Modelling Short-range Quantum Teleportation for Scalable Multi-Core Quantum Computing Architectures

Santiago Rodrigo
srodrigo@ac.upc.edu
Universitat Politècnica de Catalunya
Barcelona, Spain

Sergi Abadal
abadal@ac.upc.edu
Universitat Politècnica de Catalunya
Barcelona, Spain

Carmen G. Almudéver
cargara2@disca.upv.es
Universitat Politècnica de València
Valencia, Spain

Eduard Alarcón
eduard.alarcon@upc.edu
Universitat Politècnica de Catalunya
Barcelona, Spain

ABSTRACT

Multi-core quantum computing has been identified as a solution to the scalability problem of quantum computing. However, interconnecting quantum chips is not trivial, as quantum communications have their share of *quantum weirdness*: quantum decoherence and the no-cloning theorem makes transferring qubits a harsh challenge, where every extra nanosecond counts and retransmission is simply impossible. In this paper, we present our first steps towards thorough modeling of quantum communications for multi-core quantum computers, which may be considered as a middle point between the well-known paradigms of Quantum Internet and Network-on-Chip. In particular, we stress the deep entanglement that exists between latency and error rates in quantum computing, and how this affects the quantum network design for this scenario. Moreover, we show the concomitant trade-off between computation and communication resources for a set of parameters out of state-of-the-art experimental research. The observed behavior lets us foresee the potential of multi-core quantum architectures.

CCS CONCEPTS

• **Computer systems organization → Quantum computing**; **Distributed architectures**; • **Networks → Network on chip**.

KEYWORDS

Quantum Computing, Many-core Quantum Computers, Quantum Communications, Quantum Computers Scalability

The final publication is available at ACM via
http://dx.doi.org/10.1145/3477206.3477461

1 INTRODUCTION

Leveraging quantum mechanics for computing, once limited to theoretical developments and algorithms, quantum computing has started to become a reality during the last decade [14]. Unconventional properties of quantum computers allow them to solve complex problems exponentially faster than a classical supercomputer [1] and tackle classically intractable problems. Fields as important as internet security, pharmacology, complex combinatorial and optimization problem solving, big data analysis or AI could make a *quantum leap* when fully-fledged quantum computers become available.

However, extraordinary performance requires an extraordinary environment: quantum computing is highly sensitive to any type of external force. In order to preserve quantum information from corruption, quantum computers need to be kept at cryogenic temperatures and every qubit (the *alter ego* of classical bits in the quantum world) in the computer must be as isolated as possible.

These demanding requirements make building quantum computers a challenging task and compromise quantum computing scalability. In fact, the largest experimental quantum computers as of today do not exceed 100 qubits [1, 27], and monolithic single-chip approaches are expected to be limited to a few thousand qubits, due to the impracticality of integrating the required control circuits and per-qubit wiring and still maintaining a low quantum error rate [14]. Despite the impressive computational power of quantum computers, it is predicted that millions of qubits will be required in order to run practical quantum algorithms [18].

Connecting several quantum chips in a multi-core fashion has been proposed as a modular approach that may enable the scaling of quantum computers [23]. This not only simplifies control circuit requirements, but also reduces crosstalk errors and other impairments derived from a densely-packed group of qubits integrated in a single-chip quantum processor.

On the flip side, communicating quantum chips is far from being a simple task. Sharing quantum data across cores is hindered by a variety of issues such as that qubits cannot be copied due to the *no cloning* theorem or that qubits have a limited lifetime rendering communication extremely latency sensitive. Moreover, this quantum-coherent network needs to work in parallel with a chip-scale network transporting classical data, intended for assisting quantum transfers with control and synchronization messages.

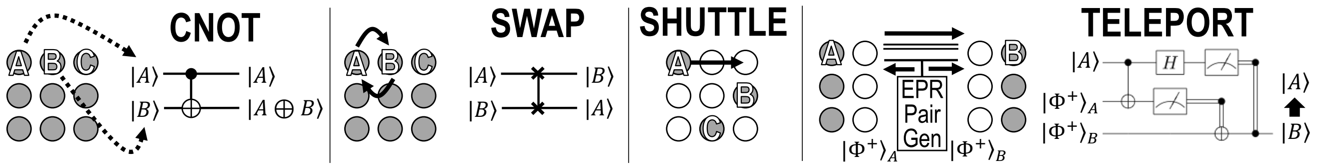


Figure 1: Fundamentals of quantum computing and communications. From left to right: controlled NOT gate between qubits A and B, swap between qubits A and B, shuttling of qubit A, and teleportation of qubit A to the position of distant qubit B (whose state becomes that of A after completing the teleportation) via an entangled pair $|\Phi^+\rangle$.

Although there are various works on this multi-core approach, for a diverse set of qubit technologies and interconnect architectures [13, 20, 23], a thorough analysis and modelling of this communication scenario, being much needed in order to understand better the interplay between computation and communications performance, is still lacking. Expertise in analogous scenarios such as Network-on-Chip (NoC), which concerns the internal communication needs of modern computer architectures [21], and the more recent Quantum Internet (QI), which refers to large-scale networks for the exchange of qubits over long distances [24], constitute strong foundations for this undertaking. Nevertheless, the distinctive characteristics of quantum communications in multi-core systems and their criticality for computation performance call for a deeper analysis of this communications context.

