Quantum random-access memory

Understand coherent data lookup and include its construction and loading costs.

Undergraduate 18 min lesson Dr. Charith Jayasekara

By the end, you'll be able to

- Trace a coherent address–data query.
- Distinguish basis encoding from amplitude encoding.
- Include loading and maintenance in a cost comparison.

Before you begin: Reversible circuits, CNOT, superposition and entanglement.

What does coherent lookup mean?

A classical memory query takes an address and returns the value stored there. A quantum memory interface must preserve a superposition of addresses while relating each address to its corresponding value. This interface is often called quantum random access memory, or QRAM.

For a fixed table D of classical bit strings, a useful reversible definition is

$$U_D|a\rangle|b\rangle = |a\rangle|b \oplus D_a\rangle.$$

The first register holds an address a. The second is a data bus b, and $\oplus$ denotes bitwise XOR. XOR makes the transformation reversible: applying the same lookup twice returns the original bus. A bus initially in zero receives the addressed table entry.

Linearity determines what happens to a superposition:

$$U_D\left(\sum_a \alpha_a|a\rangle|0\rangle\right) = \sum_a \alpha_a|a\rangle|D_a\rangle.$$

This equation describes coherent correlations. Measuring once yields one address–value pair. An algorithm gains something useful only if its subsequent transformations exploit the state before measurement.

Checkpoint. Does the bus alone contain a readable printout of every entry? No. Read the joint state and ask what a particular measurement can return. Interference and further processing determine which global properties can be extracted efficiently.

Worked example: build a small lookup exactly

Use two address qubits, written a_1a_0, and one bus qubit. Store the four-bit table

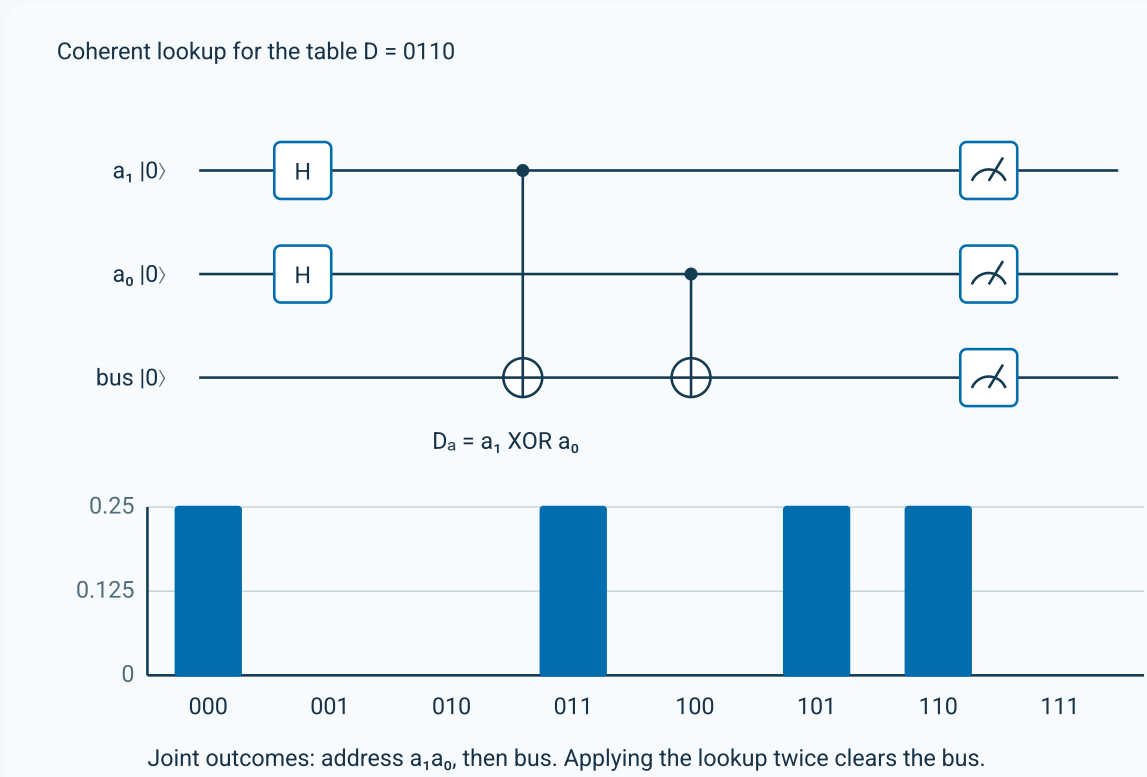
Address	00	01	10	11
D_a	0	1	1	0

Here the table is the parity function $D_a = a_1 \oplus a_0$. A CNOT from a_1 to the bus, followed by a CNOT from a_0 to the bus, implements the lookup. Two Hadamards prepare uniform addresses. The output is

