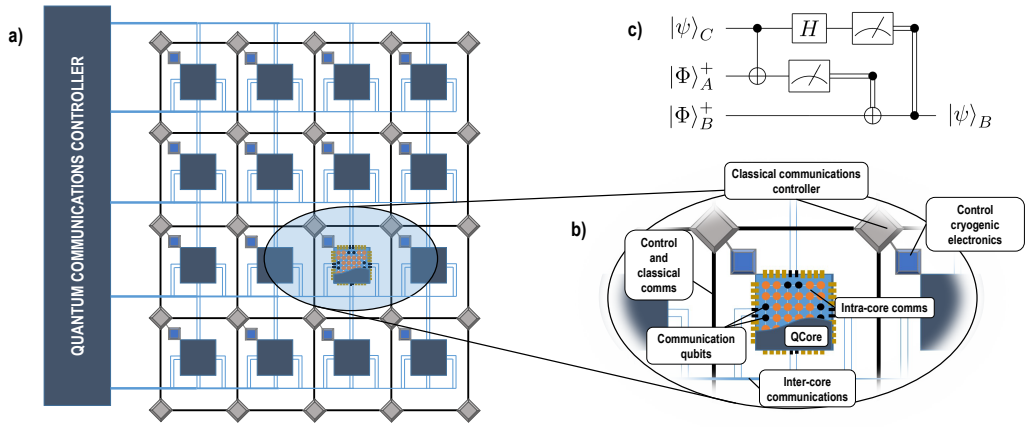

Figure 1. Multi- chip quantum computer full view. a) 2D diagram of a multi- chip architecture. Assuming quantum teleportation, the block in the left (the quantum communication controller) would also perform the distribution of entangled pairs among cores. The classical network also depicted completes the networking infrastructure. **b)** Enumeration of the components, including intra- and inter-core communications. **c)** Circuit for quantum teleportation.

to physically move the qubit. For that, both transceiver and receiver are sent one qubit out of a pair that shares an entangled state (usually implemented with photons transferred via optical fibers). Then, some basic operations between the qubit to be transmitted and the half of the entangled pair are applied, followed by a measurement. The result (a binary value) is then sent via a classical channel that connects both parties. With that information, the reception side can reconstruct the original transmitted quantum state by applying some corrections if needed. See in Fig. 1 c) the quantum circuit associated to this operation.

A key role here is performed by the light-to-matter transducers that transfer the quantum state contained in the superconducting/ion/spin qubit to the photon and viceversa. Deterministic and active light-to-matter teleportation has been realized with light pulses for atomic ensembles, photonic qubits and cavity quantum electrodynamics for various types of trapped ions, and quantum dots for solid-state systems, with increasing fidelities up to $\sim 90\%$ (although for small scenarios, with 2-3 qubits) [8].

The most notable advantages of this technique are the distance-independent latency, the decoupling of transfer into two different channels (entangled qubit pair and classical) for better protection of quantum data, and the compatibility with different qubit technologies. On the other hand, the efficiency of entangled pair distribution and network integration pose hard challenges. The interested reader can read further on the topic in [8].

In a multi-core environment (see Fig. 1 a) and b) for a full view of an example of multi-chip architecture), the described complexity of quantum communications is inserted into the already constrained quantum computation environment. That is the main reason to couple the design of computing and communications. Incurring in longer latencies, buffering waiting times or data losses, something which may be easily overcome in classical multi-core computing, may be crucial for the quantum case.

ENTANGLING COMPUTATION AND COMMUNICATIONS

Layered stacks are a powerful tool to tackle complex systems, in computing systems and communications. This type of hierarchical conceptualization has already been used in some existing proposals of layered architectures for quantum computing [9]. However, all of them focus on single- chip quantum computers, lacking a communications perspective.

