QONTOS

TECHNICAL PAPER

# A modular superconducting-photonic quantum computer.

First-machine specification and gated research vision.


ABSTRACT

This paper specifies QONTOS-1, a two-module 1,000-physical-qubit superconducting-photonic quantum computer linked by microwave-to-optical photonic interconnects, and separates its near-term engineering targets from QONTOS-2 to QONTOS-5 research scenarios. Every numerical target is anchored to a measurable subsystem acceptance gate or is explicitly labelled as a scenario or research-vision assumption.



| | |
|---|---|
| DOCUMENT | QONTOS architecture whitepaper |
| VERSION | v5.4 · May 2026 |
| AUTHORS | QONTOS Programme |
| INSTITUTION | Zhyra Quantum Research Institute · Abu Dhabi |
| STATUS | Technical draft, for external review |

CONTENTS

# Contents

*Section index for the QONTOS architecture whitepaper.*

01

# Overview

*Architecture, first machine, and what this document covers.*

---

QONTOS is a modular quantum computing architecture in which superconducting transmon modules are connected through microwave-to-optical photonic interconnects. The first machine, designated **QONTOS-1**, is specified for two modules of 500 transmon qubits each (1,000 physical qubits total), linked by a heralded Bell-pair distribution channel with a conservative base transduction threshold of $\eta \geq 0.1\%$ for link-physics validation. Higher values, $\eta \geq 0.5\%$ for aggressive validation and $\eta \geq 1\%$ for research-scale distributed operation, are explicitly not treated as QONTOS-1 acceptance requirements. A software runtime, circuit ingest, partitioning, scheduling, executor binding, decoding, and result aggregation, is currently operational against simulator and provider backends, and forms the control plane for the first hardware integration.

This document describes the QONTOS architecture, the engineering specification of QONTOS-1, the photonic interconnect physics, the quantum error correction strategy, the software runtime, the gated engineering programme that brings QONTOS-1 to first logical-qubit operation, and the five-generation **family arc**, QONTOS-1 through QONTOS-5, that extends this architecture toward datacenter-scale fault tolerance. Each major claim is anchored to a measurable subsystem specification with explicit acceptance criteria.

## What this whitepaper is not

- Not a marketing document. The audience is engineering and scientific reviewers.
- Not a complete physics tutorial. Familiarity with superconducting qubits, surface codes, and photonic interconnects is assumed.
- Not a vendor benchmark. Targets are stated as engineering objectives gated by subsystem milestones, not measured performance.
- Not a calendar. The family arc is a gated engineering map, not a delivery schedule. Generation transitions are conditional on subsystem milestones, not dates.

## Notation and conventions

Throughout: $\eta$ denotes microwave-to-optical transduction efficiency; $p_{phys}$ denotes per-gate physical error rate; $\varepsilon_L$ denotes logical error rate per logical-qubit operation; $d$ denotes surface-code distance.

Numerical targets are quoted with their evidence basis: **measured** (from operational hardware or platform), **simulated** (from digital twin or device modelling), **target** (engineering objective with stated gate criteria), **scenario** (committed for a future generation conditional on the prior generation's acceptance), or **research vision** (long-range architecture study requiring multiple unresolved research programmes to mature together).

02

# System architecture

*Compute, interconnect, and control as one machine.*

---

The QONTOS architecture comprises three engineering domains operated as a single system: compute (superconducting transmon modules), interconnect (microwave-to-optical transducers and a low-loss optical channel), and control (room-temperature electronics with a real-time decoder farm). The three domains are coordinated by a software runtime that exposes a single execution model to the application layer, independent of how a circuit is partitioned across modules.

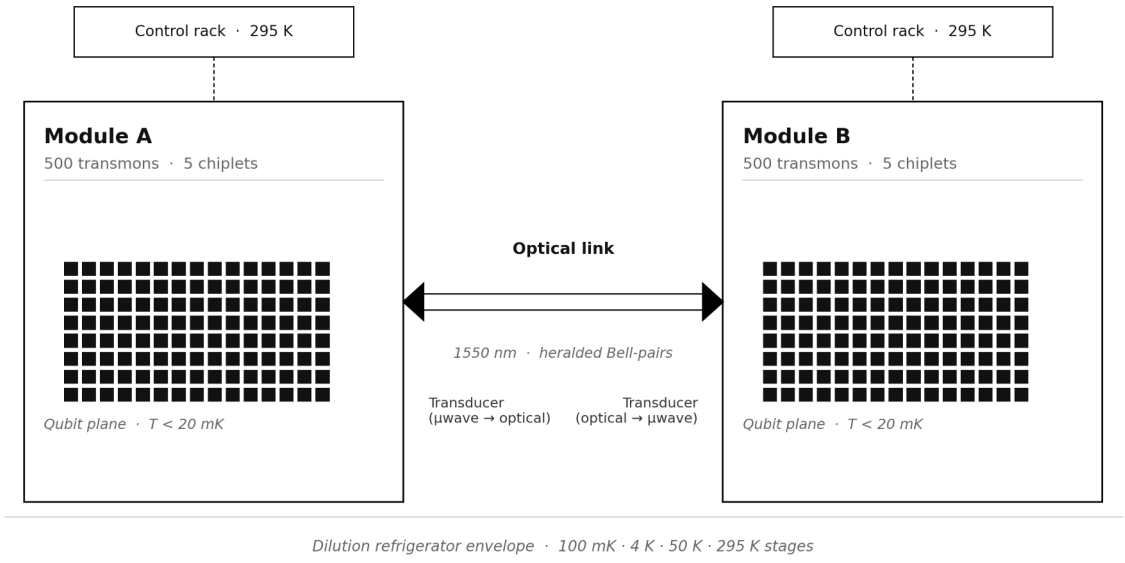



**Figure 1.** *QONTOS-1 system layout. Two superconducting transmon modules, each operating below 20 mK in a dilution refrigerator, are linked by an optical channel carrying heralded Bell pairs at 1550 nm. Each module is driven by a room-temperature control rack containing arbitrary waveform generators, digitisers, and an FPGA sequencer.*

## Compute domain

Each module houses five superconducting chiplets in a flip-chip package, with 100 tantalum-on-silicon transmon qubits per chiplet. Intra-chiplet connectivity follows a heavy-hex lattice; chiplet-to-chiplet routing is achieved through superconducting through-silicon interconnects. The qubit chip operates on the mixing chamber stage of the dilution refrigerator at a base temperature below 20 mK.

Single-qubit gates are implemented as resonant microwave drive pulses with target gate error $1 \times 10^{-4}$. Two-qubit gates are realised as tunable-coupler CZ operations with target gate error $5 \times 10^{-3}$. Dispersive readout proceeds through a chain of cryogenic amplifiers and room-temperature digitisers with target assignment fidelity 99% at 1 µs integration time.

## Cryogenic stack

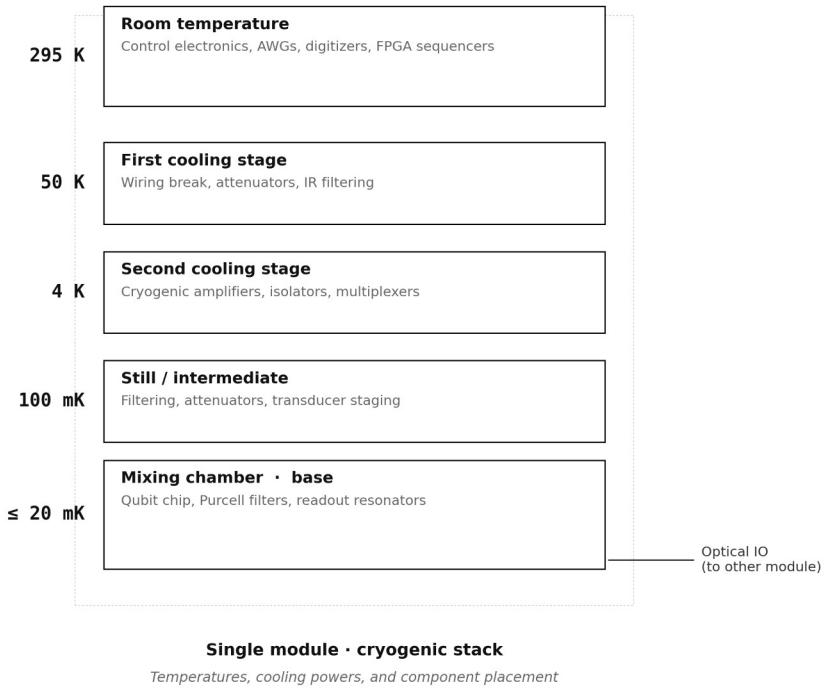



**Figure 2.** *Cryogenic envelope for a single module. The 100 mK still stage carries the photonic transducer; the mixing chamber stage at ≤ 20 mK carries the qubit chip and readout resonators.*

Each module operates within a commercial dilution refrigerator with the following stage specification: 50 K and 4 K stages for wiring thermalisation and primary amplification; 100 mK still stage for transducer staging and intermediate filtering; mixing chamber stage at ≤ 20 mK with target cooling power 500 µW at

100 mK and ≥ 25 µW at 20 mK. Wiring density is constrained to maintain a passive thermal budget below 1 µW at the mixing chamber.

### Control domain

Each module is driven by one room-temperature control rack housing AWGs with 1 ns sample-rate resolution, digitisers with synchronised trigger distribution at sub-ns jitter, and an FPGA sequencer implementing the real-time loop described in §5. The control rack is connected to the cryogenic stack through a wiring break at the 50 K stage and through Purcell-filtered drive lines at the qubit plane.

0 3

# Superconducting compute module

*Transmon physics, lattice topology, gates, readout.*

---

Each QONTOS-1 compute module is a five-chiplet superconducting processor housing 500 fixed-frequency transmon qubits. This section describes the device physics, the chip-level topology, the gate set, the dispersive readout chain, and the calibration workflow that brings these elements into a coherent computational substrate. The objective is a per-cycle error floor that brings the QONTOS-1 module to within striking distance of the surface-code threshold without optimistic assumptions about device physics. [1, 2]

## 3.1 Transmon device physics

The transmon is a weakly anharmonic oscillator formed from a Josephson junction shunted by a large capacitance. In the two-level approximation it is described by the Hamiltonian:

$$\hat{H}_q / \hbar = \omega_q a^\dagger a - (\alpha / 2) a^\dagger a^\dagger a a \quad (1)$$

*$\omega_q$ is the qubit transition frequency; $\alpha$ is the anharmonicity that separates $|1\rangle \to |2\rangle$ from $|0\rangle \to |1\rangle$ and enables a closed two-level operation under shaped microwave drives.*

QONTOS-1 transmons are designed in the EJ / EC ≈ 50 regime, with $\omega_q / 2\pi = 5.0 \pm 0.2$ GHz and $\alpha / 2\pi \approx -300$ MHz. Fabrication is on high-resistivity silicon with tantalum capacitor pads and aluminium Josephson junctions, a materials stack that has demonstrated $T_1$ in excess of 300 µs in published device characterisations.[3] The QONTOS-1 device specification targets a process-window-aware design point: $T_1 \geq 200$ µs, $T_2$_echo ≥ 100 µs, and frequency yield ≥ 90% within a ±20 MHz window after one round of laser-trimming.

## 3.2 Heavy-hex lattice topology

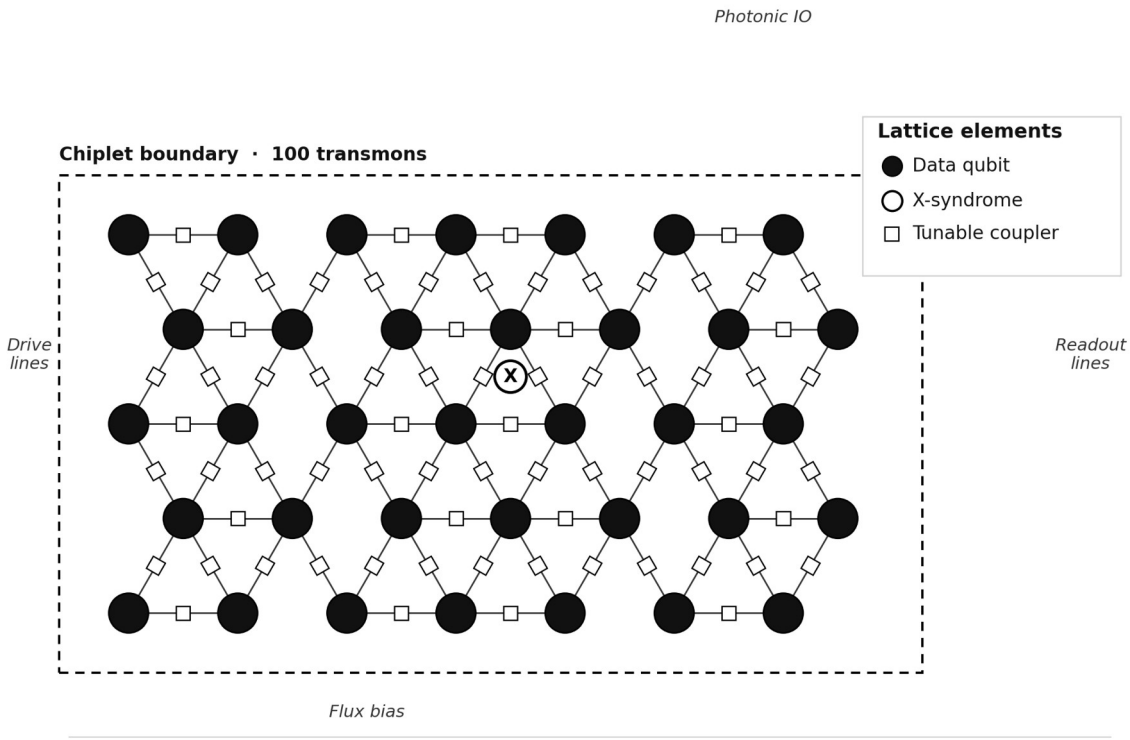



**Figure 3.** *Single-chiplet heavy-hex topology. Data qubits (filled circles) are connected by tunable couplers (small squares) in a heavy-hex pattern; X-syndrome ancillas (open circles) occupy the centres of plaquettes. Each chiplet hosts 100 transmons; five chiplets per module yield 500 physical qubits.*

QONTOS-1 adopts the heavy-hex connectivity demonstrated on IBM Eagle and Condor processors.[4] The heavy-hex graph has vertex degree two or three, sparser than the square-grid topology assumed by the canonical surface code but matched to a rotated-surface-code variant with explicit ancilla qubits.[5] The benefit is a substantial reduction in frequency-collision risk: with one tunable coupler per edge, the lattice supports independent calibration of each two-qubit gate without the all-to-all frequency-matching constraint of fixed-coupler architectures.

Chiplet boundaries are crossed by superconducting through-silicon interconnects, with the loss budget specified at less than 1 dB per crossing at 6 GHz. A five-chiplet module therefore presents an effective lattice of 500 transmons with at most four chiplet boundaries in any logical-qubit patch, well within the loss budget for a distance-7 surface code.

Frequency yield is treated as an integration risk rather than a post-fabrication assumption. The ±20 MHz laser-trim target is paired with a collision-mitigation plan: qubits are assigned to data, measure, and parking roles after wafer-level spectroscopy; tunable couplers absorb residual detuning where possible; and

unusable collision clusters are mapped out by the compiler. The acceptance criterion is not raw frequency yield alone but calibrated heavy-hex edge yield after role reassignment and coupler-bias optimisation.

The remaining defective edges are handled explicitly rather than averaged away. QONTOS-1 compiler qualification includes simulated heavy-hex yield maps with up to 15% random edge loss; the surface-code mapper must place and route logical patches around those defects at the supported distances, and modules that cannot preserve patch connectivity are binned as yield failures rather than counted toward the calibrated-edge-yield acceptance set.

