Quality of Service in Quantum Networks

Claudio Cicconetti, Marco Conti, and Andrea Passarella

Abstract

In the coming years, quantum networks will allow quantum applications to thrive thanks to the new opportunities offered by end-to-end entanglement of qubits on remote hosts via quantum repeaters. On a geographical scale, this will lead to the dawn of the quantum internet. While a full-blown deployment is yet to come, the research community is already working on a variety of individual enabling technologies and solutions. In this article, with the guidance of extensive simulations, we take a broader view and investigate the problems of quality of service (QoS) and provisioning in the context of quantum networks, which are very different from their counterparts in classical data networks due to some of their fundamental properties. Our work leads the way toward a new class of studies that will allow the research community to better understand the challenges of quantum networks and their potential commercial exploitation.

Introduction

Decades after it was theorized, quantum computing (QC) has finally emerged as a viable solution with the potential to revolutionize science, industry, and society [1]. Today, QC is being offered as-a-service by very few large companies, led by IBM, but we can expect that in the future quantum computers will become more affordable and undergo mass adoption. When this happens, their interconnection via a global quantum network, the Quantum Internet, will unlock further possibilities, as envisioned by the Internet Research Task Force (IRTF) Quantum Internet Research Group (QIRG), which is leading the standardization efforts toward the definition of a widely-adopted architecture and the identification of use cases of practical interest.

The basic unit of computation in QC is the qubit, which unlike a classical bit can be in a superposition of independent states and can be entangled with other qubits. Qubits can never be copied, but only teleported from one system to another through a process that unavoidably destroys the source state as soon as it is re-created on the destination. Teleporting is the first step toward the realization of end-to-end entanglement. The ultimate goal of a quantum network is: enabling a qubit on a local host to be entangled with another on a remote host, with the two hosts interconnected by quantum repeaters. A vibrant research community has been working toward the realization of quantum repeaters since their inception more than 20 years ago [2], but the task has proved very challenging. Only recently, technology developments have made it possible to obtain encouraging results using both optical fibers [3] and free-space via satellite links [5]. In particular, the latter is envisioned as a practical solution to interconnect quantum computers on geographical distances [5].

Even though quantum repeaters are not yet commercially available, in the QC community there is a widely accepted roadmap that foresees their gradual development through a series of upcoming generations [6]. However, quality of service (QoS) in quantum networks needs significantly different approaches with respect to classical networks, as key enabling concepts (such as throughput and delay) have to be re-defined, while others (such as fidelity, introduced later) are completely new. Furthermore, network provisioning is greatly affected by the fragility of qubits and the inherent impossibility to copy them. Despite the pivotal role these aspects will play to the commercial success of the first Quantum Internet deployments, they have received very little attention so far. To fill this gap, in this work we tackle QoS and network provisioning in quantum networks in a wide range of conditions, considering the costs induced by quantum repeaters in terms of capacity or fidelity. Our analysis is carried out with the help of many simulated scenarios, playing the role of surrogates for real-world deployments in their absence.

The rest of this article is structured as follows. In the next section we introduce our system model and assumptions, followed by a review of the essential state of the art. Then we analyze the key aspects of QoS provisioning in quantum networks as emerging from a selection of broad results obtained with simulation. A summary of the findings is presented, accompanied by a discussion on important open research directions.

Background

In this section we briefly survey some fundamental aspects of quantum computing and communication, and then we introduce the system model of quantum networks used in this work.

