

QPU Networking for High Performance Computing

Aliro



QPU Networking for High Performance Computing

Summary	1
Introduction	1
The Evolution of High Performance Computing	1
Transistors	1
Data Centers	1
Supercomputers	2
Quantum Computers	2
The New Challenges to High Performance Computing	3
Efficiency in Quantum Computing	5
The Quantum + AI Advantage	6
Progress in Quantum Computing is Paved by Quantum Networking	7
How to Network QPUs	9
Key Components of a QPU Network	10
A Sample QPU Networking Program	11
Building a Software-Defined, Reliable Quantum Network: A Vision for QPU Networking	12
Today's Entanglement-based Networks	13
References	14

Summary

Quantum computing will revolutionize industries—from drug discovery and weather modeling to logistics optimization and AI. Reaching this potential will require fault-tolerant quantum computing, a milestone dependent on scalable systems. Entanglement-based quantum networks can pave the way to powerful, scalable quantum computation by interconnecting quantum processing units (QPUs). This white paper outlines the current state of QPU networking, key technological advancements, and the milestones that will shape the future of quantum computing, highlighting a roadmap toward scalable, fault-tolerant systems that will redefine the limits of computation.

Introduction

Quantum computing promises to solve problems of unprecedented complexity and scale, but to realize that potential we first have to address scalability challenges. One way to resolve these challenges is to interconnect quantum processing units (QPUs) and create larger, more powerful systems through entanglement-based quantum networking. These networks enable distributed QPUs to share quantum information, effectively expanding computational capacity and paving the way toward fault-tolerant quantum systems.

The networked integration of QPUs into data centers and high-performance computing environments is no longer a distant vision for the future; it's unfolding today. This progress reflects a broader industry shift: quantum computing has a recognized potential to increase computational power while minimizing the energy consumption required to perform complex computations, like the massive energy consumption required to continue to develop artificial intelligence.

The Evolution of High Performance Computing

Computing has had an amazing journey through the years. The first computers were enormous machines – the size of rooms or even the size of gymnasiums. You've likely seen pictures of people plugging cables and moving punch cards around in these types of early computers. While sizable, they were very weak compared to the computing power available today.

Transistors

The computing revolution really took off with the invention of the transistor at Bell Labs, showing that a binary state could be held in the charge on a capacitor, and it could be controlled quickly. From there, there was scaling up from individual transistors on a single chip to putting many transistors on a single chip. In 1971, the first integrated circuit allowed for an increase in density at a lower cost. Eventually this evolved into the first personal computers.

Data Centers

The next big leap in computation was the creation of data centers in the early 1990s. Data centers were built to offer fast, low cost and reliable computation for millions of people at a time. Data centers are best used for large data applications, with large memory requirements, for doing matrix multiplications and performing similar

computational tasks over and over again. For example, scrolling on Instagram, sending a message, watching a movie streamed to you wherever you are. Data centers excel at performing a lot of simple computations that happen repetitively, many computations that are using the same data that's being stored in the cloud and providing this low cost compute around the world.

The Internet dot com boom was enabled by this technology, providing low cost compute at-scale. Today there are thousands of data centers in the U.S. and around the world. In a data center, optical fiber, switches, and routers perform networking, allowing classical data to flow quickly between the different processors, which are a mixture of processors, but in particular GPUs (graphics processing units), and CPU's (computational processing units), which are the main chips powering the AI boom today. Today, the industry is seeing blurring of the lines between data centers and supercomputers, with data centers beginning to implement blades of processors that are actually co-processors. NVIDIA's A100 GPU and the Blackwell chip, which is actually a combination of GPUs and CPUs on a chip that they call a compute resource are examples of how this technology is being integrated.

Supercomputers

Supercomputers became much more accessible in the late 1990s, and came into the field targeting large data, high speed networking applications that were required for performing complex tasks like weather modeling and simulations of nuclear physics. Supercomputers are also big data devices, but unlike data centers supercomputers are purpose-built and are not general-purpose machines.

Supercomputers have often been built with custom-designed processor units, which are integrated into blades that fit into racks. These systems feature enormous amounts of memory, thousands or millions of processors, and high-speed optical networking, capable of transferring data at tens or even hundreds of gigabits per second. Modern supercomputers combine GPUs and CPUs in carefully optimized configurations in order to achieve maximum performance.