We introduce a general-purpose (i.e no specific qubit or interconnect technology is assumed) layered stack specific to multi- chip

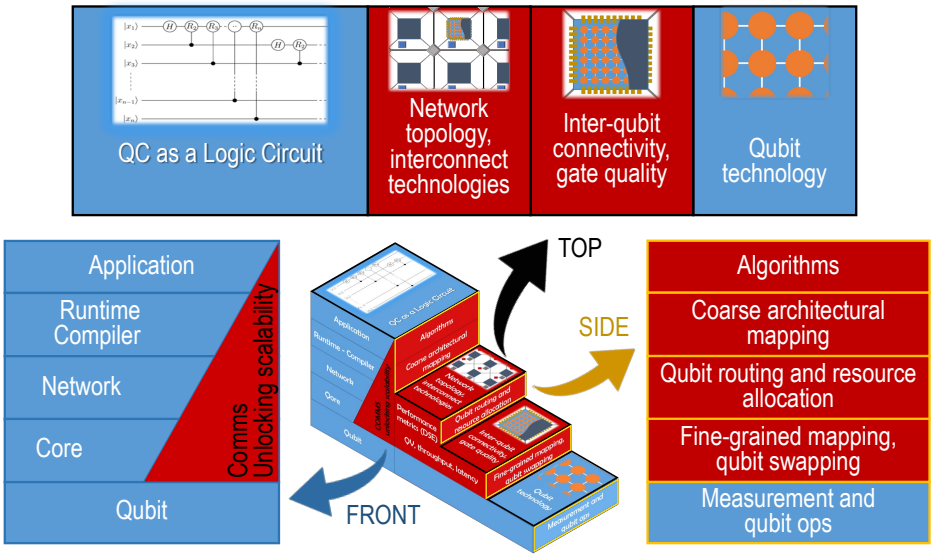

Figure 2. A double full-stack multi- chip quantum computer vision. The different abstractions of the quantum computer at each of the layers are included in the *stairway*: the step treads correspond to elements that configure that specific layer and the step risers its key functions.

quantum computing. We call it a *double full-stack* as it merges the traditional computing stack (application, runtime/compiler, micro-architecture, hardware) with communication stack (routing qubits among cores, qubit reservation and swapping, etc): quantum data transfers in multi-core quantum computers affect all the way from the code to the most basic two-qubit gates operations performed locally at a core.

Although there exist some stack proposals extending quantum computers to connected environments, these approaches come from a Quantum Internet perspective, i.e. do not integrate the quantum computation process with communications. They are network stacks rather than computer architecture stacks [10].

The full-stack layered architecture vision for multi- chip quantum computers that we propose is presented in Fig. 2. The whole network layer and the elements included in the red "wedge" correspond to the multi-chip implementation-specific kernel of the stack.

In the following, we will briefly explain each of the layers, focusing on the role of communications (and thus existing challenges for the realization of multi-core architectures) in each of them.

Qubit layer

This layer is the foundation of the quantum computer, composed by each one of the qubits that can be individually controlled and read out. Decoherence processes, together with measurement and gate performance, are the main aspects here (see Section IV in [2]). They are highly dependent upon the qubit technology (e.g. ion traps, superconducting qubits or quantum dots). and its maturity stage (see, e.g. Section 5 of [1]).

The qubit layer is not directly related to any communication process. However, it imposes some limits on latencies and qubit transfer rates of upper layers communication processes. Particularly, the coherence time (τ_c) sets a fundamental limit on the maximum time we can operate, read out the state, or transfer the qubit before the quantum information (see corresponding subsection in Section VI of [2]) is degraded irremediably due to decoherence.

Core layer

It performs the fine-grained qubit mapping inside the core as well as inter-core I/O operations control. The core layer's view is reduced to a set of qubits integrated into a single core capable of inter-operate using one and two-qubit gates. If the core is integrated into a multi-core archi-

tecture, some of its qubits will be responsible for interconnecting the core with one or several cores, acting as transducers or communication ports (see e.g. *Linking atomic qubits with photons* subsection in [4]).