$$\frac{|00\rangle|0\rangle + |01\rangle|1\rangle + |10\rangle|1\rangle + |11\rangle|0\rangle}{2}.$$

All four joint outcomes have probability $1/4$. The bus alone gives zero or one with equal probability. Its correlation with the addresses identifies the table.

This table has a short algebraic rule, which is why two CNOTs suffice. An arbitrary four-entry table may need a different circuit. Scaling this tiny parity example does not establish a cheap implementation of arbitrary large datasets.



A two-CNOT coherent lookup circuit and its four nonzero joint outcome probabilities.

The plot writes outcomes in wire order a_1, a_0, b. Its four nonzero bars, 000,011,101,110, each have probability $1/4$. These are the allowed address–bus pairs after one lookup.

Use and uncompute the data

Often the data is needed only temporarily. Suppose the bus controls a phase operation, after which the lookup is applied again. The second call clears the bus while retaining a useful phase on the addresses. With the bus prepared in $|-\rangle$, phase kickback gives the shorter identity

$$U_D|a\rangle|-\rangle = (-1)^{D_a}|a\rangle|-\rangle.$$

For the table above, this assigns signs $+, -, -, +$ to the uniform address state. Applying H again to both address qubits yields $|11\rangle$. The known parity rule has become a measurable interference pattern. This worked example links a memory interface to the oracle model used in algorithms.

Architecture and data preparation are different costs

QRAM architecture proposals include routing networks in which an address coherently directs a bus through a tree. An address of n bits refers to $N = 2^n$ cells. A short route through a tree does not remove the physical cost of supplying those cells, loading their contents, controlling switches, maintaining coherence and managing errors. The original bucket-brigade proposal distinguishes the number of elements installed from the elements activated during a query.

A compiled quantum read-only memory circuit, often called QROM, provides another way to implement a fixed classical lookup through gates. Its table-dependent gate count, auxiliary space and depth must be accounted for. Coherent access to arbitrary stored quantum states raises additional questions about the allowed memory operation and whether values are moved, swapped or otherwise accessed.

Also distinguish lookup from amplitude encoding. Preparing

$$|x\rangle = \frac{1}{\|x\|} \sum_a x_a |a\rangle$$

requires amplitudes proportional to data values. A lookup that writes x_a into a separate register is one ingredient in some preparations, rather than automatically the completed amplitude-encoded state.

Exercises

Change one table entry in a simulator and rebuild the reversible circuit. Check all four basis addresses before using superpositions. Then run a superposition, apply the same lookup twice and verify that the bus returns to zero. Basis-address correctness and coherent reversibility test different parts of the implementation.

1. For address 10 and bus 1, what does the example lookup return?
2. A speedup assumes unit-cost access to a million-entry table. What must an end-to-end comparison add?

> Worked answers
>
> 1. address 10 and bus 0, since $1 \oplus D_{10} = 0$.
> 2. data ingestion, memory construction, query implementation, updates, error management and the final output task.

Check your understanding

PAUSE & THINK

What does one coherent query guarantee?

- ☐ The defined address–data transformation, on every superposed branch
- ☐ Every table entry appears as readable classical output in one shot
- ☐ Loading and maintaining the memory costs nothing

Answer: The defined address–data transformation, on every superposed branch. Subsequent processing decides what information the computation extracts.

QRAM is a coherent memory interface. A credible algorithm comparison states how that interface is built and how its data reaches the machine.

Sources

Giovannetti, Lloyd and Maccone, QRAM proposal; the authors' architecture analysis. The parity table, gate circuit, uncomputation exercise and amplitude-encoding comparison above are original teaching material.