



Communication-efficient distributed inverse Quantum Fourier Transform[☆]

F. Javier Cardama^{a,*,}, Jorge Vázquez-Pérez^{c, ID}, Tomás F. Pena^{a,b, ID}, Andrés Gómez^{c, ID}

^a Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, 15782, Spain

^b Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, Santiago de Compostela, 15782, Spain

^c Galicia Supercomputing Center (CESGA), Avda. de Vigo S/N, Santiago de Compostela, 15705, Spain

ARTICLE INFO

Keywords:

Distributed quantum computing
Quantum Fourier Transform
Quantum algorithms
Quantum communication

ABSTRACT

The scalability of quantum computing is currently limited by physical, technological, and architectural constraints that hinder the integration of a large number of qubits within a single quantum processor. Distributed quantum computing (DQC) has therefore emerged as a viable alternative, aiming to interconnect multiple smaller quantum processing units (QPUs) to jointly operate on a global quantum state. While this paradigm enables scalable architectures, it introduces significant communication overhead due to the cost of non-local quantum operations across distant nodes. In this work we propose a distributed formulation of the iQFT over a quantum network composed of P nodes, each hosting Q qubits, enabling the execution on a global register of size $n = P \cdot Q$. Furthermore, we introduce a communication-efficient variant based on a threshold-driven pruning strategy, referred to as a *communication horizon*, which exploits the exponentially decreasing significance of controlled-phase rotations to safely omit remote gates with negligible impact. By reducing the number of inter-node quantum interactions, the proposed approach significantly lowers the quantum communication requirements of the distributed iQFT while operating within a strictly bounded approximation error. Crucially, we show that this approach fundamentally alters the scaling of the algorithm: the entanglement resource consumption per node saturates to a constant value, reducing the global communication complexity from quadratic $\mathcal{O}(P^2)$ to linear $\mathcal{O}(P)$. As the iQFT constitutes a critical building block in many quantum algorithms, the techniques presented in this paper directly contribute to improving the practicality and scalability of distributed quantum computation.

1. Introduction

Distributed quantum computing (DQC) has emerged as a promising architectural approach to overcome the scalability limitations of current monolithic quantum processors [1,2], which are fundamentally constrained by qubit noise, decoherence, and the increasing complexity of control and calibration as the system size grows [3–5]. Rather than relying on a single large quantum processor, DQC interconnects multiple smaller quantum processing units (QPUs) through quantum entanglement and classical communication channels, enabling the creation of a global quantum state spanning several physical devices [6–8]. By entangling qubits across a network of P interconnected QPUs, each hosting Q local qubits, a collective global register of size $n = P \cdot Q$ can be realized. This modular paradigm allows quantum systems to scale beyond the physical limits of individual processors [9,10].

Regarding quantum computing itself, it represents a fundamental shift in the computational paradigm, with several cornerstone quantum algorithms demonstrating its advantage, being the well known Shor's

integer factorization algorithm [11] the most representative. Shor's algorithm leverages the Quantum Phase Estimation (QPE) [12] algorithm to factor large integers in polynomial time, with the inverse Quantum Fourier Transform (iQFT) playing a central role in it, due to the QPE encoding its final results in the relative phases of quantum states expressed in the Fourier basis. The iQFT acts as a decoding operation that transforms this phase information back into the computational basis, enabling the extraction of classical outcomes through measurement. Consequently, the efficiency and scalability of the iQFT directly impact the performance of the QPE and, therefore, the Shor's algorithm—and all of those employing these oracle as part of their operations.

However, the standard implementation of the iQFT on an n -qubit register requires $\mathcal{O}(n^2)$ two-qubit controlled-phase gates and exhibits an all-to-all interaction pattern, as each qubit must, in principle, interact with every other qubit. This quadratic gate complexity and dense connectivity pose significant challenges for both monolithic and distributed quantum architectures. In distributed settings, the cost of non-local two-qubit gates is particularly high, as they require entanglement

[☆] This article is part of a Special issue entitled: 'Adv. Quantum Vol. IV' published in Future Generation Computer Systems.

* Corresponding author.

E-mail address: javier.cardama@usc.es (F.J. Cardama).

generation, quantum communication, and classical synchronization between remote nodes [13–15].

Because of these demanding communication requirements, the distribution of the QFT and iQFT serves as a benchmark challenge in the field. Standard approaches rely on topology-aware qubit mapping [16], remote-gate batching (such as the partitioning methods proposed by Kaur et al. [17]), and optimized circuit scheduling to efficiently route inter-node communications [18,19]. While these strategies minimize the practical execution time of network operations, they preserve the full logical structure of the exact circuit, meaning they do not reduce the fundamental asymptotic complexity of the communication overhead.

Recently, the theoretical limits of these exact implementations were formalized by Ebneenasir [20]. In this work, the author established the theoretical lower bounds for the communication costs of distributing the exact QFT on clique networks (quantum network with clique topology; i.e., complete graph, where any pair of quantum machines have a direct quantum link between them), rigorously proving that any exact execution necessitates an aggregate communication complexity of $\Omega(P^2)$ across P processing nodes.

In this work, we address these challenges by proposing a communication-efficient distributed implementation of the iQFT. The main contributions of this paper are twofold:

- **Distributed iQFT:** we present a systematic decomposition of the iQFT across a quantum network composed of P nodes, each with Q qubits, enabling the execution of the transform over a global register of size $n = P \cdot Q$.
- **Communication-efficient:** we introduce a communication-optimized variant of the distributed iQFT based on a well-known threshold-driven pruning strategy [21]. By establishing a maximum communication horizon, this approach successfully reduces the global entanglement complexity of the network from quadratic $\mathcal{O}(P^2)$ to linear $\mathcal{O}(P)$ scaling.

The remainder of this paper is organized as follows. Section 2 reviews the theoretical background of the iQFT and distributed quantum computing primitives. Section 3 introduces the proposed distributed iQFT formulation and the communication horizon optimization strategy. Section 4 analyzes the impact of the proposed approach in terms of communication cost and accuracy. Finally, Section 5 conclude the paper and outlines directions for future research.

2. Background

2.1. Inverse quantum fourier transform

The Quantum Fourier Transform (QFT) is a linear unitary transformation analogous to the classical discrete Fourier transform. For a quantum system of n qubits, let $N = 2^n$ be the dimension of the Hilbert space. The QFT maps the orthonormal computational basis states $\{|x\rangle\}_{x=0}^{N-1}$ to a new set of orthonormal states, which we refer to as the *Fourier basis* $\{|\tilde{x}\rangle\}_{x=0}^{N-1}$. This operation is defined as:

$$\text{QFT}|x\rangle = |\tilde{x}\rangle = \frac{1}{\sqrt{N}} \sum_{y=0}^{N-1} e^{2\pi i \frac{xy}{N}} |y\rangle. \quad (1)$$

Each state $|\tilde{x}\rangle$ in the Fourier basis is an equal superposition of all computational basis states, distinguished only by their relative phases. Consequently, the Inverse Quantum Fourier Transform (iQFT) serves as the decoding operation that maps a state from the Fourier basis back to the computational basis:

$$\text{QFT}^{-1}|\tilde{x}\rangle = |x\rangle. \quad (2)$$

Mathematically, the iQFT is the adjoint of the QFT operator ($\text{QFT}^\dagger$). Its action on an arbitrary computational basis state $|y\rangle$ is given by introducing a negative sign in the complex exponential:

$$\text{QFT}^{-1}|y\rangle = \frac{1}{\sqrt{N}} \sum_{x=0}^{N-1} e^{-2\pi i \frac{xy}{N}} |x\rangle. \quad (3)$$

This transformation is fundamental in quantum algorithms such as Quantum Phase Estimation (QPE) and Shor's algorithm, where the solution is encoded in the relative phases of the qubits (Fourier basis) and must be transformed back to the computational basis to be measured with high probability.

2.1.1. Circuit implementation

The decomposition of the iQFT into a quantum circuit consists of a structured sequence of Hadamard (H) gates and two-qubit Controlled-Phase (CP) rotations, followed by a SWAP network to reverse the order of the qubits. The specific circuit implementation considered in this work is illustrated in Fig. 1.

The rotation angle applied to a target qubit $i^\square$, controlled by qubit $i^\star$, is determined by the logical index difference k between them, denoted as:

$$\theta(i^\star, i^\square) = \theta_{i^\square - i^\star} = -\frac{\pi}{2^{i^\square - i^\star}}. \quad (4)$$

$CP(\theta_k)$, applies a diagonal phase shift depending on the index difference between the qubits ($k = i^\square - i^\star$). For the inverse transform, the rotation angle is negative:

$$CP(\theta_k) = \text{diag}\left(1, 1, 1, e^{-i\frac{\pi}{2^k}}\right), \quad (5)$$

where k represents index difference between the control and target qubits in the register. The exact implementation requires $\mathcal{O}(n^2)$ such gates. However, as k increases, the rotation angle approaches zero.

$$\lim_{k \rightarrow \infty} \theta_k = 0. \quad (6)$$

This property is crucial for distributed implementations, as it allows for the truncation of small-angle rotations (approximate iQFT) to reduce inter-node communication, a concept we detail in subsequent sections.

To illustrate the specific connectivity pattern by these operations, Fig. 2 describes the interactions between an arbitrary pair qubits, $i^\star$ as control and qubit $i^\square$ as target. The angle θ applied follows the Eq. (4):

2.1.2. Gate complexity and scaling

The computational cost of the exact iQFT is dominated by the number of two-qubit controlled-phase (CP) rotations required to encode the relative phases. For an n -qubit register, the standard decomposition involves a sequence of operations where the j th qubit (for $j = 1, \dots, n$) acts as a control for $n - j$ target qubits.

The total number of controlled-phase gates, denoted as N_{CP} , is determined by the summation of these interactions across the entire circuit:

$$N_{CP}(n) = \sum_{j=1}^{n-1} j = \frac{n(n-1)}{2}. \quad (7)$$

Consequently, the gate complexity scales quadratically with the system size, i.e., $N_{CP} \in \mathcal{O}(n^2)$. While the number of Hadamard gates scales linearly as $\mathcal{O}(n)$ and the final SWAP network requires $\mathcal{O}(n)$ operations, the quadratic growth of the entangling CP gates presents the primary bottleneck for scalability.

