METHOD STUDY

QAOA Noise-Ladder Study

How much of a trained circuit's concentration survives noise - and does a calibrated simulation predict the real device? Depth 1 to 7 and an adaptive circuit on a real 14-qubit portfolio book, replayed at three noise levels, on a calibrated simulation of ibm_kingston and on ibm_kingston itself. A method study behind the Portfolio Optimization, Credit Allocation and Collateral Allocation RQPs.

August 2026

qubit-lab.ch

qubit-lab.ch
contact@qubit-lab.ch

Contents

The study on one page

For the reader in a hurry. The quantum optimisation tools of qubit-lab.ch are run on exact simulators, where every circuit can be inspected without noise. Real quantum processors add noise with every gate, and the question a bank asks before trusting a device result is simple: how much of what the circuit found on the simulator is still visible when the same circuit runs on hardware - and can we predict that from the hardware's published error rates before paying for the run? This study answers both on a real 14-asset portfolio book, with the product's certified referee proving the optimum so that every probability is measured against a known answer.

What we did. Five trained circuits on the same book - fixed depth p = 1, 3, 5, 7 and the adaptive circuit that builds its own 12 layers - were replayed, unchanged, on four kinds of noise: three synthetic levels keyed to the two-qubit gate fidelity (99.9%, 99.5%, 99.0%), a simulation built from ibm_kingston's calibration data of the day, and ibm_kingston itself (IBM Heron r2, 156 qubits; 20,000 shots per circuit, 44 seconds of processor time in total). Every run went through the product as a customer would run it; every number is re-derived from the committed raw results; the design was written down before the first run and its amendments are disclosed in Appendix A.

Findings.

1. **Without noise, depth stops paying at about five layers** on this book: the probability of the certified optimum rises from 0.065% (p = 1) to 0.170% (p = 5) and falls back at p = 7. These circuits start from the product's classical seed, which is wrong on this book (three assets off), and they stay locked to it; only the adaptive circuit escapes, putting 10.1% on the optimum and making it the single most likely readout.
2. **Noise does not simply erase a seed-locked circuit - it smears it.** At moderate noise the optimum's probability rises (retention ratios of 1.0-1.5) because the wrong seed's peak spills onto its neighbours, the optimum among them. This is a warning sign, not good news: the readout is getting flatter, and the mass of the ten most likely states - the clean indicator - falls monotonically on every circuit and every rung.
3. **Genuine concentration decays exactly as feared.** The adaptive circuit (2,190 two-qubit gates) keeps the optimum as its most likely readout at 99.9% fidelity (2.15%, retention 0.21) and is fully depolarised at 99.5% and below. On this book no measured configuration delivers a usable readout below 99.9%.
4. **On the device, the first number is the routing overhead.** Mapping each circuit onto the processor's heavy-hex lattice multiplied the two-qubit gate count by about 2.5 (182 -> 440, 546 -> 1,302, 2,190 -> 5,454). The simulated rungs were keyed to logical gate counts; the device therefore behaves like a circuit 2.5 times deeper than any rung assumed. Only p = 1 retained anything measurable (an upper bound of 0.03% on the optimum, seed peak 8.5% -> 0.8%); from p = 3 on, the device readout is statistically uniform, and the adaptive circuit's 10.1% is erased completely (0 hits in 20,000 shots).
5. **A calibration-based prediction is optimistic by a factor of about 3.** Simulating the routed circuits with ibm_kingston's own reported gate and readout errors predicted 2.6-3.5 times more concentration than the device delivered (top-10 mass 9.4% predicted vs 3.7% measured at p = 1; 2.9% vs 1.0% at p = 3). The published error rates do not account for everything the device does to a 14-qubit circuit - idle decoherence, crosstalk, coherent errors and drift are not in a Pauli model built from them.
6. **Measuring the device is cheap; predicting it is the expensive part.** The whole device rung - five circuits, 100,000 shots - took 44 seconds of processor time with queue waits under ten seconds. The calibrated prediction of the same five circuits took 17-26 minutes of simulation each on the product's standard worker and still overstated the device by a factor of three. On this class of circuit the measurement is both cheaper and more informative than the prediction.

What it means for practice. Noise-free depth on a seed-locked circuit is cheap and worthless under noise: nothing past three layers survives realistic error rates. The adaptive circuit's ability to escape a wrong seed is real and is worth paying for only on hardware whose two-qubit fidelity is at the best-pair level (about 99.9%) after routing - which, on today's Heron processors and this 14-qubit all-to-all problem, is not yet available. The products now print the noise ladder next to the certified gap, flag retention above one as a seed-lock warning, and treat a calibrated prediction as an optimistic bound, not a forecast.

qubit-lab.ch, Daniel Hug. Simulated rungs run 21-22 August 2026 on Portfolio Optimization RQP Pro v9.7.2; device and calibrated rungs 22 August 2026 on v9.7.4 and ibm_kingston. Document version 1.0, 22 August 2026.

1. Management summary

Written for decision-makers; the measured basis for every sentence follows in sections 2-6. The three questions below are the questions this study answers for anyone weighing a hardware run.

Q1 - Does trained circuit depth survive a real device? Today, essentially no. Mapping the circuits onto the device multiplied every two-qubit gate by about 2.5, and only the shallowest circuit retained any structure on ibm_kingston; from three layers upward the readout is statistically uniform, and the adaptive circuit that puts the optimum on top in simulation left no trace of it in 20,000 device shots (by the rule of three, its optimum probability on the device is below 0.015% at 95% confidence - against 10.1% noise-free). Depth that is free in simulation is worthless under today's noise.

Q2 - Can a calibrated simulation replace the device? Only as an optimistic ceiling. Predictions built from the device's own calibration data were a factor 2.6-3.5 too rosy (a reading robust to shot noise: the 95% intervals stay above 2.3 on every circuit that retains structure) - and measuring the real device cost 44 seconds of processor time, less than the 17-26 minutes per circuit the calibrated prediction itself consumes. When a device run is this cheap, measure; treat every calibrated prediction as a best case and label it so, as the products now do.

Q3 - What hardware quality would change the answer? Roughly 99.9% two-qubit fidelity after routing on this problem - not yet available for a 14-qubit all-to-all portfolio on today's processors. Until then, the products print the noise ladder next to the certified gap, flag the seed-lock warning signature, and treat noise-free depth claims with the skepticism this study measured into them.

Every probability in this study is measured against a certified optimum, on one real portfolio book, raw and unmitigated - and the study says so. It is part of qubit-lab.ch's method-study series, alongside the concentration study whose trained circuits it replays and the QML advantage-island pair - all on qubit-lab.ch/evidence.