Basics of Quantum Computing and Networking

The unit of computation of QC is the qubit, which represents any superposition of two different states (e.g., “0” and “1”) with given (complex) probabilities depending on its initial preparation and subsequent operations. When the qubit is measured, its state collapses to a well-defined state, which is the outcome of the measurement, that is a classical bit “0” or “1,” and it is not possi-
ble to restore the stochastic state it had before this operation. A qubit can be entangled with another, which means that they share an intrinsic correlation: the state of one qubit cannot be described independently from that of the other one, even if they become separated by an arbitrarily large distance. Indeed, the goal of a quantum network (or the Quantum Internet, at a global scale) is to enable the entanglement of qubits in different locations by means of devices called quantum repeaters. The latter are quantum computers, with a limited set of operations, which can transfer the state of a qubit across different media:

- Quantum computers perform local operations on so-called matter qubits made of circuits of superconducting material. Matter qubits are stored in quantum memories, which are interconnected by logic that makes them interact via quantum gates for the execution of local algorithms.
- Quantum communication devices physically transfer from one system to another flying qubits, which most often are made of single photons encoding the state in the polarization (horizontal vs. vertical), which can be carried over fiber optic cables at normal temperatures, even though some cooling is required especially on the receiving side to reduce noise effects [7].

Overall, a quantum network looks like the example in Fig. 1 (top part), with two hosts (C1 and C2) interconnected by quantum repeaters (R1–R3). Let us see what happens when a customer requires end-to-end entanglement between qubits on C1 and C2, say matter qubits $q_{Am}^1$ and $q_{Am}^2$, assuming entanglement path C1–R1–R2–C2:

1. Two EPR pairs $\langle q_{Am}^1 q_{C1}^1 \rangle$ and $\langle q_{Am}^2 q_{C2}^2 \rangle$ are shared between the repeater R1 and the two end hosts C1 and C2, respectively. An EPR pair is a system of two qubits whose states are maximally entangled, that is, informally, they are as strongly correlated as possible, the name being a tribute to Einstein, Podolsky, and Rosen.
2. R1 performs entanglement swapping on $q_{C1}^1$ and $q_{C1}^2$, which is a local operation, involving a measurement, with the effect of entangling the other qubits in the two EPR pairs, that is, $q_{Am}^1$ and $q_{Am}^2$. The process produces classical bits, which are transmitted to C1 (or equivalently C2).
3. C1 entangles locally its matter qubit $q_{Am}^2$ with $q_{C1}^2$. The viability of this step has been demonstrated in laboratory conditions, for example, in [3] using diamond Nitrogen-Vacancy (NV) centers. The same happens on C2 for $q_{Am}^1$ and $q_{C2}^2$.
4. C1 (or equivalently C2) performs local operations on $q_{Am}^1$ based on the classical bits produced by the repeater in step 2.

End-to-end entanglement exploits quantum teleportation, which can transfer a quantum state from one qubit to another with the only constraint that the state on the original qubit is destroyed, thus without violating the “no-cloning theorem.” The step 2 above can be repeated as many times as needed along a path. For example, if the alternative (dashed) path R2–R3 in Fig. 1 is used, the procedure is the same except that both R2 and R3 have to perform entanglement swapping and provide the resulting classical qubits to C1 (or C2). The behavior described is compatible with so-called first generation (1G) quantum repeaters [6] that do not implement any form of quantum error correction (QEC), which is still in its infancy.

**System Model**

We represent a quantum network as a directed graph, with each node representing a quantum computer or repeater, and each edge representing a physical link for quantum communication, as illustrated in the bottom part of Fig. 1. The edges are characterized by two attributes: the capacity, which is the number of EPR pairs that can be generated in the unit of time for end-to-end entanglement; and the generation fidelity, which is a measure of the “quality” of the EPR pair upon generation. The fidelity is a metric between 0 and 1 that summarizes in a quantitative manner how close the state of one qubit is to another, where 1 means perfection.

**QoS Definition:** Quantum applications express their requirements in terms of: requested (end-to-end) capacity, in EPR pairs/s, and minimum fidelity.

In practice, all the processes involved in end-to-end entanglement introduce some undesirable effects. In this work we focus on two aspects that have a most significant impact on the performance of quantum networks with 1G repeaters, which results in a model widely adopted in the literature (e.g., [8]).

