A Framework for Digital Twins of Future Quantum Clouds

Waylon Luo[1], Travis Humble[2], Jiapeng Zhao[3], Tong Zhan[4], Qiang Guan[1]
[1]Kent State University, [2]Oak Ridge National Laboratory, [3]Cisco, [4]Meta

SC25 ST. LOUIS NOV 16–21

GUAN SYSTEM LABORATORY

Introduction

- **A quantum cloud** is a portal where anyone with credentials can access a backend quantum computer via the internet.
- **A digital twin of a quantum cloud** represents the digital state of a quantum cloud in real time.
- Our **motivation** is to resolve scheduling and orchestration problems in quantum clouds at the administrative level.
- We developed a **discrete event simulation framework** that simulate hardware setup and workflow of quantum clouds.
- The framework is modular and extensible. Our framework users can implement custom workflow, scheduling and allocation algorithm.

Demonstration & Use Case

- IBM Quantum Cloud Setup [1]
- Monitored job throughputs and average wait time of jobs.
- Adaptive Job Scheduling using reinforcement learning [2]
- HPC + QPU hybrid cloud simulation for workload management [3]

Ecosystem of the Framework

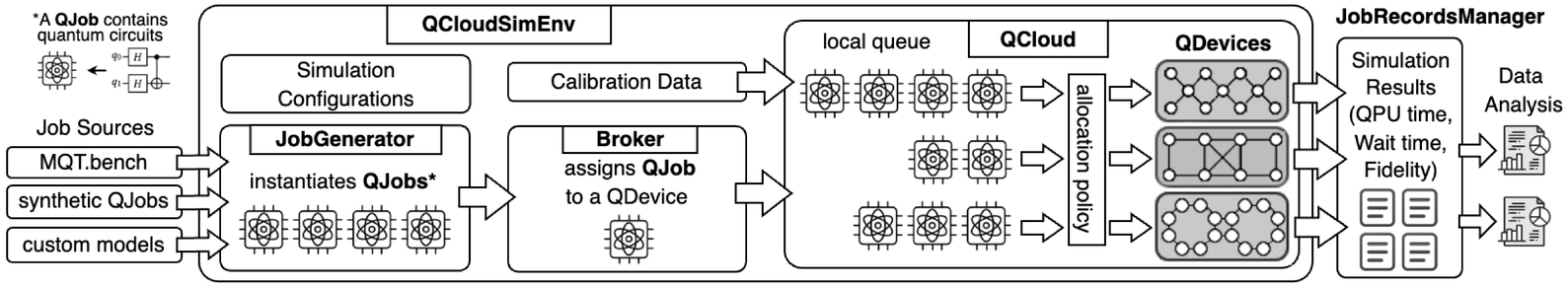

Figure 2. An overview of the simulation workflow illustrates the lifecycle of a quantum job. The job pipeline starts on the left and ends on the right.

Simulation Architecture

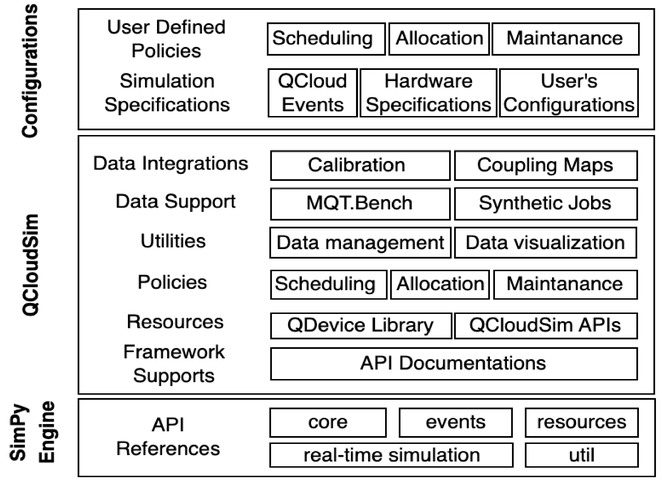

Figure 3. System Design Layers [1, 2]

Example code snippet

```
from QCloud import *

ibm_kyiv = IBM_Kyiv(env=None, name="ibm_kyiv", printlog=True)
qcloudsimenv = QCloudSimEnv(devices=[ibm_kawasaki],
                broker_class=ParallelBroker,
                job_feed_method="generator",
  job_generation_model=lambda: random.expovariate(lambd=2))
qcloudsimenv.run(until=100)
```