Supercomputers really excel at performing a large number of computations very quickly. They are particularly effective for tasks involving extensive iterations or recursive functions. With high-speed networking enabling efficient data routing, the process is seamless and well-coordinated.

Unlike data centers, which serve millions of people at a time, users book time on a supercomputer: it could be multiple users at the same time, or a single user could use the entire facility. Users will have planned in advance what computations and algorithms are run: it could be for risk assessment in finance, nuclear physics modeling, or even modeling climate change. There are a wide range of these kinds of tasks that supercomputers are best suited for.

While very powerful, supercomputers have a finite lifetime. They generally operate for 3 to 5 years and then they're retired, recycled, and the next generation of supercomputers comes online.

Quantum Computers

Quantum computing was first visioned in 1980 by Feynman. Development accelerated in 2010-2020 with rapid scaling of the number of qubits, or quantum bits, on one processor. Google recently announced is latest quantum processor, Willow, with 105 qubits. There's a number of companies around the world creating quantum computers, and these machines are poised to revolutionize how computation is performed, in what is likely the largest revolution since the transistor. The quantum piece of high performance computing evolution

has only accelerated, and what is unfolding now is a merger and a unification of multiple coprocessors working together.

Quantum computers are not a large data application, because of the limited fan in / fan out from the qubits on the chip, a couple hundred for example. Quantum computing is really impactful when there is an exponential scaling of the mathematical equations describing physics, which could be a chemistry problem or an optimization problem, etc. They are extremely useful for simulations. Quantum simulations are very natural for quantum computers: indeed they are a quantum device. Physicist Richard Feynman proposed that quantum computers could efficiently simulate quantum systems in ways that classical computers could not in his 1981 paper, "Simulating Physics with Computers." Beyond modeling physics, quantum computers are quite helpful in cases where there is no known mathematical model, but there is highly correlated complex data in a small data set.

The New Challenges to High Performance Computing

There are some similarities between data centers, supercomputers, and quantum computers—but each has its particular strengths and specializations. Depending on the data and the computation required, users will choose to use a data center, a high performance compute facility, or a quantum computer.

In reality, algorithm developers and application developers will want to be able to access all of these resources, and combine them in novel ways to solve a problem as fast and cheaply as possible. Over the last 30 years, an incredible amount of improvement and scaling of transistors and processors has created a competitive and diverse computational environment. Moore's Law is an observation made by Gordon Moore, the co-founder of Intel: the number of transistors on a microchip doubles approximately every two years, while the cost per transistor decreases. At the same time, the frequency of the clocks of the processors increased as well, from hertz to megahertz to gigahertz. It's now holding stable from 2 to 3 gigahertz for common processors.



Moore's Law powered classical compute scaling Aliro

As discussed previously, this has driven exponential improvements in computing power, performance, and efficiency over time. Instead of the number of transistors on a single chip increasing, over the last ten years the scaling up of computation has been due to parallelization: having multiple processors in a line that essentially provides more transistors per processing unit.

Moore's Law not only tracked the improvement in processor capability, there was also growth in energy consumption. The chart below shows the electricity consumption on the vertical axis and the year on the horizontal axis showing a fairly gradual, stable increase: predictable energy consumption in terawatt hours per year.



But energy consumption turned Moore's Law Into Moore's Achilles Heel

There's a jump around 2010 as parallelization of processors is used to increase computational power, and as a result there is more heat generated, more cooling required, all resulting in a jump in energy consumption. And so, after about 40 years of steady improvements, Moore's Law has slowly become more of a challenge: how to provide enough power to to scale up the chips. This is currently unfolding in the news around AI and energy consumption. In the last few years there has been an exponential increase in energy consumption due to the scaling of the number of chips, the density of transistors, the power required, the heat removal required to power these parallelized chips.

This energy problem that data centers and supercompute designers are facing is not going to be solved with today's power grid or energy sources. In the past, data centers would invest in cheaper energy sources: solar, hydro, etc. These options are no longer sufficient to bridge the energy consumption gap.