First, the entanglement swapping operation performed by quantum repeaters is a stochastic process that succeeds with a given probability ($q$). With current technology the failure probability is quite high, for example, if linear optics components are used then $q$ is smaller than 50 percent [9]. In a chain of hops, when one entanglement swapping fails, even at a single repeater, the overall end-to-end entanglement fails and all the intermediate EPR pairs are wasted. Thus, the success probability decreases exponentially with the number of intermediate hops. Second, also the fidelity of the end-to-end entangled qubits decreases exponentially with the number of hops due to noise introduced by imperfect operations [10]. Later we analyze in quantitative terms the implications of the above features on QoS through simulations in varied conditions. Our goal is to raise awareness in the research community of the profound differences between quantum and classical networks, for what concerns not only physical-layer technologies but also higher level aspects such as the provisioning of network resources.

**State of the Art**

In the literature the problem of “quantum routing” has been well investigated. It is commonly formulated as follows: find the best path to establish end-to-end entanglement between two nodes under given application constraints, usually the expected rate, in EPR pairs/s, and a minimum fidelity. Many of the works have adapted known results from classical data networks to the quan-
In this section we introduce the simulation methodology, then delve into the analysis of the results in two types of experiment, aimed at studying aspects related to admission control function and network provisioning.

**Simulation Tool and Methodology**

We have developed a custom C++ simulator of the system model described earlier, released as open source on GitHub together with the scripts to fully reproduce the experiments (https://github.com/cicciconetti/quantum-routing — git tag v1.4, experiment label 003). The tool simulates a dynamic system where new flows request admission following a Poisson distribution with a given arrival rate. A traffic flow is characterized by the source and destination nodes and its QoS, defined in terms of the requested rate (r, in EPR pairs/s) and the minimum end-to-end fidelity (F). Four classes of QoS are considered, as given by all the combinations of r ∈ {1, 10} and F ∈ {0.7, 0.9}.

We assume that q and Finit are the same for all links, in particular it is q = 0.5 and Finit = 0.95 unless specified otherwise. Under this assumption, the selection of the best path of an incoming traffic flow within the available residual capacity of the network can be greatly simplified. In fact, selecting the shortest path from source to destination minimizes the gross rate and maximizes the fidelity at the same time. The gross rate is defined as the actual capacity that must be reserved to meet the minimum requested rate considering that entanglement swapping at intermediate nodes may fail. Based on this observation, we have implemented the following admission control procedure in the simulator for a given new flow (which uses a temporary network graph of edge capacities):

1. Select the shortest path from source to destination.
2. Determine the gross rate along the path selected.
3. The shortest path selected in the previous step may or may not have sufficient capacity to serve the requested rate r:
   - If the capacity is sufficient, then go to next step 4;
   - Otherwise, remove the edge with minimum capacity from the temporary graph, to make sure that the same path is not selected again, and restart from step 1.
4. Compute the fidelity of the end-to-end entangled pair:
   - If it is greater or equal than the minimum requested value, then the flow can be admitted and all the capacities of the edges along the path are updated by removing the gross rate of the new flow;
   - Otherwise, the admission control fails.

The algorithm includes a loop between steps 1 and 3, which in the worst case may have to be executed a number of times equal to the number of edges. For our purposes, its efficiency is irrele-
Thus, the use of multipath, which in classical data networks is typically discouraged as it leads to increased complexity and out-of-order delivery of packets, can be revamped in quantum networks to reduce the penalty of large traffic flows compared to small ones, as mentioned in [13].

Third, for all the traffic flows, including \( r = 1, F = 0.7 \), the admission rate decreases exponentially with the arrival rate; note that the x-axis in Fig. 3 is in logarithmic scale. In other words, as an increasing number of traffic flows become active in the network, it becomes more and more difficult to admit new flows. We speculate that this non-linearity would make it difficult in practice to provision the resources, for example, to estimate how many traffic flows can be served under a given target admission rate: small deviations from the planned use of resources may have a dramatic impact on the service offered.

Even worse is the fact that this happens when the overall network resources are largely underutilized. Consider Fig. 4, which plots the average admission rate over all the QoS classes vs. the capacity used, as the load increases. The capacity used is defined as the average amount of resources reserved by active traffic flows divided by the total network capacity. As can be seen, the admission rate drops steeply already when a small fraction of the available capacity is used: with a mere 10 percent of the capacity used, the admission rate is already about 50 percent, and even with the maximum load considered, when the admission rate drops to a little more than 10 percent, the capacity used is just 5 percent. This is in contrast with what happens in classical data networks, where the admission rate is usually a smooth function of the load under nominal conditions.

