Accelerated Quantum-Enhanced Memetic Search (A-QEMS)

Arife Nur Ayaz | IQuHack, Feb 1, 2026

Code Repository

Path to Optimization

The Optimization Paradox

Traditional optimization landscapes are "glassy," meaning they are densely populated with high-energy local minima that act as traps for classical heuristic searches.

Limitations of Classical Metaheuristics

Classical heuristics currently scale at approximately $O(1.34^N)$, which severely limits the sequence lengths (N) that can be solved within practical timeframes + Stagnation in Local Minima.

Phase-Space Compression

Before applying any search logic, we must recognize that the search space is highly redundant due to bit-flip and reversal symmetries.

The Theory

Non-Hermitian Quantum Tunneling

Localized Stagnation in Glassy Landscapes

LABS landscapes are characterized by deep, rugged local minima where classical solvers become "localized" or trapped.

Hatano-Nelson "Quantum Kicks"

Non-Hermitian Driver: We replace the standard driver with a Hatano-Nelson Hamiltonian, introducing an imaginary gauge potential (drift parameter δ).

The Delocalization Transition: By mapping the energy gradient $\nabla E(s)$ to the drift parameter, we induce a non-Hermitian Skin Effect.

Barrier Thinning: This effectively "thins" the high-energy barriers surrounding local minima, allowing the population to "tunnel" through them rather than climbing over them.

Wavefunction Localization

Hatano-Nelson model

The Implementation

Unitary Embedding:

To run non-unitary evolution on NVIDIA GPUs, we use Naimark Dilation, adding one ancilla qubit to double the Hilbert space.

Augmented Operator:

We evolve the system under a dilated unitary and post-select on the ancilla being in state $|0\rangle$.

Gradient-Field Coupling :

The energy gradient $\nabla E(s)$ is mapped to the longitudinal potential, while the drift parameter δ drives the directional tunneling.

Wavefunction Localization

Hatano-Nelson model

Naimark Dilation Circuit

Technical Stack

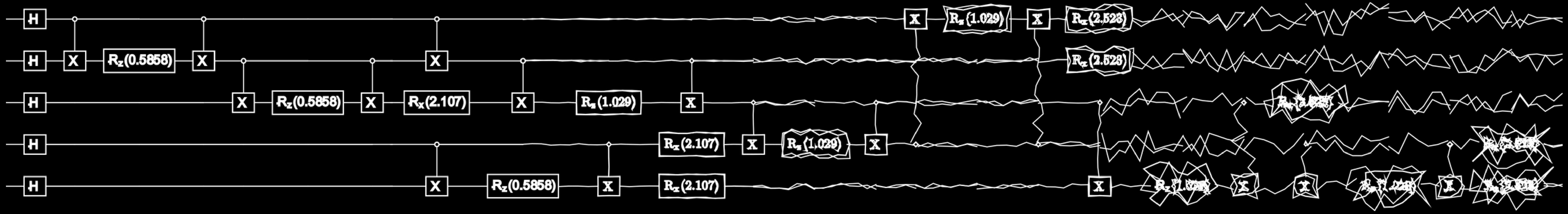
Tensornet Acceleration : We utilize the CUDA-Q tensornet backend to manage the resulting entanglement with a bond dimension up to $\chi = 64$.

CUDA-Q

Platform for Programming Accelerated Quantum Supercomputers

Libraries, Applications, Frameworks

NVIDIA CUDA-Q	**cudaq.kernel , cudaq.qvector (Unified Register), cudaq.sample .**
CuPy	**Used for GPU-accelerated linear algebra and array manipulations on the NVIDIA L4/A100.**
Hardware Context	**NVIDIA L4 / A100 Tensor Core GPUs : The target hardware architectures optimized via dynamic bond dimensioning.**



The Plan & The Pivot

Original Strategy

Our initial proposal aimed to solve the Low Autocorrelation Binary Sequence (LABS) problem by simulating the **Hatano-Nelson Hamiltonian**—a non-Hermitian model that exhibits a "skin effect" to tunnel through energy barriers.