The good news is that there is a multi-pronged solution:

Establish new energy sources, make classical chips more energy efficient, create more efficient algorithms, and use quantum computing. Quantum computers have already been proven to be more efficient for some algorithms than classical computers.

Aliro

Efficiency in Quantum Computing

In 2019, Google performed their quantum supremacy demonstration. There are two metrics that Google looked at in that demonstration: Time to solution advantage and energy efficiency advantage.

The 53-qubit Sycamore quantum processor performed a specific computation in 200 seconds. To perform this same computation with the Summit supercomputer was estimated at approximately 10,000 years to completion. The energy advantage of how much energy was required to run the Sycamore (quantum) computation versus energy required for using the Summit Supercomputer was 2000 times higher: 5 kilowatts versus 13 megawatts is an example, for a figure of merit.



10^10 time to solution advantage 2,000x higher energy advantage (5 kW vs 13 MW)

Quantum computers have demonstrated computations beyond the reach of classical supercomputers. However, advancements in classical algorithms and hardware have closed that gap, sometimes removing the claimed quantum advantage. Yet, quantum computing continues to progress, producing new results. One advantage of quantum computing that continues to endure is the energy efficiency advantage. As shown in the scaling estimates below, quantum algorithms can consume significantly less energy for certain tasks, even when accounting for error correction. Classical computing's energy consumption scales exponentially with problem size N, but quantum computers exhibit more favorable scaling: linear scaling in energy consumption relative to number of qubits but exponential scaling in computation complexity. It's been demonstrated that reliable quantum computations can achieve the same results with much lower energy consumption—a promising development for the future of High Performance Computing.



Problem size (N)

The integration of quantum processing units (QPUs) into data centers and high-performance computing (HPC) environments is on the horizon. Europe and the U.S. are making progress, with initiatives like the Department of Energy's proposal to incorporate QPUs into HPC centers by 2026–2030. These developments aim to leverage quantum computers as accelerators, focusing not only on speed but also on energy efficiency. With the rapid advancements in both quantum and classical technologies, hybrid systems are set to reshape the computational landscape within the next five years.

The Quantum + AI Advantage

Rapid advancements in quantum computing present an opportunity to integrate these systems into hybrid computational environments. This is a call to action: accelerate the adoption of quantum processors to complement classical supercomputing. Quantum computers aren't faster for every type of computation: they excel in specific tasks. For example, a hybrid approach might involve running 90% of an algorithm on a classical supercomputer and the remaining 10% on a quantum processor. This could provide advantages in time-to-solution as well energy efficiency.

An example of how this would be useful is in addressing the AI scaling problem. Recently, two major issues have surfaced regarding the challenges of scaling Artificial Intelligence:

- Data scarcity. Models like ChatGPT and Grok are trained on vast amounts of digital text, images, and videos. As data sources become exhausted, companies are digitizing old records or even using Al-generated data to train new models. This recursive training raises questions about data quality and limits scalability.
- Energy demands. The explosive growth of AI has led to equally explosive energy consumption. News headlines warn of a pending energy crisis. Companies like Google and Microsoft are now turning to nuclear power to sustain the continuing maturation of AI. These developments underscore the critical need for sustainable energy solutions.

Quantum computing could offer a powerful solution to AI's challenges:

- 1. Energy efficiency. Quantum computation can reduce energy consumption compared to classical systems. This could alleviate the energy bottleneck that limits Al's scalability.
- 2. Natural AI. There's a strong hypothesis that suggests quantum computing could offer a fundamentally better way to simulate and replicate the processes of the human brain, particularly its memory and learning mechanisms, due to quantum phenomena observed in biological systems. A quantum computer could train neural networks more efficiently and naturally, reducing the reliance on massive datasets and high energy use.
- 3. Exponential speedup. Quantum computers are capable of radically faster computations for specific algorithms, offering a level of computational power that can't be reached with classical systems alone.

Quantum computing addresses real world challenges, such as AI's scaling challenges, by offering energy efficiency and delivering exponential speedups for specific algorithms.

Progress in Quantum Computing is Paved by Quantum Networking

A large-scale quantum computer will be necessary to perform the kinds of computations discussed previously. Entanglement-based networks are required to build such a large-scale, fault-tolerant quantum computer and also to integrate it into existing and new infrastructure in data centers and in HPC facilities. Many recent developments indicate that progress in quantum computing is paved by quantum networks.