2. Question and frame

2.1 Why a noise ladder

The RQPs train every quantum circuit on an exact statevector simulator and score its readout against a certified classical referee (here SCIP branch-and-bound, proving the optimum per run). The concentration study [1] established what the trained readout looks like without noise: how much probability the circuit puts on the proven optimum, whether that optimum is the most likely state, and how many measurement shots it would take to see it. Those are simulator numbers. A real processor applies an error with every gate; a circuit that concentrates 10% on the optimum in simulation may put nothing there on the device.

The product's answer is the noise ladder: the trained circuit is never retrained under noise - it is **replayed**, unchanged, under a chosen noise model, and the same concentration metrics are computed again. The ratio noisy / exact P(optimum) is the **retention**. This study sweeps the ladder over circuit depth and over three kinds of noise - synthetic, calibrated, and the real device - to answer three questions:

1. How does concentration degrade with depth under noise, and is there a noise-optimal depth?
2. Does a simulation built from a backend's calibration data predict the backend?
3. What does a circuit that genuinely concentrates (the adaptive ansatz) look like on today's hardware?

2.2 Definitions

Throughout: **P(optimum)** is the probability the readout assigns to the certified optimum. **x-uniform** is that probability divided by 1/2^n (n = 14 qubits, 16,384 states). **Top-10 mass** is the summed probability of the ten most likely states. **Modal** means the certified optimum is the single most likely state. **Shots to 99%** is ln(0.01) / ln(1 - P(optimum)). **Retention** is noisy P(optimum) divided by exact P(optimum). **Seed peak** is the probability of the classical seed state the circuit was started from - on this book a wrong answer, three assets away from the optimum.

F2 is the two-qubit gate fidelity that parameterises the synthetic rungs. **Logical** gate counts are those of the circuit as trained (all-to-all connectivity); **routed** counts are those of the same circuit after transpilation onto the device's coupling map, which inserts SWAP operations.

Statistical floor. A noisy rung whose optimum-hit count is below 100 is marked † and read as an upper bound; no retention ratio is interpreted from it. This rule was fixed in the design before the first run.

3. Design

3.1 Instrument and instance

Portfolio Optimization RQP Pro, as deployed - licensed runs through the product, every number from its reports and result files. The instance is the 14-qubit adaptive-demo book of the resource room: a dense real portfolio book with a certified optimum proven by SCIP on every run, and a classical heuristic seed that is wrong by three assets (certified gap 1.60%). The wrong seed was kept deliberately: it matches the finding of the concentration study that the product's heuristic missed the optimum on every real book, and it makes the two circuit families behave differently under noise in an instructive way (Section 4).

3.2 Circuits

Five trained circuits, one training each (the trained circuit is the object under test; a second training seed is a disclosed limitation):

Circuit	Ansatz	Two-qubit gates (logical)	Training
p = 1, 3, 5, 7	standard QAOA, CVaR objective (alpha 0.10), classical seed start (epsilon 0.25)	182, 546, 910, 1,274	150 iterations x 3 restarts, rng seed 42
adaptive, cap 12	ADAPT-style, CVaR, classical seed start, 12 layers used	2,190	250 iterations per layer, rng seed 42

3.3 Noise rungs

Rung	Definition	Circuit basis
exact	statevector, no noise (the run itself)	logical
F2 = 99.9 / 99.5 / 99.0%	parametric: two-qubit depolarising with error 1 - F2; one-qubit error (1 - F2)/10; readout error 3(1 - F2); Qiskit Aer statevector trajectories	logical
calibrated	Aer noise model built from ibm_kingston's calibration snapshot of the day: per-gate depolarising from the reported gate errors, per-qubit readout errors; thermal relaxation excluded (Section 6.1)	routed for ibm_kingston
device	ibm_kingston (IBM Heron r2, 156 qubits), 20,000 shots per circuit, standard basis (fractional gates not exposed on the day), Qiskit Runtime Sampler	routed for ibm_kingston

The backend was chosen on the day as the best of the three Open-plan Heron processors by median CZ error (ibm_kingston: CZ median 0.213%, best quartile 0.157%, readout median 0.87%; ibm_fez 0.272%; ibm_marrakesh 0.302%).

3.4 Shots and the floor

Parametric rungs: 24,000-200,000 trajectories per rung, chosen so that the optimum's hit count clears 100 wherever the circuit allows it; the product's wall-time guard reduced the deeper arms (recorded per arm in the tables). Device: 20,000 shots per circuit, the largest comparison pool the product requests. Calibrated twins: 32,768 shots at p = 1 down to 2,048 for the adaptive circuit, sized to the worker's 40-minute guard (Section 6.3); their statistics are correspondingly coarser, and every calibrated and device row is printed next to the value a perfectly uniform readout would show at the same shot count, so that count fluctuations are not mistaken for structure.

3.5 Pre-specification

The design - question, instrument, arms, rungs, floor rule, metrics, five numbered predictions and the analysis plan - was locked on 21 August 2026 before the first run and is reproduced in Appendix A together with the six amendments recorded along the way. The hardware and calibrated rungs were deferred in the lock and executed on 22 August under amendments d-f.

4. The simulated ladder

4.1 Depth without noise saturates at about five layers

The exact column of Table 1 is the depth sweep without noise. P(optimum) rises from 0.065% at p = 1 through 0.140% and 0.170% to 0.139% at p = 7 - saturation around p = 5, as predicted (prediction 1). All four standard circuits remain **seed-locked**: the wrong seed state is their most likely readout at 8.5-9.2%, the optimum sits at 10-28 times uniform but is never modal, and it would take 2,700-7,000 shots to observe it once with 99% confidence. The adaptive circuit is the only one that escapes the seed: 10.08% on the optimum, modal, 1,651 times uniform, 44 shots to 99%. That was the concentration study's result on this book, reproduced here exactly (the training is deterministic given the seed).

Circuit	CX (logical)	exact P(opt)	F2 99.9%	F2 99.5%	F2 99.0%
p = 1	182	0.065%	0.067% (ret 1.03) †	0.098% (ret 1.50)	0.087% (ret 1.33)
p = 3	546	0.140%	0.159% (ret 1.13)	0.140% (ret 1.00) †	0.043% (ret 0.30) †
p = 5	910	0.170%	0.171% (ret 1.01) †	0.054% (ret 0.32) †	0.023% (ret 0.14) †
p = 7	1,274	0.139%	0.188% (ret 1.35) †	0.040% (ret 0.29) †	0.005% (ret 0.04) †
adaptive, cap 12	2,190	10.076%	2.147% (ret 0.21)	0.004% (ret 0.00) †	0.008% (ret 0.00) †

Table 1 - P(certified optimum) across the simulated ladder; retention in parentheses; † = fewer than 100 optimum hits (upper bound only).

