ICS 35.240.01
CCS L 70

NATIONAL STANDARD OF THE PEOPLE'S REPUBLIC OF CHINA

GB/T 47177.2-2026

Quantum computing service platform - Part 2: Performance evaluation

量子计算服务平台 第2部分：性能评估

Issued on: April 30, 2026 **Implemented on: November 1, 2026**

**Issued by: State Administration for Market Regulation;
Standardization Administration of the PRC**

Table of Contents

5 Performance Indicators

The performance of a quantum computing service platform is jointly characterized by the performance of the quantum computer system, the performance of the classical simulator for quantum resources, and the service performance. In this context, the performance of quantum computer systems and classical simulators for quantum resources corresponds to the quantum computing infrastructure layer in GB/T 47177.1-2026. Physical resources and service performance correspond to the platform services of the platform layer in GB/T 47177.1-2026, as shown in Figure 1. Quantum computer system performance. This measures the characteristics of the quantum computer system upon which a quantum computing service platform is based, including qubits. Indicators such as number of operations, connectivity, bit coherence, operation fidelity, and operation time. Quantum resource classical simulator performance. This measures the characteristics of the classical simulator resources upon which the quantum computing service platform is based, including classical... Indicators such as the number of simulators, bit size, maximum circuit depth, and quantum circuit simulation efficiency. Service performance. Used to measure the characteristics of cloud computing services that a quantum computing service platform can provide, including task response time, task throughput, and other performance metrics. Metrics include load balancing recovery time, number of available services, throughput, and new connection rate.

6.1 Performance Testing of Quantum Computer Systems

6.1.1 Performance Indicators The performance requirements of quantum computing service platforms for quantum computer systems mainly include the following indicators.

a) Number of qubits.

b) Connectivity.

c) Bit coherence.

---Longitudinal relaxation time;

---Lateral relaxation time.

d) Operational fidelity.

---Read fidelity;

---Single-bit gate fidelity;

---Two-bit gate fidelity.

e) Operation time.

---Single-bit gate operation duration;

---Two-bit gate operation duration.

6.1.2 Test Method According to GB/T 47468-2026.

6.2 Performance Testing of the Classical Simulator for Quantum Resources

6.2.1 Number of Classic Emulators The number of classical simulators refers to the number of classical simulators that a quantum computing service platform can provide for parallel execution. This is defined here. This refers to the number of full-amplitude simulators supported. The testing steps for the number of classic simulators are as follows:

a) Under a limited simulation computation method, evaluate the highest quantum level that each individual simulator in the evaluation system can support when performing quantum simulations. The number of bits (basically determined by the upper limit of memory capacity), denoted as Nm;

b) Due to limitations in system scheduling and storage capacity, the standard for the number of running qubits is set to the number of qubits that can be executed in parallel. Eye, denoted as Nm-2;

c) Open multiple emulators and run them simultaneously on random lines;

d) Gradually increase the number of parallel processes, ensuring the system can run stably simultaneously without memory errors or resource queuing. The maximum number of successful parallel runs is the number of classic simulators. Example. A classic simulator can simulate a maximum of 35 qubits at full amplitude by default. During testing, 33 qubits will be activated in parallel. The full-amplitude simulation calculation is performed using a randomly sampled circuit library. The maximum number of parallel executions the system can support at this point is equal to the number of classic simulators. This is limited by... Due to system scheduling and storage limitations, the actual number of classic emulators that can run may be lower than the theoretical maximum. Theoretically, it can support four classic emulators simultaneously, but in practice... Typically, a maximum of two classical simulators are supported. Therefore, if several classical simulators are deployed within a quantum computing service platform, they can all perform parallel testing simultaneously. The number of concurrently running simulators on the final system is the maximum number of classic simulators.

6.2.2 Bit Scale The bit size refers to the highest number of qubits that a classical simulator provided by a quantum computing service platform can accept for a task. The approximate solution method that does not support problem partitioning or subspace partitioning is not supported. For a given classical simulator, the maximum number of qubits supported is... The larger the value, the stronger the computing power. The test steps for bit-scale are as follows:

a) Select a single type of classical simulator provided by the quantum computing service platform as the test object;

b) Construct a randomized benchmark quantum circuit with circuit parameters of number of qubits N and circuit depth D, where N starts from an initial value N0. Initially, increment by step size DeltaN, while keeping D at a fixed value of D0;

c) Start and initialize the classic simulator under test, select the execution backend (CPU or GPU), and disable optimizations such as approximation, truncation, and chunking. Options;

d) Perform quantum circuit simulation for each value of N and output the final state sampling result, recording the running state;

e) Gradually increase N until the first "out of memory" or "allocation failure" error occurs. Record the maximum number of bits without an error as. Nmax is the bit size.

6.2.3 Maximum Line Depth Maximum line depth refers to the maximum depth of the simulated lines supported by the classical simulator provided by the quantum computing service platform. The test procedure for maximum line depth is as follows:

a) Select a single type of classical simulator provided by the quantum computing service platform as the test object;

b) Construct a randomized benchmark quantum circuit with the parameters of N qubits and D layers, where N remains a fixed value. N0,D starts from the initial value D0 and increases by a step size DeltaD;

6.3 Service Performance Testing

6.3.1 Task Response Time Task response time refers to the time taken by a quantum computing service platform to resolve, schedule, and execute computing tasks. The shorter the time, the better. The stronger the performance of the service platform, the better. The steps for testing task response time are as follows:

a) Submit a computational task to the quantum computing service platform;

b) The platform responds to the task request and records the time point t1 at this point;

c) The platform processes and executes the task until it is completed, and records the time point t2 at this point;

d) Extract the task parsing time (Tparse), task scheduling time (Tschedule), resource switching time (Thandoff), and actual time from the tool logs. The raw data of Toperation for task execution time;

e) Calculate the task response time according to formula (1).

6.3.2 Task Concurrency Task concurrency refers to the ability of a quantum computing service platform to process multiple quantum computing tasks simultaneously, reflecting the system's ability to handle multiple tasks at the same time. The maximum number of tasks that can be handled within a given time period. The steps for testing task concurrency are as follows:

a) Use log analysis methods to collect the service platform's operational logs, including task request logs, system operation logs, etc., and statistically analyze them. The number of task requests processed within a given time period, and the task execution time, etc.

b) Analyze the service platform's operation logs, outputting the peak number of task requests processed per unit of time (QPSpeak), and the actual number of requests executed per unit of time. Find the average time T.

c) Calculate the task concurrency according to formula (2).

6.3.3 Load Balancing Recovery Time Load balancing recovery time refers to the time a quantum computing service platform can recover from node failures in tasks involving high concurrency and resource allocation. Or, under conditions of uneven load, the time required for the system to recover to a balanced state. The test steps for load balancing recovery time are as follows:

a) Use the load balancing monitoring tool built into the quantum computing service platform to pre-configure monitoring parameters (including but not limited to load balancing). (Data such as node CPU utilization, memory usage, task request response latency, and backend server liveness status) ensures the tool can function effectively. Collect and record the above data in a timely manner;

b) Simulate node failure manually and record the time point Tfp of the failure occurrence using monitoring tools;

c) Continuously monitor the system status until the load balancing state is restored to normal system operation, and record the time point Trp at this point;