In this paper, we present our first steps towards this modelling of quantum communications for multi-core quantum computers. In particular, we have shown via simulation with values and models from state-of-the-art experimental research, how communication latencies and waiting times affect quantum communication quality, and explored the optimal share of qubits among the computation and communication roles for multi-core quantum architectures for a random algorithm. To these ends, we first provide a short tutorial on quantum computing and communications in Sec. 2. We then analyze the main similarities and differences between the quantum multi-core context and the NoC and QI scenarios in Sec. 3. This analysis drives the modeling and simulation of quantum communications at the chip scale in Sec. 4 and 5, respectively. We conclude the paper in Sec. 6.

2 BACKGROUND

Basic notions of quantum computing are needed in order to fully understand the implications of quantum communications performance, as well as to identify correctly the particularities that define quantum networks. For a deeper look into quantum computing and communications, the interested reader may refer to [2, 14].

2.1 Computing with qubits

The qubit constitutes the basic unit of computation in the quantum world. The quantum information contained in a qubit (a quantum state) can take, as in the classical world, the logical values of 0 and 1. These are usually represented as $|0\rangle$ and $|1\rangle$, also called *ket notation.* However, by virtue of the *quantum superposition*, the actual quantum state is described as a linear combination of both,

$$|\phi\rangle = v_0 |0\rangle + v_1 |1\rangle, \qquad (1)$$

where $v_0, v_1 \in \mathbb{C}$ and $|v_0|^2 + |v_1|^2 = 1$.

Extracting the information from a qubit at the macro world means measuring its physical quantum state. Due to quantum mechanics' postulates, this gives us only a partial view of it, as the $|\phi\rangle$ state collapses into the measurement basis, e.g. either $|0\rangle$ or $|1\rangle$. In other words, the measurement leads to either $|0\rangle$ with probability $|v_0|^2$ or to $|1\rangle$ with probability $|v_1|^2$, in a process that *destroys* the quantum state of the qubit. With two qubits A and B, the quantum state before measurement is the superposition of four possible values v_{00}, v_{01}, v_{10} and v_{11} corresponding to the relative probabilities of the qubits taking the $|0\rangle$ or $|1\rangle$ states after measurement. This can be generalized to any number of qubits.

The power of quantum computing comes from the operation of qubits and their probabilistic states. Quantum logic circuits are capable of altering the quantum state of qubits either with single-qubit gates, which affect the values v_0 and v_1, or gates that combine the quantum state of two or more qubits. A controlled NOT (CNOT), shown in Fig. 1, is a clear example of a two-qubit gate: a NOT is applied to a qubit B only when the control qubit A is $|1\rangle$, thus modifying the values of v_{00}, v_{01}, v_{10} and v_{11} accordingly. Quantum algorithms can be therefore described as *quantum circuits*, i.e. a sequence of quantum gates applied to the qubits in the computer.

These powerful properties are the foundations of quantum computing, but also imply some difficulties. In particular, the *no-cloning theorem* states that it is not possible to create an identical copy of any given quantum state. As a consequence, qubits are not only an abstract unity of information, but also the physical entity containing the information: if the qubit is measured or corrupted, the quantum state is lost. This is an issue because qubits nowadays are noisy and prone to *quantum decoherence*, which arises from the interactions of a qubit with the surroundings and leads, over time, to unwanted modifications of the qubit's state.

In order to improve their isolation and minimize decoherence, qubits are operated and measured in-place. However, when two qubits are required to interact by means of a two-qubit gate, their quantum state needs to be moved to adjacent positions in the computer. For instance, if we want to apply a CNOT gate between qubits A and C of Fig. 1, we will have to exchange the position of qubits B and A or C. This movement, which we refer to qubit communication, is a key operation in quantum computing and its efficiency in space and time is crucial for the whole quantum computer performance.

2.2 Communicating qubits

Quantum communication refers to the transmission of quantum state from one place to another. This can take place via the physical movement of the qubit or by the transfer of its quantum state. Here,