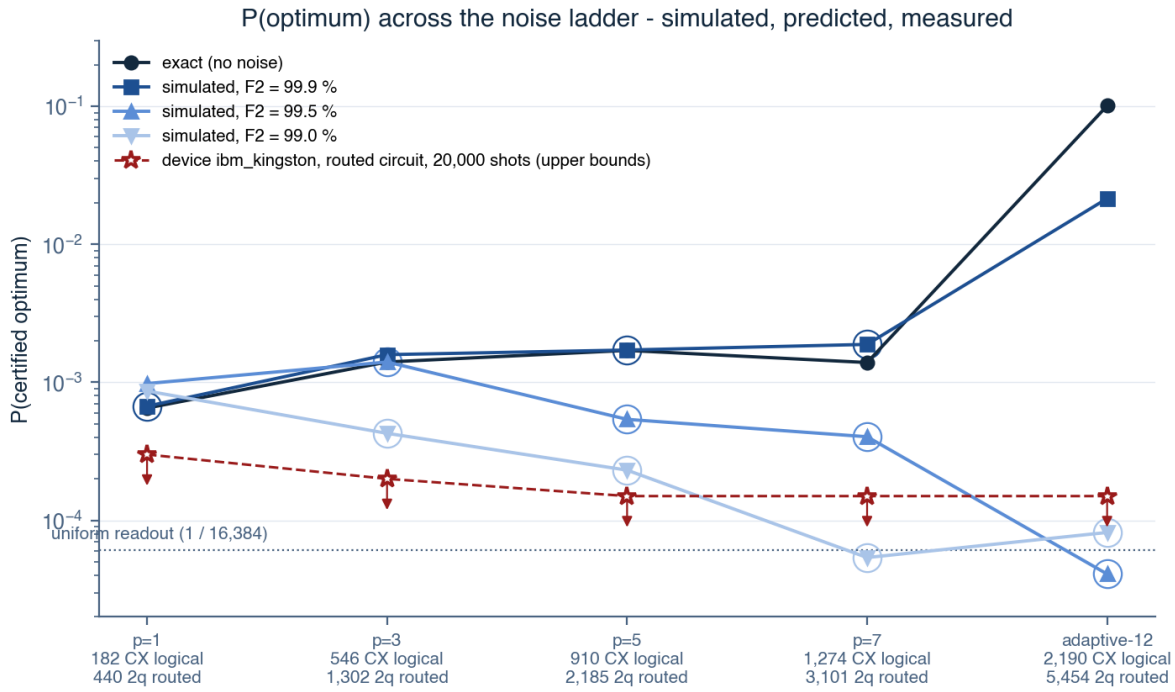

Figure 1 - P(optimum) per circuit for every rung. Simulated rungs use the logical circuit; the device ran the routed circuit (both gate counts in the axis labels). Open rings mark simulated rungs below the 100-hit floor; the device values are upper bounds (all below the floor; zero hits are drawn at 3/shots) and point downward.

4.2 Noise smears a seed-locked circuit before it erases it

Prediction 2 - geometric decay of P(optimum) with depth, about 0.83x per layer at 99.9% and 0.40x at 99.5% - was wrong in an instructive way. On the standard circuits, retention sits **above one** wherever the total circuit fidelity is still moderate: at p = 1 on all three rungs (1.03, 1.51, 1.33) and at 99.9% for every depth (1.13, 1.01, 1.35). The decay law only takes over once that regime ends: the 99.5% column falls 1.51 -> 1.00 -> 0.32 -> 0.29 and the 99.0% column 1.33 -> 0.30 -> 0.14 -> 0.04.

The mechanism: the readout of a seed-locked circuit is a tall peak on the wrong seed with small mass elsewhere. Pauli errors inside the entangling layers redistribute that peak across the seed's Hamming neighbourhood; the optimum is three flips away and catches part of the spill, and that part exceeds its tiny noise-free probability. The effect was challenged as a possible simulation error and verified with an independent trajectory sampler (different codebase and random-number stream, identical channel definitions): 0.094% against Aer's 0.098% on the p = 1 / 99.5% arm, seed peak 3.6% against 3.8% - the effect is physical. It is also useless: "noise as accidental escape" lifts P(optimum) from 0.065% to at best 0.098%, and the readout stays uninformative.

4.3 The indicator that does not lie: top-10 mass

Table 2 prints every metric for every simulated rung. The **top-10 mass** decays monotonically with noise on every circuit - p = 1: 21.3% -> 19.2% -> 11.0% -> 6.3%; p = 7: 19.3% -> 9.8% -> 1.0% -> 0.4% - including on the arms where P(optimum) rises. The retention-above-one rungs are therefore visible as "P(optimum) up, top-10 down": a flattening readout whose spill happens to reach the optimum. This is the signature the products now flag: **retention above one together with low exact concentration is a seed-lock warning, not evidence of robustness.**

Circuit / F2	P(opt) exact / noisy	x-uniform e/n	top-10 mass e/n	modal e/n	shots-to-99 % e/n	shots
p = 1 / 99.9%	0.065% / 0.067% †	11x / 11x	21.3% / 19.18%	no / no	7,070 / 6,857	131,072
p = 1 / 99.5%	0.065% / 0.098%	11x / 16x	21.3% / 10.99%	no / no	7,070 / 4,697	200,000