Listing 1. Example code

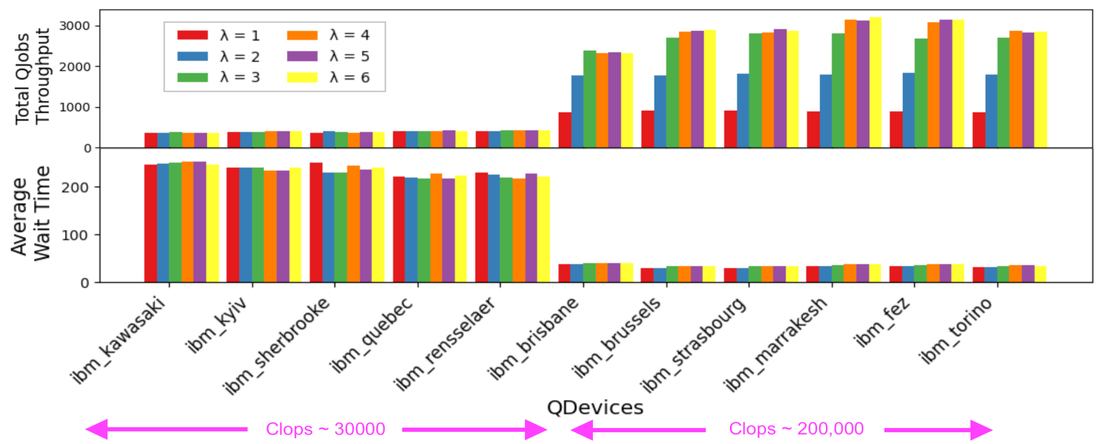

Figure 6. Total jobs processed and comparison of average wait time across various IBM quantum devices with different job arrival rates Lambda λ based on the exponential distribution random.expovariate(λ). The top plot illustrates that the total jobs processed also rise with higher λ values, with λ = 5 leading to the most jobs processed in each quantum device. The bottom plot shows that as λ increases, the average wait time increases significantly across all quantum devices. [1]

A Digital Twin of IBM Quantum Cloud

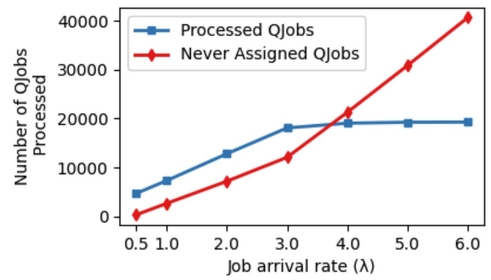

Figure 5. Comparison of processed and not processed jobs for different job arrival rates (λ). [1]

Adaptive Job Scheduling Using Reinforcement Learning for Large Quantum Circuits

- Large circuits that require multiple QPU devices can be handled by RL scheduling methods.
- During training, the scheduling action is executed and a reward is granted, depending on the state of the quantum cloud and the job parameters.
- Rewards can be fidelity, processing speed or both.

Mode	T_{sim} (s)	$\mu_F \pm \sigma_F$	T_{comm} (s)
speed	108 775.38	0.65332 ± 0.01438	5 707.80
fidelity	209 873.02	0.68781 ± 0.02605	3 822.74
fair	108 978.16	0.64373 ± 0.01478	5 685.30
rlbase	106 206.21	0.62087 ± 0.01301	6 105.52

Table 1. Performance of allocation strategies on 1,000 large circuits [2].

SCAN ME

All My Links

References (Papers):
[1] Luo, Waylon, Betis Baheri, Travis Humble, Jiapeng Zhao, Tong Zhan, Rajan Maharjan, and Qiang Guan. "A Digital Twin of Scalable Quantum Clouds." In 39th ACM SIGSIM Conference on Principles of Advanced Discrete Simulation, pp. 165-175. 2025. https://dl.acm.org/doi/10.1145/3726301.3732296
[2] Luo, Waylon, Jiapeng Zhao, Tong Zhan, and Qiang Guan. "Adaptive Job Scheduling in Quantum Clouds Using Reinforcement Learning." In 54th International Conference on Parallel Processing. 2025. https://dl.acm.org/doi/10.1145/3754598.3754641
[3] Waylon Luo, Cheng-Chang Lu, Tong Zang, and Qiang Guan. "A Simulation Framework for Workload Management in Hybrid Quantum-HPC Cloud System" In Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2025.