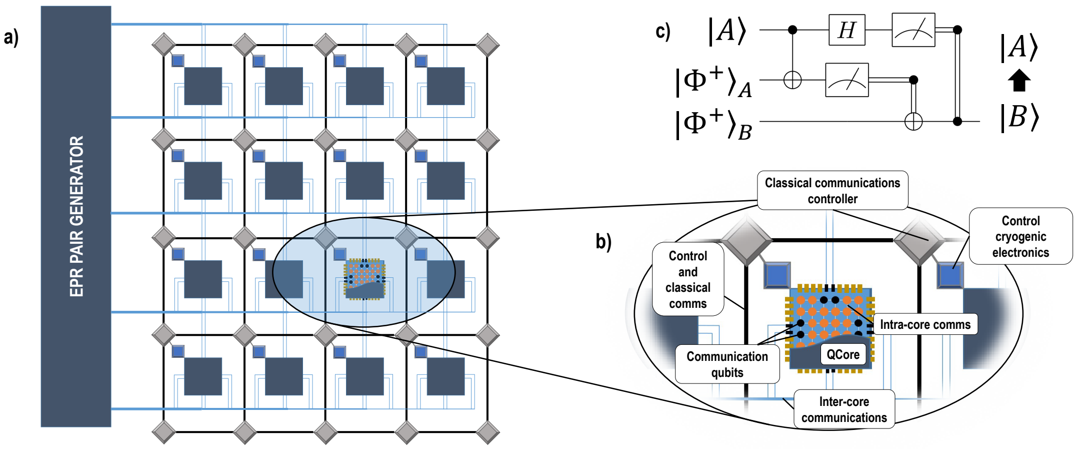


Figure 2: Multi-chip quantum computer full view. a) 2D diagram of a multi-chip architecture. The classical network also depicted completes the networking infrastructure. b) Enumeration of the components, including intra- and inter-core communications. c) Circuit for quantum teleportation.

we describe three methods depicted in Figure 1: SWAP gates, qubit shuttling, and quantum teleportation.

SWAP gates. The most basic form of communication in quantum computers is the SWAP gate. A SWAP gate can only be applied to two physically adjacent qubits, which exchange their state to one another. Thus, to move a qubit state to an arbitrary position, a chain of SWAPs can be applied. However, this implies interacting with every qubit along the way.

Qubit shuttling. Another form of communication at the single-chip quantum computer level, and specific to the ion trap qubit technology, is qubit shuttling. In this case, electromagnetic fields are used to move the qubit physically across a chip space intentionally left devoid of qubits (shown as blank positions in Fig. 1).

Qubit teleportation. A more versatile yet indirect quantum communication technique is quantum teleportation. This technique exploits the property of *quantum entanglement*, which refers to the ability of having two or more qubits containing states that cannot be described independently of each other. For two qubits X and Y, being in an entangled state $|\Phi\rangle^+$ is defined as

$$|\Phi^+\rangle = \frac{1}{\sqrt{2}}\Big(|0\rangle_X \otimes |0\rangle_Y + |1\rangle_X \otimes |1\rangle_Y \Big) = \frac{|00\rangle + |11\rangle}{\sqrt{2}}, \quad (2)$$

implying that, when measured, both qubits collapse to the same state, either $|0\rangle$ or $|1\rangle$. That is, if we measure one qubit, we can be completely sure about the other qubit's state no matter how far apart they are placed. Therefore, qubit teleportation is applicable for communication at any distance, from the chip to planetary scales.

Qubit teleportation uses a pair of entangled photons, also called EPR pair [9], and a classical channel to transfer the quantum information of a qubit without moving it physically. For that, as shown in Fig. 1, both transmitter A and receiver B are sent one qubit out of a pair that shares an entangled state, which we name $|\Phi\rangle^+_A$ and $|\Phi\rangle^+_B$. These are completely independent from the state $|A\rangle$ to be transferred. Then, some basic operations involving the qubit to be transmitted and the entangled qubit are applied, followed by a measurement. The result (a binary value) is then sent via a classical channel. With that information, the receiver can reconstruct the original transmitted quantum state by applying some corrections, turning $|B\rangle$ into $|A\rangle$. Note that by being measured, the original state of qubit A is lost and hence the no-cloning theorem is respected.

2.3 Quantum networks

In the context of quantum technologies, the challenges of quantum communications are several. First, a dedicated physical infrastructure is required to realize the quantum channels. For instance, EPR generators and optical fibers capable of transmitting entangled photons are required to perform quantum teleportation. A second challenge, posed by the no-cloning theorem, is the need to minimize the noise as qubit retransmissions are not possible. Hence, quantum networks are extremely latency-sensitive, since qubits tend to decohere as time passes, which clearly affects protocol design.

These challenges are being addressed in large-scale networks with solutions that already reached industry. Quantum cryptography, and more specifically Quantum Key Distribution (QKD), provides an unconditionally secure way to encrypt communication [25]. In this case, a string of qubits is directly transmitted using photons in dedicated optical networks and used to produce the random secret keys for the secure communication. Thanks to the no-cloning theorem, this key distribution protocol is able to detect an eavesdropper in the channel. Such a process has been deployed at city-wide scale [16] and even using free-space optics with a satellite as the secure middle-point producing the qubits [26].