Circuit / F2	P(opt) exact / noisy	x-uniform e/n	top-10 mass e/n	modal e/n	shots-to-99 % e/n	shots
p = 1 / 99.0%	0.065% / 0.087%	11x / 14x	21.3% / 6.30%	no / no	7,070 / 5,322	200,000
p = 3 / 99.9%	0.140% / 0.159%	23x / 26x	21.6% / 14.91%	no / no	3,278 / 2,900	65,536
p = 3 / 99.5%	0.140% / 0.140% †	23x / 23x	21.6% / 4.26%	no / no	3,278 / 3,279	65,536
p = 3 / 99.0%	0.140% / 0.043% †	23x / 7x	21.6% / 1.27%	no / no	3,278 / 10,777	65,536
p = 5 / 99.9%	0.170% / 0.171% †	28x / 28x	19.6% / 11.98%	no / no	2,702 / 2,684	51,907
p = 5 / 99.5%	0.170% / 0.054% †	28x / 9x	19.6% / 1.99%	no / no	2,702 / 8,535	51,907
p = 5 / 99.0%	0.170% / 0.023% †	28x / 4x	19.6% / 0.58%	no / no	2,702 / 19,918	51,907
p = 7 / 99.9%	0.139% / 0.188% †	23x / 31x	19.3% / 9.81%	no / no	3,307 / 2,444	37,173
p = 7 / 99.5%	0.139% / 0.040% †	23x / 7x	19.3% / 0.98%	no / no	3,307 / 11,411	37,173
p = 7 / 99.0%	0.139% / 0.005% †	23x / 1x	19.3% / 0.43%	no / no	3,307 / 85,592	37,173
adaptive, cap 12 / 99.9%	10.076% / 2.147%	1651x / 352x	49.6% / 11.87%	yes / yes	44 / 213	24,458
adaptive, cap 12 / 99.5%	10.076% / 0.004% †	1651x / 1x	49.6% / 0.49%	yes / no	44 / 112,631	24,458
adaptive, cap 12 / 99.0%	10.076% / 0.008% †	1651x / 1x	49.6% / 0.41%	yes / no	44 / 56,315	24,458

Table 2 - All concentration metrics, exact / noisy, for the fifteen simulated rungs. "Shots" is the trajectory count actually sampled after the guard's reductions.

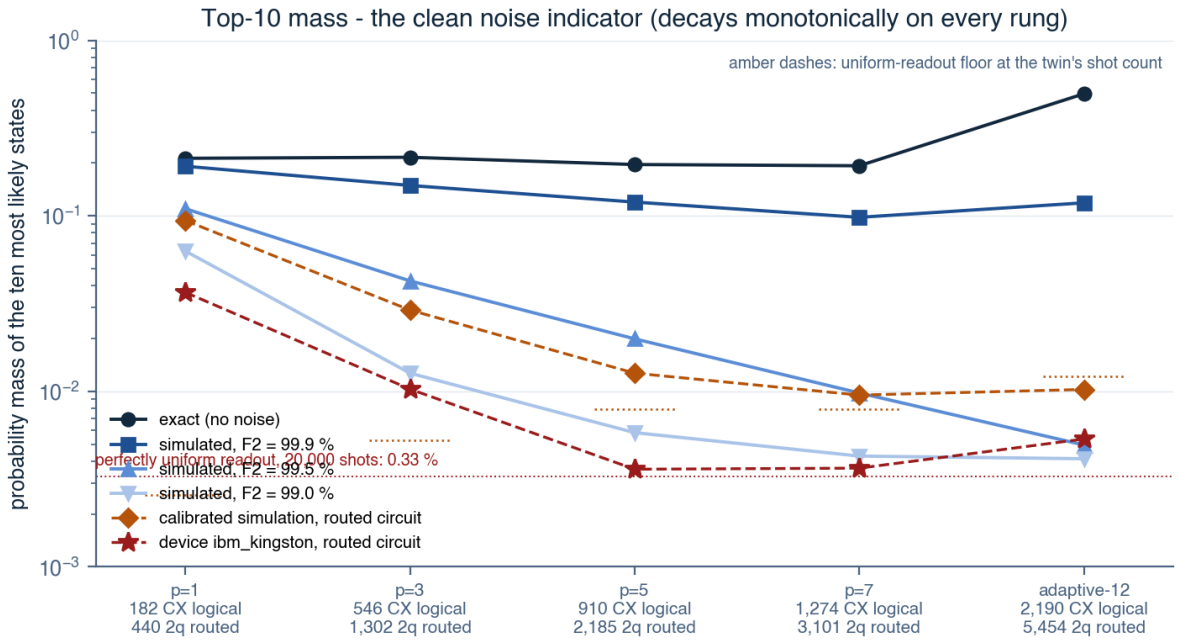

Figure 2 - Top-10 probability mass per circuit and rung, including the calibrated prediction and the device. The red dotted line is the top-10 mass a perfectly uniform readout shows when sampled with 20,000 shots; the amber dashes are the same floor at each calibrated twin's shot count.

4.4 Genuine concentration decays as feared

The adaptive circuit is the only one with something to lose, and it loses it in the predicted order. At F2 = 99.9% it keeps the optimum as its most likely readout: 2.15% (retention 0.21, 525 hits - the one deep rung that clears the floor comfortably), 352 times uniform, 213 shots to 99%. At 99.5% and 99.0% it is fully depolarised: 0.004-0.008% on the optimum, within a factor of about one of uniform, the top-10 mass at 0.4-0.5%. Its 2,190 two-qubit gates are the price of escaping the wrong seed; at today's median fidelities that price erases the readout entirely.

Prediction 4 (measured retention exceeds the crude product-of-fidelities) held where it could be tested: 0.21 measured against 0.098 expected for adaptive / 99.9%. Prediction 5 (the modal flag survives longer than the mass) held on the same rung - the optimum is still modal at 2.15% while the top-10 mass has fallen from 49.6% to 11.9%.

4.5 Noise-optimal depth

Read off as the argmax of noisy P(optimum) per rung, with the floor rule applied: at 99.9% the standard circuits are indistinguishable between p = 3 and p = 7 (all within the smear regime, all but p = 3 below the floor) and the adaptive circuit dominates; at 99.5% and 99.0% no configuration delivers a readout that differs usefully from uniform. Prediction 3 ("p = 3 at 99.9%, p = 1 below") is therefore confirmed in spirit and moot in practice - the honest answer for this book is "adaptive on best-pair hardware, otherwise do not trust the sampler".