- 1. Notable industry advances, including:
 - PsiQuantum: Plans to network individual photonic computer chips optically. Operating at the telecom wavelength (1.5 microns), PsiQuantum benefits from low-loss fibers, more easily developed lasers, and reliable sources. They've secured a significant contract with the Australian Government to build their first fault-tolerant quantum computer.
 - IonQ: Over the past six months, demonstration of entanglement between ions and photons at the telecom wavelength and ion-to-photon-to-ion entanglement—the building blocks for quantum networks.
 - IBM: Showcased a three-QPU cluster connected via superconducting wires and vacuum tubes.
 While this resembles a multichip wiring setup rather than a robust quantum network, it reflects their commitment to scaling and the need for networking.
- 2. Government and Industry Support
 - DOE: The U.S. Department of Energy has requested proposals for integrating QPUs into supercomputer facilities, signaling that opportunities and programs to network QPUs will be forthcoming.
 - AFRL: The Air Force Research Laboratory launched a program to network different types of QPUs (e.g., IonQ and Rigetti) via entanglement. Aliro is contributing to this effort with simulation and orchestration software to manage the networks.

With these developments, the High Performance Computing industry can expect a surge in announcements related to networking QPUs over the next year, as quantum networking progresses and QPUs become integrated into advanced computing systems.

Achieving a fault-tolerant quantum computer with practical applications will require millions—potentially tens or hundreds of millions—of physical qubits. The exact number depends on the quality of the qubits: their lifetimes and gate fidelities. There are several ways this might be achieved:

- 1. Single chips. Unfortunately, single-chip quantum processors are nearing their practical limits, currently supporting around 1,000 qubits in superconducting, trapped ion, or neutral atom systems.
- 2. Multi-chip integration. Multiple dies can be put on a single chip by using a carrier wafer. This method can increase capacity to around 2,000 qubits.
- 3. Networking for scale. To reach millions of qubits, quantum computers will need coherent networking. Just as classical systems have networks of CPUs and GPUs, future quantum systems will connect hundreds or thousands of quantum processing units (QPUs) via entanglement.

Optical networking required to build fault tolerant quantum computer

Example: Superconducting



True fault-tolerant quantum computing requires logical qubits and error correction at scale. Exciting progress has been made. Companies like Quera and Google have demonstrated multiple logical qubits and are working toward error-corrected systems. Increasing the number of qubits on a chip can also reduce error rates, an important threshold for error correction in quantum computing. However, quantum networking is just as critical to the scaling of quantum systems to the size needed for useful applications. It's an exciting time, with significant progress across the quantum ecosystem. Now is the moment for companies to focus on scaling and networking, laying the groundwork for commercially valuable quantum computing systems.

Aliro

How to Network QPUs

Networking QPUs is critical for scaling quantum computers, but the process depends heavily on the type of QPU and its operating wavelength. There are several major types of QPUs:

- Photonic QPUs (e.g., PsiQuantum)
 Photons are naturally suited for networking due to their ability to travel through optical fibers with low loss. Once high-performance single and two-qubit gates have been demonstrated with hundreds to thousands of qubits, scaling and networking these systems will be relatively straightforward.
- Trapped Ion QPUs (e.g., IonQ, Quantinuum)
 Trapped ions operate at wavelengths between 1.0 and 1.8 microns (e.g., barium and ytterbium), which are well-suited for networking. These systems are easier to network than superconducting qubits, given their optical compatibility.
- Superconducting QPUs (e.g., IBM, Rigetti)
 Superconducting qubits, like aluminum-based systems, operate at far-infrared wavelengths (~6,000 microns), requiring transduction to optical photons for networking. Current transduction efficiency is very low (10⁻⁶ to 10⁻⁸), presenting a significant challenge. Progress is being made, but superconducting QPU networking remains the most complex.