Along these lines, the concept of *Quantum Internet* [10] has been conceived as large-scale quantum network capable of both working in parallel with the classical internet to enhance its applications (starting with QKD) [24], as well as of interconnecting distant quantum computers to distribute computation [4]. This has opened a wide and fertile research field, where the networking challenges are being tackled. In particular, the need for quantum

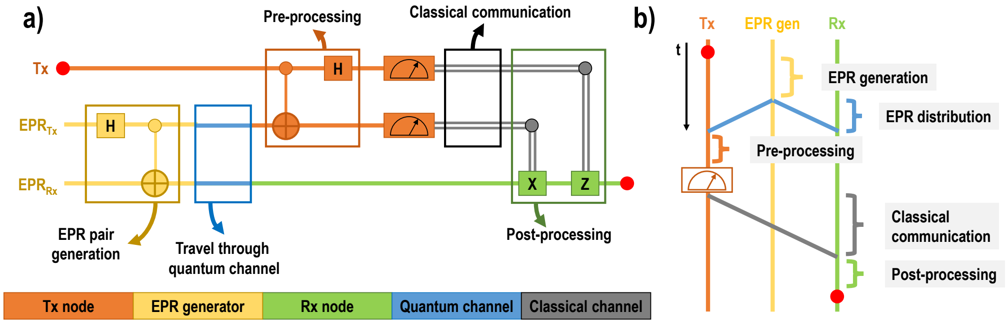


Figure 3: Teleportation circuit and time sequence diagrams, color-coded by the locations where each phase is performed.

repeaters to allow long-distance quantum teleportation [8], quantum communications protocol stack design and implementation [6, 17], and modeling of communications at this scale [2] are being investigated.

3 BETWEEN THE QUANTUM INTERNET AND NETWORK-ON-CHIP

While recent years have seen an explosion of research in quantum communications and networking for the QI, less attention has been placed on chip-scale communications for the scaling of quantum computers. Here, we provide a first context analysis of such scenario, taking base on the similarities and differences with two analogous applications. On the one hand, for its role as an interconnect among computing cores within or across chips, quantum interconnects may recall the classical concept of NoCs. On the other hand, for its quantum nature and the need to transfer quantum data using mainly quantum teleportation, it may be compared with its *big brother*, the QI.

3.1 Comparison with the Quantum Internet

Research on the QI is one step ahead of multi-core quantum computers and its main focus is on quantum teleportation. which is also a strong candidate for the multi-core scenario. Hence, the QI could seem as a good source of inspiration for models and protocols of quantum multi-cores, yet with the following non-trivial differences.

Distance: In quantum communications, quantum states get decohered with time and interaction with the environment. This means that long fiber lines cause photon qubits to irremediably lose its information. Usually, this is fixed using quantum teleportation instead of physically sending the qubits. This allows to distribute the entangled pairs and, if needed, amplify them with quantum repeaters [8, 24]. In the multi-core scenario, however, distances are in the millimeter scale, and hence there is no need (and no desire, due to stringent resource constraints) to deploy repeaters. This may simplify the link and network layer protocols.

Temperature: Going into the outside world along a country-wide inter-network makes impossible to keep low temperatures to facilitate quantum coherence. However, inter-chip networks in multi-core architectures can be kept at cryogenic temperatures, thus increasing the coherence time for communicating qubits. Thus, unlike in the QI, all the circuitry used to control and communicate the qubits in quantum computers must be prepared for operating at cryogenic temperatures. Moreover, given the limited cooling power of refrigerators, protocols need to be simplified to reduce the energy footprint. Such a constraint is not present in the QI.

Qubits as a scarce resource: Multi-core quantum computers are a resource-constrained environment, both in area and in power consumption, as opposed to the QI. On the one hand, every qubit that is sent to another core is unique: due to the no-cloning theorem, the only way to reproduce a qubit state is to repeat the computation all over again. In the QI scenario, we may have sufficient quantum computational power to recalculate some small parts of the computation, and we are dealing with larger amount of qubits, but in multi-core architectures, every qubit is unvaluable and hence all qubit transfers must be operated with the highest of guarantees and ultra-low latency.

Communications overhead: Given the resource-constrained nature of this environment, the number of qubits that we can integrate on each chip is limited. In the QI, in order to protect qubits from decoherence, Quantum Error Correction (QEC) techniques are applied. This implies that multiple physical qubits encode a logical qubit, which implies a very significant overhead. In multi-core quantum computers, qubits are scarce and, therefore, we may face the trade-off between dedicating qubits to QEC or to computation, having to sacrifice some protection along the way.

3.2 Comparison with Network-on-Chip

The NoC paradigm accompanied the rise of multi-core processors in the classical domain. Even though its divide-and-conquer approach is similar to that proposed in multi-chip quantum computing, and while some of its design principles might be applicable also here, interconnecting quantum nodes has some particularities, namely:

Latency as an error: In NoC, latency is critical because it essentially delays the computation [21], but it is generally not tied to loss of accuracy. In contrast, qubits tend to decohere as time passes, which implies that the communication latency not only delays the computation in multi-core quantum computers, but also degrades its accuracy, to the point of completely disrupting it when a particular latency threshold is exceeded. This fundamentally impacts flow and congestion control protocols, if any.