Beyond gate count, the iQFT imposes demanding connectivity constraints on the underlying hardware topology. The algorithm dictates that every qubit $i^\star$ must interact directly via a two-qubit gate with every other qubit $i^\square$ (where $i^\square > i^\star$) to apply the necessary phase shifts.

From a graph-theoretic perspective, the interaction graph of the iQFT corresponds to a complete graph K_n , where an edge exists between every pair of vertices.

This disparity between the algorithmic requirement (all-to-all) and physical constraints significantly increases the effective circuit depth and the total operation count in realistic implementations.

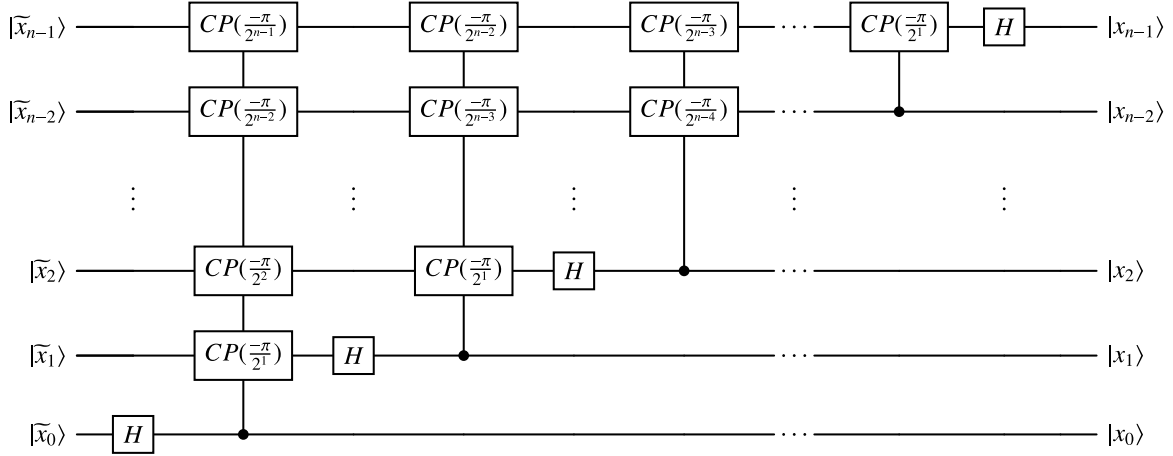


Fig. 1. Circuit realization of the n -qubit Inverse Quantum Fourier Transform (iQFT). The circuit maps the input state from the Fourier basis $|\tilde{x}\rangle$ to the computational basis $|x\rangle$ using a sequence of Hadamard gates (H) and controlled phases $CP(\theta_k)$ where $\theta_k = \pi/2^k$.

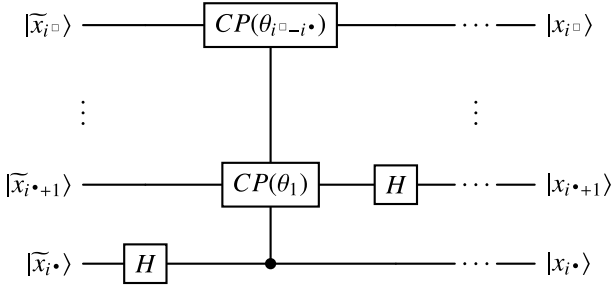


Fig. 2. Interaction between the i^* -th qubit as control and the $i^\square$ -th qubit as target within the iQFT sequence. The diagram highlights the Hadamard gate applied to $|\tilde{x}_{i^*}\rangle$ and the conditional phase rotation controlled by $|\tilde{x}_{i^*}\rangle$ acting on $|\tilde{x}_{i^□}\rangle$.

2.2. Distributed quantum computing: teledata and telegate

Scaling quantum computation beyond the qubit limits of a single processor requires the interconnection of multiple Quantum Processing Units (QPUs) into a distributed network. In such an architecture, the execution of a monolithic quantum circuit is partitioned across distinct nodes. To realize non-local interactions between qubits residing on physically separated QPUs, the system must rely on quantum communication channels and the distribution of entanglement [22–24].

We assume the network links are capable of generating and sharing Einstein–Podolsky–Rosen (EPR) pairs, specifically the Bell state $|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, between nodes. Utilizing this shared resource, two primary communication primitives are established to support distributed logic: *Teledata* [25] and *Telegate* [26,27].

2.2.1. Teledata: state teleportation

The Teledata strategy, based on the standard quantum teleportation protocol [25], physically relocates the quantum state of a qubit from a source node to a destination node. As depicted in Fig. 3, this protocol consumes exactly one shared EPR pair and requires one round of classical communication. The fundamental advantage of this primitive is that once the state is reconstructed at the destination, it becomes entirely local; consequently, an arbitrary number of subsequent quantum gates can be executed on the migrated qubit without incurring

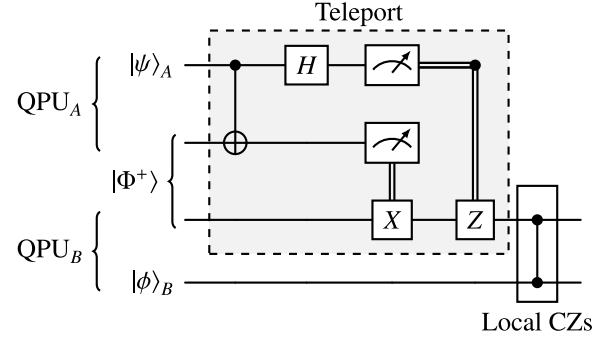


Fig. 3. Circuit schematic of the *Teledata* protocol. The quantum state $|\psi_A\rangle$ is teleported from QPU_A to QPU_B utilizing a shared EPR pair. Upon receiving the classical measurement outcomes for Pauli corrections (X, Z), the state is reconstructed at the destination. This allows the subsequent interaction with $|\phi_B\rangle$ (depicted here as a CZ gate) to be executed as a local operation, avoiding non-local gate overhead.

further inter-node communication penalties. However, if the algorithmic logic demands that the qubit act as a control or target back in its original node later, a “round-trip” teleportation is necessary, effectively doubling the entanglement and latency costs.

2.2.2. Telegate: remote gate application

The Telegate protocol enables the execution of a non-local controlled unitary U between a control qubit in node A and a target qubit in node B without transporting the quantum information of the control qubit itself [26,27]. As illustrated in Fig. 4, the protocol proceeds in three high-level phases: first, an entanglement stage (Cat-Ent) temporarily extends the superposition of the control across the network to a remote “proxy” qubit via a shared EPR pair; second, the intended controlled operation is executed locally at the destination node using the proxy as the control; and finally, a disentanglement stage (Cat-DisEnt) decouples the channel and restores the original control qubit via Pauli corrections. This sequence realizes the remote interaction at a total resource cost of exactly one EPR pair and two bidirectional rounds of classical communication.

Compared to state teleportation, Telegate is highly resource-efficient for isolated non-local interactions, consuming only half the entanglement required for a Teledata “round-trip” (1 EPR pair versus 2).

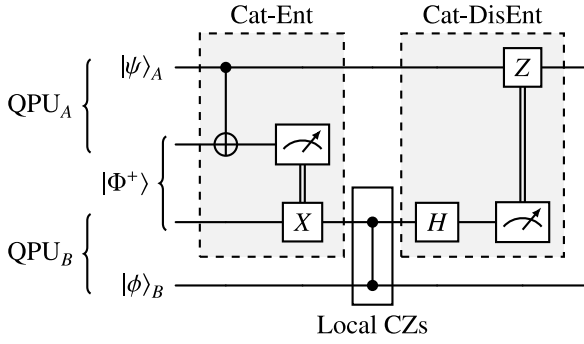


Fig. 4. Circuit implementation of the *Telegate* protocol. This primitive executes a non-local controlled-Z gate between a control qubit $|\psi_A\rangle$ in QPU_A and a target $|\phi_B\rangle$ in QPU_B using a single shared EPR pair. The process involves an entanglement stage (Cat-Ent) to link the control to the channel, followed by the local interaction, and a disentanglement stage (Cat-DisEnt). Note the bidirectional classical communication required for the X and Z Pauli corrections.

However, this efficiency comes at the expense of higher synchronization overhead. The protocol necessitates a bidirectional exchange of classical information (a signal to initiate the remote sequence and the return of measurement results to finalize the gate) which exposes the circuit to the latency of two distinct classical communication events.

2.3. Approximate QFT^{-1} with phase fraction ε

The exact formulation of the QFT^{-1} entails a complete interaction graph, requiring a two-qubit controlled-phase rotation $CP(\theta_k)$ for every pair of qubits, where k denotes the index difference between the control and target qubits ($i \square - i^*$). As detailed in Section 2.1, the rotation angle is given by:

$$\theta_k = -\frac{\pi}{2^k}. \quad (8)$$

A critical observation for optimization is the exponential decay of these rotation angles as k increases. The magnitude of the phase shift is reduced by half for each increment. Consequently, the total phase accumulated by the last target qubit $n-1$, assuming all possible control qubits are in the state $|1\rangle$, follows a geometric progression:

$$\varphi(n-1) = \sum_{k=1}^{n-1} \theta_k = \sum_{k=1}^{n-1} \frac{\pi}{2^k} = \pi \left(1 - \frac{1}{2^{n-1}}\right). \quad (9)$$

As $n \rightarrow \infty$, this sum converges to π , representing a maximum phase shift of 180° . However, the contributions from distant qubits (large k) become vanishingly small. This behavior is illustrated in Fig. 5, which visualizes how the significance of the rotation diminishes rapidly with k .

To reduce the gate count, we can employ an approximate QFT^{-1} by truncating these negligible rotations. To bound the algorithmic error introduced by this truncation, we define a fractional phase tolerance ε , guaranteeing that the total accumulated phase of the discarded gates strictly does not exceed $\varepsilon\pi$. Based on this tolerance, we introduce a threshold parameter t , such that all $CP(\theta_k)$ gates with $k > t$ are discarded. The unitary operator for the approximate transform, denoted as $\tilde{U}$, differs from the exact operator U by the omission of these small-angle gates.

To rigorously formalize the algorithmic error introduced by this truncation strategy, we encapsulate the analytical tracking of the discarded rotations into a formal phase-bound framework.