**Takeaway Messages:**

- The exponential increase of the gross rate/exponential decrease of the fidelity with the path length practically limit the distance between communicating parties in a quantum network: in the near future, dense small networks will be less preferable than large sparse ones, possibly using satellite links to extend coverage.
- Aggregated requests (elephants) are penalized compared to smaller requests (mice): multipath can be investigated to address this issue.
- Admission control is more erratic than in classical networks: operators should consider the non-linear/steep function response and devise service plans accordingly.

**Network Provisioning**

We now focus on quantum network provisioning. To this aim, we use artificial topologies, like in [14], generated via a Poisson Point Process (PPP) by dropping a number of nodes, with given average, in a flat square grid with edge size 100 km. A link is then added between two nodes if their Euclidean distance is smaller than 15 km, which is a realistic value based on current quantum communication network technology [18]. Examples of the resulting topologies are illustrated in Fig. 2 (top right and bottom). We consider uniform traffic compositions, with a single QoS class. The source and destination pairs are assigned in one of the following ways:
• **Uniform**: the source and destination pairs are selected randomly from the set of nodes in the network in a uniform manner, like in the first set of experiments;

• **Weighted**: each node has a probability to be a source (or a destination) of a traffic flow that is proportional to the sum of the capacities of the links connected to that node. The weighted assignment is used to mimic a system where the network infrastructure was provisioned based on an estimate of the future (average) traffic demands.

In Fig. 5 we show the admission rate of three QoS classes (left: $r = 1$, $F = 0.7$; center: $r = 10$, $F = 0.7$; right: $r = 10$, $F = 0.9$) for the two assignment strategies. The arrival rate is, on average, 100 flows/s, with increasing link capacity (x-axis) and number of nodes (y-axis). In all the plots an increasing x-axis or y-axis value corresponds to a higher density of network infrastructures, hence higher deployment and operational costs. On the other hand, the color maps to the admission rate from red (low) to blue (high), which in turn significantly affects the customer experience.

The admission rate severely degrades as the application QoS requirements become more demanding, both requested rate (from $r = 1$ to $r = 10$) and fidelity (from $F = 0.7$ to $F = 0.9$). Indeed, reasonable admission rates (around 50 percent) can only be achieved with many nodes and a high average link capacity, regardless of whether a uniform or weighted assignment is used, where the latter only slightly improves the performance. With more relaxed application QoS requirements, that is, $r = 1$, $F = 0.7$ (left plots in Fig. 5), the admission rate is significantly better, especially with a weighted assignment policy. It is interesting to note that this metric improves regularly at each step in the two axes. In other words, from the point of view of the applications, it is the same if the link capacity is increased by 50 EPR pairs/s (x-axis) or 20 nodes are added to the network (y-axis). However, the provisioning cost along the two directions can be very different for the quantum network operator:

• Adding a new node means that a site equipped with quantum repeaters is added to the network. In the simplest case, this can be as straightforward as adding a new rack-mountable piece of equipment to a data center or point of presence (PoP) that is already interconnected to a fiber optic infrastructure used for quantum and classical communications. However, the transmission of
qubits is in general more fragile than that of classical information. Thus, we can also expect that quantum repeaters will be needed in new sites to be provisioned specifically for quantum communication, which would hugely increase the per-node deployment cost.

- The link capacity depends on the amount of physical carriers used, for example, fiber optic cables, and on the number of devices used to generate EPR pairs and to entangle flying and matter qubits, most likely involving single photon detection. This suggests that deployment and operation costs should scale linearly with link capacity, but the statement will have to be confirmed when mass production of 1G quantum repeaters begins.