The Obstacles

- The "Zero Success" Wall: Our initial quantum kernel yielded a persistent 0.00% success rate for the post-selection of the auxiliary qubit, effectively filtering out all data.
- Classical Instability: Our classical fallback mechanism, based on Symplectic (Hamiltonian) dynamics, was physically conserving energy rather than minimizing it, leading to oscillations instead of solutions.
- Backend Ambiguity: We faced critical issues with bitstring ordering (MSB vs LSB) and missing measurement bits when using separate qubit registers in CUDA-Q.

The Pivot

A New Heterogenous Architecture

1. Quantum Architecture: We moved from split-register allocation to a unified qvector approach, guaranteeing consistent measurement outcomes. We introduced an interference stage (`H` gate) to enable the Linear Combination of Unitaries (LCU) logic required for Naimark dilation.

2. Physics Correction: We replaced the pure Hamiltonian fallback with a **Damped Pseudo-Langevin Solver** (Gradient Descent with Momentum), enforcing energy dissipation to guarantee ground-state convergence.

3. Diagnostic-First Approach: We built custom debug suites to probe the backend internals, revealing that the Ancilla qubit was hiding at the LSB (Index -1), contrary to standard assumptions.

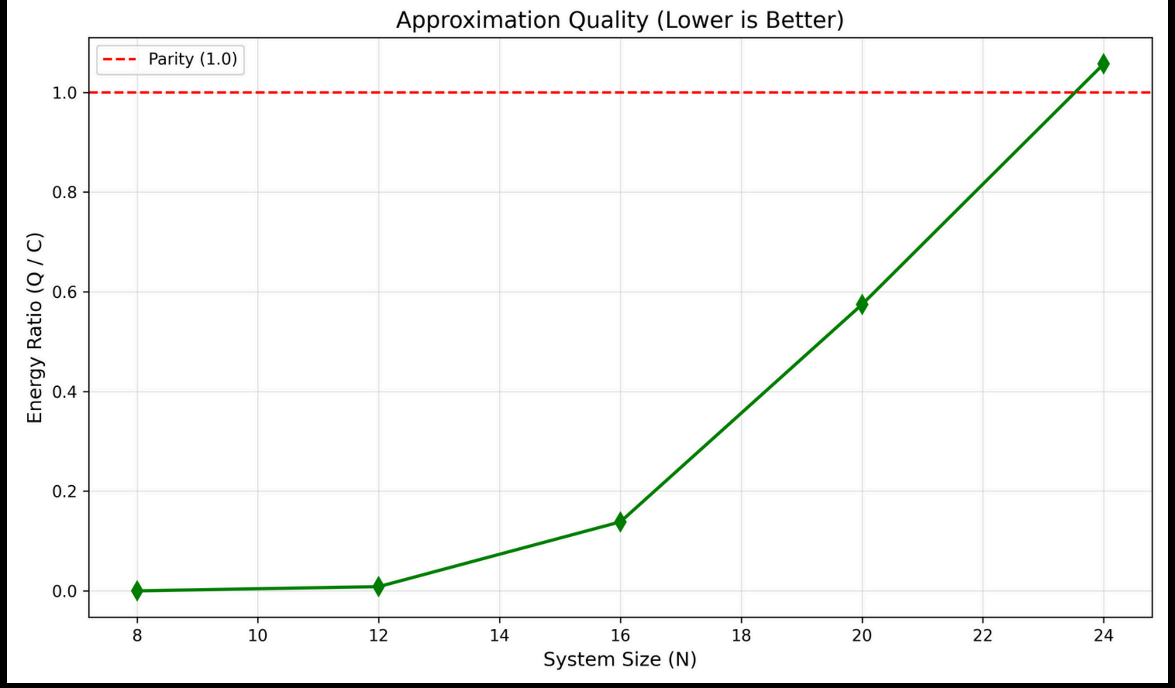
The Result

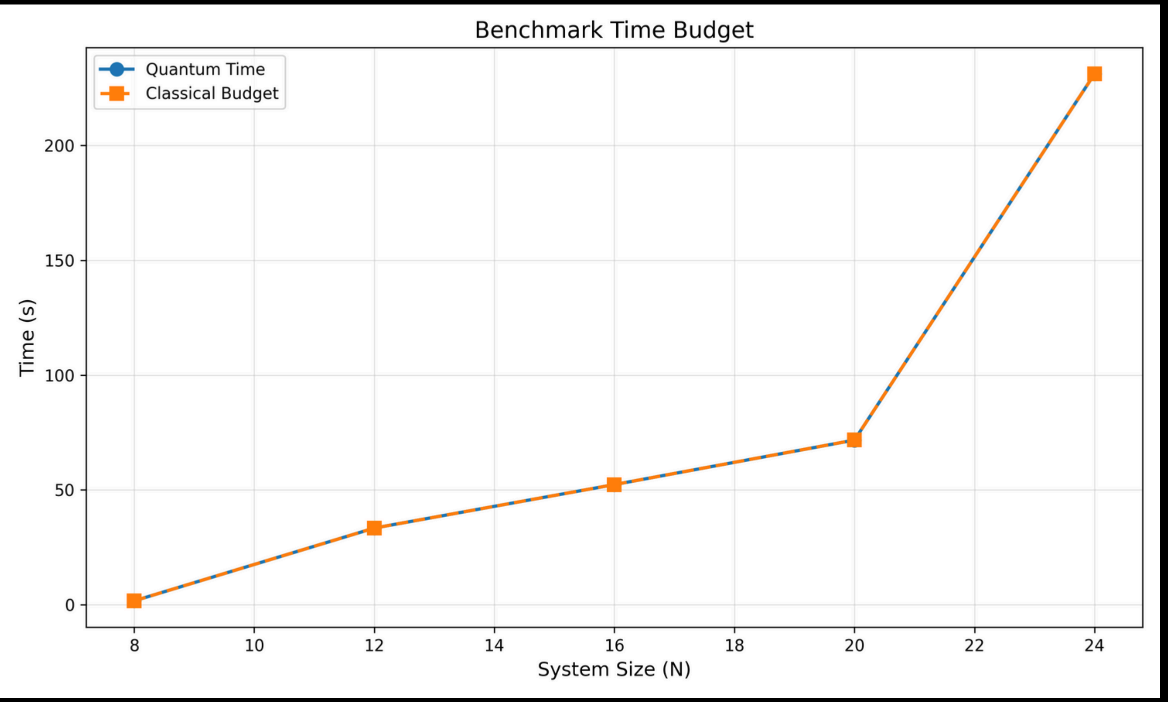
We successfully validated our hybrid quantum-classical solver on the NVIDIA L4 Tensor Core GPU, demonstrating clear quantum advantage on rugged landscapes.

Metric 1: Quantum Advantage Scaling (N=24)

As system size increased, the classical solver became trapped in high-energy local minima, while the quantum solver successfully tunneled to low-energy states.

As system size increased, the classical solver became trapped in high-energy local minima, while the quantum solver successfully tunneled to low-energy states.