First QPUs to be networked likely laser coupled: Aliro[®] ions and atoms

	QPU Type	Emission Wavelength
High Efficiency	photonic QC	optical (1.5 µm)
	trapped ions (e.g. Ba, Yb)	1.0, 1.8 μm
	neutral atoms (e.g. Rb, Sr)	0.3-0.6 µm
Low Efficiency	superconducting (e.g. Al)	6,000 µm

Key Components of a QPU Network

Early QPU networking efforts will focus on homogeneous QPU networks, which involve linking similar types of quantum processors. The long-term goal is to connect different types of quantum processors. This method leverages the strengths of various QPU architectures. The heterogeneous model recognizes that different quantum processors will specialize in distinct tasks, making their integration essential for unlocking the full potential of quantum computing. The components of a quantum network that interconnects QPUs are:

- Entanglement. Entanglement is at the heart of quantum networking. Unlike classical networking, which connects independent devices, entanglement enables QPUs to operate as a single, larger quantum system. This allows for distributed quantum gates and circuits across multiple processors. Without entanglement, a networked quantum system cannot scale into a fault-tolerant quantum computer.
- Controller. The controller is a critical component that handles the distribution and synchronization of entanglement across the network. It ensures that all processors remain coherently linked and manage the execution of quantum operations seamlessly.
- Optical Fiber. Optical fibers are a straightforward part of the network, as long as they operate at low-loss wavelengths. These fibers facilitate the exchange of photons, which carry entanglement between QPUs.
- Transduction. Transduction involves converting qubit signals to optical photons and vice versa. For heterogeneous networks, efficient transduction is essential to link QPUs operating at different wavelengths. While current transduction efficiencies are low, progress is being made in improving this technology.



A Sample QPU Networking Program

This sample program of how networking QPUs might be accomplished emphasizes key steps and components, starting with heterogeneous qubit platforms, such as trapped ions, and leveraging simulation tools to design and refine networks before physical implementation.

Simulation is a critical first step in planning a quantum network or building a quantum computer. It enables researchers to evaluate the performance of individual components and the system as a whole. Key elements to simulate include:

- Qubits. Assess the performance of qubits on a chip.
- Transducers. Model how efficiently transducers convert signals between microwave, optical, and telecom domains.
- Optical fibers. Simulate loss and fidelity in photon transmission through fibers connecting the QPUs.

Simulators are powerful design tools when used to guide the selection, design, and fabrication of components. Simulators can also provide helpful insights into whether a QPU network is feasible, given the specific challenges of a project.

After simulation and design, the next step is to implement and operate the network of QPUs. This stage requires setting up the physical infrastructure, establishing entanglement, and ensuring the system can reliably execute quantum operations and quantum gates. As quantum networks become more sophisticated, orchestration plays a critical role in transforming diverse quantum devices into a cohesive system. This includes managing entanglement distribution, coordinating quantum gates, and ensuring reliable operation. Orchestration provides a scalable network infrastructure, just like the classical Internet. With a scalable network in place, the focus then shifts to solving real-world problems. The ability to reliably run circuits and algorithms across multiple QPUs opens the door to computing for complex, industry-specific challenges like optimization, cryptography, and quantum simulations.

Building a Software-Defined, Reliable Quantum Network: A Vision for QPU Networking

Creating a successful network of QPUs with an entanglement-based quantum network requires a system that is software-defined, reliable, and adaptable to changes in the system, such as loss of entanglement or fluctuations in photon generation rates. To achieve this, network operators must be able to view, manage, and set up entanglement across QPUs effectively.

The first step in QPU networking is ensuring robust and consistent entanglement across multiple nodes. This involves setting up and orchestrating high-quality entanglement across the network, and implementing transducers to convert signals from optical frequencies (used for networking) to other domains, such as microwave frequencies for superconducting qubits. Each QPU could require multiple transducers in order to maintain efficient communication. Achieving this milestone will be critical for executing gates and running initial quantum circuits.



Once entanglement is established, the next milestone will be demonstrating basic quantum operations across a networked system. Running basic circuits across multiple QPUs will serve as a proof of concept, akin to a "Hello, World" demonstration for quantum networks. This demonstration is likely about 12 to 18 months away, signaling rapid progress in the field.



After demonstrating these basic operations, the focus shifts to scaling the network for more complex algorithms and real-world applications. Milestones here would include:

- Ensuring gates operate with high fidelity and reliability over time,
- Running application-specific algorithms across the network to solve meaningful problems.
- Showing that multiple QPUs can work together to achieve significant speedups and computational advantages over standalone systems.