To complete our analysis, in Fig. 6 we show the admission rate obtained when considering quantum repeaters that exceed the technology limitations expected for near-term devices, that is, with entanglement swapping success probability \( q \) increasing from 0.5 to 1 and fidelity of freshly generated EPR pairs \( F_{\text{init}} \in \{0.95, 0.97, 0.99\} \). We do so in a PPP network with the same characteristics as in the previous scenario, but with a fixed maximum link capacity (300 EPR pairs/s) and average number of nodes (120); furthermore, we only consider traffic flows with \( r = 10, F = 0.9 \), which are the most challenging QoS requirements. The plot shows that \( F_{\text{init}} = 0.95 \) hampers the admission rate for all values of the entanglement swapping success probability. Already with \( F_{\text{init}} = 0.97 \) the admission rate is allowed to grow from 25 percent to about 35 percent, but still with a sub-linear trend. Instead, with \( F_{\text{init}} = 0.99 \) the admission rate is not constrained noticeably by the minimum fidelity requested \( F_q \); hence it is limited only by \( q \). In particular, when \( q = 1 \), the provisioning process becomes the same as with classical communications, and hence can be considered an upper bound of the achievable performance.

**Takeaway Message:** The customer experience, in terms of the blocking probability, depends significantly on the network resources provisioned. The latter can be expanded along two directions, with similar effect from the users’ perspective: adding new nodes or increasing the capacity of existing links, which likely will have different costs and practical implications. The evolution of quantum repeaters will soften the network provisioning difficulties, but only in the long-term, that is, when the entanglement swapping success probability and generation fidelity will be close to 1.

**Summary and Open Issues**

In this article we have investigated some fundamental aspects of QoS and provisioning in quantum networks with 1G repeaters. Our work extends the state of the art on quantum routing by providing a new perspective from the point of view of the network operator, which in the future will have to cope with the definition of suitable business models and service plans based on the estimated revenues and costs. Constraints are met, due to the exponential growth of the network resource demands and the exponential decrease of the fidelity with the path length. Because of the fragility of qubits and the impossibility to copy them, network resource demands increase exponentially with the path length, while the fidelity decreases exponentially.

Under these conditions, simulation results have shown that, even when the network is lightly loaded, there is an unexpectedly high fraction of traffic flows that cannot be served because their QoS constraints are not met. This becomes even worse when high network loads, where the unused capacity is difficult to reach, both with completely random source/destinations and when nodes with greater capacity are more likely to originate traffic flows. Requirements, and in general it suggests that quantum network resources have to be greatly overprovisioned, which can dramatically hamper the deployment of the Quantum Internet. The evolution of technology, for example, to increase the generation fidelity and reduce the probability of failed entanglement swapping in quantum repeaters, will alleviate the problem, but this cannot be expected to happen soon.

In our work, we have unveiled only the tip of a new iceberg of research with many questions left unanswered:

- What is the impact of noise models (e.g., for decoherence of flying qubits over distance) on top of the effects considered in this article?
- Distillation is a procedure that increases the fidelity of end-to-end qubits, at the cost of reducing the capacity further: is this an appli-
culation-level feature or should the network provide distillation as a service?

• Based on the results obtained, multipath may help to better use the network resources, but what is the downside in terms of control/management complexity?

• We have considered a pure online model, where each new traffic flow is served immediately or dropped, but what if the quantum network operator wished to pre-reserve paths based on expected traffic demands?

• Does the definition of QoS capture well the requirements of real applications and what are their associated business models?

• Some applications, for example, distributed QC, may require a variable rate of EPR pairs: what is the impact in terms of provisioning and can different types of application co-exist peacefully in the same network?

• What is the role of time synchronization in the network?

• In principle, EPR pairs can be prepared and stored in quantum memories for later use: how to best use such reservoirs of resources that have no classical counterpart?

To conclude, inherent properties of quantum networks may greatly affect QoS requirements and provisioning. In particular, the potential unfairness between traffic flows with different characteristics has to be addressed, and in general the operators may have to cope with the need of greatly over-provisioning resources. If not addressed properly, these features can dramatically hamper the deployment of the Quantum Internet.