## 3.3 Single-qubit gates

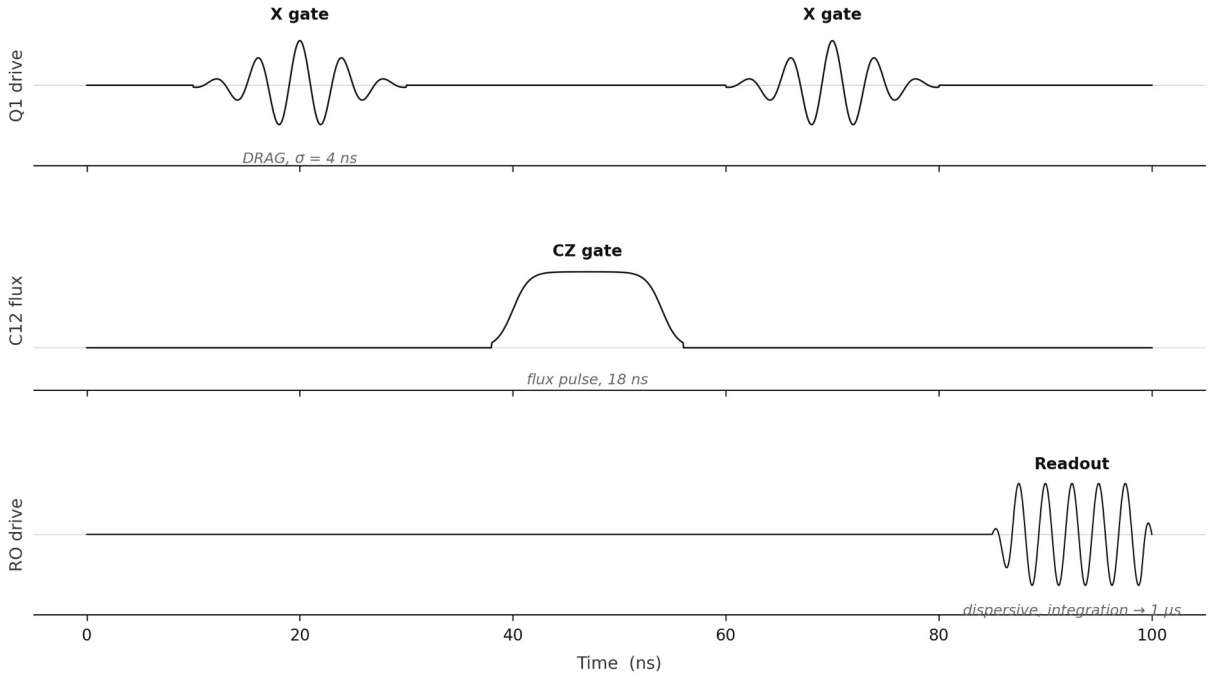



**Figure 4.** *Pulse sequence for one logical operation on a two-qubit subsystem: two single-qubit X gates on Q1 (DRAG-shaped, σ = 4 ns), one CZ gate via flux-pulsed tunable coupler C12, and a dispersive readout drive on the readout resonator. All channels share a common 1 ns FPGA sample-rate grid.*

Single-qubit gates are implemented as resonant microwave drives with derivative-removal-by-adiabatic-gate (DRAG) envelopes,[6] applied through Purcell-filtered drive lines at the qubit plane. The DRAG pulse has the form:

$$\Omega(t) = \Omega_x(t) \cdot \cos(\omega_q t) - (\Omega_y(t) / \alpha) \cdot \sin(\omega_q t) \quad (2)$$

*Ω_x(t) is a Gaussian envelope of width σ ≈ 4 ns; the Ω_y(t) term suppresses leakage to |2⟩ by applying a derivative correction scaled by the anharmonicity α.*

With σ = 4 ns and total pulse length 4σ = 16 ns, the QONTOS-1 target single-qubit gate error is $1 \times 10^{-4}$ as measured by randomised benchmarking, dominated in roughly equal parts by coherent control imperfection and incoherent decay during the pulse window. Calibration of single-qubit gates is performed every 4 hours via a closed-loop optimisation that updates amplitude, drag coefficient, and detuning to track slow drift in qubit frequency and AWG output.

## 3.4 Two-qubit gates

Two-qubit entangling gates are implemented as controlled-Z (CZ) operations via a tunable coupler placed between every nearest-neighbour data-qubit pair.[7] The coupler is itself a transmon, with its frequency biased by a flux pulse to selectively activate the $|11\rangle \rightarrow |02\rangle$ avoided crossing for approximately 18 ns. Net coupling rates are $J / 2\pi \approx 5$ MHz at the activated bias, with idle coupling suppressed below 100 kHz. The QONTOS-1 target CZ error of $5 \times 10^{-3}$ is consistent with published two-qubit-gate fidelities on similar devices,[8] with leakage-out-of-computational-subspace targeted below $5 \times 10^{-4}$ per gate.

Two-qubit calibration is the most expensive step in the system calibration workflow: a full lattice characterisation runs at system bring-up and is refined every 24 hours, with online interleaved-RB checks every 4 hours to detect drift outside the operational envelope. Calibration drift events that exceed the drift envelope trigger an automatic re-baseline of affected couplers within a maintenance window.

## 3.5 Dispersive readout

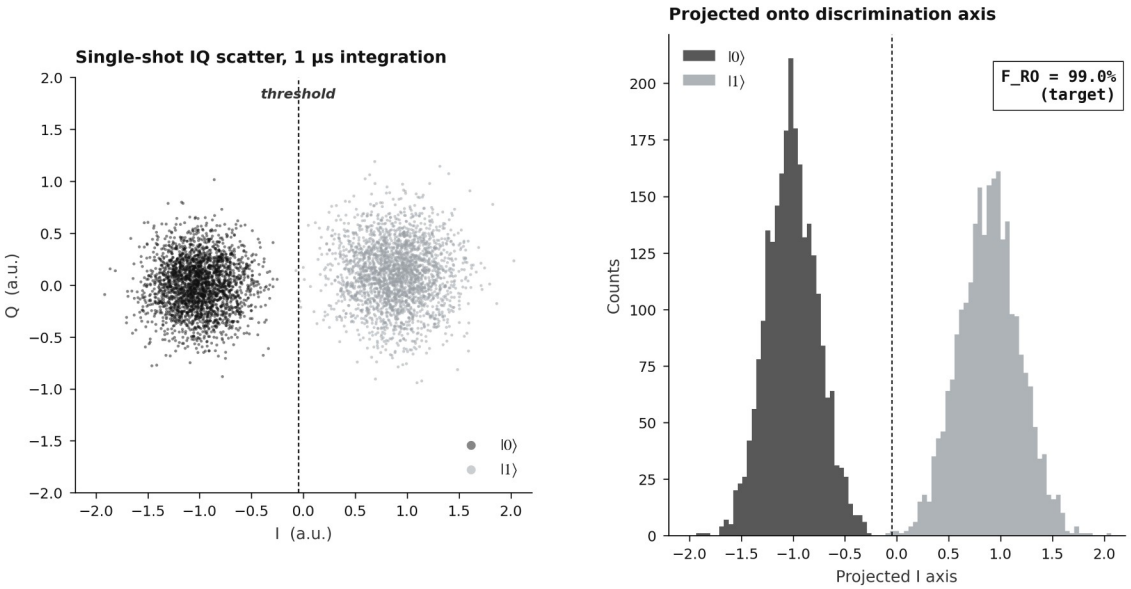



**Figure 5.** *Single-shot readout discrimination. Left: |0⟩ and |1⟩ states form distinct clusters in the (I, Q) signal plane after 1 µs of dispersive integration. Right: projected onto the optimal discrimination axis, the two distributions are well separated with target readout assignment fidelity F_RO = 99.0%.*

Readout is performed via the dispersive shift of a separate linear resonator coupled to each transmon. The qubit state translates into a state-dependent phase shift of a probe tone:

$$\chi = g^2 / (\omega_q - \omega_r) \quad (3)$$

*The dispersive shift χ; g is the qubit-resonator coupling and ω_r is the readout-resonator frequency. Detuning is set to keep | ω_q − ω_r| ≈ 1 GHz, well into the dispersive regime.*

Probe tones are amplified by a Josephson travelling-wave parametric amplifier at the 4 K stage and digitised at room temperature. The QONTOS-1 readout target, F_RO = 99.0% assignment fidelity at 1 µs integration time, is consistent with the noise temperature, dispersive shift, and resonator linewidth selected in the chiplet design. Frequency multiplexing of eight readout resonators per amplifier line constrains the wiring density at the 4 K stage to a level consistent with the cryogenic envelope described in §2.

The 8× readout-multiplexing target carries an explicit crosstalk budget. Resonator spacing, linewidth, TWPA compression margin, and digital matched filtering must keep correlated assignment error below 0.5% per resonator pair. If the measured crosstalk exceeds that value, the fallback is a 4× multiplexing mode with additional 4 K line count accepted before claiming the 99.0% readout target.

04

# QONTOS-1 specification

*The first machine, in numbers.*

---

QONTOS-1 is the first-generation hybrid machine. Its purpose is to validate the architecture end-to-end: two-module operation, photonic interconnect at full system scale, software runtime bound to native hardware, and the first distributed circuits across modules. QONTOS-1 is not designed to deliver fault-tolerant computation; it is designed to retire the architectural risks that stand between the current platform and a fault-tolerant successor.

**COMPUTE**

| | |
|---|---|
| Modules | 2 |
| Transmons per module | 500 (5 chiplets × 100) |
| Total physical qubits | 1,000 |
| Connectivity | heavy-hex lattice (intra-chiplet) · superconducting routing (chiplet-to-chiplet) |

| | |
|---|---|
| Qubit substrate | tantalum-on-silicon transmon |
| Base temperature | ≤ 20 mK |

### GATE AND READOUT TARGETS

| | |
|---|---|
| Single-qubit gate error | **target** $1 \times 10^{-4}$ |
| Two-qubit gate error | **target** $5 \times 10^{-3}$ |
| Readout assignment error | **target** $1 \times 10^{-2}$ (1 μs integration) |
| Stabilizer cycle time | **target** 1 μs |
| $T_1$ coherence | **target** ≥ 200 μs |
| $T_2$ coherence | **target** ≥ 100 μs |

### PHOTONIC INTERCONNECT

| | |
|---|---|
| Optical wavelength | 1,550 nm (telecom C-band) |
| Transducer architecture | electro-optic / piezo-optomechanical (under selection) |
| Transduction efficiency η | **target** ≥ 0.1% (base validation) · ≥ 0.5% (aggressive validation) · ≥ 1% (research threshold) |
| Heralding probability | **target** 0.05 – 0.10, measured under thermal closure |
| Bell-pair fidelity (raw) | **target** ≥ 0.85 for G3 tomography; ≥ 0.98 purified for logical operations |
| Input-referred added noise | **target** n_add < 1 photon for quantum-enabled operation; < 0.05 for high-fidelity Bell operations |
| End-to-end link latency | **target** ≤ 25 μs (0–5 m inter-module distance) |

### SOFTWARE AND CONTROL

| | |
|---|---|
| Runtime | **operational** QONTOS orchestration platform, bound to native executor |
| FPGA sequencer resolution | 1 ns per sample |
| Decoder | real-time minimum-weight perfect matching · syndrome ingest ≤ 5 μs |

| | |
|---|---|
| Feed-forward correction | conditional Pauli frame update · ≤ 10 µs total loop |

0 5

# Real-time loop

*Timescales from nanoseconds to seconds.*

---

Quantum error correction operates only when each stabilizer cycle, decoder pass, and feed-forward correction completes within the qubit coherence window. The QONTOS control loop is specified as a hierarchy of nested timescales, each with measurable acceptance criteria.

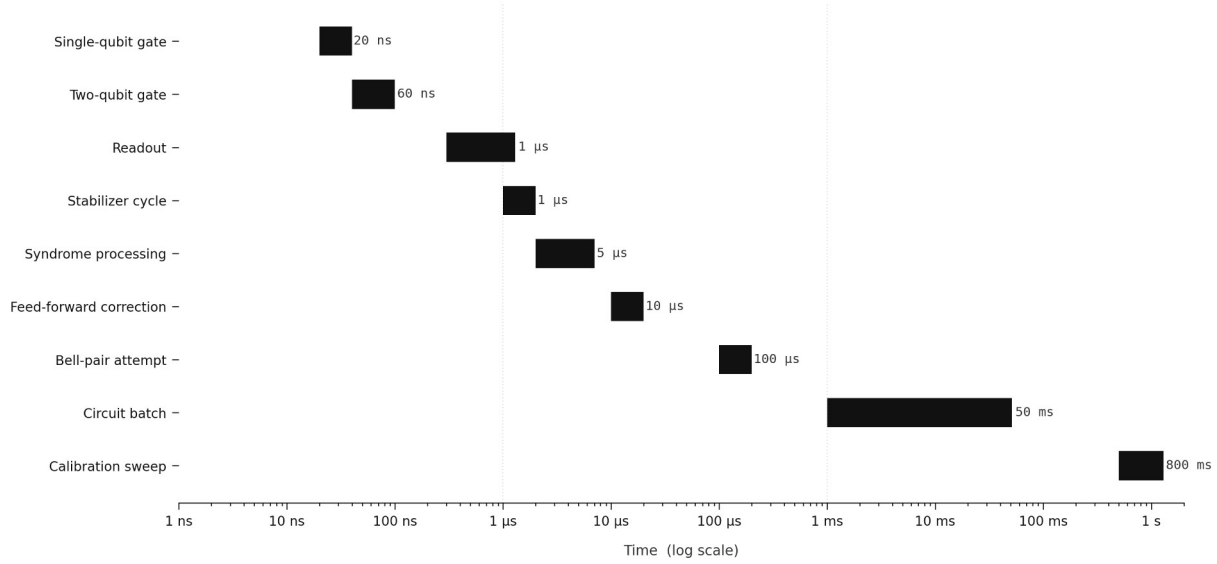



**Figure 6.** *Timescale hierarchy of the QONTOS real-time loop. Individual gates execute in tens of nanoseconds; the stabilizer cycle closes in 1 µs; the decoder ingests syndromes and emits feed-forward corrections within 10 µs; circuit batches and calibration sweeps occupy the millisecond-to-second band.*

The decoder latency budget is the critical specification. With a stabilizer cycle time of 1 µs and a decoder ingest target of 5 µs, the buffer must absorb a sustained 5-cycle backlog before feed-forward correction becomes possible. This sets the minimum throughput requirement on the matching engine and the minimum $T_2$ requirement on the qubits.

### Software runtime

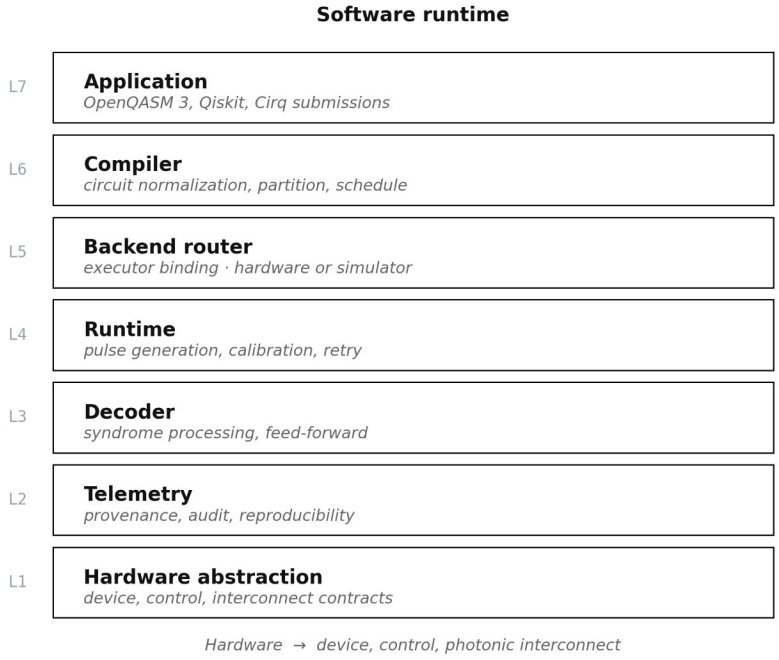



**Figure 7.** *Layered organisation of the QONTOS software runtime. Layers L7–L1 are operational against simulator and provider backends today; L1 (hardware abstraction) gains native executor bindings at the first-module bring-up gate.*

The runtime accepts circuits in OpenQASM 3, Qiskit, and Cirq formats and produces three artifacts per execution: result data, a provenance record (compiler version, calibration epoch, backend identity, sample counts), and a replay envelope sufficient to reproduce the execution on a different backend. The replay envelope is the mechanism by which results from simulator, provider, and native hardware are made directly comparable.