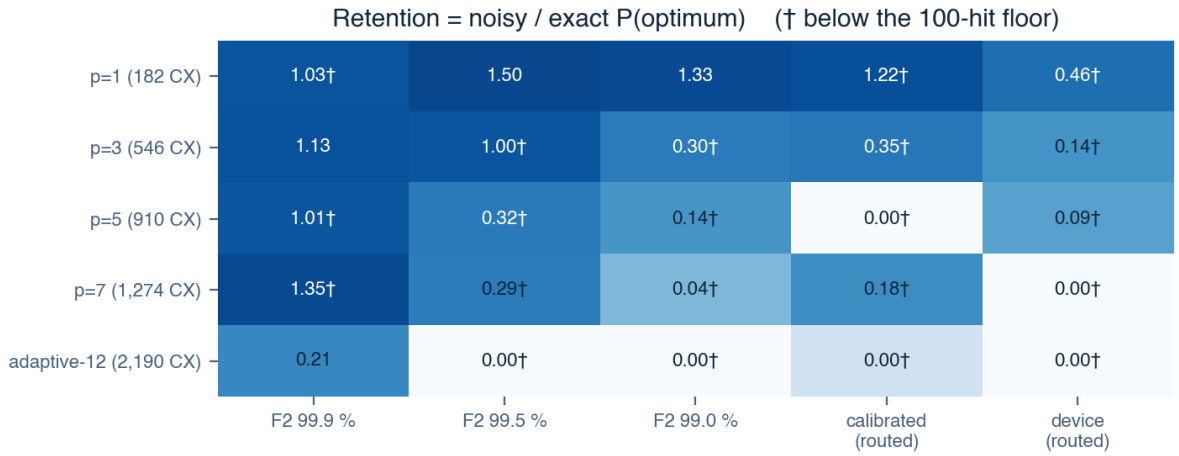

Figure 4 - Retention (noisy / exact P(optimum)) for every circuit and rung; † marks values below the 100-hit floor.

5. The device rung

5.1 The first number is the routing overhead

The circuits were trained for all-to-all connectivity; ibm_kingston's heavy-hex lattice connects each qubit to two or three neighbours. The product's transpiler (Qiskit preset pass manager, optimisation level 1) inserted the SWAP operations needed, and the two-qubit gate count grew by a factor of 2.3-2.5 on every circuit (Table 3). The 14-qubit all-to-all cost layer is the reason; it is inherent to dense portfolio QUBOs and not a property of this book. The simulated rungs were keyed to the logical counts, so the device should be read against a rung about 2.5 times deeper than any simulated - which places ibm_kingston's 99.79% median CZ fidelity in the territory of the simulated 99.0% rung for these circuits.

5.2 Results

Circuit	2q gates logical -> routed	2q depth routed	exact P(opt)	device P(opt) (hits of 20,000)	most likely state: exact -> device	top-10 mass: exact -> device	distinct strings in 20,000 shots	device time (s)
p = 1	182 -> 440	139	0.065%	0.030% (6) †	8.48% -> 0.80%	21.3% -> 3.67%	8,170	57
p = 3	546 -> 1,302	438	0.140%	0.020% (4) †	9.17% -> 0.14%	21.6% -> 1.03%	10,536	226
p = 5	910 -> 2,185	615	0.170%	0.015% (3) †	8.99% -> 0.04%	19.6% -> 0.36%	11,268	225
p = 7	1,274 -> 3,101	824	0.139%	0.000% (0) †	9.18% -> 0.04%	19.3% -> 0.36%	11,199	16
adaptive, cap 12	2,190 -> 5,454	1,404	10.076%	0.000% (0) †	10.08% -> 0.06%	49.6% -> 0.53%	10,871	10
uniform readout, 20,000 shots (sampling floor)			0.006%	0.006% (1.2)	0.04%	0.33%	~11,800	

Table 3 - The device rung on ibm_kingston, 20,000 shots per circuit. "Most likely state" is the wrong seed on the standard circuits and the optimum on the adaptive circuit (exact). The last row is what a perfectly uniform readout shows at 20,000 shots: the sampling floor. Device time is the Sampler execution time reported by Qiskit Runtime.

Three readings:

- **p = 1 is the only circuit that retains anything measurable.** Six hits on the optimum (0.03%, an upper bound; Wilson 95% interval 0.014-0.065%), the seed peak down from 8.5% to 0.8%, top-10 mass from 21% to 3.7% (Wilson 95% interval 3.4-3.9%), and 8,170 distinct bitstrings in 20,000 shots. The distribution is recognisably the trained one, flattened by about a factor of five.
- **From p = 3 the readout is statistically uniform.** Top-10 mass 1.0% at p = 3 and 0.36% at p = 5 and p = 7 against a 0.33% sampling floor; most-likely-state probability 0.04% against a 0.04% floor; 10,500-11,300 distinct strings against about 11,800 for pure uniform sampling.
- **The adaptive circuit's 10.1% is erased completely.** 5,454 routed two-qubit gates, zero hits on the optimum, top-10 mass 0.53%. The circuit that escapes the wrong seed in simulation leaves no trace of it on the device.

5.3 What the device rung adds to the simulated ladder

Nothing in Table 3 contradicts Section 4 - it locates the device on the ladder. The simulated 99.0% rung (logical circuits) gave top-10 masses of 6.3% (p = 1), 1.3% (p = 3), 0.6% (p = 5), 0.4% (p = 7) and 0.4% (adaptive); the device gave 3.7%, 1.0%, 0.36%, 0.36% and 0.53% on circuits 2.5 times longer. The device sits at or slightly below the simulated 99.0% rung for every circuit. Forty-four seconds of processor time bought that placement; the simulated ladder had predicted it.

6. Prediction versus device

6.1 What "calibrated" means here

For each circuit the product replayed the routed circuit on Aer under a noise model built from ibm_kingston's calibration snapshot of the day (09:54 UTC; the device runs followed between 08:48 and 09:40 UTC on the same calibration): per-gate depolarising errors from the reported CZ and single-qubit gate errors and per-qubit readout assignment errors. Thermal relaxation (T1/T2 decay during idle time) is excluded by design: it enters a trajectory simulator as Kraus channels on every gate, which cost orders of magnitude more to sample than Pauli channels and would have made the deeper twins infeasible within the product's run budget; the consequence - an optimistic prediction - is measured in Section 6.2. This is the same channel family as the parametric rungs with the device's own numbers in place of a single F2, on the circuit the device actually executed.

6.2 Results

Circuit (routed 2q)	shots (twin)	top-10 mass: exact / calibrated / device	most likely state: exact / calibrated / device	uniform floor at twin shots (top-10 / max)	calibrated : device (top-10)
p = 1 (440)	32,768	21.3% / 9.41% / 3.67%	8.48% / 3.00% / 0.80%	0.26% / 0.03%	2.6x
p = 3 (1,302)	8,192	21.6% / 2.91% / 1.03%	9.17% / 0.61% / 0.14%	0.53% / 0.06%	2.8x
p = 5 (2,185)	4,096	19.6% / 1.27% / 0.36%	8.99% / 0.15% / 0.04%	0.79% / 0.10%	3.5x
p = 7 (3,101)	4,096	19.3% / 0.95% / 0.36%	9.18% / 0.10% / 0.04%	0.79% / 0.10%	2.6x
adaptive, cap 12 (5,454)	2,048	49.6% / 1.03% / 0.53%	10.08% / 0.15% / 0.06%	1.22% / 0.15%	1.9x

Table 4 - Calibrated prediction against the device on the routed circuits. The "uniform floor" column is what a perfectly uniform readout shows at the twin's shot count; values at or below it are indistinguishable from uniform. The last column is the prediction's optimism on the top-10 mass; the p = 7 twin is the re-run described in Section 6.3.