Theorem 1 (Accumulated Phase Error Bound). Let $\varphi(n-1) = \sum_{k=1}^{n-1} |\theta_k|$ be the absolute phase accumulated by the final target qubit within an n -qubit iQFT register (assuming a worst-case scenario where all controls are $|1\rangle$),

and let $\varphi(t) = \sum_{k=1}^t |\theta_k|$ be the implemented accumulated phase under a pruning threshold t . The absolute phase deviation resulting from omitting all controlled-phase interactions where $k > t$ is strictly upper-bounded by:

$$\varphi(n-1) - \varphi(t) \leq \frac{\pi}{2^t}. \quad (10)$$

Proof. By partitioning the exact accumulated phase summation at the threshold cutoff parameter t , the residual phase contribution of the truncated controlled-phase gates can be isolated as follows:

$$\varphi(n-1) = \varphi(t) + \sum_{k=t+1}^{n-1} |\theta_k|. \quad (11)$$

Consequently, the deviation between the exact and the implemented rotation is determined entirely by the sum of the truncated angles. By substituting the specific absolute rotation values $|\theta_k| = \pi/2^k$, we obtain:

$$\varphi(n-1) - \varphi(t) = \sum_{k=t+1}^{n-1} \frac{\pi}{2^k} = \pi \sum_{k=t+1}^{n-1} 2^{-k}. \quad (12)$$

Using the geometric series property $\sum_{k=t+1}^{n-1} 2^{-k} = 2^{-t} - 2^{-(n-1)}$, the analytical expression for the approximation error evaluates exactly to:

$$\varphi(n-1) - \varphi(t) = \pi \left(\frac{1}{2^t} - \frac{1}{2^{n-1}} \right). \quad (13)$$

In the asymptotic limit as the global register scales ($n \rightarrow \infty$), the negative component $1/2^{n-1}$ vanishes. This establishes the strict, network-independent upper bound:

$$\varphi(n-1) - \varphi(t) \leq \frac{\pi}{2^t}. \quad \square \quad (14)$$

To ensure that the approximation does not exceed the specified tolerance $\varepsilon\pi$, we require the bounded deviation to satisfy $\varphi(n-1) - \varphi(t) \leq \varepsilon\pi$. Solving for the threshold t , we obtain the cutoff condition:

$$2^t \geq \frac{\pi}{\varepsilon\pi} \implies t \geq \log_2 \left(\frac{1}{\varepsilon} \right) = -\log_2(\varepsilon). \quad (15)$$

Thus, to maintain an accuracy of ε , it is sufficient to retain only those controlled-phase gates where

$$k \leq t = \lceil -\log_2(\varepsilon) \rceil. \quad (16)$$

This transforms the gate complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \cdot t)$, offering a significant reduction for large n while preserving algorithmic fidelity.

3. Methods and methodology

3.1. Distributed iQFT across P nodes

To distribute the iQFT across a networked architecture, we must restructure the monolithic circuit into a sequence of operations that distinguishes between intra-node (local) and inter-node (distributed) dependencies.

To formalize the distributed architecture, we define a quantum circuit characterized by the following parameters:

- **Network dimension:** The system consists of P nodes, indexed by the identifier $p \in \{0, 1, \dots, P-1\}$. To explicitly distinguish functional roles during distributed quantum operations, we introduce the modifiers p^* to denote a node acting as a control and $p^\square$ to denote a target node.
- **Node capacity:** We assume a homogeneous distribution where each node hosts exactly Q local qubits. Consequently, the total global register size is $n = P \cdot Q$.
- **Qubit state notation:** We introduce the notation $|\tilde{x}_q^{(p)}\rangle$ to denote the state of the q th qubit residing in the p th node. Here, $q \in \{0, 1, \dots, Q-1\}$ represents the local index of the qubit within its host node. Similar to the node notation, we use q^* and $q^\square$ to designate a control and target qubit, respectively. To simplify

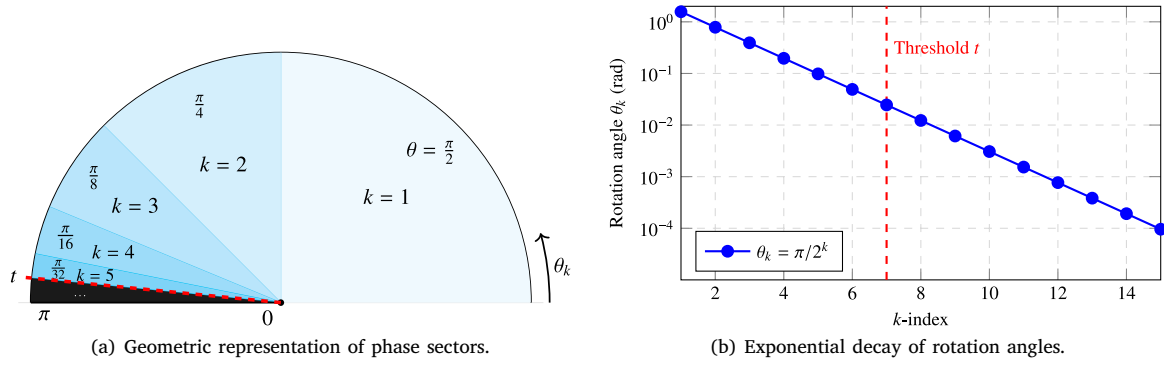


Fig. 5. Representation of the accumulated phase in the iQFT. **(a)** Geometric visualization showing the sectors for $k = 1$ to $k = 5$. The threshold cutoff t indicates the truncation point; interactions beyond this limit contribute a negligible phase shift bounded by $\varepsilon\pi$. **(b)** Semilogarithmic plot quantifying the rapid exponential decay of the rotation angles $\theta_k = \pi/2^k$ as a function of the k parameter, explicitly highlighting a generic truncation threshold $t = 7$.

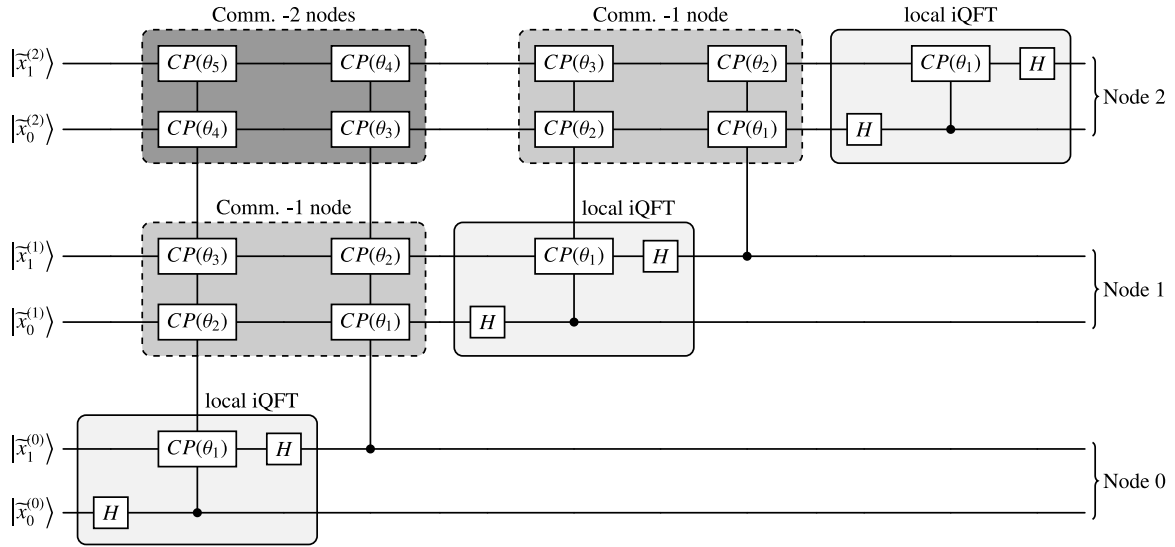


Fig. 6. Distributed decomposition of a 6-qubit iQFT across 3 nodes ($p = 0, 1, 2$). The circuit is restructured into Local iQFT blocks (light gray), which require no external communication, and Communication Blocks (darker gray), where Node p interacts with qubits from previous nodes $p' < p$. The phase angles follow $\theta_k = \pi/2^k$.

complex mathematical expressions and improve readability, we define the shorthand notation $|\tilde{x}_q^{(p)}\rangle^*$ as strictly equivalent to $|\tilde{x}_q^{(p^*)}\rangle$, concisely indicating that both the specific qubit and its host node serve the control role. The same equivalence applies to the target notation $|\tilde{x}_q^{(p)}\rangle^\square \equiv |\tilde{x}_q^{(p^\square)}\rangle$.

- **Global mapping:** The global index of any given qubit within the complete n -qubit register can be recovered through the linear mapping function $I(q, p) = p \cdot Q + q$.

The standard iQFT decomposition (as seen in Section 2.1) requires applying Controlled-Phase (CP) rotations between every pair of qubits. In a distributed setting, these interactions fall into two distinct categories:

1. **Local blocks:** Interactions where both the control and target qubits reside within the same node p (so, $p^\square = p^*$). These operations can be executed entirely within the QPU without utilizing the quantum network.
2. **Communication blocks:** Interactions where the control qubit resides in node p^* and the target qubit resides in node $p^\square$ (with $p^\square \neq p^*$). These require entanglement-assisted communication primitives to realize the non-local phase shifts.

Fig. 6 illustrates this segmented architecture for a specific case of $P = 3$ nodes with $Q = 2$ qubits each ($n = 6$). The structure reveals a distinct execution pattern: each node p must first apply a series of *Communication blocks*, interacting sequentially with nodes $p' < p$. Only after these non-local phases are applied can the node proceed to execute its *Local block*. Crucially, this local component can be identified as a lower-order iQFT instance; specifically, it corresponds to a local iQFT acting on the Q qubits assigned to the node, executed without external dependencies.