0 6

# Software platform and orchestration

*Circuit ingest, partitioning, scheduling, distributed execution, and verifiable result aggregation.*

---

The QONTOS software platform is the developer-facing interface to the hardware described in this paper. It is a Python SDK that compiles user circuits, partitions them across an evolving compute substrate, schedules each partition onto a backend, executes the assignments concurrently, aggregates the results, and emits a cryptographic proof of execution. The platform is

operational today against simulator and external-provider backends; the same surface becomes the production runtime for QONTOS-1 at first-module bring-up. Public source, examples, benchmarks, and digital twin are available at github.com/qontos under the Apache-2.0 license.

## 6.1 Design principles

Three architectural choices shape the software platform. **Separation of planning from runtime**: circuit normalisation, partitioning, scheduling, and backend assignment complete before any hardware tick, leaving the real-time loop of §5 to execute the prepared plan without further planning overhead. **Provider-agnostic execution**: every backend (simulator, IBM Quantum, Amazon Braket, future native QONTOS hardware) is addressed through a single ExecutorContract interface, so circuits compile once and run anywhere. **Cryptographic execution integrity**: every run emits a three-layer SHA-256 proof chain that binds the circuit, the plan, and the result into a single verifiable artefact.

The platform is organised as seven layers (L7 to L1), shown in the runtime stack shown in §5. L7 to L5 handle planning (ingest, compilation, routing); L4 is the real-time runtime described in §5; L3 to L1 handle decoding, telemetry, and hardware abstraction. Each layer publishes a stable contract; layer implementations can be swapped independently as the hardware substrate evolves.

## 6.2 Circuit ingest and normalisation

The ingest layer accepts circuits in four representations: OpenQASM 2.0, OpenQASM 3.0, native Qiskit QuantumCircuit objects, and PennyLane JSON tape descriptions. Each is parsed into a canonical intermediate representation (CircuitIR) with an explicit qubit count, depth, gate list, measurement schedule, and a provenance record naming the source format and parser version. Validation runs at ingest: gate-set compatibility, qubit-index bounds, measurement-target consistency, and deferred-classical-control resolution all fail at ingest rather than at execution.

The CircuitIR is the single object passed between every layer below the ingest. No downstream layer needs to understand OpenQASM, Qiskit, or PennyLane semantics directly. This is what makes the runtime hardware-agnostic: a circuit written in Qiskit and a circuit written in OpenQASM 3.0 become the same object after ingest and execute on the same backend through the same scheduling logic.

## 6.3 Partition planning

Circuits that exceed a single module's qubit envelope are decomposed into partitions, each of which can execute on a single module with cross-module entanglement realised through the photonic interconnect of §7. Three partitioning strategies are supported, selected automatically by circuit size unless the user overrides:

- **GreedyPartitioner** (O(n)), used for circuits below ~20 logical qubits; assigns gates to partitions in topological order, fast iteration during development.
- **SpectralPartitioner** (O(n log n)), the default for larger circuits; constructs the gate-connectivity graph and uses spectral clustering on its Laplacian to minimise the number of inter-partition gates (which translate into Bell-pair-distributed CNOTs at runtime).
- **ManualPartitioner** (O(1)), used when the user has supplied an explicit qubit-to-module mapping (typical for benchmarking or replay-from-proof workloads).

The partition plan is itself an immutable artefact: it carries the strategy used, the qubit-to-partition mapping, the inter-partition gate list, and a hash that binds it to the source CircuitIR. This hash propagates into the execution proof of §6.6.

## 6.4 Scheduling and execution routing

Once partitioned, each partition is scored against every available backend using a four-factor weighted sum: gate fidelity match (weight 0.60), backend queue depth (0.15), normalised cost per shot (0.10), and qubit capacity fit (0.15). Backends are ordered by score; the top assignment for each partition is emitted as a scheduled task. Weights are configurable per-policy through the ScoringWeights record, allowing fidelity-first, cost-first, or latency-first scheduling profiles depending on the workload.

The scheduler emits a list of tasks rather than dispatching directly; this keeps the planning stage purely deterministic and side-effect-free. Dispatching is the responsibility of the execution router, which holds the live connection state to each backend and handles authentication, submission, polling, and cancellation. A new backend is added to the platform by implementing the ExecutorContract interface (validate, submit, poll, cancel, normalise_result, normalise_error) and registering the implementation; no scheduler or planner changes are required.

## 6.5 Result aggregation

Partitioned execution produces one result per backend. The result aggregator reconstructs the full-circuit distribution from the partition results using the inter-partition gate structure recorded in the partition plan. Aggregation is mathematically grounded: for product-state partitions, the joint distribution is the tensor product of the partition distributions; for entangled partitions, the inter-partition Bell pairs introduce known classical-shadow correlators that are inverted at aggregation. Shot-count mismatches between partitions are handled through importance-weighted resampling with a Bessel-corrected variance estimator.

Aggregated results carry an explicit provenance record: the list of backends used, the compiler version, the partition strategy, the shot counts per partition, the time of each dispatch, and the calibration epoch active at each backend at dispatch time. Together these constitute the replay envelope described in §5.

## 6.6 Execution integrity

Every run emits a three-layer cryptographic proof chain. **Layer 1** binds the source circuit and the manifest (input format, parser version, validation result) into a single SHA-256 hash. **Layer 2** binds the partition plan, the scheduling decision, and the chosen backend assignments. **Layer 3** binds the result, the provenance record, and the calibration epoch active during execution. Each layer's hash incorporates the previous layer's hash, so any tampering at any stage invalidates the chain.

The proof chain serves three purposes. First, **reproducibility**: given the proof, the same run can be replayed on any backend, and the replayed result can be verified to match (within statistical tolerance) the original. Second, **auditability**: for regulated workloads (financial modelling, pharmaceutical discovery, cryptographic analysis), the proof chain provides a machine-checkable record of which backend produced which result and under what conditions. Third, **comparability**: results from a simulator, from an external provider, and from native QONTOS hardware become directly comparable because their proof chains share the same canonical structure.

## 6.7 Pre-hardware operability and the open ecosystem

The platform is operational today against three classes of backend: local simulators (Qiskit Aer and the qontos-sim digital twin), IBM Quantum hardware, and Amazon Braket. This means the L7 to L4 layers of the runtime are exercised end-to-end against real workloads ahead of any QONTOS-1 hardware integration, retiring software risk on a timeline independent of the hardware bring-up gates of §10. At first-module bring-up, the platform gains a fourth backend class (native QONTOS hardware) through the same ExecutorContract interface that already addresses the simulator and external-provider backends.

The platform and its supporting tooling are open. Six repositories at **github.com/qontos** carry the public source: the **qontos** SDK (the platform described in this section); **qontos-sim** (simulators, digital twin, and tensor-network models for pre-hardware validation); **qontos-benchmarks** (reproducible evidence, methodology, and regression validation); **qontos-examples** (runnable notebooks and adoption paths); **qontos-research** (whitepapers, roadmap, and technical publications including this paper); and the **.github** organisation-standards repository. All public repositories are Apache-2.0 licensed.

The platform's role in the engineering programme is to make QONTOS-1 hardware integration the **only** novel risk at first-module bring-up. By the time the first transmon module ships, the orchestration layer, the scheduling logic, the result aggregation, and the proof chain will all have run against simulator and

provider backends at scale, with their performance characterised against the qontos-benchmarks suite. What QONTOS-1 contributes is a fourth backend class; the software around it is already in production.

07

# Photonic interconnect

*Microwave-to-optical transduction and heralded Bell-pair distribution.*

---

The QONTOS interconnect distributes heralded Bell pairs between modules using microwave-to-optical transduction. A microwave photon emitted from a transmon in Module A is up-converted to 1,550 nm by an electro-optic resonator, propagated through a low-loss fibre to Module B, down-converted back to microwave, and absorbed into a target transmon. The process is heralded, successful Bell-pair generation is signalled by a coincidence click on a superconducting nanowire single-photon detector (SNSPD), so failed attempts do not corrupt the logical state.[20,21]

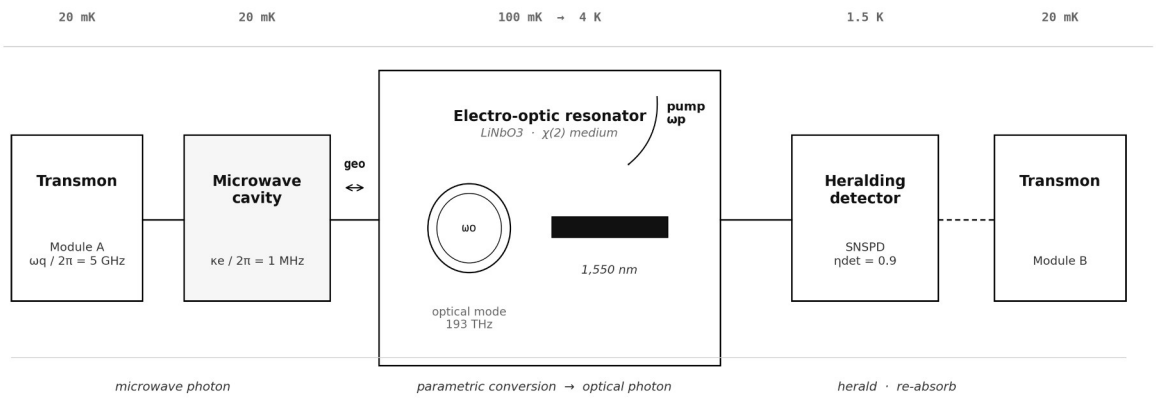



**Figure 8.** *Microwave-to-optical transducer subsystem. The transmon in Module A is dispersively coupled to a microwave cavity which shares an electro-optic $LiNbO_3$ resonator with an optical mode at 1,550 nm. A pump tone at ω_p mediates parametric conversion. The optical photon propagates to the heralding SNSPD on the 1.5 K stage; a coincidence click triggers the joint Bell state with the receiving transmon in Module B.*

## 7.1 Transduction physics

Electro-optic transduction proceeds through three coupled modes: a microwave cavity at ω_e (≈ 5 GHz), an optical mode at ω_o (≈ 193 THz), and a strong classical pump at ω_p = ω_o − ω_e. The effective microwave-optical coupling rate is g_eo = g_0 · √n_p where g_0 is the single-photon vacuum coupling and

$n_p$ is the intracavity pump photon number.[22] In the resolved-sideband regime ($\omega_e \gg \kappa$), the transduction efficiency is:

$$\eta = \eta_{ext} \cdot \eta_{det} \cdot \frac{4 g_{eo}^2 / \kappa_e \kappa_o}{(1 + 4 g_{eo}^2 / \kappa_e \kappa_o)^2} \quad (6)$$

*$\eta_{ext}$ is the external-coupling efficiency at each port; $\eta_{det}$ is the detector efficiency; $\kappa_e$ and $\kappa_o$ are the total cavity linewidths. Cooperativity $C = 4 g_{eo}^2 / \kappa_e \kappa_o = 1$ is the impedance-matched optimum.*

QONTOS-1 specifies $\kappa_e / 2\pi = 1$ MHz, $\kappa_o / 2\pi = 10$ MHz, $\eta_{ext} = 0.5$ at each port, and $\eta_{det} = 0.9$ as design variables, not measured performance. The revised QONTOS-1 base threshold is $\eta \geq 0.1\%$, which is the acceptance level for link-physics validation. An aggressive validation band of $\eta \geq 0.5\%$ may be used for repeated tomography and low-rate entanglement experiments. The $\eta \geq 1\%$ regime is now treated as a research threshold for distributed logical operation, not as a QONTOS-1 gate requirement. Reaching that threshold requires either substantially higher $g_{eo} / 2\pi$, lower optical and microwave linewidths, or improved single-photon coupling $g_0$ while keeping absorbed optical power within the cryogenic budget.[22, 25]

## 7.2 Heralded Bell-pair protocol

Bell-pair distribution follows the Cabrillo-style scheme.[20] Each module weakly excites its local transducer to emit at most one photon per attempt; the photons interfere on a beam splitter co-located with the SNSPD; a single click projects the two transmons into a Bell state of definite parity. The end-to-end Bell-pair generation rate is bounded by the thermally allowed attempt rate at the transducer input, the square of the transduction efficiency (since both ends must succeed), and the heralding probability:

$$R_{BP} = R_{src} \cdot \eta^2 \cdot p_{herald} \quad (7)$$

*Bell-pair generation rate as a function of pump rate $R_{src}$, transduction efficiency $\eta$ (squared because both ends transduce), and heralding probability $p_{herald}$.*

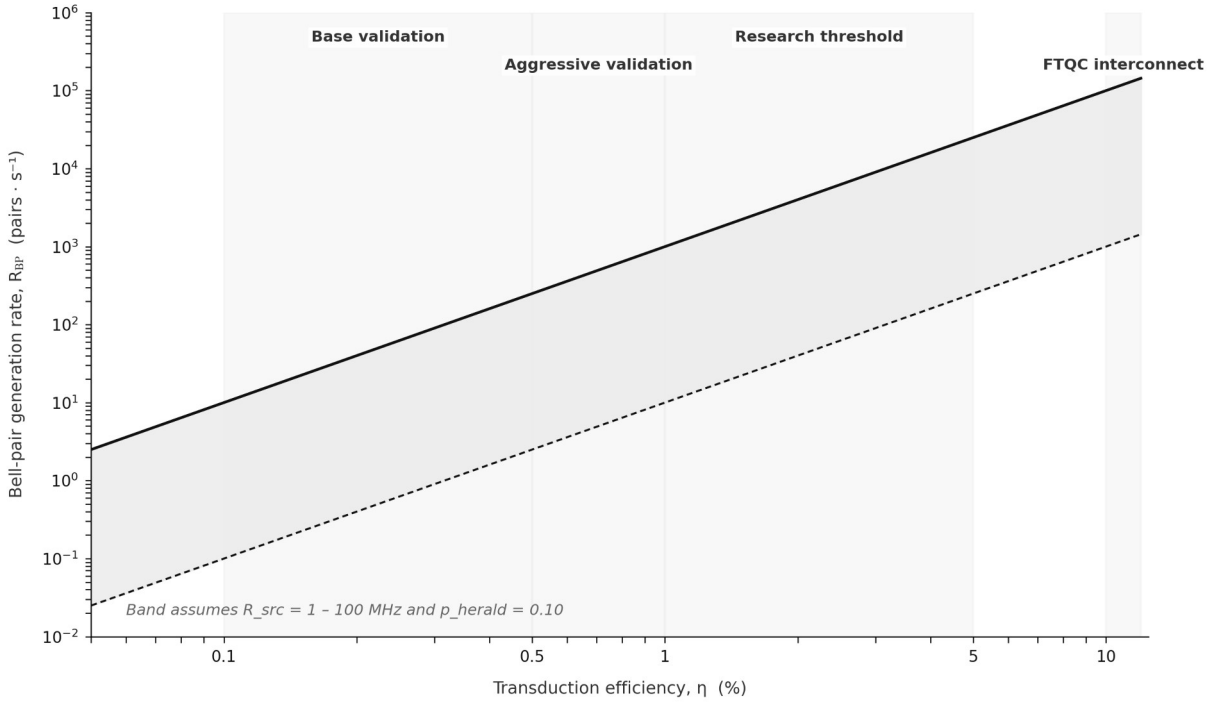



**Figure 9.** *Bell-pair generation rate as a function of transduction efficiency. The plotted curve is retained as a scaling law; QONTOS-1 acceptance uses thermally closed attempt rates in the 1 – 100 MHz band rather than assuming GHz repetition at the cold transducer.*

## 7.3 Operating regimes

| REGIME | H | RAW RATE | PURIFIED RATE | USE CASE |
|---|---|---|---|---|
| Base validation | ≥ 0.1% | 0.1 – 10 $s^{-1}$ | n/a; raw tomography | Link physics and dark-count calibration. |
| Aggressive validation | ≥ 0.5% | 2.5 – 250 $s^{-1}$ | < 1 $s^{-1}$ diagnostic | Bell tomography; exploratory swapping. |
| Research threshold | ≥ 1% | 10 – 1,000 $s^{-1}$ | 10× – 100× raw-pair cost | Low-duty-cycle distributed logical experiments. |
| FTQC interconnect | ≥ 10% plus purification | $10^4$ – $10^6$ $s^{-1}$ | after purification/repeater overhead | Routine cross-module lattice surgery, research vision. |

The purified-pair column is a scenario assumption, not a QONTOS-1 acceptance claim. BBPSSW-style recurrence purification consumes multiple noisy raw Bell pairs per accepted higher-fidelity pair and may require one or more rounds depending on measured raw fidelity, dark-count fraction, memory error, and classical-feed-forward latency.[30] Cross-module logical-gate timing is therefore reported only with an explicit purified-pair rate; raw Bell-pair generation rate alone is insufficient to budget lattice surgery.