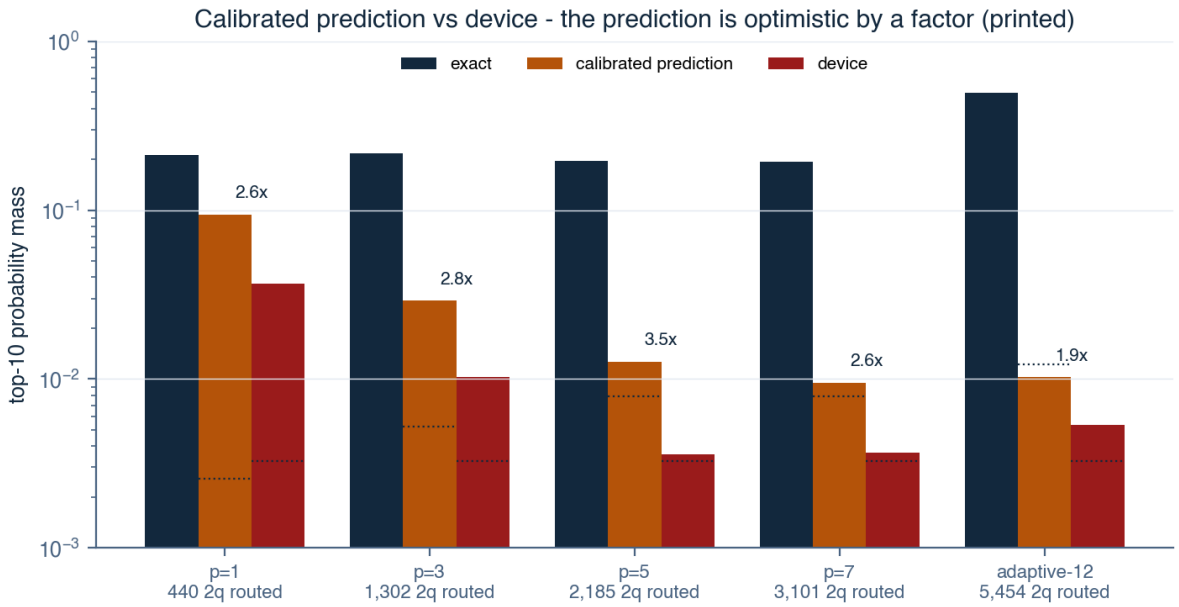

Figure 3 - Top-10 mass: exact, calibrated prediction and device per circuit, with the prediction's optimism factor printed. Dotted marks are the uniform-readout floors at the respective shot counts.

The prediction is optimistic by a factor of 2.6 at p = 1, 2.8 at p = 3 and 3.5 at p = 5 on the top-10 mass (shot-noise 95% intervals 2.4-2.8 / 2.3-3.4 / 2.4-5.0, holding the observed top-10 sets fixed; the intervals overlap, so the apparent growth of the factor with depth is not resolved at these shot counts - the factor itself is: even the interval edges stay above 2.3), and by 3.3-4.2x on the most-likely-state probability (re-derived from the raw counts; Revision 1.2 printed 3.8-4.4x, an artefact of computing the ratios from the table's rounded percentages - at these low counts the per-circuit most-likely-state ratios carry wide shot intervals and the top-10 factors are the robust reading). At p = 7 and at the adaptive circuit both prediction and device are at their respective uniform floors, so the factors there (2.6x, 1.9x) carry no information and no intervals are attached. The direction is robust: on every circuit where either side retains structure, the device is flatter than its published error rates imply.

The residual is what a Pauli model built from calibration data does not contain: decoherence during the idle time that routing and scheduling insert (thermal relaxation was excluded for cost, Section 6.3), crosstalk between simultaneously driven pairs, coherent (non-Pauli) errors that do not average like depolarising noise over a single circuit, and drift between the calibration snapshot and the run. Which of these dominates was not measured and is a limitation (Section 8). For practice the factor itself is the result: **a calibration-based prediction is an optimistic bound, to be read with a margin of about 3 on this class of circuit.**

6.3 Cost of prediction versus measurement

The device rung cost 44 seconds of processor time for five circuits and 100,000 shots (57, 226, 225, 16 and 10 seconds of execution; queue waits of 5-10 seconds); the account's monthly Open-plan allowance of 600 seconds covers a dozen such rungs. The calibrated prediction of the same circuits cost 1,582 / 1,211 / 1,028 / 1,397 / 1,110 seconds of trajectory sampling on the product's four-vCPU worker for 32,768 / 8,192 / 4,096 / 4,096 / 2,048 shots - seventeen to twenty-six minutes per circuit for a fraction of the device's shot count, and that with thermal relaxation excluded (Section 6.1). One prediction run was repeated after a transient error on IBM's account service interrupted the download of the calibration snapshot; the repeated run is the one reported.

The comparison is lopsided in a way worth stating plainly: for a 14-qubit circuit the device answers in under a minute with ten times the shots, and the answer is the ground truth the prediction is trying to approximate. A calibrated simulation earns its place when the device is unavailable or the question is "what if the error rates were better"; as a forecast of a specific backend on a specific day it is an optimistic bound, to be read with the factor of Section 6.2.

7. Synthesis - what changes in the products

The study sharpens three rules of the RQPs into numbers:

1. **Depth without noise is cheap and worthless.** On a seed-locked circuit nothing past three layers buys anything that survives realistic noise, and the noise-free gain from p = 3 to p = 5 (0.140% -> 0.170%) is inside the smear. Default depth stays at 3; the manuals say why.
2. **The adaptive circuit's escape is worth paying for only on best-pair hardware after routing.** Its 10.1% survives at 99.9% two-qubit fidelity on the logical circuit (2.15%, still modal) and at nothing below; on today's Heron processors, with a 2.5x routing overhead on a 14-qubit all-to-all problem, it reads uniform. The manuals' "exact-regime only, replay to check" rule now carries a number: **about 99.9% after routing**.
3. **Retention above one is a warning, not a result.** The products' noise ladder now flags retention > 1 together with low exact concentration as a seed-lock signature, and the top-10 mass is printed as the primary noise indicator.
4. **A calibrated prediction is an optimistic bound.** The products label the IBM-calibration noise source accordingly: it excludes thermal relaxation, and on this class of circuit it over-predicted the device by a factor of about 3. The device run, at 44 seconds for 100,000 shots, is both cheaper and closer to the truth than the prediction.