Communications play also here a remarkable role, as two-qubit operations inside a quantum core are usually constrained to contiguous locations (see e.g. *Qubit plan organisation* subsection in Section VI of [2]). Therefore, qubit movement or swapping –the most basic form of quantum communication– is a constant for almost every computation. Also, in the multi- chip case, the core receives and sends quantum states from and to other cores. Gate quality metrics are hence key for communications performance. First, gate latency: the time spent in performing a certain quantum operation (such as a SWAP gate for intra-core communication). And second, gate fidelity, which represents the accuracy of a given quantum operation. Long gate latencies and low gate fidelities will affect the time and number of transfers a qubit may be able to support before losing the quantum information it stores.

Note also that the performance of this communication process will be affected by the qubit interconnection topology, the number of qubits per core, and the inter-qubit spacing e.g. a large processor with an uneven topology may need on average longer travels.

Network layer

This layer is fully responsible for interconnecting cores, implementing the qubit-to-core mapping defined by the upper layer and optimizing inter-core communication (i.e. involving qubit transfers among cores).

The specific inter-core topologies and interconnect technologies (e.g. ion shuttling, qubit teleportation...) that define the connection among cores are key. These will determine the *inter-core connectivity* in terms of core-to-core distances, inter-core communication latencies and qubit transfer rates, along with other technology-specific parameters such as e.g. number and output fidelity of EPR generators for qubit teleportation, or trapped voltage and segment size for ion shuttling [8], [6].

Communications are crucial at this layer, as they are ubiquitous in every action performed at this level, from qubit routing to remote gates. That implies particularly resource reservation protocols and network scheduling.

Runtime/Compiler Layer

It is in charge of compiling the code to quantum assembly and coordinate the execution of the instructions together with the coarse architectural mapping (i.e. partitioning of the algorithm among the existing cores, in analogy with the *mapping* process in classical many-core computer architectures), always in pursuit of optimized processing. At this layer, we *see* the quantum computer as a set of connected quantum cores (i.e. "processing units").

Inter-core communications, as well as some details on the capabilities and topology of the multi- chip platform, are implied in the coarse mapping process. However, qubit transfers are not directly controlled by this layer.

Application Layer

The upper-most layer corresponds to the code description of the quantum algorithm to be run on the quantum computer. This layer is hardware agnostic, meaning that low-level architectural details or constraints are not considered. In any case, the code might include some compiler directives enabling optimized qubit distribution and instructions execution, as it is already done in multi-core classical computing.

ON MULTI-CORE SCALABILITY PERFORMANCE AND COMMUNICATION BOUNDS

Having quickly reviewed how communications are deeply entangled into a multi-chip quantum computation platform, we should also design these architectures in a communications-conscious way. Being the underlying technology still in its infancy, we are now at an early design stage, where thorough analysis of these systems are key to demonstrate their theoretical potential and define design guidelines and bounds on the resource overheads and computational costs that are assumed.

Exploring such a complex and unknown space requires powerful tools and well-defined aims. That is why we have chosen Design Space Exploration (DSE), which is a fit candidate for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MM.2021.3092706, IEEE Micro