Uniqueness of data: Due to the no-cloning theorem, data cannot be copied and distributed to multiple cores. Data is physically moved around, and therefore scheduling the communication operations is of critical importance to minimize qubit movements and consolidate interactions with a given qubit in the minimum amount of time. NoCs are generally not bound to scheduling, although efforts in real-time embedded systems or machine learning accelerators also advocate for it in the classical domain [15, 22]. In any case, this aspect is at the frontier between the network and the architecture,

Welcome back to circuit switching: Quantum teleportation uses both a classical channel and a quantum channel to transmit the information: the measurement output at the Tx node (2 bits), and the entangled photon qubit pair. While the classical channel works as expected (plain information travelling through a wire), the quantum channel is not used to transmit directly quantum information, but a quantum resource: entanglement. This resource is used thereafter to *teleport* the qubit. In this way, we could describe the entangled pair as a channel itself, being the EPR pair generator a shared resource. Similarly to what is done in circuit switching technologies, the two communicating parties are interconnecting by means of a channel (entanglement) reservation. This makes multi-core quantum networks depart from the traditionally packet-switched NoC paradigm: the topology among cores is virtually configured as needed, and there is a shared resource (the EPR pair generator, which may also be decentralized) that enables communication.

4 A COMMUNICATIONS-CENTRIC MODEL OF QUANTUM TELEPORTATION

Current research on quantum teleportation revolves around improving link quality and robustness, as well as on developing further supporting technology. In contrast, the importance of modelling quantum teleportation as a communications system has only recently been highlighted. This kind of models enable a deeper understanding of this technology from the communications standpoint, which facilitates and provides guidelines for protocols and system design [6, 12]. Models for quantum teleportation have been mostly developed for the QI scenario [3]. Because of the differences previously mentioned between multi-core quantum communications and QI, a specific modelling of the former is needed. This is of crucial importance because the performance of a multi-core quantum computer will depend on its communications.

Let us assume a quantum computer composed of N_C chips/cores that communicate through quantum teleportation (Fig. 2). Therefore, three elements are needed: a classical network connecting the cores, a generator of entangled pairs, and an optical network connecting every node with the generator. Each of the cores integrates N_Q^C physical qubits, and has N_P light-to-matter ports for the quantum interconnection allowing for the same number of parallel transfers. Each of the ports has a qubit buffer attached to it. The total number of qubits dedicated to buffer/communication tasks is N_Q^{COMM}, leaving $N_Q^{COMP} = N_Q^C - N_Q^{COMM}$ for computation.

The quantum teleportation process can be represented using the circuit diagram from Fig. 3a), where each color represents a different location in the network. Furthermore, Fig. 3b) shows the corresponding time diagram which breaks down the delays of the teleportation process. Latency is indeed crucial, as it is directly correlated to decoherence of qubits. In our model, a qubit transfer proceeds as follows: when a given qubit q holding a quantum state $|\phi\rangle$ has to be moved to a different core, the controller checks whether any of the N_P communication ports is free. If not, the qubit needs to wait in the communication buffer as long as there is any available position. Otherwise, we assume that the communication of that qubit is not possible due to excessive decoherence.

4.1 Qubit teleportation delay

When a communication port becomes available, the controller triggers the entangled pair generator in order to start the EPR pair distribution. The EPR pair generation takes a non-deterministic time with mean T_{EPR} [3], after which the entangled photons can be sent to the Tx and Rx nodes. We assume an ideal optical channel with no photon loss and with a fixed delay $T_{DIST} = d/c'$ related to the distance between the EPR generator and the nodes, assumed $d = 4mm$, and the speed of light in the optical medium $c' = 2 \cdot 10^8$. Usually, the EPR pair generator is a shared resource, hence T_{EPR} will depend also on its utilization by the other cores in the network.

On the Tx side, the pre-processing involving the entangled qubit and the qubit to be transferred takes a fixed delay T_{PRE}, typically composed by the delay of applying a CNOT gate, a Hadamard gate, and the measurement on both qubits. The resulting classical bits are sent to the Rx node, in a process assumed error-less and taking a fixed time T_{CLAS}. At Rx, these bits guide the modifications to be made to the received entangled qubit with Pauli X and Z gates. This postprocessing takes on average $T_{POST} = 1/2 \cdot (T_X + T_Z)$.

Therefore, a single qubit transfer will take T_{TX}, which corresponds to the critical path on the time diagram in Fig. 3,