8. Limitations

- **One book, one training per circuit.** The sweep is a single 14-qubit instance with a known-wrong seed; each circuit is one trained object, so training-seed variance is not measured. The seed-lock behaviour is expected to generalise (it reproduces the concentration study's real-book findings); the specific retention numbers are not.
- **Synthetic rungs on logical circuits.** The parametric ladder did not include routing; the 2.5x overhead was discovered on the device. A simulated ladder on the routed circuits would place the device more precisely than the "at or below the 99.0% rung" reading of Section 5.3.
- **The calibrated prediction omits thermal relaxation** (Section 6.1), so part of the factor-3 residual may be recoverable with the full model at much higher cost. The decomposition of the residual (idle decoherence, crosstalk, coherent errors, drift) was not attempted.
- **Statistics at depth.** Every device and calibrated row is below the 100-hit floor on P(optimum); the device statements rest on the top-10 mass and the most-likely-state probability against explicit uniform floors. The calibrated twins at p >= 5 ran with 2,048-4,096 shots and are themselves at their floors.
- **No error mitigation.** The device ran raw (no twirling, no readout mitigation, no dynamical decoupling beyond the transpiler's default scheduling), which is how the product runs and the fair comparison to the raw simulations - but not the best a Heron can do.
- **One calibration, one day, one backend.** The optimism factor is a single measurement on ibm_kingston on 22 August 2026.

9. Reproducibility

Every run is a licensed product run with a job identifier; the committed raw results (**results_ladder.committed.jsonl** for the simulated ladder, **results_hardware_rung.json** and the per-run result files for the device and calibrated rungs), the design lock with amendments, the figure and table scripts and the independent retention verifier are held in the study folder and are available on request (contact@qubit-lab.ch). IBM job identifiers for the device rung: da4m4jk3jnrc73agd8kg (p = 1), da4m66s3jnrc73agda6g (p = 3), da4m993otlns739b191g (p = 5), da4mcsk3jnrc73agdgdg (p = 7), da4mfak3jnrc73agding (adaptive). The tables in this document are generated from the committed files by script; no number was typed by hand.

[1] qubit-lab.ch, QAOA Concentration Study, August 2026 - https://qubit-lab.ch/QAOA_Concentration_Study_2026-08.pdf

Appendix A - The pre-specified design (DESIGN.md, 21 August 2026) and its amendments

Question. How does readout concentration degrade with noise as circuit depth grows — and is there a measurable noise-optimal depth? Secondary: does a calibrated noise simulation predict the real device (hardware rung, run later when a token is at hand)?

Design is locked before any run. Amendments will be recorded here with dates.

A.1 Instrument

Portfolio Optimization RQP Pro (V9 9.7.0, api rev 00082) as deployed — licensed runs through the product, every number from its certified reports. Replay-only noise (the product's rule): circuits are trained on the exact simulator, then replayed under the noise model. No training under noise.

A.2 Instance

The 14-qubit adaptive-demo book (**QuantumPortfolioOptimizer_demo_14_adaptive.xlsx**, resource-room copy): dense real book, certified optimum proven by SCIP per run, heuristic seed known-wrong by 3 assets (gap 1.60%) — the classical seed start is therefore seeded from a wrong answer on this book; predictions below account for it. AMENDMENT OPTION (decide at first run, record here): if the wrong seed suppresses exact P(optimum) below the statistical floor at any depth, switch the sweep to **qaoa_warm_start_reference_bits**-free demo book demo_7/demo_16 equivalent — NOT exercised unless needed.

A.3 Arms

Depth sweep (one variable = depth; same family, same seed, same recipe):

- Ansatz standard, CVaR objective (alpha 0.10), classical seed start (epsilon 0.25), p in {1, 3, 5, 7}; qaoa_maxiter 150, restarts 3, layer-wise angle warm start off, rng seed 42 (one training per depth; the trained circuit is the object under test).

Separate exhibit (different family, not part of the sweep):

- Adaptive ansatz, cap 12, CVaR, classical seed start, 250 iterations/layer, seed 42.

Noise rungs per circuit (parametric, primary knob = 2-qubit gate fidelity F2; 1-qubit error = (1-F2)/10, readout = 3*(1-F2); honest depolarizing conversion):

- none: exact statevector (the run itself)
- low: F2 = 99.9%
- medium: F2 = 99.5%
- high: F2 = 99.0%
- calibrated: NoiseModel.from_backend of the chosen IBM backend (prediction rung)
- hardware: same backend, real device (measurement rung) — DEFERRED to a later session with Daniel's IBM token; not part of this run set.

One licensed run per (circuit, noise rung): 5 circuits x 3 parametric rungs = 15 runs plus the 5 exact columns come free with each run. Calibrated rung: only if a token is available this session, else deferred with the hardware rung (recorded here).

A.4 Budgets and statistical floor

Noisy replay shots: 16,384 at p <= 3; 65,536 at p >= 5, at F2 99.0%, and for the adaptive circuit. FLOOR: a noisy rung whose optimum-hit count is below 100 is reported as an upper bound only, never as a retention ratio (the single-shot trap, recorded 2026-08-21). Worker: small; medium if the 65,536-shot replays exceed the 900 s guard (the guard's automatic shot reduction is recorded if it fires).

A.5 Metrics per rung (all from the product's noise ladder / reports)

P(optimum), x-uniform, top-10 mass, modal flag, shots-to-99%, retention vs exact, 2-qubit gate count and depth, expected circuit fidelity (crude product-of-errors).

A.6 Predictions (stated before running)

1. Exact P(optimum) saturates by p ~ 3 (phase-2 depth result) - on THIS book the wrong seed may cap it low at all depths; whatever it is, it will not rise past p=3 by more than noise on the training.
2. Noisy P(optimum) falls roughly geometrically with depth: each layer adds ~180 two-qubit gates, so retention per layer ~ (F2)^180 - about 0.83x/layer at 99.9%, 0.40x/layer at 99.5%, 0.16x/layer at 99.0%.
3. Therefore the noise-optimal depth is the saturation point (p ~ 3) at 99.9%, and p = 1 at 99.5% and below - deeper never pays under noise on this instance.
4. Measured retention will EXCEED the expected-circuit-fidelity product (as in the first credit ladder: 0.086 vs 0.019), because depolarized shots still land on the optimum by chance and partial coherence survives; the gap shrinks as F2 falls.
5. The modal flag survives noise longer than the mass does (credit ladder precedent).