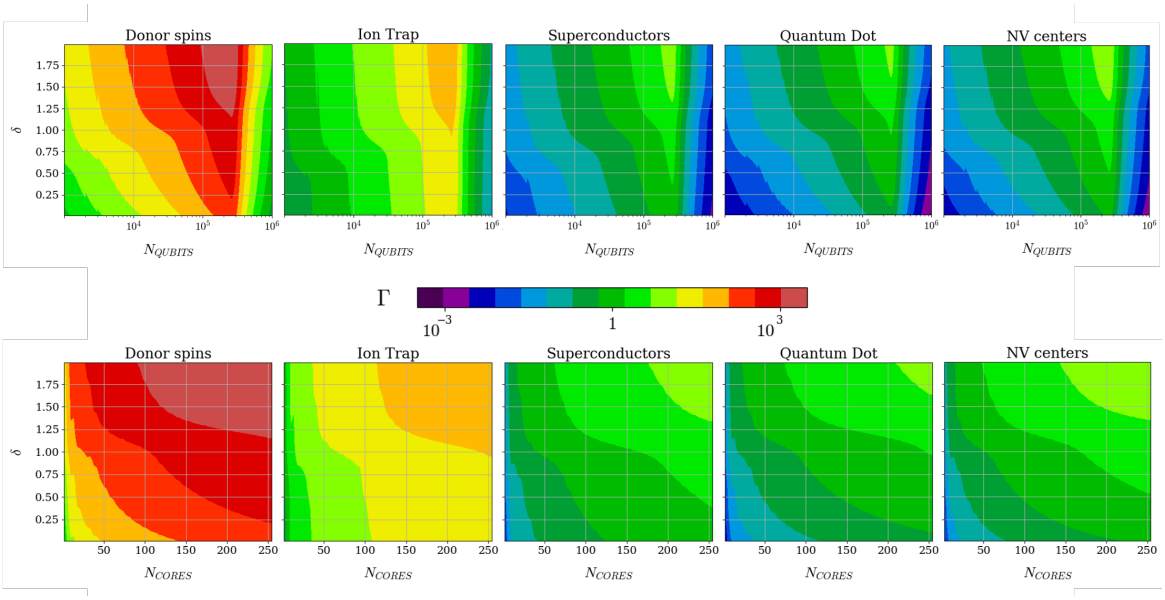

Figure 3. Quantitative qubit technology gap analysis Quantum computer's performance is plotted on a wide range of δ (technology expected improvement, a multiplicative factor applied to each of the components of Γ computed as a linear prediction of current state of the art parameters and their evolution over the past years), for existing qubit technologies. Being Γ an aggregate metric, its absolute values are not as important as its relative trends. The steep drop off in top figures occurs as the number of qubits integrated in a single chip reaches the limit, thus aggravating the overall performance. Equivalent performance can be obtained with a notably lower number of qubits if the technology is improved.

unknown design spaces without prior knowledge and for extracting design trends and guidelines. The double full-stack layered architecture just presented helps us as a simplified architecture model for the purpose of DSE.

In this exploration, we have aimed at answering some key questions: will the multi-core approach unlock the current monolithic single-core quantum computers' scalability bottlenecks? How do the existing qubit technologies compare as candidates for multi-core quantum computing? Our first results, summarized below, are promising.

First, we developed a lightweight analytical model adequate for our aim. For a specific scenario and a set of requirements (core-to-core communication latency, gate fidelity upper-bound, coherence time τ_c range, etc.) we studied the evolution of the design performance while varying architecture configurations and sizes. The aggregated performance metric (Γ) is a simplified behavioral model based on a weighted product capturing several computation and communications key elements, in normalized figures: number of qubits (J_{Qb}), qubit coherence time relative to the mean gate latency (J_{QF}), gate fidelity (J_F), qubit integration limitations(J_I, which includes cross-talk and other physical impairments), and core-to-core communications overhead (J_C). We refer the reader to [11] for further details. More specifically:

$$\Gamma = \frac{w_{Qb}J_{Qb} \cdot w_{QF}J_{QF}}{w_F J_F \cdot w_I J_I \cdot w_C J_C} \quad (1)$$

where $w_i \in (0, 1]$ correspond to the weights applied to each metric (in the presented results all of them are set to 1).

Using the same model we could also perform the first quantitative technology gap analysis (see some results reproduced in Fig. 3), i.e. a performance comparison of the existing qubit technologies and their evolution in the next years. This analysis opens a window to the future, letting us know which technologies may provide higher return after a certain research investment. Under the used assumptions and models, we could al-

ready draw some conclusions on the comparison of existing technologies, e.g. donor spins in Si seem to be the best performing technology, with still much room for growth.