$$T_{TX} = T_{EPR} + T_{DIST} + T_{PRE} + T_{CLAS} + T_{POST}. \tag{3}$$

4.2 Maximum qubit transfer rate

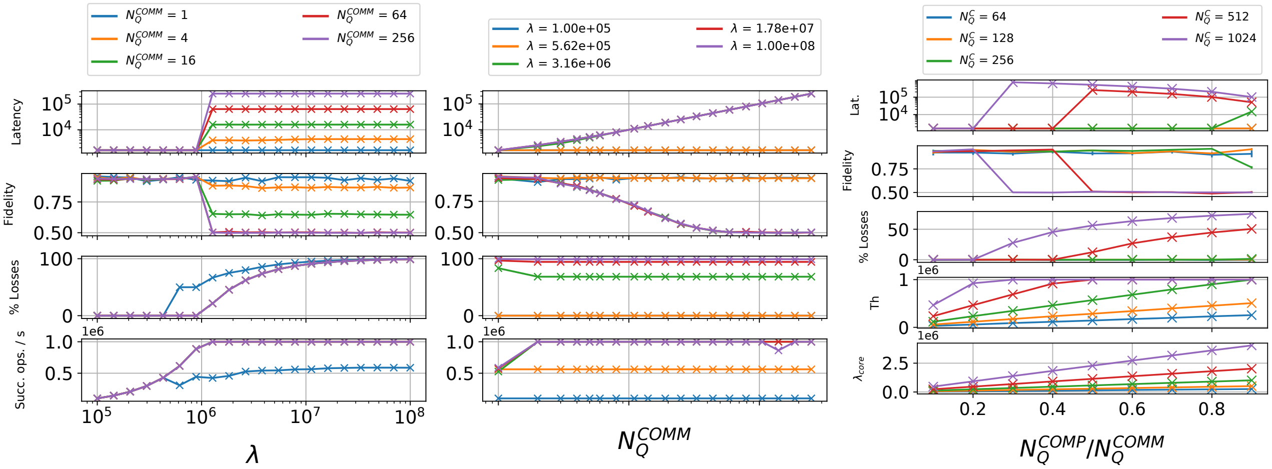
This communication process consists on various consecutive operations on a middle-node (the EPR generator), and both the Tx and Rx nodes. From a communications perspective, it is essential to compute the maximum qubit transfer rate, i.e. the maximum number of transfers that can be performed if the system is operated continuously. This will depend on the ratios among the different timing components in Eq. (3).

The EPR pair generation and distribution acts as the quantum channel. Thus, the rate at which such generation happens is the first and fundamental bottleneck, hence having a transfer rate upper-bounded by the time to generate an EPR, so that

$$r_{MAX} \leq \frac{1}{T_{EPR}}. \tag{4}$$

However, the equality will hold only when the pre- and post-processing do not become a bottleneck. For the case $N_P = 1$:

- **Pre-processing bottleneck:** Both Tx and EPR_{Tx} qubits have to be operated and measured before a new entangled pair is received, hence we will have no bottleneck iff $T_{PRE} < T_{EPR}$.
- **Post-processing bottleneck:** The Rx qubit has to wait for the classical bits and operate two single-qubit gates, hence we will have no bottleneck iff $T_{PRE} + T_{CLAS} + T_{POST} < T_{EPR}$.



(a)

(b)

Figure 4: Modeling communication VS computation trade-off in multi-core quantum architectures. a) Stress test of a quantum teleportation channel at multi-core scales. The 2D input design space formed is swapped both in traffic (λ) and size of the quantum buffer (N_Q^{COMM}) b) A multi-core system with 256 cores executing a random algorithm with 80% two-qubit gates is explored while changing the computation to communications qubits ratio.

Table 1: Notation, symbol definitions and values used.

Notation	Meaning	Value
N_C	Number of cores/chips	256
N_Q^C	Number of qubits per chip	1024
N_P	Number of connectors per chip	1
T_{EPR}	Mean of EPR pair generation time	10^3 ns
T_{DIST}	EPR pair distribution time	0.01 ns
T_{PRE}	Pre-processing time	390 ns
T_{CLAS}	Classical transfer time	0.02 ns
T_{POST}	Post-processing time	30 ns
T_{TX}	Total time of quantum transfer	1420 ns
r_{MAX}	Maximum qubit transfer rate	10^6 qbps

5 SIMULATION RESULTS

Using this communications perspective for multi-core quantum networks, we have performed a set of simulations in order to: *i)* validate the effects of delays and losses on the teleportation fidelity and error rate using experimental values, and *ii)* explore the behavior of a simulated multi-core environment executing a randomly distributed quantum circuit under varying ratios of communication and computation qubits.

For this task, we have used the NetSquid simulator for quantum networks [5], which has allowed us to simulate a realistic setting, with quantum memories and processors, EPR generators as well as classical and quantum links, taking into account qubit decoherence, quantum gates latencies, communication delay models, etc.

In the first set of simulations, we have modelled two independent quantum cores connected to an EPR generator and interconnected through a classical link. Apart from the processing and networking capabilities, we have implemented a quantum FIFO buffer for the qubit transfer, in order to explore the effects of long qubit waiting times on the fidelity of the qubit communication. We have also implemented the teleportation protocol (pre-processing and post-processing phases) on each node, adding the buffer management. Both quantum nodes and EPR generator work with parameters found in Table 1, taken from transmon superconducting qubit technology used in [11] and [7].

In order to stress the system for high communication rates, we have introduced a constant rate (λ) qubit arrival process at the Tx node (Alice). In Fig. 4a), an exploration over the λ input qubit rate and the buffer size at the Tx node is shown. From top to bottom, the successful number of teleportations per second, mean communication latency, mean fidelity, and the % qubit losses and are shown. See how the system saturates for $\lambda > 10^6$, as it is the maximum teleportation rate (which, as the system complies with the pre-processing and post-processing inequalities, is equal to the EPR generation rate), except for the case where there is no buffer: in that case, the absence of buffer implies losing any arrival while a teleportation is in process. Of course, having larger buffers does not make the difference in terms of losses, but it does in terms of fidelity: for a saturated system, longer buffers mean longer average waiting times, and that implies lower fidelity. Therefore, for this simple system, it is sufficient to have a minimum buffer to avoid as much losses as possible without losing too much fidelity.