A.7 Analysis plan (fixed)

One figure: P(optimum) vs p, one line per rung (exact / 99.9 / 99.5 / 99.0), the adaptive circuit as detached points at its layers-used; hit counts printed per point; retention table; noise-optimal depth read off per rung as argmax of noisy P(optimum). Negative results reported. Memo + figure committed under this folder. Publication (evidence page / B-post / newsletter #3) only after Daniel reviews the memo.

Amendment 2026-08-21 (after the first arm, before any analysis) The wrong seed caps exact P(optimum) at ~0.07% at p=1 (11 noisy hits at 16,384 shots

- below the floor). Instance is KEPT (the wrong-seed cap matches the concentration study and is part of the story); the shot ladder is raised instead: p=1 -> 131,072 shots on all rungs; p>=3, adaptive, and all F2=99.0% rungs -> 65,536. Rungs whose noisy hit count still lands below 100 are reported as upper bounds ("< x"), per the floor rule - expected for deep circuits at F2 <= 99.5%, and that inability to measure is itself the result. The first p=1/99.9% arm (16,384 shots, job_239c92af) is superseded by its 131,072-shot rerun and kept in the raw data.

Amendment 2026-08-21b (after the batching fix) Production qiskit-aer segfaulted (signal 11) on single noisy runs well above ~16k shots; v9 9.7.1 samples in batches of <= 16,384 (summed counts, per-batch seeds recorded) - validated on the previously crashing arm. p=1 arms raised to 262,144 shots so their optimum-hit counts clear the 100 floor (the completed 131,072-shot p=1/99.9% arm stands, reported with its hit count of 88 and CI).

Amendment 2026-08-21c (before resuming) Shot branching disabled in v9 9.7.2 (container segfault, reproduced in the image; plain trajectories stable at ~2.4x cost). Wall guard on the medium worker raised to 2,400 s via env for the deep arms (p=7 at 65,536 no-branch trajectories ~ 21 min); default stays 900 s elsewhere. Sweep resumes with the previously crashing arm (p=1, F2 99.5%, 262,144 shots) as validation.

Amendment 2026-08-21d (hardware rung scope, Daniel; extended 2026-08-22) The deferred hardware rung runs FOUR circuits: p=1 (control, near noise-free by construction), p=3 and p=5 (the measurable middle - 546 and 910 CX; two mid points make the device's per-layer decay a fitted slope rather than a guess), adaptive-12 (the exhibit - likely depolarized; the measurement is the point). Same four circuits on the calibrated-sim rung (same backend, prediction vs device). Runs executed by Daniel through the product page with his IBM token (token never leaves the session); analysis from the job results.

Amendment 2026-08-22e (hardware rung = full parity, Daniel) Hardware and calibrated-sim rungs run the SAME FIVE circuits as the simulated rungs (p=1, 3, 5, 7, adaptive-12). Backend chosen on the day by best CZ median error among the Open-plan Herons: ibm_kingston (CZ median 0.213%, best quartile 0.157%, readout median 0.87% on 2026-08-22). Hardware shots 20,000 per circuit (the product ties hardware shots to the candidate-pool request in exact mode). Fractional gates not exposed on these backends on the day -> CZ decomposition; transpiled two-qubit counts recorded per run. Runs submitted through the MCP connector with the token injected server-side (rqp-mcp fd98297).

Amendment 2026-08-22f (hardware rung executed; calibrated rung definition) Hardware rung DONE on ibm_kingston, 20,000 shots per circuit, five circuits (IBM jobs da4m4jk3jnrc73agd8kg, da4m66s3jnrc73agda6g, da4m993otlns739b191g, da4mcsk3jnrc73agdgdg, da4mfak3jnrc73agding). Routing on the heavy-hex lattice multiplied the two-qubit count by ~2.5 (182->440, 546->1,302, 910->2,185, 1,274->3,101, 2,190->5,454 CZ) - the simulated rungs

were keyed on LOGICAL gate counts; this is recorded as the first hardware finding. Hardware concentration metrics are computed from the product's returned counts against the certified optimum (same definitions as the simulated ladder). Calibrated rung DEFINED as: per-gate depolarising from ibm_kingston's calibrated gate errors + calibrated readout, on the routed circuit, thermal relaxation excluded (Kraus sampling is orders of magnitude more expensive than Pauli sampling and would make the deeper twins infeasible; same channel family as the parametric rungs). Twins at 32,768 requested shots, reduced at depth to fit the worker's guard; the shots actually sampled are recorded per arm. One twin (p = 7) repeated after a transient IBM account-service error; the repeat is the reported run.

Series terminology note: pre-specified, not pre-registered - the design was locked in writing before execution, but registration-grade public chronology is proven only from the VQC island study onward (public design tag before the first run); future studies add an independent registry deposit.

Revision 1.2 (August 2026): presentation-level revision - management summary added, pre-specification terminology aligned with the series, series cross-references. All measurements and verdicts are unchanged from v1.1.

Revision 1.3 (August 2026): finite-shot intervals added to the headline statistics, recomputed from the committed raw counts (each published figure was reproduced exactly before an interval was attached). Conventions: optimum-hit proportions carry Wilson 95% intervals; the zero count carries a rule-of-three upper bound; the prediction-optimism factors carry binomial-bootstrap 95% intervals with the observed top-10 sets held fixed (re-selecting the set per resample is upward-biased near the uniform floor); no intervals are attached where both sides sit at the uniform floor. These intervals quantify shot noise only - the one-calibration, one-day, one-backend scope of Section 8 is unchanged, and the routing overhead is a deterministic transpiler count that carries no interval. One correction: the most-likely-state optimism range is 3.3-4.2x on the raw counts; the previously printed 3.8-4.4x was an artefact of computing ratios from rounded table entries. All verdicts are unchanged.

Disclaimer

This document reports measurements made with qubit-lab.ch's rapid quantum prototyping tools on one portfolio instance, three synthetic noise models, one calibration snapshot and one quantum processor on one day. It is a method study, not investment advice and not a claim of quantum advantage. All probabilities are measured against a certified classical optimum; all device results are raw, unmitigated readouts. Numbers are reproducible from the committed raw results described in Section 9.