## 7.4 Thermal and noise budget

The interconnect is now budgeted as a heat-and-noise-limited subsystem. Optical pump energy at 1,550 nm is approximately $1.28 \times 10^{-19}$ J per photon; with $10^4$ to $10^5$ intracavity pump photons and thermally allowed attempt rates in the 1 – 100 MHz band, launched optical power spans roughly 0.01 nW to 1.3 µW before insertion losses and stage absorption. QONTOS-1 acceptance requires a measured thermal closure report showing absorbed power by cryostat stage, not just a photon-number estimate. Recent quantum-enabled transduction work demonstrates that added-noise and rate must be reported together; below-one-photon added noise is not by itself sufficient for high-fidelity Bell operations.[25,26]

| BUDGET ITEM | QONTOS-1 LIMIT | ACCEPTANCE MEASUREMENT |
|---|---|---|
| Optical pump launched at cold transducer | ≤ 1.5 µW burst; ≤ 150 nW sustained base mode | Calibrated optical power at cryostat input and return-loss corrected in-device estimate. |
| Absorbed power at transducer stage | ≤ 10 nW sustained | Temperature-rise measurement during pump burst train. |
| Parasitic load coupled to mixing chamber | ≤ 100 nW incremental; passive load remains < 1 µW | Mixing-chamber thermometer and qubit $T_1/T_2$ drift under pump-on/off protocol. |
| Input-referred added noise | n_add < 1 photon for quantum-enabled readout; < 0.05 for Bell operations | Noise-calibrated upconversion/downconversion measurement with pump leakage and Raman background separated. |
| SNSPD dark-count contribution | < 1% false-herald fraction in G3 run | Dark-count and accidental-coincidence subtraction over ≥ 1 hour. |

### 7.5 Latency budget and phase-lock

For modules separated by 0 – 5 m of optical fibre, the end-to-end link latency is dominated by phase-lock acquisition (≈ 10 µs) and decoder ingress (≈ 5 µs). Optical propagation contributes ≤ 25 ns. A dedicated phase-lock subsystem maintains coherence between the pump tones of the two transducers using a shared 10 MHz reference distributed over the same fibre as the Bell-pair photons (in a different wavelength band) and a feedback loop with sub-100 fs RMS jitter. The 25 µs latency target is achievable across the full 0 – 5 m range in the base regime.

0 8

# Quantum error correction

*Surface code, distributed logical operations, resource estimates.*

---

QONTOS-1 is specified below the two-qubit-gate error threshold of the rotated surface code but not yet at a regime where useful fault-tolerant quantum computation is achievable. This section describes the error model, the chosen code family, the real-time decoder pipeline, distributed logical operations across modules via lattice surgery, the magic-state distillation budget, and the resource estimate for a representative algorithmic target.[9,10]

### 8.1 Error budget

The effective per-cycle physical error rate is the sum of seven named contributors: single-qubit gate, two-qubit gate, readout, idle decoherence, calibration drift, decoder latency overhead, and cross-module Bell-pair distribution. The QONTOS-1 target sums to 1.75%, which exceeds the rotated-surface-code threshold of approximately 1% and is therefore not yet in the fault-tolerant operating regime. Closing the 0.75 percentage-point gap is the central engineering objective of the device, control, and interconnect programmes.

The consequence is stated explicitly: at the G4 first-logical-qubit gate, a distance-5 encoded memory is expected to demonstrate stabilizer operation, syndrome extraction, real-time decoding, and measured logical-error attribution, but it is not expected to beat the physical error rate. In this regime $\varepsilon_L$ may be several times $p_{phys}$, depending on leakage, decoder backlog, and correlated readout error. Logical-qubit advantage, defined as $\varepsilon_L < p_{phys}$ under sustained operation, is moved to QONTOS-2.

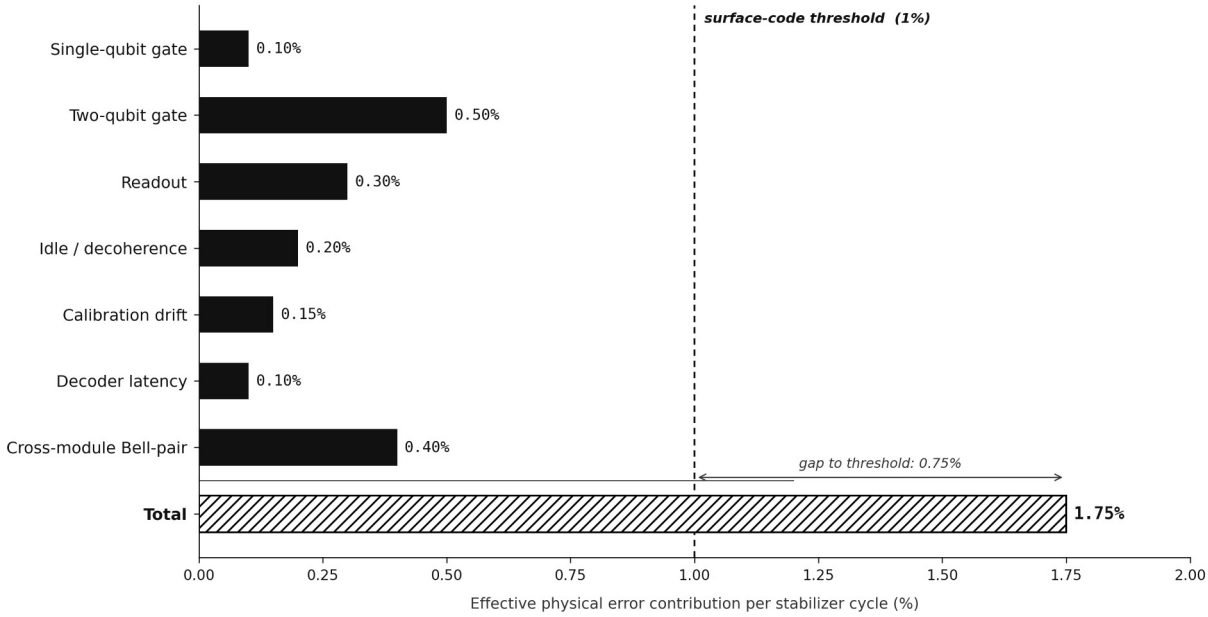



**Figure 10.** *Per-cycle error budget for QONTOS-1 target operation. The seven named contributors must collectively fall below the surface-code threshold for fault-tolerant operation. The current target sums to 1.75%, leaving 0.75 percentage points to close through device, control, and interconnect improvements.*

## 8.2 Rotated surface code

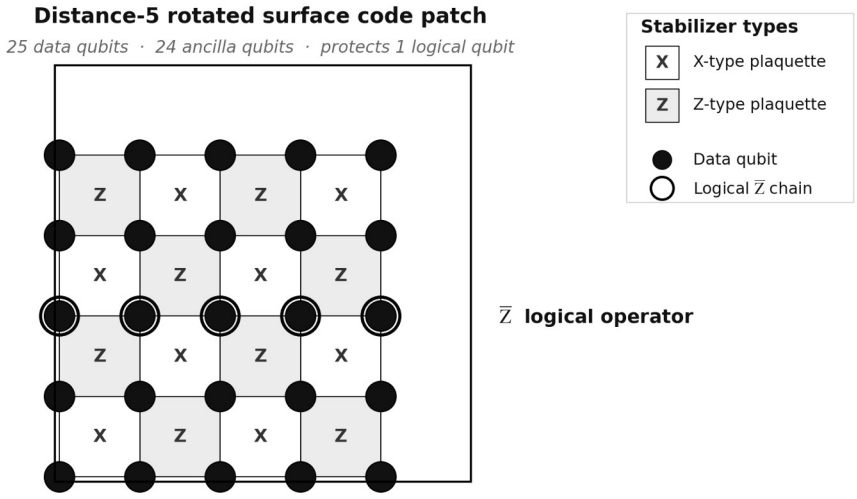


**Figure 11.** *Distance-5 rotated surface code patch on QONTOS-1. Filled circles are data qubits; open circles are syndrome ancillas. X-type plaquettes (white) and Z-type plaquettes (light grey) alternate on a checkerboard, with a logical $\bar{Z}$ operator running as a chain of data qubits across the patch.*

QONTOS adopts the rotated surface code as the primary near-term error-correcting code.[5] In the regime well below threshold, the logical error rate per code cycle scales approximately as:

$$\varepsilon_L = a \cdot (p_{phys} / p_{th})^{((d+1)/2)} \quad (4)$$

*a ≈ 0.03 is a code-family prefactor; p_th ≈ 1% is the threshold for circuit-level depolarising noise; d is the code distance, rounded up to the nearest odd integer.*

Solving for the required distance to reach a target logical error rate ε_L given p_phys yields:

$$d_{req} = 2 \cdot \log(\varepsilon_L / a) / \log(p_{phys} / p_{th}) + 1 \quad (5)$$

*Physical-qubit overhead per logical qubit is then approximately $2d^2$ for the rotated surface code, plus a similar number of ancilla qubits.*

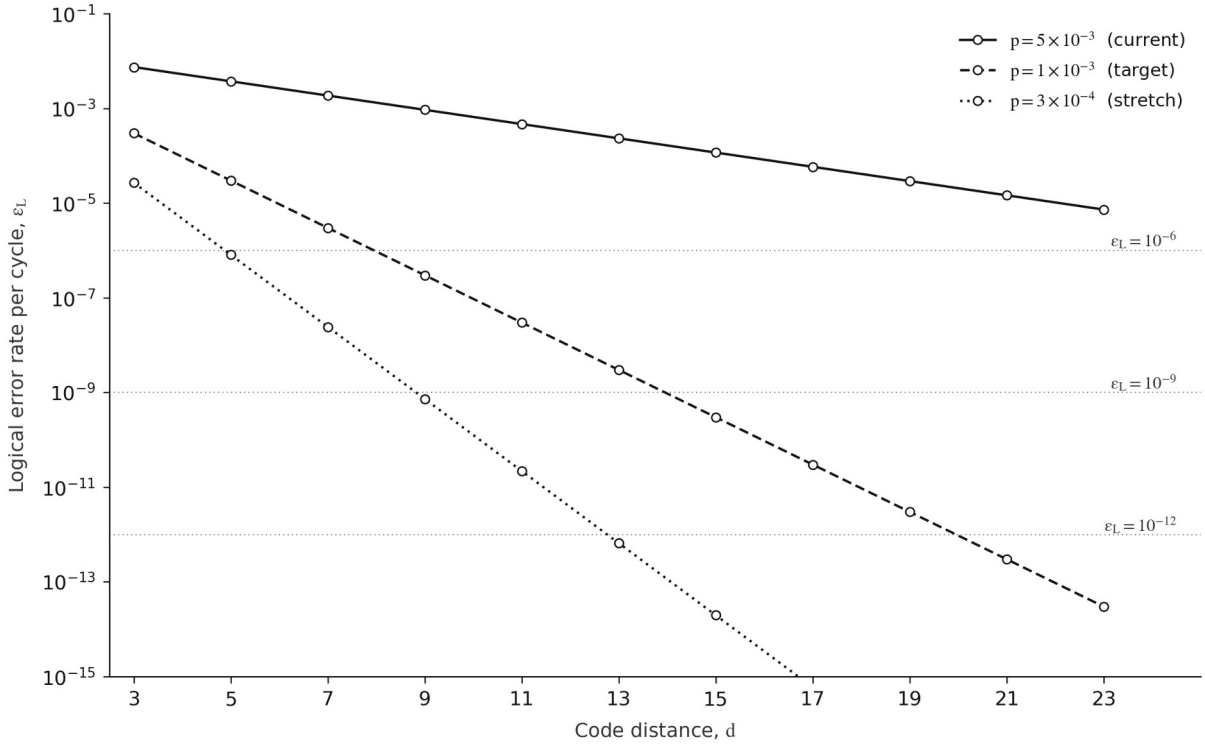



**Figure 12.** *Logical error rate per cycle as a function of code distance, for three physical-error-rate operating points. Exponential suppression with distance enables the QONTOS-1 target ($p_{phys} = 1 \times 10^{-3}$) to reach $\varepsilon_L = 10^{-9}$ at d = 15. The current operating point ($p_{phys} = 5 \times 10^{-3}$) reaches only $\varepsilon_L = 10^{-5}$ at the same distance, the engineering gap.*

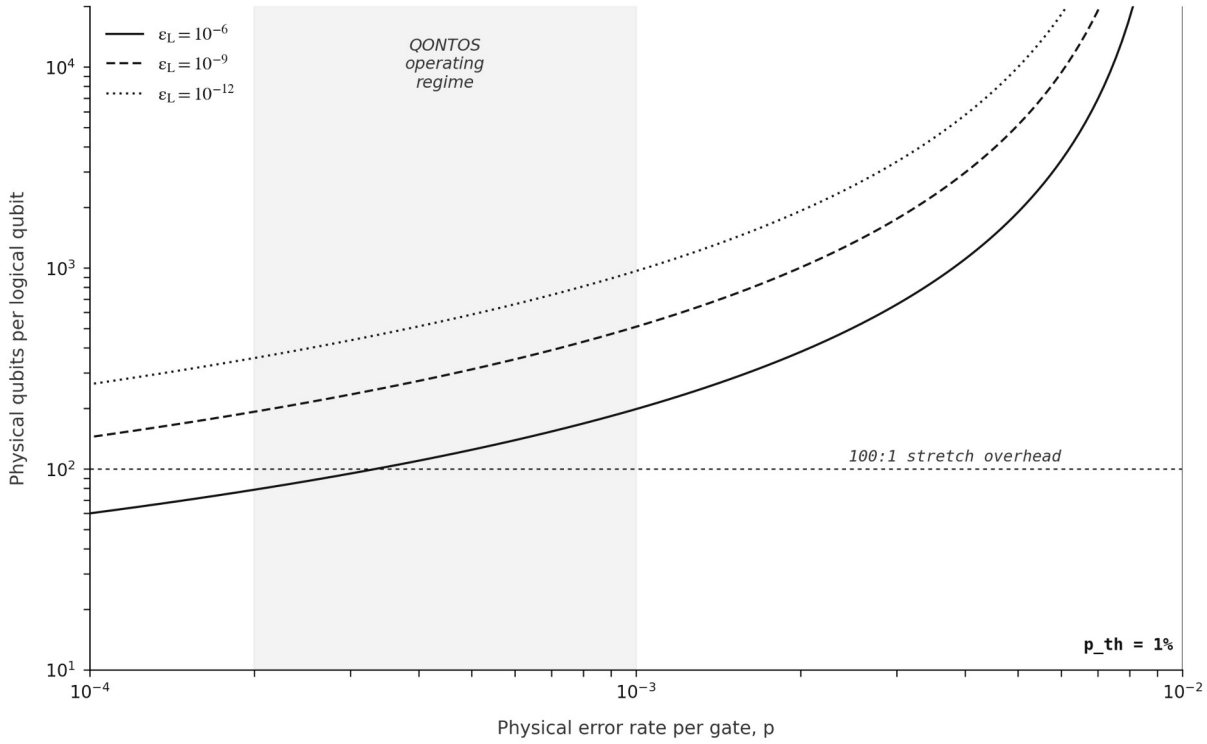



**Figure 13.** *Physical-to-logical overhead as a function of physical error rate, for three logical-error targets ($10^{-6}$, $10^{-9}$, $10^{-12}$). The QONTOS operating regime ($p_{phys}$ between $2 \times 10^{-4}$ and $10^{-3}$) crosses the 100:1 stretch overhead line only at the strict lower edge of the band.*