Taking a step further, we performed a more realistic modeling using well-known quantum algorithms as benchmarks, as well as more accurate accountability of communications overhead. We also designed a multi- chip-specific quantum compiler, based upon Qmap and OpenQL [12]. This let us compare the actual overhead of executing a given algorithm on different quantum computers, with varying numbers and sizes of cores. Using QFT and Grover's search, as well as randomly generated quantum circuits, we tested a large number of different configurations (from 1 to 16 cores and a number of qubits per core ranging from 16 to 1024), being able to determine the "break-even point" for a key multi-core technology parameter: the intra-core communications latency (i.e. the value for which every multi-core configuration's performance supersedes that of an equivalent single-core quantum computer).

These results are reproduced in Fig. 4. The performance metric used for these experiments (Γ') is now derived from simulated measurements from the compiled code:

$$\Gamma' := \frac{\#\text{ gates} \times \#\text{ qubits required}}{\left(\frac{\text{Latency (multi-chip)}}{\text{Latency (single-chip)}}\right)} \quad (2)$$

In the left-hand side plots, the "break-even points" for inter-core communications latency are represented as the crossing point of the single-core performance line (which is flat, as it does not depend on the inter-core communication latency) with the performance curves of different architectures, using four different benchmarks. In the right plot, the "break-even" curves comparison for different number of cores and benchmarks is shown.

MULTI-CORE OPEN CHALLENGES

The possibility of having thousands of qubits working in the same quantum computer while being able to perform per-qubit control and maintaining high qubit isolation as we separate them into clusters of reduced size are clear advantages for multi- chip quantum computing success. However, the challenges that involve this approach are not negligible, and shall require bringing together expertise from researchers working at the bottom-most layer (qubit technology candidates, cryogenic control circuitry), classical communications and Network-on-Chip experts, and quantum computer architects.

Quantum core Input/Output ports

Each of the cores composing a multi-core device is not a reduced size quantum standalone chip only and it needs communication ports for collaborating with other cores acting as the interface between the core and the network layers. Because of the particularities of quantum computing, these ports must be qubits themselves (i.e. capable of storing quantum states). Moreover, they may have to act also as matter-to-photon transceivers, as most probably the inter-core communication will employ photons in optical waveguides, such as in the case of quantum teleportation [4], [8]. The challenge is finding the right combination of qubit technology and quantum state transfer, knowing that they need to be compatible: current research is working on different solutions, ranging from light pulses for atomic ensembles, to quantum dots for solid-state systems [8].

Standardize inter-core communication technology

Communicating quantum information among two separate quantum chips has already been experimentally demonstrated [7]. However, the different available technologies (i.e. ion shuttling, qubit teleportation) are still in their infancy and qubit technology-dependent. The challenge is enabling chip-to-chip communications satisfying upfront highly demanding requirements: apart from analyzing which technology should be used, the communication latencies, overall fidelity, and qubit rates need further improvement in order to reduce communication overhead. Indicative bounds such as the ones obtained in our previous works may help in this task.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MM.2021.3092706, IEEE Micro