The generalized interaction logic is depicted in Fig. 7. For a control qubit $|\tilde{x}_i^{(p)}\rangle^*$ in node p^* and a target qubit $|\tilde{x}_j^{(p)}\rangle^\square$ in a remote node $p^\square$, the rotation angle depends on the global index difference between them. Assuming the node index $p^\square$ represents the higher-order significance bits, the index difference k between the control qubit q^* of node p^* and the target qubit $q^\square$ of node $p^\square$ is derived from their global indices:

$$I^* = p^* \cdot Q + q^*, \quad I^\square = p^\square \cdot Q + j^\square. \quad (17)$$

The rotation index k corresponds to the difference $|I^\square - I^*|$. In the standard decomposition where controls are applied from distant nodes $p < p'$, the specific rotation angle applying Eq. (4) is given by:

$$\theta(I^*, I^\square) = \frac{-\pi}{2^{Q(p^\square - p^*) + (q^\square - q^*)}}. \quad (18)$$

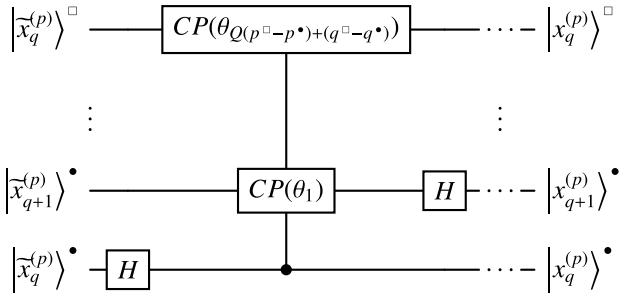


Fig. 7. Generic interaction between a target qubit $q^\square$ in node $p^\square$ and a control qubit q^* in a distant node p^* . The rotation angle is determined by the global index difference, formalized as θ_k where $k = \frac{-\pi}{2^k}$.

This formulation allows us to map the abstract algorithm directly to physical link requirements. The communication blocks effectively group these interactions.

To formally demonstrate that the intra-node operations constitute a standard iQFT of dimension Q (local block), we apply the locality condition $p^* = p^\square$ to the generalized rotation formula derived in Eq. (18). When both control and target qubits reside in the same node, the term $Q(p^\square - p^*) = 0$. Consequently, the rotation angle θ simplifies to depend solely on the local indices q^* and $q^\square$:

$$\theta_{p^\square=p^*} = \frac{-\pi}{2Q(0)+(q^\square-q^*)} = \frac{-\pi}{2q^\square-q^*}, \quad \text{for } q^\square > q^*. \quad (19)$$

This expression is mathematically identical to the phase shift definition of the monolithic iQFT shown in Fig. 7 and Eq. (4), specifically for a difference $k = i^\square - i^*$. Since the indices i range from 0 to $Q-1$, the set of operations within this block generates exactly the interaction graph of a Q -qubit iQFT. Thus, the *Local Block* is equivalent to an iQFT(Q), verifying that no inter-node communication is required for these specific gates.

Fig. 8 depicts the generalized execution schedule for an arbitrary node p within the network. The quantum circuit is abstracted into a sequence of communicative layers followed by a local computation (as shown in Fig. 6).

Instead of treating non-local interactions as isolated gates, the interaction with preceding nodes is structured into communicative blocks denoted by $\text{CP}(\Theta_d)$. Here, d represents the logical distance to the control node (i.e., interacting with node $p^\square = p^* + d$). Each block encapsulates the complete set of remote controlled-phase rotations required between the Q qubits of the target node $p^\square$ and the Q qubits of the control node $p^\square - d = p^*$.

Mathematically, Θ_d is defined as the set of all phase rotations to be applied within that specific block. Following the generalized rotation derived in Eq. (18), this set is strictly formulated as:

$$\Theta_d = \{\theta(I(q^*, p^*), I(q^\square, p^\square)) \mid p^\square = p^* + d\}, \quad (20)$$

where $q^\square$ denotes the qubit index in the target register (node $p^\square$) and q^* the index in the control register (node $p^\square - d = p^*$). This formulation ensures that all cross-node dependencies are satisfied for the given distance.

As shown in the circuit, the sequence proceeds from the most distant interactions ($d = p$, corresponding to node 0) to the nearest neighbor ($d = 1$, corresponding to node $p-1$).

3.1.1. Entanglement resource allocation profile

Entanglement cost is strictly tied to the inter-node migration of the quantum state, rather than the number of remote gates. When a control qubit $|\tilde{x}_q^{(p)}\rangle^*$ must interact with multiple targets in a remote node $p^\square$, the Telegate protocol consumes exactly one EPR pair to create a cat-entangled proxy at the destination. Once this proxy is established in the remote node, it can be applied to an arbitrary number of target qubits with zero additional entanglement cost.

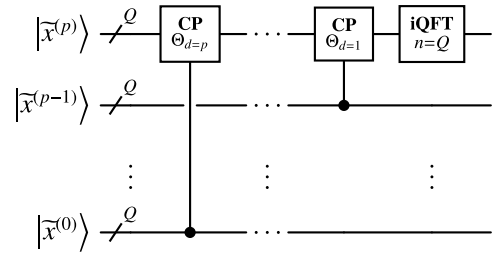


Fig. 8. Generic interaction between nodes, execution model for an arbitrary node p . The circuit illustrates the sequential application of communication blocks CP , starting from the most distant dependency (Node 0, $d = p$) up to the nearest (Node $p-1$, $d = 1$). The execution of this sequence assumes that all previous nodes $0, \dots, p-1$ have already initiated their corresponding operations to act as controls. The process concludes with a local inverse QFT on the Q qubits of node p .

3.2. Communication efficiency: block pruning strategy

While the standard approximate iQFT reduces the total gate count by discarding small-angle rotations, its impact is particularly transformative in a distributed setting. In monolithic architectures, removing a gate merely saves execution time. However, in a distributed architecture, the primary bottlenecks are the generation of entanglement (EPR pairs) and the latency of classical communication rounds. Therefore, the optimization goal shifts from simply reducing N_{CP} to minimizing the number of active inter-node links.

We leverage the geometric decay of the rotation angles derived in Section 2.1 to define a *communication horizon*. Recall that for a target accuracy ϵ , any controlled-phase rotation $\text{CP}(\theta_k)$ with distance $k > t$ can be approximated as the identity, where $t = \lceil -\log_2(\epsilon) \rceil$ (Eq. (16)). We apply this threshold to the coarse-grained communication blocks $\text{CP}(\Theta_d)$ (Eq. (20)) defined in Section 3.1.

Based on this truncation condition, we can formalize the maximum interaction range required in the distributed architecture.

Proposition 1 (Communication Horizon). For a distributed quantum register of node capacity Q and a target algorithmic tolerance ϵ , the maximum required interaction distance between nodes, denoted as the communication horizon $d_{\max}$, is strictly bounded by:

$$d_{\max} = \left\lceil \frac{\lceil -\log_2(\epsilon) \rceil - 1}{Q} \right\rceil + 1. \quad (21)$$

Proof. A communication block $\text{CP}(\Theta_d)$ encapsulates interactions between a target node $p^\square$ and a control node at distance d . To determine the relevance of the entire block, we consider the *maximum interaction strength* within it, which corresponds to the smallest qubit distance $k_{\min}$ (i.e., the largest rotation angle). This occurs between the least significant qubit of the target node ($q^\square = 0$) and the most significant qubit of the control node ($q^* = Q-1$). Applying Eq. (18):

$$k_{\min}(d) = Q \cdot (p^\square - p^*) + (0) - (Q-1) = Q(d-1) + 1. \quad (22)$$

If this strongest interaction falls below the significance threshold (i.e., if $k_{\min}(d) > t$), then every other interaction in the block will also satisfy $k > t$. Consequently, the entire communication block becomes negligible and can be pruned from the schedule. Setting the retention condition $k_{\min}(d_{\max}) \leq t$, we solve for $d_{\max}$:

$$Q(d_{\max} - 1) + 1 \leq t \implies d_{\max} \leq \frac{t-1}{Q} + 1. \quad (23)$$

Substituting t from Theorem 1 (Eq. (16)) and applying the floor function for integer node distances completes the proof. $\square$

This structural limitation directly alters the asymptotic scaling of the algorithm's quantum network footprint.

Corollary 1 (Linear Global Entanglement Scaling). *In the communication efficient distributed iQFT, the global entanglement complexity scales linearly $\mathcal{O}(P)$ with respect to the total number of nodes P , rather than the quadratic $\mathcal{O}(P^2)$ scaling of the exact distributed implementation.*

Proof. In the exact distributed iQFT architecture, each node must establish entanglement links with all its preceding nodes, yielding an all-to-all communication topology where the total number of inter-node blocks grows as $\sum_{i=1}^{P-1} i \in \mathcal{O}(P^2)$. By Proposition 1, the pruning strategy bounds the connectivity degree of any node p to a maximum of $d_{\max}$ nearest neighbors. Since $d_{\max}$ depends solely on the localized parameters Q and ϵ and is independent of P , the maximum number of required communication links per node is bounded by a constant $\mathcal{O}(1)$. Summing this over all P nodes reduces the aggregate global communication complexity to $\mathcal{O}(P)$. $\square$

3.2.1. Granularity of the truncation strategy

Pruning operates strictly at the block level. If a block is retained ($k_{\min}(d) \leq t$), all its internal gates are executed—even those where $k > t$. Since the EPR pair required to migrate the control qubit is already consumed by the block's significant interactions, executing the remaining weak rotations incurs zero additional communication cost. Retaining them maximizes algorithmic fidelity without any entanglement overhead.

To illustrate the practical impact, consider a network composed of nodes with capacity $Q = 4$ qubits. Targeting a high algorithmic fidelity with $\epsilon = 10^{-5}$, the phase truncation threshold is $t \approx 17$. Applying Eq. (21), the communication horizon is bounded by $d_{\max} = 5$. This implies that in a large-scale network (e.g., 100 nodes), Node 99 is not required to establish costly long-range entanglement links with Nodes 0 through 93; it strictly needs to coordinate with its 5 nearest predecessors (Nodes 94–98). Interactions beyond this horizon contribute negligible phase shifts and are safely discarded, drastically reducing the demand on quantum repeater networks.

3.3. Metrics evaluated

To rigorously assess the impact of the proposed distributed architecture and optimization strategies, we quantify performance using the following set of metrics. These metrics are designed to evaluate the trade-offs between algorithmic fidelity, communication resources, and circuit locality.