Quantum computation and communications entanglement [19] is even more important at multi-core scale. In particular, the scarce resources available cause a trade-off on where to assign qubits: computation or communication.

To validate this, we have done a first analysis on the issue simulating the following scenario: a quantum algorithm is executed on a multi-core quantum computer. We fix the number of gates per qubit N_G^Q and the percentage of two-qubit gates N_G^{2Q}. The algorithm is completely random, both in the qubit interactions (two-qubit gates) and in the distribution of the gates along the execution time. Because of the randomness of the algorithm, the two-qubit gates are uniformly distributed among all qubits and along time, so every qubit has an equal amount of communication requirements. Therefore, we can compute the equivalent λ_{core} (the inter-core qubit rate required from any of the cores when executing the algorithm).

We assume to have only one EPR generator connected to all cores. Having a constant λ_{core} for all cores, the optimal sharing strategy is equivalent to a Time Division Multiplexing (TDM). Therefore, each core sees the EPR generator as a source with an observed T_{EPR} equal to N_C times the actual T_{EPR}.

To explore the computation/communications ratio, we have assigned a fixed role to each of the qubits in a core: either it is used for computation or for the buffer. Therefore, changing the ratio of N_Q^{COMP}/N_Q^{COMM} effectively changes the λ_{core}, as well as the buffering capabilities.

In Fig. 4b), the corresponding exploration over the communication to computation qubit ratio, for a fixed percentage (80%) of two-qubit gates and number of cores N_C in the system is shown. This implies that the actual size of the algorithm being executed gets smaller as we increase the ratio. The bottom plot corresponds to the λ_{core} for each ratio. On the rest of the plots, from top to bottom, the successful number of teleportations per second, mean communication latency, mean fidelity, and the % qubit losses and are shown. For each value of N_Q^C (total number of qubits per core), the optimum ratio would correspond to the maximum ratio for which the system does not saturate. For instance, for 512 qubits per core, we should have 40% of the qubits reserved for computation, while the rest would go to communication tasks.

6 CONCLUSIONS

Quantum communications pose completely new challenges that are fostering innovative research, specially on large-scale communications, such as QI applications. However, more attention should be paid to multi-core quantum computing, being a key approach for quantum computing scalability and hence its ultimate success. In this paper, we have started to investigate the critical trade-offs, particularities and optimal designs for these architectures. We have shown how waiting times and latencies can greatly affect quantum communication quality, and have tested how a quantum algorithm (the random case) may behave on a many-core scenario. In particular, due to the scarce amount of qubits available, we have explored and obtained the optimal share of qubits among the computation and communication roles. Future work will include studying behavior of structured well-known algorithms, the impact of multi-core algorithm mapping, and exploring environments with larger number of EPR generators and inter-core ports.

REFERENCES

[1] F Arute et al. 2019. Quantum supremacy using a programmable superconducting processor. *Nature* 574, 7779 (2019), 505–510.

[2] A. S. Cacciapuoti, M. Caleffi, F. Tafuri, F. S. Cataliotti, S. Gherardini, and G. Bianchi. 2019. Quantum internet: Networking challenges in distributed quantum computing. *IEEE Network* 34, 1 (2019), 137–143.

[3] A. S. Cacciapuoti, M. Caleffi, R. Van Meter, and L. Hanzo. 2020. When entanglement meets classical communications: Quantum teleportation for the quantum internet. *IEEE Transactions on Communications* 68, 6 (2020), 3808–3833.

[4] M. Caleffi, A. S. Cacciapuoti, and G. Bianchi. 2018. Quantum Internet: From communication to distributed computing!. In *Proceedings of the 5th ACM International Conference on Nanoscale Computing and Communication*. 1–4.

[5] Tim Coopmans, Robert Knegjens, Axel Dahlberg, David Maier, Loek Nijsten, Julio Oliveira, Martijn Papendrecht, Julian Rabbie, Filip Rozpędek, Matthew Skrzypczyk, et al. 2020. NetSquid, a discrete-event simulation platform for quantum networks. *arXiv preprint arXiv:2010.12535* (2020).

[6] Axel Dahlberg, Matthew Skrzypczyk, Tim Coopmans, Leon Wubben, Filip Rozpedek, Matteo Pompili, Arian Stolk, Przemysław Pawełczak, Robert Knegjens, Julio de Oliveira Filho, et al. 2019. A link layer protocol for quantum networks. In *Proceedings of the ACM Special Interest Group on Data Communication*. 159–173.

[7] Christian Dickel, JJ Wesdorp, NK Langford, S Peiter, Ramiro Sagastizabal, Alessandro Bruno, Ben Criger, F Motzoi, and L DiCarlo. 2018. Chip-to-chip entanglement of transmon qubits using engineered measurement fields. *Physical Review B* 97, 6 (2018), 064508.