## 8.3 Real-time decoder pipeline

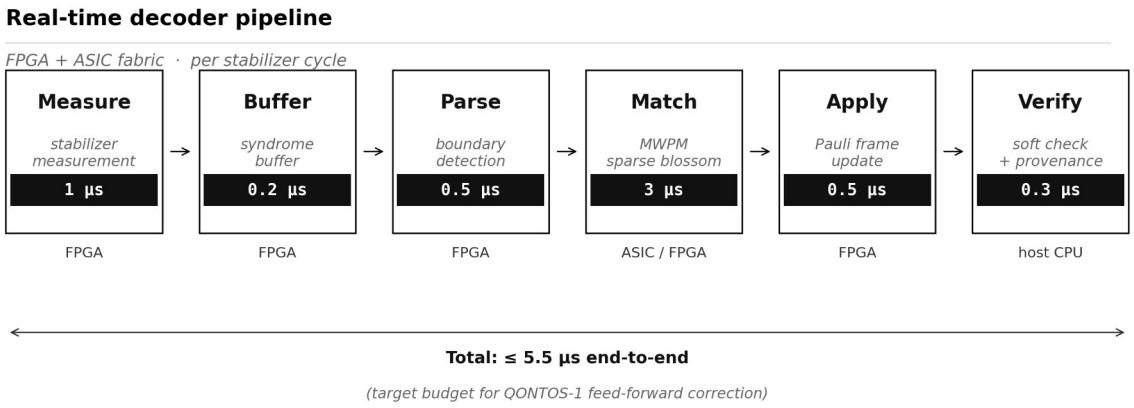



**Figure 14.** *Real-time decoder pipeline. Each stage is implemented on FPGA fabric or dedicated ASIC; the total end-to-end latency from stabilizer measurement to Pauli-frame update is specified at ≤ 5.5 µs, matching the feed-forward budget set by the qubit coherence time.*

QONTOS-1 uses a minimum-weight perfect matching (MWPM) decoder implemented in the sparse-blossom variant on dedicated FPGA fabric.[11] Stabilizer measurements at 1 µs cycle time stream into a syndrome buffer; boundary detection and graph construction occupy 0.5 µs; sparse-blossom matching completes in ≤ 3 µs for a distance-5 patch; the resulting Pauli-frame update is applied within 0.5 µs of decoder completion. A final 0.3 µs is reserved for soft consistency checks and provenance recording.

The 5.5 µs budget assumes one decoder instance per logical qubit, replicated across the lattice, and is a QONTOS-1 / distance-5 to distance-11 target only. At larger code distances ($d \geq 21$) the matching graph grows quadratically and worst-case MWPM-style matching is not compatible with a fixed 5.5 µs latency envelope. QONTOS-4/5 therefore require a different decoder architecture: spatial and temporal windowing, hardware parallelism, and workload-dependent logical-cycle timing. Those future systems are not assumed to keep the QONTOS-1 decoder latency unchanged.[12]

## 8.4 Distributed logical operations via lattice surgery

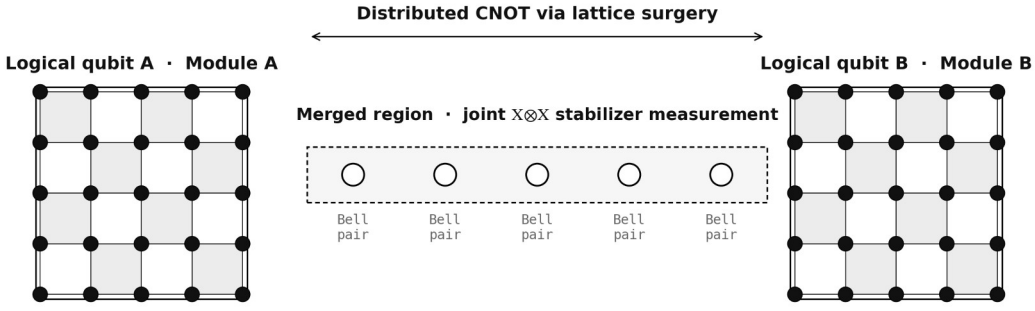



**Figure 15.** *Distributed CNOT between logical qubits in different modules via lattice surgery. The merged region implements a joint X⊗X stabilizer measurement using Bell pairs distributed by the photonic interconnect. The merge/split protocol completes in d code cycles plus the photonic latency overhead.*

Two-qubit logical gates across module boundaries are implemented via lattice surgery,[13] in which a strip of data qubits between two patches is measured as a joint stabilizer. In QONTOS-1 the joining qubits are not physical transmons but Bell pairs distributed by the photonic interconnect: a distance-d merge consumes approximately d Bell pairs per code cycle for d code cycles, or $O(d^2)$ Bell pairs per merge. A distance-5 merge therefore consumes roughly 25 Bell pairs before retries, purification, and feed-forward overhead. At

the revised QONTOS-1 base Bell-pair rate this is a slow validation experiment, not a useful logical-gate primitive.

Raw Bell-pair fidelity also compounds across a merge: if 25 independent Bell pairs each have fidelity 0.90, the uncorrected joint success envelope is far below the level required for a logical operation. QONTOS-1 therefore separates raw tomography (G3, $F \geq 0.85$) from logical-operation readiness, which requires either purified Bell pairs or much higher raw fidelity, with an effective pair fidelity of at least 0.98 before a cross-module lattice-surgery demonstration is claimed. In QONTOS-2 and QONTOS-3 schedules, the logical merge time is computed from the measured purified-pair rate after recurrence failures and memory decay, not from the raw herald rate.

## 8.5 Magic-state distillation

Non-Clifford gates (notably the T gate) are implemented via injection of distilled magic states. QONTOS-1 adopts the (15, 1, 3) distillation protocol[14] as the baseline, which produces one high-fidelity magic state from fifteen low-fidelity inputs with output infidelity scaling as $35 \cdot (p_{in})^3$. At an input infidelity of $10^{-3}$ this yields an output infidelity below $4 \times 10^{-8}$, sufficient for algorithmic applications requiring T-gate counts up to approximately $10^7$.

Magic-state factories occupy a dedicated region of the lattice during their distillation cycle. The often-quoted 200-physical-qubit number is an unprotected distance-5 layout estimate and is not used as a protected-factory resource claim. A protected factory must include encoded injection, routing, syndrome extraction, factory I/O, and idle-logical storage; even at small distance this moves the physical-qubit footprint into the hundreds to low thousands, and at $d \geq 11$ into many thousands. QONTOS-1 treats magic-state distillation as an integration experiment, not an application-throughput claim.

## 8.6 Resource estimation: representative algorithm

As a sanity check on the resource budget, consider Shor's factoring algorithm applied to a 2048-bit RSA modulus. Recent compilation work[15] estimates the algorithm requires approximately $2 \times 10^9$ Toffoli gates, equivalent to approximately $1.5 \times 10^{10}$ T gates after standard decomposition. At a logical clock frequency of 1 kHz (one logical operation per millisecond, including magic-state injection), this is a 4,800-hour computation requiring approximately 4,000 logical qubits at distance $d \geq 27$ to maintain a sub-unity total logical error budget.

This paragraph is a scale reference, not a QONTOS-5 commitment. The Gidney-Ekerå 2021 estimate is explicitly a 20-million-noisy-qubit, $p = 10^{-3}$, 1 µs-cycle, 10 µs-reaction-time benchmark for an 8-hour run; a later 2025 Gidney estimate reduces the count to below one million noisy qubits only by changing the algorithmic construction and accepting a run time below one week.[29] QONTOS therefore brackets

RSA-2048 as an external resource-estimate landmark. It does not claim that a $10^6$-physical-qubit QONTOS-5 configuration can reproduce the 2021 eight-hour estimate without the algorithmic and factory-overhead assumptions being re-derived for the QONTOS architecture.

09

# Verification and benchmarking

*How performance is measured, attributed, and reproduced.*

---

Every numerical target in this whitepaper is anchored to a specific measurement protocol with a defined gate set, sample size, and reproducibility envelope. This section describes the benchmarking ladder that QONTOS uses to validate subsystem and system-level performance, from single-qubit randomised benchmarking to end-to-end mirror circuits at multi-module scale.[16]

## 9.1 Single-qubit randomised benchmarking

Standard Clifford randomised benchmarking (RB)[17] is used for single-qubit gate characterisation. Sequences of length $m \in \{1, 2, 4, ..., 512\}$ are sampled from the single-qubit Clifford group; each sequence is inverted by its computed final Clifford and the survival probability is measured. The exponential decay of survival with m yields the average gate error per Clifford. Interleaved-RB variants attribute error to specific gates (X, X/2, Y, Y/2). The QONTOS calibration framework runs RB every 4 hours per qubit and triggers re-calibration when the drift envelope exceeds $2 \times 10^{-4}$ from the previous epoch.

## 9.2 Two-qubit benchmarking

Two-qubit characterisation combines three complementary methods. Interleaved RB on the two-qubit Clifford group provides per-gate CZ error. Cross-entropy benchmarking (XEB)[18] on random circuits of depth 10 – 30 provides circuit-level fidelity with good statistical efficiency. Mirror benchmarking[19], applying a random Clifford followed by its inverse, provides a single-number fidelity estimate that is robust to state-preparation-and-measurement errors. The QONTOS-1 two-qubit gate target ($5 \times 10^{-3}$ error) is defined as the worst-case value across all three methods on any coupler in the lattice.

## 9.3 System-level benchmarks

System-level benchmarks exercise the full software stack and the decoder pipeline. Two are particularly important. Repeated stabilizer-cycle experiments measure the logical-error rate as a function of code distance and number of cycles, directly validating the QEC suppression relation (Equation 4). Cross-

module Bell-pair tomography exercises the interconnect at system scale and reports the joint two-qubit state fidelity after distribution and entanglement swapping. The QONTOS-1 acceptance criterion at gate G3 is a Bell-pair fidelity of $F \geq 0.85$ across a sustained 1-hour run.

### 9.4 Reproducibility and provenance

Every reported number, internal, external, or in this whitepaper, is bound to a provenance record that captures the calibration epoch, the AWG and digitiser firmware versions, the decoder weight matrix, the compiler version, the device temperature time-series, and a hash of the raw measurement data. Re-running the same protocol against the same calibration epoch reproduces the result to within statistical uncertainty. Re-running against a different calibration epoch is treated as a distinct measurement and reported separately.

10

# Engineering programme

*Gates, deliverables, and the path to QONTOS-1.*

---

The QONTOS programme is organised as five concurrent engineering tracks, software runtime, device and control, photonic interconnect, cryogenic and systems, QEC pilot, synchronised by four formal gates (G1 – G4). Each gate is a measurable subsystem milestone, not a calendar event. Subsequent phases are conditional on gate passage.

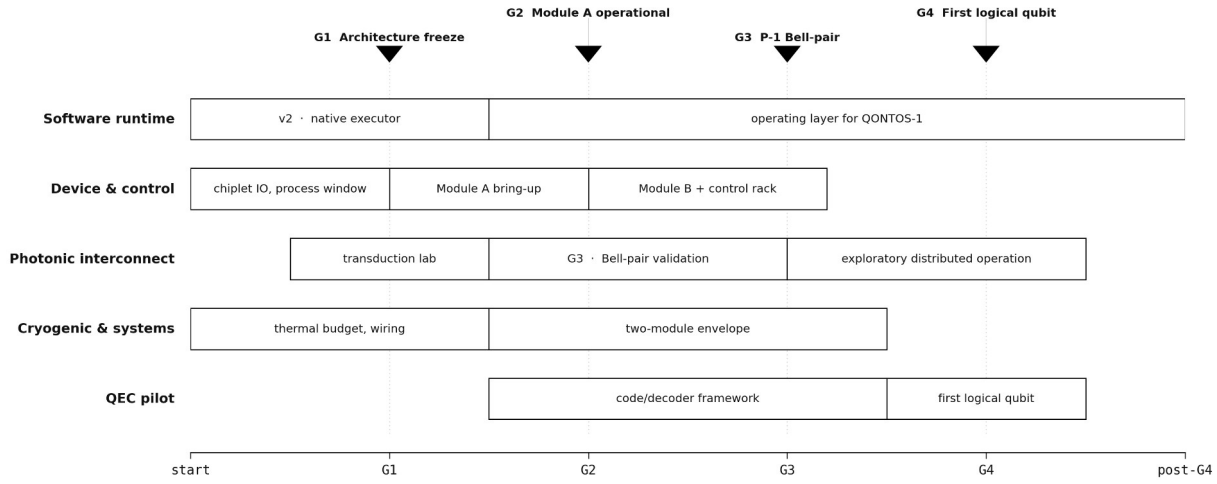



**Figure 16.** *Five-track engineering programme synchronised by four formal gates: G1 architecture freeze, G2 first module operational, G3 P-1 Bell-pair validation, and G4 first logical qubit. The horizontal axis is a dependency order, not a delivery calendar.*

## Gate definitions

| GATE | TRIGGER | ACCEPTANCE CRITERIA |
|---|---|---|
| **G1** Architecture freeze | Entry gate | Chiplet IO and package-loss budget closed; control-density and wiring budgets within cryogenic envelope; software runtime v2 deployed with native-executor stubs. |
| **G2** Module A operational | After G1 | Module A bring-up complete with RB-measured single- and two-qubit gate errors at target; readout assignment at target; control loop closes at 1 μs stabilizer cycle. |
| **G3** P-1 Bell-pair validation | After G2 | Heralded Bell pairs distributed between Module A and Module B at base $\eta \geq 0.1\%$; cross-module tomography with raw $F \geq 0.85$; latency and thermal budgets closed over a 1-hour run. |
| **G4** First logical qubit | After G3 | One encoded distance-5 surface-code logical qubit operational with real-time decoding and measured $\varepsilon_L$; $\varepsilon_L < p_{phys}$ is not required until QONTOS-2. |

## Beyond QONTOS-1

The acceptance criteria for G1 – G4 close out the QONTOS-1 programme. The successor generations, QONTOS-2 through QONTOS-5, extend this architecture toward datacenter-scale fault tolerance. Each generation is gated on the prior generation's acceptance plus a specific set of technology shifts in device, control, photonics, and cryogenics. The family arc is described in §11.

1 1

# QONTOS family: successor generations

*Five generations from architecture validation to datacenter-scale fault tolerance.*

QONTOS-1 is the first machine in a planned family of five hybrid superconducting–photonic systems. Each generation is defined by a specific engineering role and is conditional on the prior generation's acceptance gates having passed. The family is presented as a single architectural arc with attenuating confidence: QONTOS-1 is a **target** (the engineering subject of this whitepaper); QONTOS-2 and QONTOS-3 are **scenarios** (the engineering shape is plausible but numerical envelopes are conditional on the prior generation's acceptance); QONTOS-4 and QONTOS-5 are **research vision** cases (long-range architecture studies requiring multiple unresolved research programmes to land together).

No generation in the family carries a calendar date. Each generation carries a gating expression, a specific set of subsystem milestones that must close before the generation enters engineering scope.