3.3.1. Fidelity vs. Threshold

We evaluate the accuracy of the approximate distributed iQFT by computing the state fidelity, defined as:

$$F = |\langle \psi_{\text{ideal}} | \psi_{\text{approx}} \rangle|^2, \quad (24)$$

where $|\psi_{\text{ideal}}\rangle$ is the output state of the exact transform and $|\psi_{\text{approx}}\rangle$ is the output of the pruned circuit. This metric is analyzed as a function of the pruning threshold t . The objective is to validate the theoretical error model derived in Section 2.3 by comparing the fidelity obtained from full state-vector simulations against the theoretical lower bound implied by the chosen accuracy parameter ϵ .

What does it prove? This serves to empirically validate the theoretical error bound and cutoff condition derived in Section 2.3 (Eq. (16)).

3.3.2. Expected state probability vs. Threshold

While state-vector simulations provide exact fidelity calculations under ideal conditions, assessing the architecture's viability in the NISQ era requires modeling stochastic hardware errors. To facilitate the introduction of realistic physical noise profiles, such as those provided by IBM Qiskit Runtime FakeBackends, we utilize a measurement-based error metric compatible with shot-based simulators (which output discrete measurement counts rather than exact state amplitudes).

The evaluation protocol involves initializing the system in a known Fourier basis state $|\bar{x}\rangle$, which corresponds to the ideal QFT transformation of a computational basis state $|x\rangle$. Upon applying the approximate distributed iQFT, we measure the output in the computational basis. In an ideal execution, the probability mass should concentrate entirely on the state $|x\rangle$. We define the empirical error E_{counts} as the fraction of measurement shots that fail to collapse into the expected computational state:

$$E_{\text{counts}} = 1 - \frac{N_x}{N_{\text{total}}}, \quad (25)$$

where N_x is the count of shots matching the expected bitstring x , and N_{total} is the total number of shots executed.

What does it prove? This metric allows us to evaluate how the execution of the circuit is affected by applying different pruning thresholds in noisy systems, providing a practical method to assess algorithmic resilience in large-scale distributed architectures where direct fidelity computation is intractable.

3.3.3. Scalability of entanglement resources (EPR per node)

This metric quantifies the aggregate cost of inter-node communication required to execute the algorithm. We measure the total count of EPR pairs consumed across the entire network during the circuit execution. We analyze the behavior of this cost as a function of the pruning threshold t , comparing the baseline distributed iQFT against the optimized implementation.

What does it prove? This metric directly demonstrates the global communication savings achieved by the pruning strategy described in Section 3.2.

3.3.4. Accuracy under connectivity constraints

Suppose hardware limitation of a fixed communication horizon $d_{\max}$ (the maximum distance between interacting nodes) and a specific node capacity Q , we analyze the achievable algorithmic accuracy. By inverting the relationship established in the block pruning strategy, we determine the minimum error ϵ that can be guaranteed when the communication topology is restricted.

What does it prove? This analysis confirms the feasibility of the bounded interaction model and the communication horizon ($d_{\max}$) proposed in Section 3.2.

3.3.5. Coupling ratio

We define the Coupling Ratio η as the quotient between the number of remote controlled-phase gates (inter-node) and local controlled-phase gates (intra-node). This metric serves as a proxy for the degree of independence of the nodes. We analyze how η evolves with respect to the node capacity Q , the total network size P , and the approximation tolerance ϵ .

$$\eta(Q, P, \epsilon) = \frac{N_{CP}^{\text{remote}}}{N_{CP}^{\text{local}}} \quad (26)$$

The objective is to try to achieve a $\eta \rightarrow 0$ in the distributed algorithm.

What does it prove? This demonstrates the effectiveness of the distributed decomposition in Section 3.1 in clustering operations locally to minimize network latency.

3.3.6. Communication overhead

To quantify the operational penalty introduced by the network protocols, we define the *Communication Overhead* (γ). This metric is calculated as the ratio between the number of physical operations dedicated exclusively to quantum communication (N_{comm}) and the number of logical computational gates (N_{comp}):

$$\gamma = \frac{N_{\text{comm}}}{N_{\text{comp}}} \quad (27)$$

In the context of the *Telegate* protocol utilized in this work, N_{comm} encompasses all auxiliary operations necessary for state transfer, including Bell pair generation, entanglement swapping operations (CNOTs), and measurement-dependent corrections. Conversely, N_{comp} refers strictly to the logical controlled-phase rotations required by the iQFT algorithm. A visual decomposition distinguishing between these communication-specific overheads and the fundamental computational gates is presented in Fig. 4. It is important to emphasize that γ is a global metric, representing the aggregate overhead of the distributed execution rather than a per-node average.

What does it prove? This metric exposes the communication cost of the distributed implementation, quantifying the exact ratio of physical overhead operations required to execute a single useful logical gate.

3.4. Experimental methodology

To validate the proposed distributed architecture and the efficiency of the pruning strategy, we implemented the distributed iQFT algorithm utilizing the Qiskit software development kit. The numerical simulations were performed using the statevector simulator backend provided by Qiskit Aer. This simulator performs an ideal execution of the quantum circuit, computing the exact final quantum state vector $|\psi_{\text{final}}\rangle$ rather than a probability distribution of measurement outcomes.

The choice of a state-vector simulation is essential for this study for two primary reasons: first, it allows for the precise calculation of the algorithmic fidelity $\mathcal{F}$ by computing the overlap with the ideal mathematical output, enabling a direct verification of the theoretical error bounds and threshold equations derived in Section 2.3. Second, it permits the exact tracking of entanglement resources without the statistical noise associated with shot-based sampling.

All experiments were conducted on the High-Performance Computing cluster of the *Centro Singular de Investigación en Tecnoloxías Intelixentes* (CITIUS).¹ Specifically, the simulations were executed on a dedicated compute node designed for memory-intensive workloads, equipped with the following specifications:

- **Processors:** $4 \times$ Intel Xeon Gold 6248 CPUs @ 2.50 GHz (20 cores per socket, totaling 80 physical cores).
- **Memory:** 1 TB of RAM.

This high-memory configuration is critical for the study, as the memory requirements for state-vector simulations scale exponentially with the number of qubits (16×2^n bytes). The available 1 TB of RAM enabled the simulation of distributed systems with sufficient qubits to analyze the asymptotic behavior of the communication horizon d_{max} and the locality ratio η .

4. Results and discussion

The results presented in this work demonstrate that the communication aware optimization of the distributed iQFT significantly improves scalability without compromising algorithmic integrity. By restructuring the circuit to exploit the quadratic decay of controlled-phase rotation angles, we effectively transform a global, all-to-all interaction problem into a localized execution model. In this section, we analyze the performance of the proposed architecture, evaluating the trade-offs between algorithmic fidelity, communication resources, and circuit locality.

¹ More information about the HPC cluster: <https://wiki.citius.gal/centro-servizos:hpc>

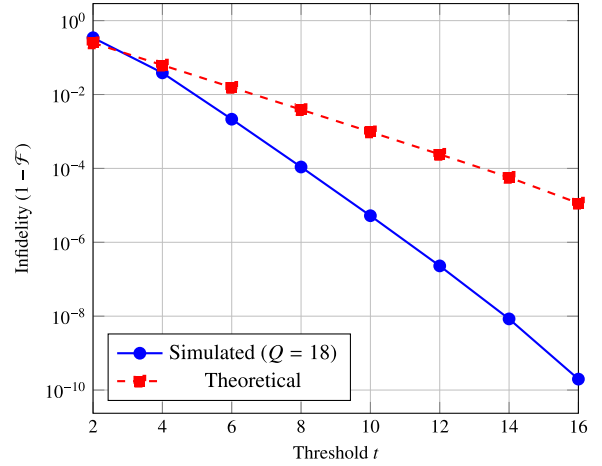


Fig. 9. Analysis of the algorithmic infidelity ($1 - \mathcal{F}$) as a function of the pruning threshold t for local register size $Q = 18$. The data is plotted on a semilogarithmic scale to visualize the **exponential decay** of the error. Solid lines represent the results obtained from full state-vector simulations, while dashed lines correspond to the theoretical error bounds derived in Eq. (16).

4.1. Fidelity and theoretical bounds

Fig. 9 illustrates the infidelity, defined as $1 - \mathcal{F}$ (Eq. (24)), as a function of the pruning threshold t (Eq. (16)) for a system size of $Q = 18$ qubits. The data is presented on a semilogarithmic scale to accommodate the wide dynamic range of the error, which spans from approximately 10^{-1} to 10^{-10} .

The linearity of the curves indicates that the infidelity decays exponentially with a linear increase in the threshold t . This behavior confirms the efficacy of the proposed pruning strategy: a moderate increase in the threshold parameter yields a reduction in the final state error by several orders of magnitude. For instance, in the $Q = 18$ case, increasing the threshold from $t = 4$ to $t = 10$ suppresses the error by nearly four orders of magnitude.

A consistent divergence is observed between the numerical simulations and the analytical predictions, where the simulated infidelity (solid lines) remains strictly bounded by the theoretical error estimates (dashed lines). This discrepancy is attributable to the conservative nature of the theoretical bound, which is derived under a worst-case scenario assumption. The theoretical upper model assumes that every pruned controlled-phase gate, $CP(\theta)$, introduces its maximal possible error (physically corresponding to the control and target qubits invariably being in the state $|1\rangle$). However, in a practical execution environment, the quantum register exists in a superposition of computational basis states. For any subspace of the wavefunction where the control qubit is $|0\rangle$, the $CP(\theta)$ gate acts as the identity operator ($\mathbb{I}$), and its truncation introduces zero error. Consequently, the actual accumulated phase error is statistically lower than the sum of the absolute magnitudes of the discarded angles, offering a safety margin where the realized algorithmic fidelity exceeds the strict design requirements.

4.2. Expected state probability under fake backends

To evaluate the resilience of the proposed architecture against physical decoherence, we analyzed the Expected State Probability (E_{counts}) across different truncation thresholds t . The algorithm was executed on an ideal state-vector simulator alongside three realistic noisy hardware models provided by IBM Qiskit Runtime (Fake Hanoi, Fake Montreal, and Fake Paris). The results are depicted in Fig. 10.