Acknowledgment

This work was co-funded by European Union, PON Ricerca e Innovazione 2014-2020 FESR/FSC, project ARS01_00734 QUANCOM.

References

[1] J. Sevilla and C. J. Riedel, “Forecasting Timelines of Quantum Computing,” arXiv:2009.05045 [quant-ph], Sept. 2020, arXiv: 2009.05045.
[2] H.-J. Briegel, W. Dr, J. I. Cirac, and P. Zoller, “Quantum Repeaters: The Role of Imperfect Local Operations in Quantum Communication,” Physical Review Letters, vol. 81, no. 26, Dec. 1998, pp. 5932–35.
[3] M. Pompili et al., “Realization of a Multi-Node Quantum Network of Remote Solid-State Qubits,” Science, vol. 372, no. 6539, Apr. 2021, pp. 259–64, arXiv: 2102.04471.
[4] J. Yin et al., “Satellite-Based Entanglement Distribution Over 1200 Kilometers,” Science, vol. 356, no. 6343, June 2017,
[5] L. de Forges de Parny et al., “Satellite-Based Quantum Information Networks: Use cases, Architecture, and Roadmap,” Technical Report arXiv:2202.01817, arXiv, Feb. 2022. arXiv:2202.01817 [quant-ph] type: article.

[6] W. J. Munro et al., “Inside Quantum Repeaters,” IEEE J. Selected Topics in Quantum Electronics, vol. 21, no. 3, May 2015, pp. 78–90.

[7] N. Tomm et al., “A Bright and Fast Source of Coherent Single Photons,” Nature Nanotechnology, Jan. 2021.

[8] K. Chakraborty et al., “Entanglement Distribution in a Quantum Network: A Multicommodity Flow-Based Approach,” IEEE Trans. Quantum Engineering, vol. 1, 2020, pp. 1–21.

[9] N. Sangouard et al., “Quantum Repeaters Based on Atomic Ensembles and Linear Optics,” Reviews of Modern Physics, vol. 83, no. 1, Mar. 2011, pp. 33–80.

[10] H.-J. Briegel et al., “Quantum Repeaters for Communication,” arXiv:quant-ph/9803056, Mar. 1998, arXiv:quant-ph/9803056.

[11] R. Van Meter et al., “Path Selection for Quantum Repeater Networks,” Networking Science, vol. 3 nos. 1–4, Dec. 2013, pp. 82–95, arXiv:1206.5655.

[12] C. Cicconetti, M. Conti, and A. Passarella, “Request Scheduling in Quantum Networks,” IEEE Trans. Quantum Engineering, vol. 2, 2021, pp. 2–17.

[13] M. Pant et al., “Routing Entanglement in the Quantum Internet,” Quantum Information, vol. 5, no. 1, Dec. 2019, pp. 25.

[14] W. Dai, T. Peng, and M. Z. Win, “Optimal Remote Entanglement Distribution,” IEEE JSAC, vol. 38, no. 3, Mar. 2020, pp. 540–56.

[15] Y. Yu et al., “Entanglement of Two Quantum Memories via Fibres Over Dozens of Kilometres,” Nature, vol. 578, no. 7794, Feb. 2020, pp. 240–45.

Biographies

Claudio Cicconetti has a Ph.D. in Information Engineering from the University of Pisa (2007). He worked for Intecs (until 2013) and MBI (until 2018); he is now a researcher in the Ubiquitous Internet group of IIT-CNR. His research interests include quantum computing and communication, and serverless edge computing.

Marco Conti is the Director of IIT-CNR. He has been the CNR PI in several projects funded by the European Commission since FP6. He has published 300+ research papers on design, modelling, and performance evaluation of computer and communication systems, and their use for decentralized solutions in self-organizing networks.

Andrea Passarella is a Research Director at IIT-CNR and Head of the Ubiquitous Internet group. He works on the design and performance evaluation of Future Internet networks, especially with self-organizing and human-centric solutions. He has been the CNR PI in several FP7, H2020, and CHIST-ERA projects.