## 11.1 Family overview

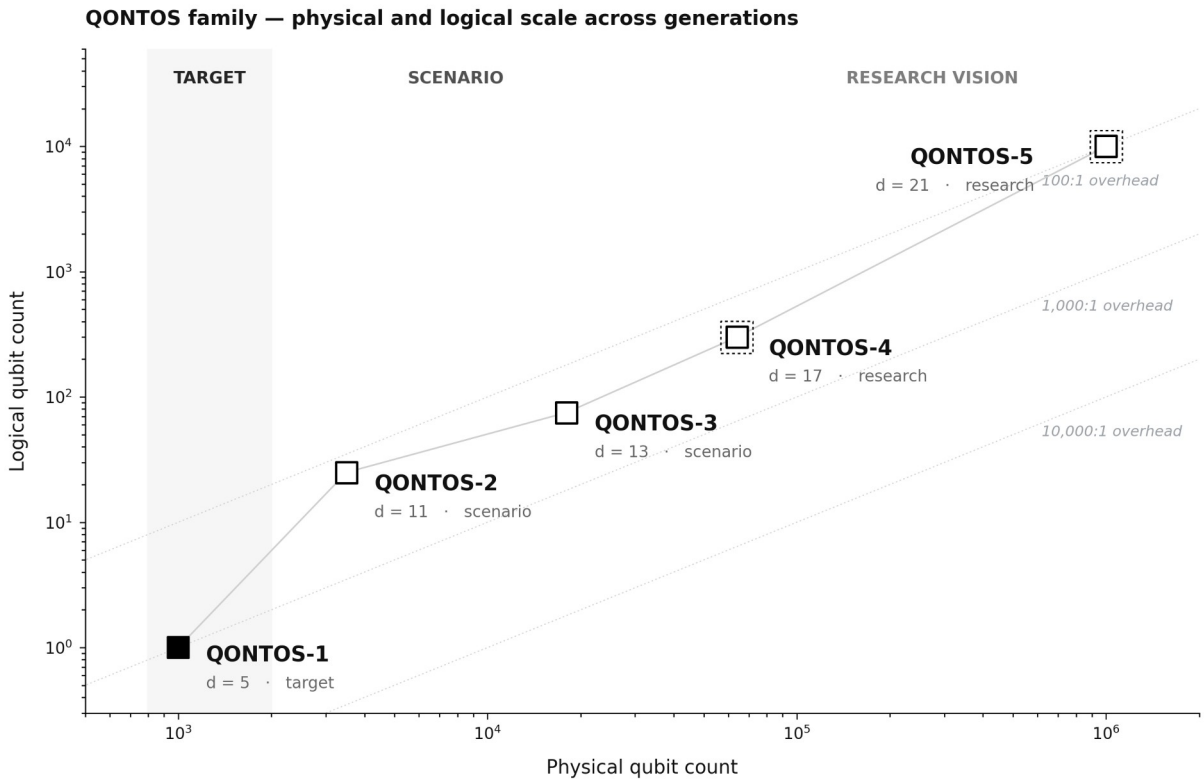



**Figure 17.** *QONTOS family, physical-to-logical qubit scale across generations, with status labels (target / scenario / research vision). QONTOS-1 (solid marker) is anchored in this whitepaper; QONTOS-2 and -3 are conditional scenarios; QONTOS-4 and -5 (dashed outer outline) are research-vision cases. Dotted reference lines mark constant physical-to-logical overhead ratios.*

| GENERATION | MODULES | PHYSICAL QUBITS | LOGICAL QUBITS | DEFINING ROLE | STATUS |
|---|---|---|---|---|---|
| **QONTOS-1** | 2 | ~ $10^3$ | first logical | Architecture validation | TARGET |
| **QONTOS-2** | 4 – 8 | 2.5 – 5 × $10^3$ | 10s | Logical-qubit advantage | SCENARIO |
| **QONTOS-3** | 16 – 32 | 1 – 2.5 × $10^4$ | 50 – 100 | First useful FTQC pilot | SCENARIO |
| **QONTOS-4** | 64 – 128 | 3.2 – 6.4 × $10^4$ | 100 – 500 | Production architecture study | RESEARCH |
| **QONTOS-5** | 500 – 2,000+ | 2.5 × $10^5$ – $10^6$+ | $10^3$ – $10^4$+ | Datacenter architecture study | RESEARCH |

Generation transitions are gated on subsystem milestones, not on calendar time. The same gating discipline that governs the four G1 – G4 gates of §10 extends across generations: each transition is conditional on specific, measurable evidence from the prior generation.

## 11.2 QONTOS-1: Architecture validation

QONTOS-1 is detailed in §2 through §10 of this whitepaper. Its purpose is to retire architectural risk: two modules of 500 transmons, connected by a base-regime photonic interconnect ($\eta \geq 0.1\%$ under closed thermal budget), validated end-to-end by an operating software runtime. The first logical qubit at distance $d = 5$ is the G4 acceptance criterion. QONTOS-1 does not claim useful fault-tolerant computation; it claims that the hybrid superconducting-photonic architecture is operable end-to-end and that its named risks (chiplet IO, control density, transduction efficiency, decoder latency, cross-module Bell-pair fidelity) retire at measurable gates.

WORKLOAD CLASS

QONTOS-1 supports physical-qubit characterisation, distributed Bell-pair tomography, and the first encoded-circuit demonstrations at $d = 5$. It is a validation platform, not a computation platform.

**Status.** TARGET. Gating: G1 through G4 acceptance criteria as defined in §10.

## 11.3 QONTOS-2: Logical-qubit advantage

QONTOS-2 raises the surface-code distance from d = 5 to d = 9 to 11 and increases module count from 2 to a band of 4 to 8, anchoring the first sustained logical-qubit advantage on a hybrid superconducting-photonic platform. Where QONTOS-1 demonstrates that the architecture is operable, QONTOS-2 demonstrates that operating it delivers a logical error rate strictly below the underlying physical error rate, sustained across continuous workloads at the lattice-cycle cadence.

### ROLE AND SCOPE

QONTOS-2 retires the below-threshold question at scale. By running multiple logical qubits at d = 9 to 11 with a measured logical error rate $\varepsilon_L \leq 10^{-6}$ across at least 24 hours of continuous stabilizer cycling, QONTOS-2 establishes that the QONTOS architecture operates below the surface-code threshold with cross-module Bell-pair distribution remaining accurate enough not to inflate the logical error rate above that threshold.

### MODULE ORGANISATION

Each module retains the 500-transmon five-chiplet flip-chip package of QONTOS-1, with refinements to the chiplet fabrication process informed by QONTOS-1 yield data. Modules are paired into supermodule couplets, with each couplet sharing a single dilution refrigerator at an expanded mixing-chamber cooling power of ≥ 50 µW at 20 mK. Up to four couplets are racked in a single equipment row, with a shared pulse-tube backbone supplying the 50 K and 4 K stages. Total physical qubit count spans 2,500 to 5,000 across the configuration band.

### PHOTONIC INTERCONNECT

QONTOS-2 requires the transducer to reach the research-threshold band, $\eta \geq 1\%$, with a measured noise and thermal budget rather than by increasing pump power alone. Since Bell-pair rate scales as $R_{BP} = R_{src} \cdot \eta^2 \cdot p_{herald}$, this moves exploratory distributed operations from rare events into the 10 to 1,000 $s^{-1}$ band under 1 – 100 MHz attempt rates. Phase-lock tightens to sub-100 fs RMS jitter through a shared 10 MHz facility reference distributed over the same fibre as the Bell-pair photons. Inter-module topology remains pair-wise: a 4-module configuration uses six pair-wise links for all-to-all connectivity, while the 8-module configuration uses a sparse 12-link cycle-plus-chord topology to bound photonic-system complexity.

### ERROR CORRECTION

The rotated surface code is operated at d = 9 to 11, with logical-qubit count in the range 10 to 20. The MWPM decoder migrates from per-logical FPGA instances (QONTOS-1) to a stitched FPGA fabric: matching graphs at module boundaries are stitched with the parallel-window protocol of Skoric et al. [12], adding approximately 0.5 µs of latency per boundary. The decoder budget remains within the 5.5 µs

envelope of §8.3. Magic-state distillation is operational at prototype cadence (10 to 30 distilled states per second from one (15, 1, 3) factory), sufficient for sparse non-Clifford workloads but not yet for sustained algorithmic operation.

#### WORKLOAD CLASS

QONTOS-2 supports encoded Clifford circuits of arbitrary depth at logical-qubit count up to ~10, with sparse non-Clifford gate injection. The flagship workload is repeated stabilizer-cycle benchmarking: running stabilizer cycles for hours at a time and measuring the logical error rate against cycle count, directly validating the QEC suppression relation of §8.2 in a multi-module distributed setting. Encoded Bell-pair tomography across module boundaries provides the second flagship measurement.

#### ACCEPTANCE CRITERIA

- Distance-11 surface-code patch operates at $\varepsilon_L \leq 10^{-6}$ sustained over ≥ 24 hours of continuous cycling.
- Cross-module distributed CNOT is demonstrated as a low-duty-cycle experiment using purified Bell pairs; no sub-200 µs target is claimed.
- Lattice runs at 1 µs stabilizer cycle for ≥ 1 hour without re-baseline.
- ≥ 5 logical qubits operate concurrently with $\varepsilon_L < p_{phys}$ (the operational definition of logical-qubit advantage).
- Cross-module raw Bell-pair fidelity F ≥ 0.90 and purified effective fidelity F ≥ 0.98 sustained over a 1-hour run.

**Status.** SCENARIO. Gating: QONTOS-1 G4 acceptance, transduction efficiency reaching the aggressive regime, decoder fabric supporting stitched matching across module boundaries, and shared-cryoplant pilot validation.

## 11.4 QONTOS-3: First useful FTQC pilot

QONTOS-3 is the first machine in the family designed to deliver useful work. With 16 to 32 modules and $10^4$ to $2.5 \times 10^4$ physical qubits, the system supports 50 to 100 logical qubits at d = 11 to 13, sufficient for representative quantum chemistry workloads (30 to 50 spin-orbital VQE or quantum phase estimation) and the first quantum-advantage demonstrations on problems of independent scientific interest.

#### ROLE AND SCOPE

QONTOS-3 retires the useful FTQC question. The acceptance criterion is not just that logical operation is below threshold (that is the QONTOS-2 milestone) but that a fault-tolerant computation completes on a problem where classical simulation is either intractable or requires resources beyond what is reasonable to

commit. The flagship workload class is electronic-structure simulation of strongly correlated molecular systems.

### MODULE ORGANISATION

The 16 to 32 modules are organised in racks of 4 modules each (4 to 8 racks total). Each rack hosts a shared cryoplant with redundant pulse-tube cold heads, and a rack-level photonic switch matrix routes Bell-pair traffic between module pairs within and across racks. The chiplet design is refined for the third time, with target physical error rate $p_{phys} \leq 5 \times 10^{-4}$ measured via per-coupler interleaved randomised benchmarking. Thermal budget at the mixing-chamber stage is bounded below 1 µW per module under sustained workload.

### PHOTONIC INTERCONNECT

QONTOS-3 introduces the multi-rack photonic mesh. Bell-pair generation is no longer pair-wise: a wavelength-division multiplexed (WDM) photonic backbone carries multiple wavelengths per fibre, and a fast photonic switch fabric routes individual Bell pairs to their destination module pair on demand. Effective Bell-pair rate per module pair is treated as an acceptance measurement, not a fixed assumption: the scenario band is $10^3$ to $10^5$ purified pairs per second, with aggregate facility throughput scaling with active wavelengths only after loss, dark counts, and purification overhead are included. Phase-lock is distributed via a facility-wide 100 MHz reference clock with sub-50 fs RMS jitter requirement at every module location.

### ERROR CORRECTION

Logical-qubit count rises to 50 to 100 at d = 11 to 13. The MWPM decoder migrates from FPGA to dedicated ASIC tiles, with each ASIC handling approximately four logical qubits and inter-tile stitching protocols handling the boundaries. Magic-state distillation factories are now production scale: a dedicated lattice region operates one or more (15, 1, 3) factories at steady-state cadence, producing > 100 distilled $|T\rangle$ states per second per factory. For higher-fidelity workloads, two-level distillation is supported, producing $|T\rangle$ states at output infidelity below $10^{-12}$.

### WORKLOAD CLASS

QONTOS-3 supports representative chemistry-class workloads: VQE on 30 to 50 spin-orbital active spaces, QPE on small to medium-sized molecules, classical-shadow tomography for verification, and the first quantum-advantage demonstrations on problems where the classical baseline (DMRG, AFQMC, classical shadows) has been carefully characterised. Optimisation workloads (max-cut on small graphs, portfolio rebalancing on a few hundred assets) are supported as secondary applications. Cryptanalytic precursor experiments at small moduli are supported but are not part of the QONTOS-3 acceptance criteria.

ACCEPTANCE CRITERIA

- First quantum-advantage demonstration on a chemistry workload, with classical-shadow verification of the result against the best available classical baseline.
- Magic-state distillation factory operates at steady state for ≥ 6 hours, producing |T⟩ states at output infidelity ≤ $4 \times 10^{-8}$.
- Distance-13 lattice surgery between modules completes within the measured logical-cycle envelope after purification overhead is included.
- ≥ 50 logical qubits operate concurrently for ≥ 1 hour at $\varepsilon_L \leq 10^{-8}$.
- End-to-end runtime overhead from inter-module routing, decoder latency, and magic-state delivery is below 20% of the logical-clock period.

**Status.** SCENARIO. Gating: QONTOS-2 acceptance, magic-state factory operability at production cadence, multi-rack photonic mesh with WDM, ASIC decoder tiles, and physical-error programme reaching $5 \times 10^{-4}$.

## 11.5 QONTOS-4: Production FTQC architecture study

QONTOS-4 scales QONTOS-3 by an order of magnitude across both module count and physical qubit count. The defining engineering transition is from pilot to production: from a single research machine to a deployable system class, with redundancy, hot-spare modules, fleet-level calibration, and economic operability over multi-month workloads. With 64 to 128 modules and $3.2 \times 10^4$ to $6.4 \times 10^4$ physical qubits under the 500-qubit module assumption, the architecture study supports 100 to 500 logical qubits depending on code distance, routing, factories, and spare capacity.

ROLE AND SCOPE

QONTOS-4 is not a committed build target in this whitepaper. It is a production architecture study that asks what evidence would be required before a pilot machine could become a deployable system class: sustained logical operation, hot-spare replacement, fleet calibration, multi-rack cryogenics, and a decoder fabric whose latency scales with code distance rather than being assumed constant.

MODULE ORGANISATION

The 64 to 128 modules occupy a building-scale facility with multiple equipment rows. Each row contains 8 to 16 modules in a shared cryoplant, with redundant helium liquefaction at the row level and at the facility level. Hot-spare modules are active and calibrated: when a module fails or drifts outside its operational envelope, the photonic switch fabric reroutes traffic around it while a maintenance procedure replaces it. The chiplet design enters its fourth refinement, with physical error rate $p_{phys} \leq 3 \times 10^{-4}$. Power

consumption is bounded below approximately 1 MW per equipment row, dominated by the cryogenic plant.

### PHOTONIC INTERCONNECT

The photonic interconnect at QONTOS-4 is a full WDM mesh: 10 to 40 simultaneous wavelengths per fibre, an arbitrary-topology photonic switch fabric, and aggregate Bell-pair rate of $10^8$ to $10^9$ $s^{-1}$ across the facility. Phase-lock is maintained sub-microsecond across all module locations through a hierarchical distribution: a primary 1 GHz facility reference clock at building level, with module-local 100 MHz boards locked to within sub-50 fs RMS jitter. The photonic switch fabric supports online reconfiguration of the inter-module topology, enabling traffic patterns to be adapted to workload.