For the ideal simulator, increasing the threshold t monotonically improves the success probability. As t grows, the algorithm transitions

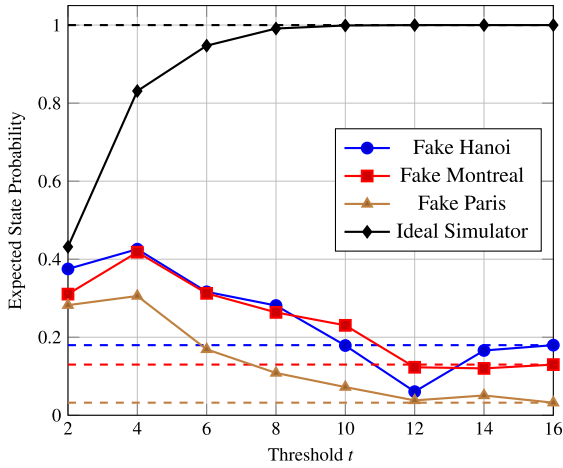


Fig. 10. Expected State Probability ($1 - E_{\text{counts}}$) as a function of the truncation threshold t for register $Q = 18$. The ideal simulator converges to 1.0, whereas noisy backends degrade due to physical circuit depth. The horizontal dashed lines represent the baseline performance of the standard iQFT without thresholding for each respective backend.

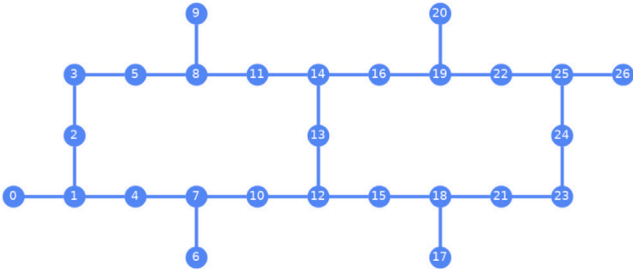


Fig. 11. Topological coupling map of the three IBM Qiskit Fake backends utilized in the simulations (Hanoi, Montreal and Paris). The sparse, restricted connectivity imposes a heavy SWAP-gate penalty for two-qubit operations.

from a highly approximate state to a more exact transformation, reaching a 100% success rate at $t = 16$. Interestingly, at the most aggressive pruning level ($t = 2$), the algorithmic error dominates the execution, yielding an ideal baseline probability that is nearly identical to those obtained from the noisy hardware models.

Conversely, the noisy backends exhibit a markedly inverse trend: as the threshold t increases, the empirical success probability steadily degrades (i.e., the error E_{counts} increases). This degradation is fundamentally driven by the severe mismatch between the algorithmic requirements of the distributed iQFT and the physical connectivity of the hardware.

As illustrated in Fig. 11, the physical hardware features a sparse and restricted topological coupling map. However, the exact iQFT theoretically demands an all-to-all connectivity graph. When higher thresholds are applied, a larger number of controlled-phase gates are retained in the circuit. Transpiling these extended logical interactions to the restricted physical topology requires the insertion of extensive SWAP-gate routing networks. This routing drastically increases the physical circuit depth, amplifying the accumulation of depolarizing and readout errors, which quickly eclipse the theoretical gains in algorithmic fidelity.

It is important to emphasize that at $t = 16$, the point at which the ideal simulator attains perfect accuracy, the circuit becomes mathematically equivalent to the conventional, state-of-the-art iQFT. The degraded performance observed on noisy backends precisely at this value of t therefore indicates the intrinsic fragility of the standard iQFT in NISQ backends, rather than any fundamental shortcoming of the

proposed pruning strategy. Instead, these findings substantiate the essential role of the introduced communication horizon. By intentionally enforcing a more conservative threshold, the architecture drastically reduces the algorithmic demand for non-local interactions. On NISQ backends characterized by sparse coupling maps, this reduction translates directly into a significantly smaller physical circuit depth by avoiding extensive SWAP-gate routing. Consequently, the communication horizon suppresses the otherwise accumulation of physical noise, providing a critical operational trade-off that maximizes the robust extraction of computational information on NISQ hardware.

4.3. Entanglement resource scaling

Fig. 12 presents a comparative analysis of the entanglement resources required for the distributed iQFT, expressed as the number of EPR pairs generated *per node*. The results are depicted as heatmaps varying with the local system size (Q) and the number of computing nodes (P) for the unbounded case (no pruning) and two pruned configurations ($t = 7$ and $t = 3$).

It is crucial to emphasize that the metric displayed corresponds to the EPR pairs generated *per individual node*. While the total entanglement consumption across the entire network would naturally scale with P , the per-node burden provides a more insightful measure of the local communication overhead. As observed in the leftmost panel (unbounded case), without pruning, the resource demand per node grows linearly with both Q and P . This leads to a quadratic aggregate complexity $\mathcal{O}(P^2)$, where a system of $P = 20$ nodes with $Q = 20$ qubits necessitates ≈ 180 EPR pairs per node.

The central and rightmost panels demonstrate the efficacy of the proposed pruning strategy. By applying a filtering threshold t , this initially quadratic growth is effectively suppressed. Consistent with the theoretical analysis in Section 3.2, the threshold acts as a cutoff for long-distance interactions. For stricter thresholds (e.g., $t = 3$), the number of EPR pairs per node saturates rapidly, becoming independent of the system size P . Even a moderate threshold of $t = 7$ caps the requirement at a maximum of approximately 7 EPR pairs per node, regardless of the network size. This transition from quadratic to linear global scaling (constant per node) transforms the all-to-all entanglement requirement into a localized resource pattern, shifting the bottleneck from entanglement generation rates to local gate fidelities and making the architecture highly scalable.

4.4. Coupling ratio and node decoupling

To assess the computational efficiency of the distributed implementation relative to the communication overhead, we analyze the *Coupling ratio* (η defined in Eq. (26)). In distributed quantum architectures, the quantum interconnect bandwidth and latency constitute the primary performance bottlenecks, often orders of magnitude slower than local gate operations. An ideal distributed algorithm should strive for $\eta \rightarrow 0$, indicating that the computational workload is dominated by local processing.

Fig. 13 illustrates the evolution of η as a function of the pruning threshold t . In the standard, unpruned iQFT, the algorithm exhibits a high dependency on remote operations, with η values reaching ≈ 35 for large systems, indicating that inter-node communication dominates the execution time.

A central contribution of this work is the establishment of the communication horizon, $d_{\text{max}}(Q, \epsilon)$ (Eq. (21)), which acts as a decisive control parameter. Since the amplitude of the controlled-phase rotation $\theta_k = \pi/2^k$ decays exponentially with the logical distance k , the pruning threshold t effectively imposes a hard limit on the interaction range. This mechanism preferentially filters out long-range remote operations typically corresponding to high k . The application of this optimization drastically suppresses the ratio. For thresholds $t \leq 7$, η approaches

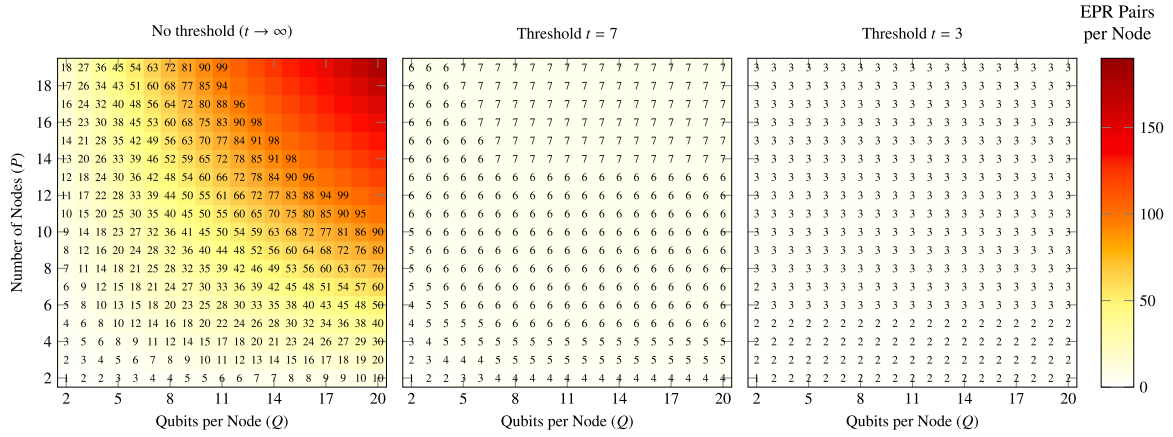


Fig. 12. Density of entanglement resources, quantified as the number of EPR pairs generated *per computing node*, as a function of the local register size Q and the total number of nodes P . Note that **higher values correspond to increased communication overhead and resource consumption**. The leftmost panel (Unbounded) shows a linear increase in local resource demand with respect to P , implying a quadratic global complexity. The central ($t = 7$) and rightmost ($t = 3$) panels demonstrate how the pruning threshold effectively saturates this cost, rendering the resource consumption independent of the network size for sufficiently large distributed systems.

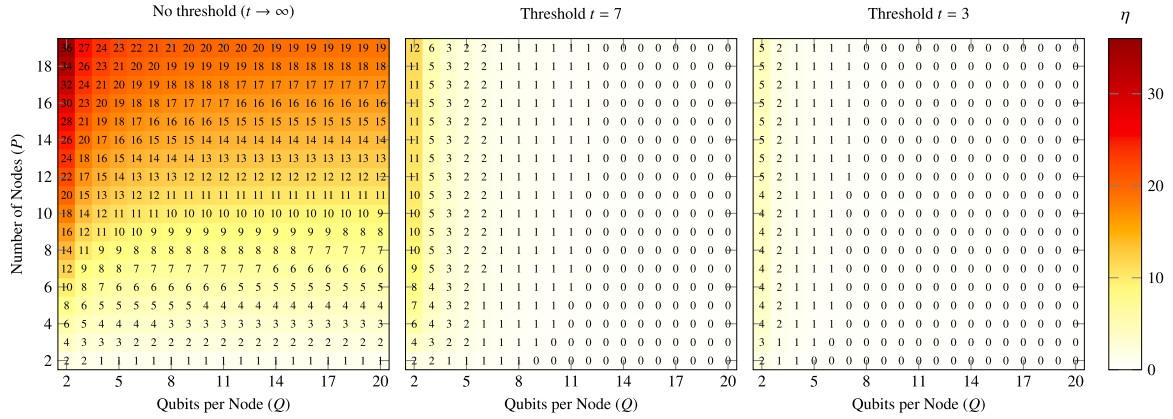


Fig. 13. Evolution of the Coupling Ratio η , defined as the ratio between remote and local controlled-phase gates ($\eta = N_{remote}/N_{local}$), as a function of the pruning threshold. Note that **higher values represent a higher dependency on inter-node communication**, which constitutes the primary bottleneck in distributed architectures. The plot demonstrates how the proposed pruning strategy successfully minimizes this ratio ($\eta \rightarrow 0$), effectively decoupling the nodes and shifting the computational load towards efficient local processing.

asymptotic zero ($\eta \rightarrow 0$). This reduction signifies a fundamental decoupling of the nodes: the computational workload is almost entirely confined to the local QPU, with inter-node links serving only for sparse, nearest-neighbor synchronization. This decoupling ensures that the latency of the quantum channel does not exponentially compound.

