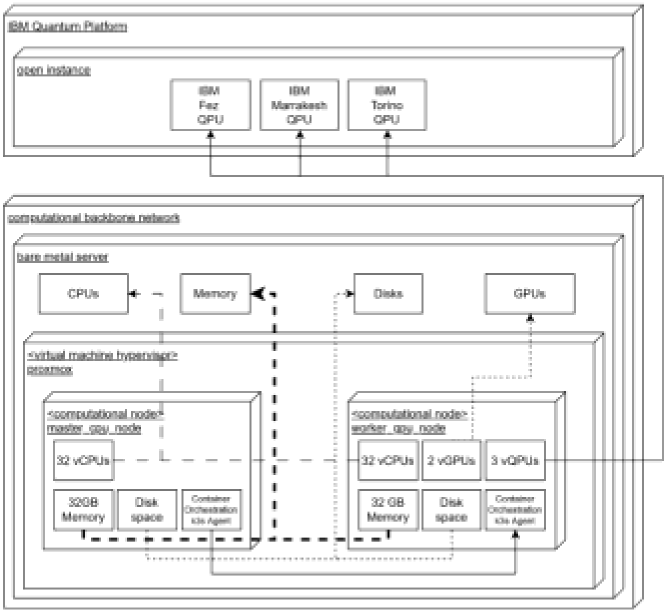
Leveraging cloud-native infrastructure for dynamic and flexible quantum-classical MLOps

Angelo Impedovo[1], **Vito Nicola Losavio**[1], **Corrado Loglisci**[1,2]

1. *Università degli Studi di Bari "Aldo Moro", Department of Computer Sciences, via E. Orabona, 4, Bari, 70125, Italy*
2. *Consorzio Interuniversitario Nazionale per l'Informatica, CINI-Laboratory HPC-KTT, via E. Orabona, 4, 70125, Italy*

Abstract: Quantum classical ML workflows demand dynamic coordination of heterogeneous resources, which traditional MLOps cannot provide. The proposed Kubernetes based framework enables elastic hybrid execution, seamless quantum resource allocation, and reproducible operations across unified cloud native infrastructures.

Quantum Machine Learning promises gains in speed, memory efficiency, and expressivity for high-dimensional tasks, but integrating quantum algorithms into real workflows requires coordinating CPUs, GPUs, simulators, and remote QPUs—capabilities missing in classical-oriented MLOps frameworks [1]. Reproducibility is further complicated by quantum noise, device variability, and hardware-dependent encoding, highlighting the need for a modular, hardware-agnostic infrastructure that unifies classical and quantum development. Such an MLOps system must support heterogeneous backends, multiple programming models including gate-level and pulse-level quantum programming, and strong data and model versioning enriched with quantum-specific metadata. It also requires experiment tracking, inference support across hardware types, concurrent pipeline orchestration, and CI/CD integration to produce reproducible containerized artifacts.



To address these needs, the proposed framework adopts a Kubernetes-based cloud-native architecture with distributed CPU/GPU nodes and remote QPU access via cloud gateways. Ephemeral containers provision hardware quotas on demand, namespaces isolate development stages, and custom Device Plugins expose virtual QPUs for transparent access to quantum backends. Prefect, Minio, and MLflow provide orchestration, storage, and experiment tracking, while quantum circuits are packaged as containerized artifacts for dynamic scheduling on quantum-capable nodes. The framework was validated on a spam-classification task using two contrasting models: the Quantum Optical Neuron and the Quantum Tree Tensor Network, which relies on quantum circuits for training and inference using angle encoding. Running on a K3s cluster with GPU-accelerated simulation and virtual QPUs linked to IBM systems, the infrastructure automatically routed tasks to classical or quantum nodes as needed, providing a consistent developer experience despite differing computational requirements.

References

[1] Impedovo, A., Losavio, V.N. & Loglisci, C. Leveraging cloud-native infrastructure for dynamic and flexible quantum-classical MLOps. *Quantum Mach. Intell.* **8**, 8 (2026). https://doi.org/10.1007/s42484-026-00360-1