### ERROR CORRECTION

Logical-qubit count reaches 500 to 1,000 at d = 13 to 17. The decoder fabric scales to a fleet ASIC: a single decoder substrate (or a small number of physically separate but logically unified substrates) addresses all logical qubits in the facility. Magic-state factories are multi-level: first-level (15, 1, 3) distillation feeds second-level distillation, producing $|T\rangle$ states at output infidelity below $10^{-15}$ for the longest workloads. The fleet decoder supports online code-distance adjustment per logical qubit, allowing the machine to balance throughput and logical fidelity at workload granularity.

### WORKLOAD CLASS

QONTOS-4 supports full-scale chemistry: FeMoco-class active spaces (100 to 150 spin-orbitals) for nitrogen fixation studies, full electronic-structure simulation of pharmacologically relevant intermediates, and material-science calculations for high-T_c superconductors. Cryptanalytic precursor workloads are supported: Shor scaling tests at 768-bit and 1024-bit moduli, with calibration data informing the QONTOS-5 resource estimate. Optimisation workloads scale to thousand-variable problems with real-world coefficients.

### ACCEPTANCE CRITERIA

- Sustained 30-day workload at ≥ 100 logical qubits, maintaining the operational envelope without facility-wide re-baseline.
- Mean-time-between-replacement of a module > 90 days, with hot-spare substitution completing in < 1 hour without halting the workload.
- Cryptanalytic precursor workloads are limited to scaling studies and resource-estimate calibration, not RSA-breaking demonstrations.
- Logical clock frequency is reported as a measured workload envelope; no fixed 1 kHz fleet-wide target is assumed at d ≥ 17.
- Multi-tenant resource allocation with logical-qubit isolation between concurrent workloads.

**Status.** RESEARCH VISION. Gating: QONTOS-3 acceptance, multi-rack cryoplant in production operation, WDM photonic mesh with hot-spare module routing, fleet ASIC decoder, and physical-error programme reaching $3 \times 10^{-4}$.

## 11.6 QONTOS-5: Datacenter-scale FTQC architecture study

QONTOS-5 is the long-range datacenter target. The system spans 500 to 2,000+ modules across a multi-building cryogenic facility. Under the 500-qubit module assumption, 500 modules provide $2.5 \times 10^5$ physical qubits; reaching approximately $10^6$ physical qubits requires roughly 2,000 modules or a larger module class. This is a research-vision architecture study for general-purpose FTQC, not a claim that RSA-2048 factoring is available at the lower end of the module-count band.

### ROLE AND SCOPE

QONTOS-5 is not an engineering commitment under this whitepaper. The defining question is whether a superconducting-photonic architecture can be operated at datacenter scale after the interconnect, cryogenic plant, decoder, factory, and maintenance problems have been independently solved. The acceptance criteria below are therefore architecture-study criteria, not programme delivery promises.

### MODULE ORGANISATION

The 500 to 2,000+ modules occupy a multi-building cryogenic facility, organised hierarchically: 8 to 16 modules per equipment row, 4 to 8 rows per equipment hall, 4 to 16 halls per building, and 2 to 4 buildings per site. Each building has its own helium plant; inter-building helium distribution provides redundancy against single-plant failure. Power consumption is dominated by the cryogenic plant (approximately 1 to 3 MW per building) and the photonic infrastructure (approximately 0.5 MW total per site). The chiplet design reaches its fifth refinement, with physical error rate $p_{phys} \leq 10^{-4}$.

### PHOTONIC INTERCONNECT

The QONTOS-5 photonic backbone is multi-building. In-facility links use kilometric optical fibres routed in cable trays alongside helium transfer lines; inter-building links use single-mode fibre in conduit. Inter-building entanglement distribution is mediated by intermediate quantum repeaters: stationary nodes at building junctions perform entanglement swapping and purification, extending the effective Bell-pair distance without sacrificing fidelity. Phase-lock across the multi-building facility is maintained at sub-100 fs RMS jitter through a hybrid optical-microwave clock distribution system referenced to a primary optical frequency standard.

### ERROR CORRECTION

Logical-qubit count reaches $10^4$+ at d ≥ 21. The decoder fabric becomes reconfigurable: code distance can be increased dynamically at runtime to support workloads with longer logical-clock cycles, and decoded syndromes are pipelined through a multi-stage classical compute fabric. Magic-state factories are continuously operational, with online quality control adjusting distillation depth based on workload requirements. The fleet decoder operates as a managed resource, with logical-qubit allocations scheduled across the facility according to user demand and workload priority.

### WORKLOAD CLASS

QONTOS-5 supports general-purpose FTQC. The flagship architecture study is an RSA-2048 resource-estimate closure exercise against §8.6, including algorithm choice, code distance, factory layout, purification overhead, and decoder latency. Other candidate workloads: large-scale QPE for chemistry and condensed-matter problems, quantum machine learning at industrially relevant scale, secure multi-party computation primitives, and the first demonstrations of quantum advantage in cryptographic primitives beyond integer factoring. Concurrent multi-user operation is supported through logical-qubit isolation in the fleet decoder.

### ACCEPTANCE CRITERIA

- RSA-2048 resource estimate is closed end-to-end for the QONTOS architecture and reconciled against both the 2021 20-million-qubit benchmark and later sub-million-qubit research estimates.
- $10^4$ logical qubits operate concurrently for ≥ 30 days at $\varepsilon_L \leq 10^{-12}$.
- Facility operations, maintenance, cryogenic replacement, and photonic rerouting procedures are documented and tested at architecture-study scale.
- Multi-user resource allocation with isolation between logical-qubit allocations to different workloads.
- Mean-time-between-facility-incident > 12 months under sustained multi-tenant operation.

**Status.** RESEARCH VISION. Gating: QONTOS-4 acceptance, device-error programme reaching $p_{phys} \leq 10^{-4}$, distance d ≥ 21 surface-code operability, multi-building photonic infrastructure with entanglement-distribution repeaters, and operational documentation sufficient for commercial deployment.

## 11.7 QONTOS-5 facility view

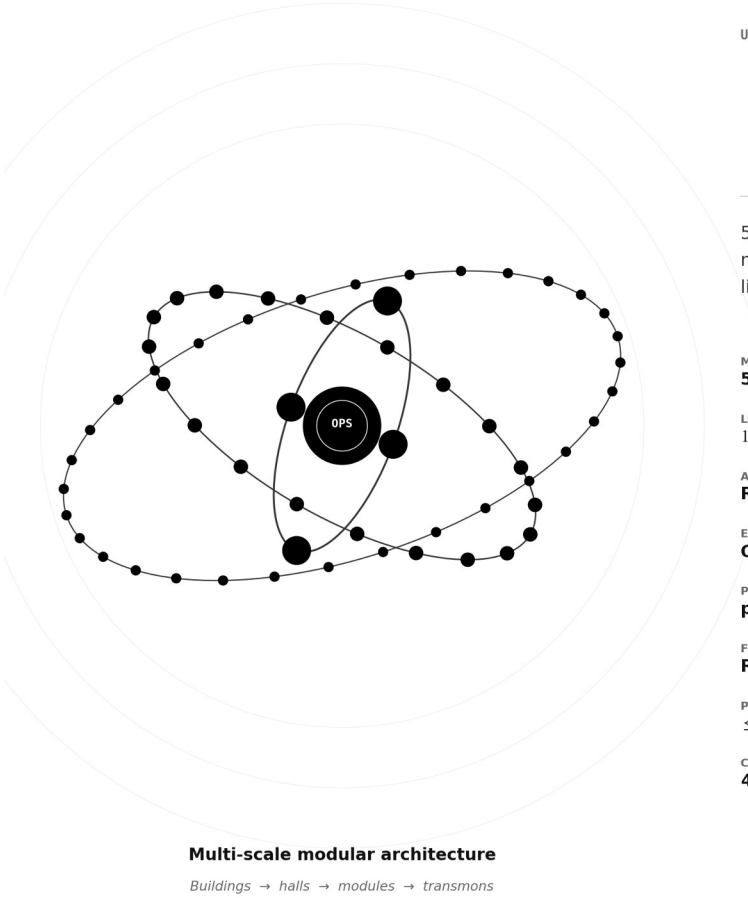
The QONTOS-5 facility view places the research-vision system as a single editorial card. The central orbital rendering represents the facility's multi-scale modular architecture: the nucleus is the orchestration core, the three orbits are the four compute buildings, the equipment halls they contain, and the modules that fill those halls; the surrounding halo is the population of transmons that reaches $2.5 \times 10^5$ physical qubits at 500 modules and up to $10^6$+ physical qubits only at the 2,000+ module upper envelope, unless a larger module class is introduced. The right column carries the headline specifications. The bottom strip locates QONTOS-5 in the QONTOS family roadmap as the fifth generation, currently a research-vision architecture study conditional on the QONTOS-1 through QONTOS-4 gates closing.

# QONTOS-5

*Datacenter-scale architecture study*

**FIG 18**

QONTOS family · generation 5 of 5




**Multi-scale modular architecture**

*Buildings → halls → modules → transmons*

GENERATION 5 OF 5

UP TO

$10^6+$ **Physical Qubits**

500 to 2,000+ modules across multiple cryogenic buildings, linked by a photonic Bell-pair backbone.

MODULE COUNT
**500 – 2,000+**

LOGICAL QUBITS
$10^3$ **–** $10^4$**+**

ARCHITECTURE
**Research vision**

ERROR CORRECTION
**Q-LDPC + surface**

PHOTONIC REGIME
**purified Bell-pair backbone**

FLAGSHIP WORKLOAD
**RSA resource-estimate closure**

POWER ENVELOPE
$\leq$ **16 MW**

CRYOGENICS
**4× redundant He plants**

**QONTOS family roadmap**

*Five generations from first machine to facility scale*

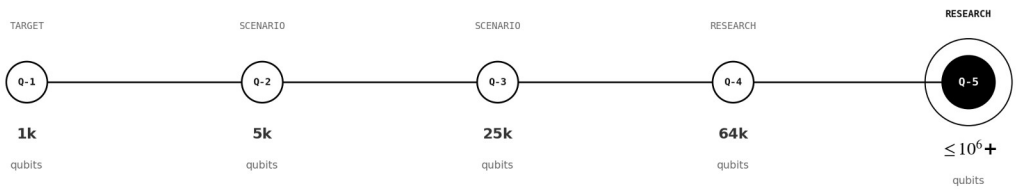



**Figure 18.** *QONTOS-5 datacenter-scale FTQC architecture study. The orbital hero shows the multi-scale facility concept; the right column summarises the 2,000+ module upper envelope required for up to $10^6+$ physical qubits, and the bottom strip places QONTOS-5 in the research-vision part of the family roadmap.*

## 11.8 Technology shifts across generations

**Technology shifts across generations**
*Each transition is unlocked by specific engineering deltas, not by calendar time.*

| Engineering domain | QONTOS-1 | QONTOS-2 | QONTOS-3 | QONTOS-4 | QONTOS-5 |
|---|---|---|---|---|---|
| **Modules** | 2 | 4 – 8 | 16 – 32 | 64 – 128 | 500 – 2,000+ |
| **Physical qubits** | $1.0 \times 10^3$ | $2.5 – 5 \times 10^3$ | $1 – 2.5 \times 10^4$ | $3.2 – 6.4 \times 10^4$ | $2.5 \times 10^5 – 10^6$+ |
| **Logical qubits** | 1 | 10s | 50 – 100 | 100 – 500 | $10^3 – 10^4$+ |
| **Code distance** | d = 5 | d = 9 – 11 | d = 11 – 13 | d = 13 – 17 | d ≥ 21 |
| **Phys. error rate** | $5 \times 10^{-3}$ | $\leq 1 \times 10^{-3}$ | $\leq 5 \times 10^{-4}$ | $\leq 3 \times 10^{-4}$ | $\leq 1 \times 10^{-4}$ |
| **Transduction η** | base ≥ 0.1% | research ≥ 1% | purified mesh | architecture study | architecture study |
| **Photonic topology** | pair-wise link | pair-wise link | multi-rack mesh | WDM mesh | building backbone |
| **Decoder fabric** | FPGA per logical | stitched FPGA fabric | tile-stitched ASIC | fleet ASIC | fleet reconfigurable |
| **Cryogenic plant** | independent fridges | shared cryoplant (pilot) | full shared plant | multi-rack plant | multi-building plant |
| **Status** | **TARGET** | **SCENARIO** | **SCENARIO** | **RESEARCH** | **RESEARCH** |

**Figure 19.** *Technology shifts across the QONTOS family. Each generation transition is unlocked by specific engineering deltas in one or more domains. Modules, code distance, transduction regime, photonic topology, decoder fabric, and cryogenic plant configuration each evolve along their own trajectories; the generation labels mark synchronisation points where multiple deltas land together.*

Reading the matrix horizontally shows the technology trajectory within a single engineering domain. Reading vertically shows the configuration of a single generation. The transitions are not additive in the sense of buying more of the same thing: each transition requires a qualitative change in at least one domain. The shared cryoplant lands at QONTOS-2, the multi-rack mesh at QONTOS-3, the fleet ASIC at QONTOS-4, and the multi-building infrastructure at QONTOS-5.

## 11.9 Commitment posture

The QONTOS family is presented as a single architectural arc with attenuating confidence as generations progress outward. Only QONTOS-1 is a commitment under this whitepaper. QONTOS-2 and QONTOS-3 are scenarios that become commitments when their gating evidence lands. QONTOS-4 and QONTOS-5 are research-vision cases that remain conditional on multiple breakthroughs across device, control, photonics, and cryogenics maturing together.

This is not a roadmap with calendar promises. It is an engineering map with gating expressions. Each generation transition is a measurable event, not a calendar event. The discipline is the same as for QONTOS-1: a target moves from scenario to commitment only when the prior generation's acceptance criteria have been demonstrated under the verification and benchmarking protocol of §9.