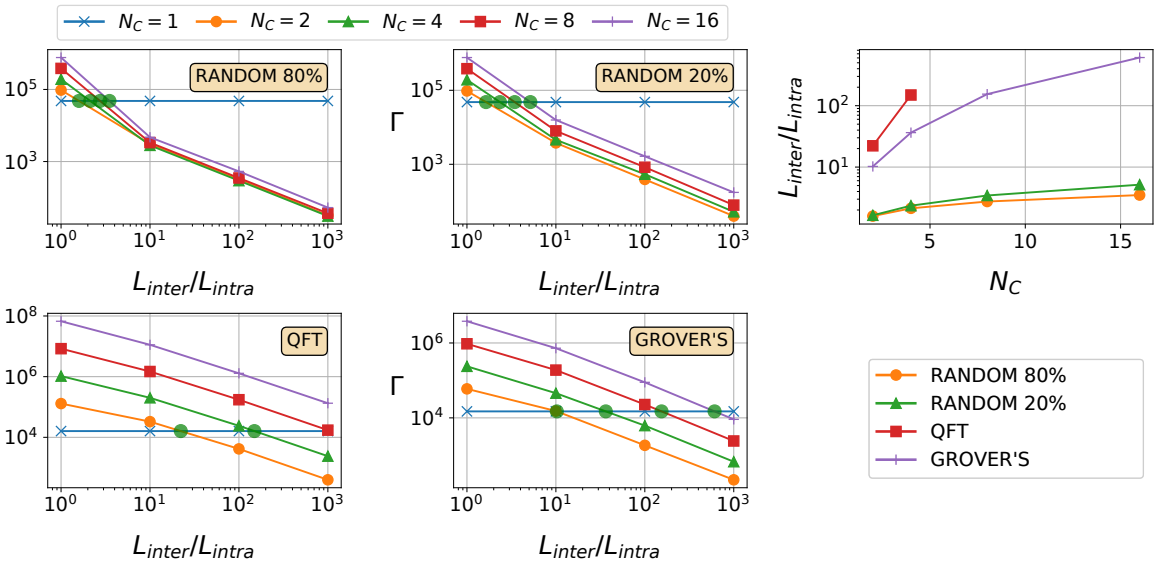

Figure 4. Inter-core latency upper bounds for several architecture configurations in number of cores (N_C) and number of qubits per core (N_Q^C). The benchmarks used are Grover's main routine, QFT, and two random quantum algorithm with both high (80%) and low (20%) interaction among qubits: percentage of gates that are two-qubit gates. L_{inter}/L_{intra} is the relative overhead of inter-core communications latency. The evaluation points lacking in the plot correspond to compilations that took an excessive amount of time to complete

Combined intra- and inter-core communication model

The complexity of quantum information transfer makes it challenging for researchers in the field of classical communications to contribute with their expertise. However, there is an urgent need for a standard communication model of both intra- and inter-core communications, that may facilitate the adaptation of existing solutions in classical comms to the quantum equivalent problems: networking protocols, resource reservation, routing algorithms, buffer management, queueing models, etc. Describing a series of quantum SWAP gates or quantum teleportation, and the related error, with these high-level concepts, enables complex analysis and powerful solutions, as it is already being doing for Quantum Internet [5].

Quantum-core-specific communication protocols

The advances on the previous challenge will facilitate the design of protocols that may benefit from the special characteristics of quantum communications, and most importantly, aim at minimizing the most limiting factor in the quantum computing world, i.e. quantum decoherence . Simple but sophisticated core I/O protocols may control the inbound and outbound traffic to avoid information losses (which in quantum communications cannot be recovered, by the no-cloning theorem) due to congestion, resource reservation protocols may help EPR distribution arrive just-in-time for scheduled communications, etc.

Quantum compilers for multi-chip devices

Mapping the qubits and scheduling gates in a single-core quantum computing is already a complex optimization task, due to the strict limitations that quantum decoherence and gate fidelity impose on expected output error rates. The case of multi-core quantum computers adds the extra complication of having to map the algorithm (with hundreds or thousands of qubits) onto separated cores, interconnected with a costly and limited shared network. Developing compilers (and hence mappers) for such architectures is an unnegotiable challenge for optimizing operations and communication overheads, and therefore for the ultimate success of this approach.

CONCLUSIONS

The presented vision for quantum computers, entangling computing and communications in a multi- chip approach, allows multi-disciplinary expertise to help to unlock the scalability issue of quantum computers. The challenges ahead are fascinating, and tackling them will help us to know better the inner workings of quantum computing. Particularly important is the role of time in multi-core quantum communications. In classical communications, long latencies may not be a big issue, and sometimes waiting time pays off in communication overall quality. However, time in quantum communications is directly translated into quantum decoherence, and thus into error. Therefore, the ability to reducing communications overhead will determine the success of large-scale quantum computing.