<SLIDE 22 bottom>

The ultimate goal is to transition from simple demonstrations to fully functional quantum computers, built from networked QPUs.

Today's Entanglement-based Networks

Entanglement-based quantum networks are being built today by a variety of organizations for a variety of use cases – benefiting organizations internally, as well as providing great value to an organization's customers. Telecommunications companies, national research labs, and systems integrators are just a few examples of the organizations Aliro is helping to leverage the capabilities of quantum secure communications.

Building entanglement-based quantum networks that use entanglement is no easy task. It requires:

- Emerging hardware components necessary to build the network.
- The software necessary to design, simulate, run, and manage the network.
- A team with expertise in the fundamental science of entanglement-based quantum networks and classical networking.
- Years of hard work and development.

This may seem overwhelming, but Aliro is uniquely positioned to help you build your quantum network. The steps you can take to ensure your organization is meeting the challenges and leveraging the benefits of the quantum revolution are part of a clear, unified solution already at work in networks like the EPB Quantum Network[™] powered by Qubitekk in Chattanooga, Tennessee.

AliroNet[™], the world's first full-stack entanglement-based network solution, consists of the software and services necessary to ensure customers will fully meet their advanced secure networking goals. Each component within AliroNet[™] is built from the ground up to be compatible with entanglement-based networks of any scale and architecture. AliroNet[™] is used to simulate, design, run, and manage quantum networks as well as test, verify, and optimize quantum hardware for network performance. AliroNet[™] leverages the expertise of Aliro personnel in order to ensure that customers get the most value out of the software and their investment.

Depending on where customers are in their quantum networking journeys, AliroNet[™] is available in three modes that create a clear path toward building full-scale entanglement-based secure networks:

(1) Emulation Mode, for emulating, designing, and validating entanglement-based quantum networks,

(2) Pilot Mode for implementing a small-scale entanglement-based quantum network testbed, and

(3) Deployment Mode for scaling entanglement-based quantum networks and integrating end-to-end applications. AliroNet[™] has been developed by a team of world-class experts.

To get started on your Quantum Networking journey, reach out to the Aliro team for additional information on how AliroNet[™] can enable secure communications.

www.alirotech.com

References

Wikipedia contributors. "Transistor Count." Wikipedia, https://en.wikipedia.org/wiki/Transistor_count

Ritchie, Hannah. "Moore's Law." Our World in Data, https://ourworldindata.org/moores-law

Latitude Media. "Mapping the Data Center Power Demand Problem in Three Charts." Latitude Media, <u>https://www.latitudemedia.com/news/mapping-the-data-center-power-demand-problem-in-three-charts</u>

Quantum supremacy using a programmable superconducting processor - Scientific Figure on ResearchGate. Available from:

https://www.researchgate.net/figure/Demonstrating-quantum-supremacy-a-Verification-of-benchmarking-metho ds fig4_336744162

Auffèves, Alexia, and Olivier Ezratty. *Image by Alexia Auffèves and Olivier Ezratty*. 2022. *Pasqal*, <u>https://www.pasqal.com/news/quantum-computing-rethinking-energy-consumption/</u>

Kindig, Beth. "AI Power Consumption: Rapidly Becoming Mission-Critical." Forbes, 20 June 2024, https://www.forbes.com/sites/bethkindig/2024/06/20/ai-power-consumption-rapidly-becoming-mission-critical

Cohen, Ariel. "Al Is Pushing the World Toward an Energy Crisis." Forbes, 23 May 2024, https://www.forbes.com/sites/arielcohen/2024/05/23/ai-is-pushing-the-world-towards-an-energy-crisis/

Gelles, David. "A.I.'s Insatiable Appetite for Energy." The New York Times, 11 July 2024, <u>https://www.nytimes.com/2024/07/11/climate/artificial-intelligence-energy-usage.html</u>

Halper, Evan. "Amid Explosive Demand, America Is Running out of Power." The Washington Post, 7 Mar. 2024, <u>https://www.washingtonpost.com/business/2024/03/07/ai-data-centers-power/</u>