[8] Wolfgang Dür, H-J Briegel, Juan Ignacio Cirac, and Peter Zoller. 1999. Quantum repeaters based on entanglement purification. *Physical Review A* 59, 1 (1999), 169.

[9] Albert Einstein, Boris Podolsky, and Nathan Rosen. 1935. Can quantum-mechanical description of physical reality be considered complete? *Physical review* 47, 10 (1935), 777.

[10] H Jeff Kimble. 2008. The quantum internet. *Nature* 453, 7198 (2008), 1023–1030.

[11] Morten Kjaergaard, Mollie E Schwartz, Jochen Braumüller, Philip Krantz, Joel I-J Wang, Simon Gustavsson, and William D Oliver. 2020. Superconducting qubits: Current state of play. *Annual Review of Condensed Matter Physics* 11 (2020), 369–395.

[12] Wojciech Kozlowski, Axel Dahlberg, and Stephanie Wehner. 2020. Designing a quantum network protocol. In *Proceedings of the 16th International Conference on emerging Networking EXperiments and Technologies*. 1–16.

[13] 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.

[14] National Academies of Sciences. 2019. *Quantum computing: progress and prospects*. National Academies Press.

[15] Saptadeep Pal, Daniel Petrisko, Matthew Tomei, Puneet Gupta, Subramanian S Iyer, and Rakesh Kumar. 2019. Architecting waferscale processors-a GPU case study. In *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 250–263.

[16] Momtchil Peev, Christoph Pacher, Romain Alléaume, Claudio Barreiro, Jan Bouda, W Boxleitner, Thierry Debuisschert, Eleni Diamanti, Mehrdad Dianati, JF Dynes, et al. 2009. The SECOQC quantum key distribution network in Vienna. *New Journal of Physics* 11, 7 (2009), 075001.

[17] Alexander Pirker and Wolfgang Dür. 2019. A quantum network stack and protocols for reliable entanglement-based networks. *New Journal of Physics* 21, 3 (2019), 033003.

[18] John Preskill. 2018. Quantum Computing in the NISQ era and beyond. *Quantum* 2 (2018), 79.

[19] Santiago Rodrigo, Sergi Abadal, Eduard Alarcón, and Carmen G Almudever. 2020. Exploring a Double Full-Stack Communications - Enabled Architecture for Multi-Core Quantum Computers. *arXiv preprint arXiv: 2009.08186* (2020).

[20] Santiago Rodrigo, Medina Bandic, Sergi Abadal, Hans van Someren, Eduard Alarcón, and Carmen G. Almudéver. 2021. Scaling of Multi-Core Quantum Architectures: A Communications-Aware Structured Gap Analysis. In *Proceedings of the 18th ACM International Conference on Computing Frontiers* (Virtual Event, Italy) *(CF '21)*. Association for Computing Machinery, New York, NY, USA, 144–151. https://doi.org/10.1145/3457388.3458674

[21] Daniel Sánchez, George Michelogiannakis, and Christos Kozyrakis. 2010. An Analysis of On-Chip Interconnection Networks for Large-Scale Chip Multiprocessors. *ACM Transactions on Architecture and Code Optimization* 7, 1 (2010), Article 4. https://doi.org/10.1145/1756065.1736069

[22] Martin Schoeberl, Florian Brandner, Jens Sparsø, and Evangelia Kasapaki. 2012. A statically scheduled time-division-multiplexed network-on-chip for real-time systems. In *2012 IEEE/ACM Sixth International Symposium on Networks-on-Chip*. IEEE, 152–160.

[23] LMK Vandersypen, H Bluhm, JS Clarke, AS Dzurak, R Ishihara, A Morello, DJ Reilly, LR Schreiber, and M Veldhorst. 2017. Interfacing spin qubits in quantum dots and donors — hot, dense, and coherent. *npj Quantum Information* 3, 1 (2017), 1–10.

[24] Stephanie Wehner, David Elkouss, and Ronald Hanson. 2018. Quantum internet: A vision for the road ahead. *Science* 362, 6412 (2018).

[25] Feihu Xu, Xiongfeng Ma, Qiang Zhang, Hoi-Kwong Lo, and Jian-Wei Pan. 2020. Secure quantum key distribution with realistic devices. *Reviews of Modern Physics* 92, 2 (2020), 025002.

[26] Juan Yin, Yuan Cao, Yu-Huai Li, Sheng-Kai Liao, Liang Zhang, Ji-Gang Ren, Wen-Qi Cai, Wei-Yue Liu, Bo Li, Hui Dai, et al. 2017. Satellite-based entanglement distribution over 1200 kilometers. *Science* 356, 6343 (2017), 1140–1144.

[27] Han-Sen Zhong, Hui Wang, Yu-Hao Deng, Ming-Cheng Chen, Li-Chao Peng, Yi-Han Luo, Jian Qin, Dian Wu, Xing Ding, Yi Hu, et al. 2020. Quantum computational advantage using photons. *Science* 370, 6523 (2020), 1460–1463.