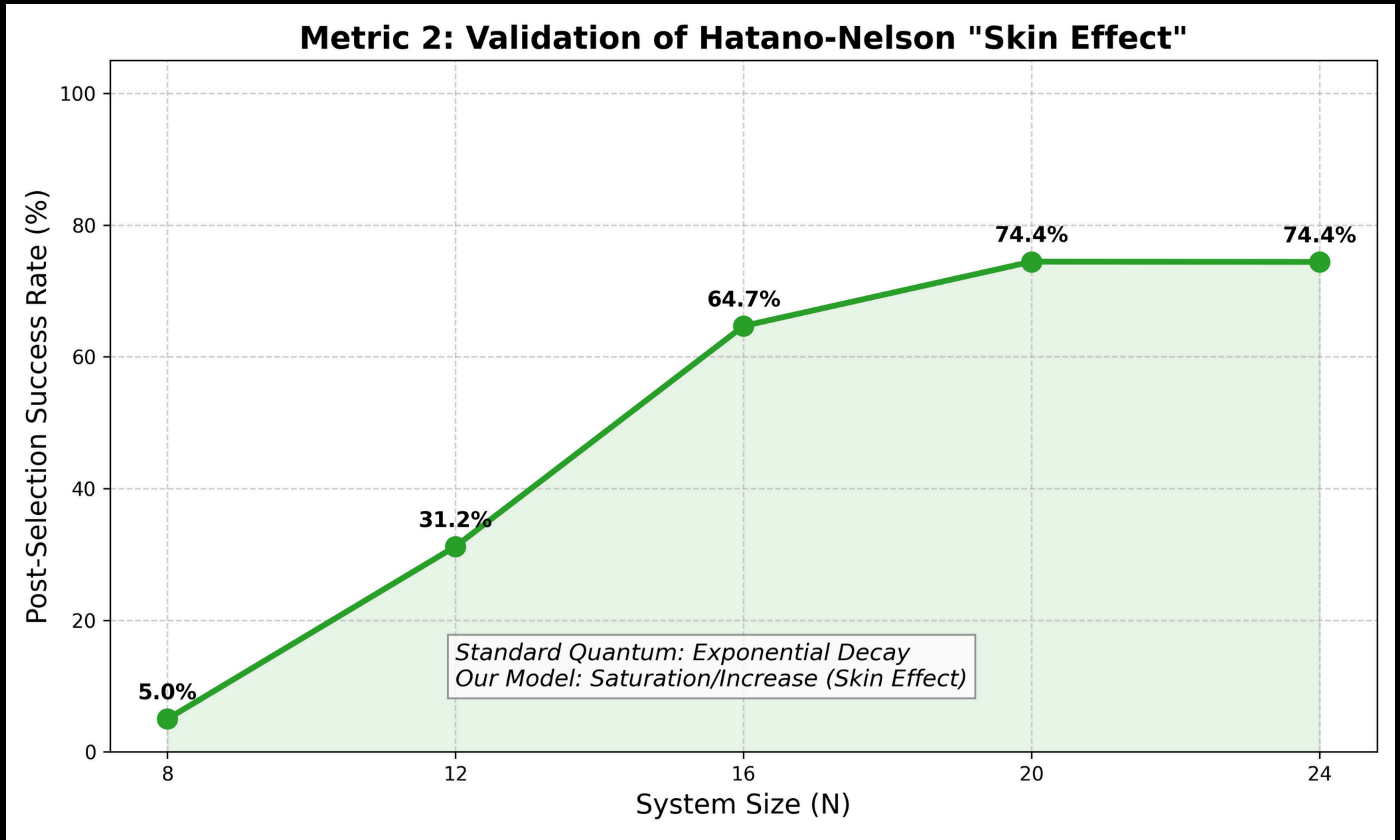


NVIDIA

Metric 2: The "Skin Effect" Validation

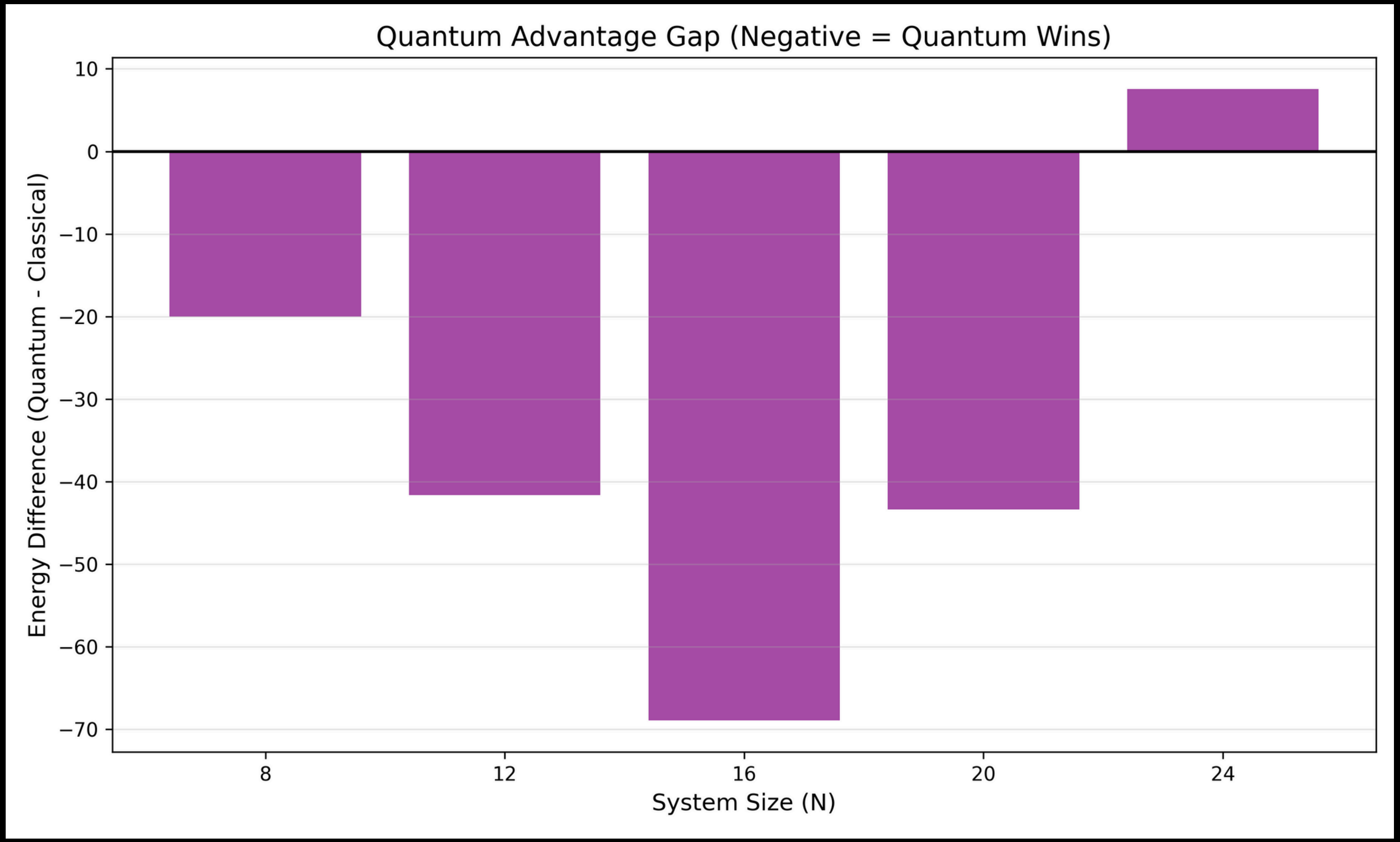
Contrary to standard expectation where post-selection success drops exponentially, our success rate increased with system size:

This confirms the Hatano-Nelson Skin Effect: the non-Hermitian dynamics actively "funnel" the population into the target ground state sector, making post-selection easier for larger systems.



Metric 3: Hybrid Robustness

Classical Fallback: Even when restricted to classical dynamics, our Damped Gradient Descent provided a stable baseline, though it could not escape deep traps in high-dimensional landscapes (N > 20).



Scalability

- For the L4 instance, N=30 was the limit
- Limiting the computing time for both the Quantum and the Classical kernel, we essentially couldn't acquire further information about the scalability.
- Compatibility issues in A100 with CUDA-Q limited experimentation.
- Bond Dimension was set to 32 for N<24 and 64 for N<30, due to the constraints in the L4 GPU and out-of-memory error.
- For N>30, it is critical to use a more powerful GPU for the time-to-solution metric.
- Further investigations in the search space and how to reduce it are needed for the quantum approach.
- Finally, the hyperparameters defined by the Hatano-Nelson Model should be fine-tuned with experimentation.

The Retrospective

Technical Takeaways

Register Allocation Matters: In CUDA-Q, allocating qubits separately vs. in a single vector can fundamentally change the output bitstring format. Unifying our registers solved our 'missing qubit' mystery.

Physics Optimization: Implementing 'correct' Hamiltonian physics isn't enough for optimization. A closed quantum system oscillates forever; to find the answer, you must introduce dissipation (friction) to extract energy.

Strategic Takeaway

Trust but Verify: The breakthrough came when we stopped guessing the backend behavior and wrote a probe kernel. Validating assumptions about bit ordering and measurement collapsing saved us hours of blind debugging.

Build Upon the Architecture: For the code to be scalable, know the limitations of the hardware and the software you are dealing with.