REFERENCES

1. M. Martonosi and M. Roetteler, "Next steps in quantum computing: Computer science's role," 2019.
2. C. G. Almudever *et al.*, "The engineering challenges in quantum computing," in *Proceedings of the DATE'17*. IEEE, 2017, pp. 836–845.
3. J. Preskill, "Quantum computing in the nisq era and beyond," *Quantum*, vol. 2, p. 79, 2018.
4. K. R. Brown *et al.*, "Co-designing a scalable quantum computer with trapped atomic ions," *npj Quantum Information*, vol. 2, no. 1, 2016.
5. S. Wehner *et al.*, "Quantum internet: A vision for the road ahead," *Science*, vol. 362, no. 6412, p. eaam9288, 2018.
6. V. Kaushal *et al.*, "Shuttling-based trapped-ion quantum information processing," *AVS Quantum Science*, vol. 2, no. 1, p. 014101, 2020.
7. D. Llewellyn *et al.*, "Chip-to-chip quantum teleportation and multi-photon entanglement in silicon," *Nature Physics*, vol. 16, no. 2, pp. 148–153, 2020.
8. S. Pirandola *et al.*, "Advances in quantum teleportation," *Nature photonics*, vol. 9, no. 10, pp. 641–652, 2015.
9. X. Fu *et al.*, "An experimental microarchitecture for a superconducting quantum processor," in *Proceedings of the MICRO-50*, 2017, pp. 813–825.
10. A. Pirker and W. Dür, "A quantum network stack and protocols for reliable entanglement-based networks," *New Journal of Physics*, vol. 21, no. 3, p. 033003, 2019.
11. S. Rodrigo, S. Abadal, E. Alarcón, and C. G. Almudever, "Exploring a double full-stack communications-enabled architecture for multi-core quantum computers," *arXiv preprint arXiv:2009.08186*, 2020.
12. N. Khammassi *et al.*, "Openql: A portable quantum programming framework for quantum accelerators," *arXiv preprint arXiv:2005.13283*, 2020. [Online]. Available: https://github.com/QE-Lab/OpenQL

Santiago Rodrigo is a PhD candidate at Universitat Politècnica de Catalunya. Santiago holds two Bachelor degrees in EE and CS and an M.Sc. from the same university. His research is focused on quantum communications and multi-core quantum architectures. Contact him at srodrigo@ac.upc.edu.

Sergi Abadal is distinguished researcher at Universitat Politècnica de Catalunya. His research interests include on-chip networks, quantum computing, and wireless computing systems. Contact him at abadal@ac.upc.edu.

Eduard Alarcón is a professor at Universitat Politècnica de Catalunya, with research interests in resource-constrained short-range wireless communications, spacecraft distributed architectures, AI chip co-processor architectures, quantum computing architectures, and structured design of complex systems. He has served in different positions at IEEE, and holds a PhD (2000) and EE engineering degree -national award- from UPC.

Medina Bandic is a PhD Student at QuTech, TU Delft. She holds Bachelor and Master studies in CS and has worked in the industry as a software engineer. Her interests include quantum computing, software development, machine learning and data science.

Hans van Someren is a researcher at the Quantum and Computer Engineering Department and QuTech, TU Delft. Until 2015 he was principal architect at ACE Associated Computer Experts. Currently his interests are in quantum computing, i.e. compilation, scheduling, optimization and mapping, architecture exploration and programming models. Contact him at j.vansomeren-1@tudelft.nl.

Carmen G. Almudéver is a distinguished researcher at Universitat Politècnica de València. Until 2021, she was with QuTech, TU Delft. Her research includes quantum programming languages and compilers, mapping of quantum algorithms, architecting and benchmarking of quantum computers, quantum error correction and fault-tolerant quantum computa-