4.5. Accuracy under connectivity constraints

Fig. 14 depicts the impact of imposing a fixed communication horizon d_{max} on the algorithmic accuracy. The heatmaps quantify the theoretical error bound defined in Eq. (12) as a function of the local node capacity (Q) and the maximum allowed communication distance (d_{max} defined in Eq. (21)).

The concept of the communication horizon emerges as a decisive parameter for distributed architecture design. As evidenced by the rapid error convergence in the heatmap, the algorithmic error decays precipitously as the communication horizon increases. Even for moderate horizon values, the fidelity of the distributed iQFT remains extremely high. For instance, in a system with local capacity $Q = 9$ qubits per node, restricting the interaction range to just $d_{max} = 2$ neighboring nodes yields an error bound of approximately 2^{-10} . Extending connectivity beyond a small neighborhood provides diminishing returns in accuracy while incurring significant communication costs.

A fundamental property revealed by this analysis is the independence of the error bound from the total number of nodes P . The accuracy is determined solely by the local register size Q and the communication horizon d_{max} . This implies that for any target precision ϵ , the connectivity graph does not need to scale with the network size. A constant connectivity radius guarantees a specific error tolerance whether the network consists of $P = 20$ or $P = 100$ nodes. This independence is a prerequisite for scalable distributed quantum computing, as it prevents the communication overhead from diverging as the system grows.

4.6. Communication overhead

Fig. 15 analyzes the Communication Overhead ratio (γ) across different system sizes and pruning configurations. A crucial observation from these results is that the overhead factor remains relatively invariant with respect to the pruning threshold t .

This stability implies that the application of the threshold does not alter the fundamental architectural efficiency of the distributed protocol. The rationale lies in the fixed operational cost of the *Telegate* mechanism: each executed remote *CP* gate necessitates a specific, proportional set of auxiliary communication operations (Bell pair generation, entanglement swapping, and corrections). Therefore, when the

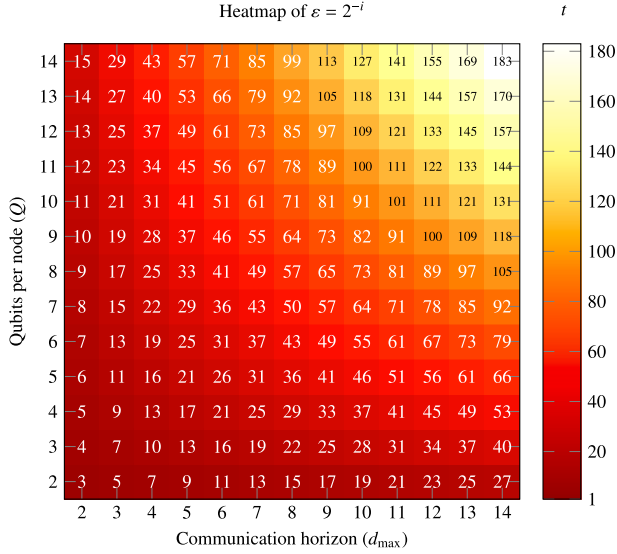


Fig. 14. Impact of the communication horizon $d_{\max}$ and local register size Q on the algorithmic accuracy. The values within the cells represent the **negated exponent i** of the theoretical error bound (such that $\epsilon \approx 2^{-i}$). Therefore, **higher values indicate lower errors and higher fidelity**. A rapid convergence is observed, showing that small horizons are sufficient to achieve high precision.

Table 1

Comparison of distributed QFT/iQFT approaches and communication scaling across P nodes.

Reference	Paradigm	Key Concept	Comm. Scaling
Standard Distributed iQFT	Exact	Baseline all-to-all mapping	$\mathcal{O}(P^2)$
Compiler Batching [17]	Exact	Remote-gate batching and scheduling	$\mathcal{O}(P^2)$
Theoretical Bound [20]	Exact	Lower bound proof for clique networks	$\Omega(P^2)$
Proposed work	Approximate	Communication horizon ($d_{\max}$)	$\mathcal{O}(P)$

pruning strategy removes a logical gate, it simultaneously eliminates the associated communication burden.

Consequently, while the *total count* of operations decreases drastically, the *ratio* between communication and computation tasks (γ) remains stable. This confirms that the overhead is an intrinsic property of the chosen quantum interconnect protocol and is not artificially inflated or distorted by the pruning approximation.

4.7. Comparison with state-of-the-art and theoretical bounds

To contextualize the impact of the proposed architecture, Table 1 summarizes the communication complexity and core strategies of existing distributed QFT and iQFT frameworks.

An analysis of the distributed quantum computing literature reveals that existing solutions operate under distinct methodological paradigms. In this context, the work of Ebneasir [20] constitutes a fundamental theoretical predecessor that establishes the strict lower bounds for the communication costs of distributing the exact QFT on clique networks. This study rigorously proves that any exact implementation is mathematically constrained to an aggregate communication

complexity of $\Omega(P^2)$ due to the dense, all-to-all logical connectivity of the circuit. The linear $\mathcal{O}(P)$ communication scaling achieved by our proposed communication-efficient iQFT is fully consistent with Ebneasir's theoretical limit and does not violate it, since his theorem applies strictly to the exact formulation. By strategically transitioning to an approximate transform bounded by a fractional phase tolerance ϵ , the introduction of the communication horizon ($d_{\max}$) deliberately circumvents this mathematical barrier. This demonstrates that long-range interactions are structurally dispensable, successfully localizing the network overhead.

Conversely, this algorithmic reduction is fundamentally complementary to the compiler-level optimization frameworks proposed by Kaur et al. [17]. The approach by Kaur et al. focuses on mitigating network latency by batching and parallelizing the mandatory inter-node interactions required by the circuit. In contrast, our pruning strategy intervenes at an earlier stage of the software stack, drastically reducing the absolute volume of non-local operations before the circuit is ever processed by the scheduler. Consequently, both methodologies are not mutually exclusive but highly synergetic: an optimized compiler can seamlessly apply remote-gate batching techniques over the circuit already simplified by our communication horizon, further maximizing the global efficiency of the distributed architecture.

5. Conclusions

In this work, we have presented a communication-aware optimization of the Inverse Quantum Fourier Transform (iQFT) designed specifically for distributed quantum architectures. By rigorously analyzing the geometric decay of the controlled-phase rotation angles, we introduced a pruning strategy that transforms the algorithm's connectivity requirements from a global all-to-all topology to a bounded nearest-neighbor interaction model.

Our theoretical and empirical analyses demonstrate that establishing a communication horizon, $d_{\max}$, allows for a drastic reduction in the consumption of entanglement resources. Specifically, we showed that the number of EPR pairs per node transitions from a linear scaling $\mathcal{O}(P)$ with respect to the network size to a constant saturation level $\mathcal{O}(1)$, effectively making the resource overhead independent of the total number of nodes in the system. Furthermore, the analysis of the Coupling Ratio (η) confirms that our method successfully decouples the processing nodes, reducing the dependency on high-latency inter-node links to near-zero values while maintaining algorithmic fidelity well within the theoretical error bounds derived.

Ultimately, this approach provides a scalable pathway for implementing large-scale Quantum Phase Estimation and other Fourier-based primitives on near-term networked quantum processors, offering a tunable trade-off between execution fidelity and communication bandwidth.

5.1. Future work

While the current study utilized state-vector simulation to validate the theoretical error models and resource scaling exactitude, several key areas for future investigation. First, it would be valuable to extend the validation of this algorithm to specialized distributed quantum computing (DQC) simulators such as NetQMPI, CUNQA, or CUDA-Quantum. These platforms would allow for a more granular analysis of network-specific effects, including link latency, generation rates, and the impact of classical control delays on decoherence.

Additionally, future work should address the integration of this optimized iQFT schedule with quantum error correction codes suitable for distributed environments, such as surface codes over noisy interconnects, to further suppress the effective error rates in the presence of realistic channel noise.

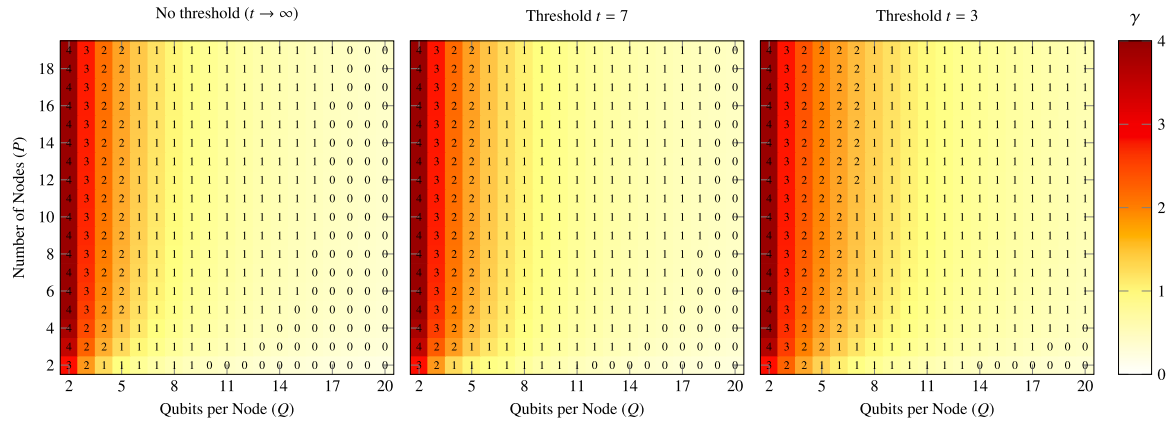


Fig. 15. Global Communication Overhead γ , defined as the ratio between the number of auxiliary communication operations (entanglement generation, swapping, and corrections) and the logical computational gates ($N_{\text{comm}}/N_{\text{comp}}$). This metric indicates the multiplicative factor by which the communication burden exceeds the useful processing work. Note that **higher values are detrimental**, representing a regime where the system execution is dominated by network maintenance rather than algorithmic progress.

CRedit authorship contribution statement

F. Javier Cardama: Writing – review & editing, Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Jorge Vázquez-Pérez:** Writing – review & editing, Validation, Supervision, Software, Methodology, Conceptualization. **Tomás F. Pena:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Andrés Gómez:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT and Gemini in order to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded by the European Union EuroHPC program with grant agreement 101194491 and by MICIU/AEI/10.13039/501100011033 with project number PCI2025-163229, the Agencia Estatal de Investigación (Spain) (PID2022-141623NB-I00), the Consellería de Cultura, Educación e Ordenación Universitaria Galician Research Center accreditation 2024–2027 ED431G-2023/04, and the European Regional Development Fund (ERDF).

Data availability

No data was used for the research described in the article.

References

- [1] M. Caleffi, M. Amoretti, D. Ferrari, J. Illiano, A. Manzalini, A.S. Cacciapuoti, Distributed quantum computing: A survey, *Comput. Netw.* 254 (2024) 110672, <https://doi.org/10.1016/J.COMNET.2024.110672>, URL <https://www.sciencedirect.com/science/article/pii/S1389128624005048>.
- [2] D. Barral, F.J. Cardama, G. Díaz-Camacho, D. Faílde, I.F. Llovo, M. Mussa-Juane, J. Vázquez-Pérez, J. Villasuso, C. Piñeiro, N. Costas, J.C. Pichel, T.F. Pena, A. Gómez, Review of distributed quantum computing: From single QPU to high performance quantum computing, *Comput. Sci. Rev.* 57 (2025) 100747, <https://doi.org/10.1016/J.COSREV.2025.100747>, URL <https://www.sciencedirect.com/science/article/pii/S1574013725000231>.
- [3] H. Bluhm, L.R. Schreiber, Semiconductor spin qubits - a scalable platform for quantum computing? *Proc. - IEEE Int. Symp. Circuits Syst.* 2019-May (2019) <https://doi.org/10.1109/ISCAS.2019.8702477>, URL <https://ieeexplore.ieee.org/abstract/document/8702477>.
- [4] J. Preskill, Quantum computing in the NISQ era and beyond, *Quantum* 2 (2018) 79, <https://doi.org/10.22331/q-2018-08-06-79>, URL <https://quantum-journal.org/papers/q-2018-08-06-79/>.
- [5] J. Cacheiro, Á.C. Sánchez, R. Rundel, G.B. Long, G. Dold, J. Friel, A. Gómez, QMIO: A tightly integrated hybrid HPCQC system, 2025, *Arxiv* 2505.19267, URL <https://arxiv.org/pdf/2505.19267>.
- [6] S. Wehner, D. Elkouss, R. Hanson, Quantum internet: A vision for the road ahead, *Science* 362 (6412) (2018) <https://doi.org/10.1126/SCIENCE.AAM9288>, URL <https://doi.org/10.1126/SCIENCE.AAM9288>.
- [7] D. Ferrari, A.S. Cacciapuoti, M. Amoretti, M. Caleffi, Compiler design for distributed quantum computing, *IEEE Trans. Quantum Eng.* 2 (2021) <https://doi.org/10.1109/TQE.2021.3053921>, URL <https://ieeexplore.ieee.org/abstract/document/9334411>.
- [8] J. Illiano, M. Caleffi, A. Manzalini, A.S. Cacciapuoti, Quantum internet protocol stack: A comprehensive survey, *Comput. Netw.* 213 (2022) 109092, <https://doi.org/10.1016/J.COMNET.2022.109092>, URL <https://www.sciencedirect.com/science/article/pii/S1389128622002250>.
- [9] P. Escofet, A. Das, S.B. Rached, S. Rodrigo, J. Domingo, F. Sebastiano, M. Babaie, B. Keskin, E. Charbon, P.H. Bolívar, M. Palesi, E. Blokhina, B. Staszewski, A. Nag, A. García-Sáez, S. Abadal, E. Alarcón, C.G. Almudéver, On the impact of classical and quantum communication networks upon modular quantum computing architecture system performance, in: 2025 IEEE International Conference on Quantum Computing and Engineering, QCE, IEEE, 2025, pp. 984–995, <https://doi.org/10.1109/QCE65121.2025.00110>, URL <https://ieeexplore.ieee.org/abstract/document/11249763>.
- [10] M. Palesi, E. Russo, G. Ascia, H. Rafique, D. Patti, V. Catania, S. Abadal, A. Das, P. Escofet, E. Alarcón, C.G. Almudéver, Assessing the role of communication in modular multi-core quantum systems, 1, 2025, *arXiv:2510.11053*, URL <https://arxiv.org/pdf/2510.11053>.
- [11] P.W. Shor, Algorithms for quantum computation: Discrete logarithms and factoring, *Proc. - Annu. IEEE Symp. Found. Comput. Sci. FOCS* (1994) 124–134, <https://doi.org/10.1109/SFCS.1994.365700>, URL <https://ieeexplore.ieee.org/document/365700>.
- [12] D. Jiang, X. Liu, H. Song, H. Xie, An survey: Quantum phase estimation algorithms, *IEEE Inf. Technol. Netw. Electron. Autom. Control. Conf. ITNEC* 2021 (2021) 884–888, <https://doi.org/10.1109/ITNEC52019.2021.9587010>, URL <https://ieeexplore.ieee.org/abstract/document/9587010>.
- [13] C.G. Almudever, R. Wille, F. Sebastiano, N. Haider, E. Alarcón, From designing quantum processors to large-scale quantum computing systems, *Proc. -Design*,

- Autom. Test Eur. DATE (2024) <http://dx.doi.org/10.23919/DATE58400.2024.10546849>, URL <https://ieeexplore.ieee.org/abstract/document/10546849>.
- [14] D. Ferrari, S. Carretta, M. Amoretti, A modular quantum compilation framework for distributed quantum computing, *IEEE Trans. Quantum Eng.* 4 (2023) <http://dx.doi.org/10.1109/TQE.2023.3303935>, URL <https://ieeexplore.ieee.org/abstract/document/10214316>.
- [15] F.J. Cardama, J. Vázquez-Pérez, C. Piñeiro, T.F. Pena, J.C. Pichel, A. Gómez, NetQIR: An extension of QIR for distributed quantum computing, *Future Gener. Comput. Syst.* 174 (2026) 107989, <http://dx.doi.org/10.1016/J.FUTURE.2025.107989>, URL <https://www.sciencedirect.com/science/article/pii/S0167739X25002845>.
- [16] S. Sang, L. Hour, Y. Han, Learning-optimized qubit mapping and reuse to minimize inter-core communication in modular quantum architectures, *Phys. Rev. Appl.* 25 (2026) 054023, <http://dx.doi.org/10.1103/rfn7-c5n6>.
- [17] E. Kaur, S. Pouryoucef, H. Shapourian, J. Zhao, M. Kilzer, R. Kompella, R. Nejabati, Optimized quantum circuit partitioning across multiple quantum processors, *IEEE Trans. Quantum Eng.* 6 (2025) 1–17, <http://dx.doi.org/10.1109/TQE.2025.3623158>.
- [18] S. Bahrani, R.D. Oliveira, J.M. Parra-Ullauri, R. Wang, D. Simeonidou, Resource management and circuit scheduling for distributed quantum computing interconnect networks, *IEEE J. Sel. Areas Commun.* XX (2024).
- [19] J. Liu, A. Zang, M. Suchara, T. Zhong, P.D. Hovland, Hardware-software co-design for distributed quantum computing, *Proc. - Des. Autom. Conf.* (2025) <http://dx.doi.org/10.1109/DAC63849.2025.11132538>.
- [20] A. Ebneenasir, Lower bounds for the communication costs of distributing quantum Fourier transform on clique networks, QCE, in: 2025 IEEE International Conference on Quantum Computing and Engineering, vol. 01, 2025, pp. 1003–1014, <http://dx.doi.org/10.1109/QCE65121.2025.00112>.
- [21] D. Coppersmith, An approximate Fourier transform useful in quantum factoring, 2002, *ArXiv:Quant-Ph/0201067*, [arXiv:quant-ph/0201067](https://arxiv.org/abs/quant-ph/0201067), URL <https://arxiv.org/abs/quant-ph/0201067>.
- [22] J.I. Cirac, A.K. Ekert, S.F. Huelga, C. Macchiavello, Distributed quantum computation over noisy channels, *Phys. Rev. A* 59 (6) (1999) 4249, <http://dx.doi.org/10.1103/PhysRevA.59.4249>, URL <https://journals.aps.org/prl/abstract/10.1103/PhysRevA.59.4249>.
- [23] F.V. Mendes, R.V. Ramos, Schemes for teleportation of quantum gates, *Quantum Inf. Process.* 10 (2) (2010) 203–212, <http://dx.doi.org/10.1007/S11128-010-0189-7>, URL <https://link.springer.com/article/10.1007/s11128-010-0189-7>.
- [24] D. Main, P. Drmota, D.P. Nadlinger, E.M. Ainley, A. Agrawal, B.C. Nichol, R. Srinivas, G. Araneda, D.M. Lucas, Distributed quantum computing across an optical network link, *Nat.* 638 (8050) (2025) 383–388, <http://dx.doi.org/10.1038/s41586-024-08404-x>, URL <https://www.nature.com/articles/s41586-024-08404-x>.
- [25] C.H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, W.K. Wootters, Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels, *Phys. Rev. Lett.* 70 (13) (1993) 1895, <http://dx.doi.org/10.1103/PhysRevLett.70.1895>, URL <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.70.1895>.
- [26] D. Gottesman, I.L. Chuang, Quantum teleportation is a universal computational primitive, *Nature* 402 (6760) (1999) 390–393, <http://dx.doi.org/10.1038/46503>, URL <http://arxiv.org/abs/quant-ph/9908010>.
- [27] D. Gottesman, I.L. Chuang, Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations, *Nat.* 402 (6760) (1999) 390–393, <http://dx.doi.org/10.1038/46503>, URL <https://www.nature.com/articles/46503>.