12

# Risks and mitigations

*Named failure modes with engineering responses.*

The architecture has eleven principal risks. Each is named, attributed to a subsystem, and bound to an engineering response. The programme is structured so that each risk either retires at a gate or triggers a documented re-plan with a fallback path.

| RISK | SUBSYSTEM | ENGINEERING RESPONSE |
|---|---|---|
| Two-qubit gate error fails to reach $5 \times 10^{-3}$ | Device, control | Re-baseline at higher error; lengthen code distance; extend G1 calibration programme. |
| Transduction efficiency stalls below 1% research threshold | Interconnect | Do not claim distributed logical operations; retain G3 as link-physics validation if $\eta \geq 0.1\%$; revisit transducer architecture (electro-optic vs piezo-optomechanical). |
| Transducer added noise or pump heating exceeds budget | Interconnect, cryogenic | Reduce duty cycle; move pump absorption to higher-temperature stage; add purification; require pump-on/off thermal and noise closure before G3. |
| Decoder latency exceeds 10 µs | Control, software | Move matching engine onto dedicated FPGA fabric; reduce stabilizer cycle frequency; trade against logical error rate. |
| Decoder scaling fails at $d \geq 21$ | Control, software | Treat QONTOS-4/5 as research vision until windowed or hardware-parallel decoder latency is measured at the required distance. |
| Wiring density saturates cryogenic envelope | Cryogenic, systems | Reduce per-qubit line count via frequency multiplexing; relocate amplification stages; revisit packaging. |

| | | |
|---|---|---|
| 500-qubit module frequency collisions reduce calibrated edge yield | Device, control | Role reassignment after spectroscopy; tunable-coupler bias optimisation; compiler maps out collision clusters; acceptance uses calibrated edge yield. |
| Readout multiplexing crosstalk exceeds assignment budget | Readout, control | Fall back from 8× to 4× multiplexing; increase resonator spacing; add matched-filter crosstalk subtraction before claiming 99% assignment fidelity. |
| Calibration drift exceeds operational window | Control, software | Increase calibration sweep frequency; deploy in-situ drift trackers; bound circuit batch length against drift envelope. |
| Photonic link timing jitter degrades Bell-pair fidelity | Interconnect, control | Tighten phase-lock loop; reduce optical path length; introduce mid-link timing reference; bound link length. |
| Magic-state factory overhead is undercounted | QEC, architecture | Report unprotected and protected factory footprints separately; do not use 200-qubit prototype layouts for protected-throughput claims. |

13

# Comparison to state of the art

*Where QONTOS-1 sits in the contemporary landscape.*

This section places the QONTOS-1 targets alongside reported specifications from the leading hardware programmes: IBM Quantum (superconducting, monolithic), Google Quantum AI (superconducting, monolithic), Quantinuum (trapped-ion, QCCD architecture), Rigetti (superconducting, modular packaging), and IonQ (trapped-ion, photonic interconnect). The comparison is deliberately conservative: QONTOS-1 is presented at its target operating point, against the best-reported performance of operating machines.[23,24]

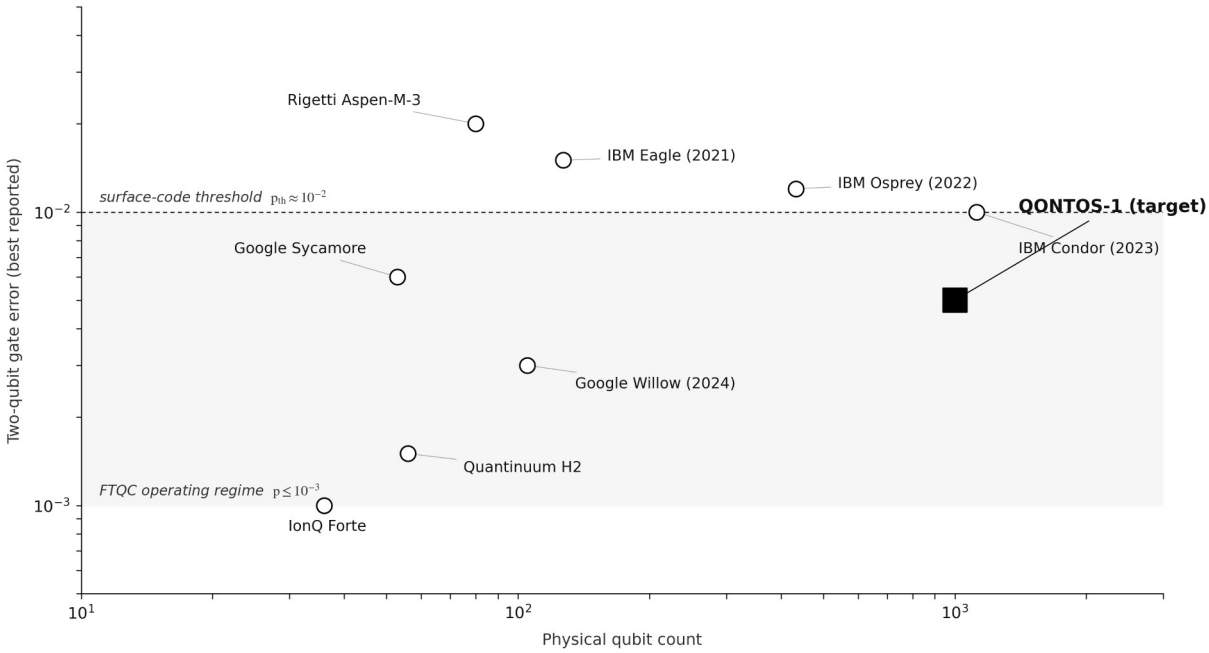



**Figure 20.** *Two-qubit gate error versus physical qubit count, across leading quantum hardware programmes. The shaded band marks the fault-tolerant operating regime ($p \leq 10^{-3}$); the dashed line is the surface-code threshold ($p_{th} \approx 10^{-2}$). QONTOS-1 (target) is shown as a filled square at 1,000 physical qubits and $5 \times 10^{-3}$ two-qubit gate error, above threshold but at a physical-qubit scale not yet reached by trapped-ion systems and a gate-error level approaching the best superconducting demonstrations.*

## 13.1 Positioning

Three observations follow from the comparison. First, QONTOS-1 is the only architecture at the 1,000-physical-qubit scale that is modular by design; existing systems at that scale (IBM Condor) are monolithic. Second, the QONTOS-1 two-qubit gate target ($5 \times 10^{-3}$) is more conservative than recent demonstrations on small-scale devices (Quantinuum H2, Google Willow), reflecting the additional engineering load of running 1,000 qubits in coordinated calibration rather than tens. Third, the photonic interconnect is the dimension on which QONTOS-1 differs from all other listed programmes, none of which currently operates a multi-module distributed logical-qubit configuration at scale.

## 13.2 What QONTOS-1 does not claim

QONTOS-1 does not claim to be at the gate-error frontier; that frontier currently belongs to trapped-ion systems (Quantinuum, IonQ) on the order of tens of qubits. QONTOS-1 does not claim to demonstrate logical-qubit advantage; G4 is an encoded-memory and real-time-decoding milestone, while logical-qubit advantage moves to QONTOS-2. QONTOS-1 does claim to be the first hybrid superconducting–photonic architecture brought to system scale, and to retire the engineering risks that stand between this architecture and a fault-tolerant successor.

### 13.3 Related work not used as direct comparators

The photonic and transduction literature is directly relevant even when the hardware modality differs from QONTOS. PsiQuantum's foundry-oriented photonic platform demonstrates high-visibility heralded photonic-qubit operations conditional on detection, while Xanadu's Aurora system demonstrates a networked modular photonic architecture and real-time feed-forward but also reports that loss remains the central gap to fault tolerance.[27,28] These programmes inform the QONTOS requirements for multiplexing, detector performance, and loss accounting. Recent piezo- and electro-optomechanical transduction demonstrations also motivate the revised QONTOS-1 interconnect posture: optical readout of a superconducting transmon has been demonstrated through a piezo-optomechanical transducer, and quantum-enabled microwave-to-optical transduction has reached below-one-photon added noise, but neither result by itself closes the efficiency, thermal, and Bell-fidelity budgets required for distributed surface-code operation.[25,26]

1 4

# Conclusion

*QONTOS-1 in one paragraph; the family in two more.*

---

QONTOS-1 is a two-module, 1,000-physical-qubit superconducting–photonic quantum computer. It is specified to be brought up across four engineering gates (G1 – G4), supported by an operational software runtime that pre-dates the hardware. The machine is not designed to deliver fault tolerance; it is designed to retire the architectural risks (chiplet IO, control density, transduction efficiency, decoder latency, cross-module Bell-pair fidelity) that stand between current superconducting practice and a modular fault-tolerant successor.

QONTOS-2 and QONTOS-3 extend the architecture into useful fault-tolerant quantum computing: QONTOS-2 demonstrates the first sustained logical-qubit advantage with 4 – 8 modules at code distance 9 – 11; QONTOS-3 delivers the first useful FTQC pilot with 16 – 32 modules, 50 – 100 logical qubits, and magic-state factories operating at workload-scale cadence. These generations are scenarios, not commitments, committed when QONTOS-1's gates land and the architectural deltas in §11 are retired.

QONTOS-4 and QONTOS-5 are research-vision architecture studies: production FTQC and datacenter-scale FTQC respectively. QONTOS-5 is no longer presented as a direct 2048-bit RSA execution target; it is presented as the scale at which the QONTOS architecture must be reconciled against external RSA resource estimates, including decoder latency, factories, purification, and operations. These cases are

conditional on multiple breakthroughs across device, control, photonics, and cryogenics maturing together. They are the research map, not the commitment.

15

# Appendix A: Notation

*Symbols and definitions used throughout.*

| SYMBOL | DEFINITION |
|---|---|
| ω_q | Qubit transition frequency (\|0⟩ ↔ \|1⟩). QONTOS-1 nominal value: 2π × 5 GHz. |
| α | Transmon anharmonicity, $\alpha = \omega_{12} - \omega_{01}$. QONTOS-1 nominal: 2π × (−300) MHz. |
| E_J / E_C | Josephson energy to charging energy ratio. Transmon regime: ≈ 50. |
| $T_1$ | Energy relaxation time of the qubit. QONTOS-1 target: ≥ 200 µs. |
| $T_2$ , $T_2$_echo | Dephasing time (free induction / Hahn echo). QONTOS-1 target: ≥ 100 µs (echo). |
| ω_r | Readout-resonator frequency, detuned from ω_q by ≈ 1 GHz. |
| χ | Dispersive shift, $\chi = g^2 / (\omega_q - \omega_r)$. |
| g | Qubit-resonator coupling, or single-photon vacuum coupling. |
| F_RO | Single-shot readout assignment fidelity. Target: 99.0% at 1 µs integration. |
| η | Microwave-to-optical transduction efficiency. |
| p_herald | Probability of heralding click given a successful Bell-pair preparation. |
| R_BP | Bell-pair generation rate, pairs per second. |
| p_phys | Per-gate physical error rate, dominant contributor to per-cycle p. |

| | |
|---|---|
| p_th | Surface-code threshold for circuit-level depolarising noise. ≈ $10^{-2}$. |
| ε_L | Logical error rate per cycle of the encoded qubit. |
| d | Surface-code distance. Logical-qubit overhead ≈ $2d^2$. |
| κ_e , κ_o | Total linewidth of the electro-optic microwave and optical resonator modes. |
| g_eo | Effective microwave-optical coupling, g_eo = g_0 √n_p. |
| MWPM | Minimum-weight perfect matching decoder. |
| DRAG | Derivative-Removal-by-Adiabatic-Gate single-qubit pulse shape. |
| RB / XEB | Randomised benchmarking / cross-entropy benchmarking. |
| FTQC | Fault-tolerant quantum computing. |

16

# Appendix B: Parameter tables

*Consolidated specification of the QONTOS-1 device, control, and interconnect targets.*

## DEVICE PARAMETERS

| PARAMETER | SYMBOL | QONTOS-1 TARGET | NOTES |
|---|---|---|---|
| Qubit transition frequency | ω_q / 2π | 5.0 ± 0.2 GHz | Fixed-frequency tantalum-Al transmon. |
| Anharmonicity | α / 2π | −300 MHz | Standard transmon E_J / E_C ≈ 50. |
| Energy relaxation | $T_1$ | ≥ 200 µs | Materials-stack target; published Ta devices exceed 300 µs. |
| Dephasing (echo) | $T_2$_echo | ≥ 100 µs | Limited by 1/f flux and TLS noise. |

| | | | |
|---|---|---|---|
| Single-qubit gate length | – | 16 ns | 4σ DRAG envelope, σ = 4 ns. |
| Two-qubit gate length | – | 18 ns | Tunable-coupler CZ flux pulse. |
| Single-qubit gate error | p_1Q | $1 \times 10^{-4}$ | Per Clifford, measured via single-qubit RB. |
| Two-qubit gate error | p_2Q | $5 \times 10^{-3}$ | Per CZ, measured via interleaved RB / XEB. |
| Readout assignment fidelity | F_RO | 99.0% | Single-shot, 1 µs integration. |
| Readout-resonator linewidth | κ_r / 2π | 5 MHz | Set by Purcell filter. |
| Frequency-multiplex factor | – | 8 | Readout resonators per amplifier chain. |
| Readout crosstalk budget | – | < 0.5% | Correlated assignment error per resonator pair; fallback to 4× multiplexing. |
| Calibrated edge yield | – | ≥ 90% | After qubit role reassignment and frequency-collision mitigation. |

### CRYOGENIC AND CONTROL PARAMETERS

| STAGE | TEMPERATURE | COOLING POWER | NOTES |
|---|---|---|---|
| Room | 295 K | n/a | AWGs, digitisers, FPGA sequencers. |
| First cooling | 50 K | ≥ 50 W | Pulse-tube cold head. |
| Second cooling | 4 K | ≥ 1.5 W | Pulse-tube; cryogenic amplifiers, TWPA. |
| Still | 100 mK | ≥ 500 µW | Filtering, attenuators, transducer staging. |

| | | | |
|---|---|---|---|
| Mixing chamber | ≤ 20 mK | ≥ 25 μW at 20 mK | Qubit chip, Purcell filters, readout resonators. |

### INTERCONNECT PARAMETERS

| PARAMETER | SYMBOL | TARGET | NOTES |
|---|---|---|---|
| Optical wavelength | λ | 1,550 nm | Telecom C-band; ≤ 0.2 dB/km fibre loss. |
| Microwave cavity linewidth | κ_e / 2π | 1 MHz | Electro-optic resonator. |
| Optical mode linewidth | κ_o / 2π | 10 MHz | Whispering-gallery $LiNbO_3$ resonator. |
| Single-photon vacuum coupling | g_0 / 2π | ≈ 1 kHz | Design-level estimate. |
| Pump photon number | n_p | $10^4$ – $10^5$ | Limited by measured cryogenic heating. |
| Detector efficiency | η_det | 0.9 | SNSPD at 1.5 K stage. |
| Transduction efficiency | η | ≥ 0.1% base · ≥ 0.5% aggressive · ≥ 1% research | QONTOS-1 acceptance uses the base threshold. |
| Raw Bell-pair generation rate | R_BP | 0.1 – 1,000 $s^{-1}$ | Thermally closed 1 – 100 MHz attempt-rate band. |
| Purified Bell-pair rate | R_pur | scenario-specific | Computed after BBPSSW-style recurrence failures, memory error, and dark-count subtraction; required for cross-module logical-gate timing. |
| Raw Bell-pair fidelity | F_BP | ≥ 0.85 | G3 tomography; logical operations require purified F ≥ 0.98. |
| Input-referred added noise | n_add | < 1 photon | Quantum-enabled operation; high-fidelity Bell operations require < 0.05. |

| | | | |
|---|---|---|---|
| Inter-module fibre length | – | 0 – 5 m | Single rack to adjacent rack. |
| End-to-end link latency | – | ≤ 25 µs | Dominated by phase-lock acquisition. |

INTERCONNECT THERMAL CLOSURE

| LOAD PATH | QONTOS-1 LIMIT | VERIFICATION |
|---|---|---|
| Cold-transducer launched optical pump | ≤ 1.5 µW burst; ≤ 150 nW sustained base mode | Calibrated optical power and duty-cycle log. |
| Absorbed power at transducer stage | ≤ 10 nW sustained | Pump-on/off thermal step response. |
| Incremental mixing-chamber load | ≤ 100 nW; passive total < 1 µW | Mixing-chamber thermometer plus qubit $T_1/T_2$ drift check. |
| False-herald fraction | < 1% during G3 run | SNSPD dark counts and accidental-coincidence subtraction. |

17

# References

*Selected technical foundations.*


[1] **Krantz, P., Kjaergaard, M., Yan, F., Orlando, T. P., Gustavsson, S., & Oliver, W. D.** A quantum engineer's guide to superconducting qubits. *Applied Physics Reviews 6, 021318 (2019).*

[2] **Koch, J., Yu, T. M., Gambetta, J., Houck, A. A., Schuster, D. I., Majer, J., Blais, A., Devoret, M. H., Girvin, S. M., & Schoelkopf, R. J.** Charge-insensitive qubit design derived from the Cooper pair box. *Physical Review A 76, 042319 (2007).*

[3] **Place, A. P. M., et al.** New material platform for superconducting transmon qubits with coherence times exceeding 0.3 milliseconds. *Nature Communications 12, 1779 (2021).*

[4] **Chamberland, C., Zhu, G., Yoder, T. J., Hertzberg, J. B., & Cross, A. W.** Topological and subsystem codes on low-degree graphs with flag qubits. *Physical Review X 10, 011022 (2020).*

[5] **Fowler, A. G., Mariantoni, M., Martinis, J. M., & Cleland, A. N.** Surface codes: Towards practical large-scale quantum computation. *Physical Review A 86, 032324 (2012).*

**[6]** **Motzoi, F., Gambetta, J. M., Rebentrost, P., & Wilhelm, F. K.** Simple pulses for elimination of leakage in weakly nonlinear qubits. *Physical Review Letters 103, 110501 (2009).*

**[7]** **Yan, F., et al.** Tunable coupling scheme for implementing high-fidelity two-qubit gates. *Physical Review Applied 10, 054062 (2018).*

**[8]** **Sung, Y., et al.** Realization of high-fidelity CZ and ZZ-free iSWAP gates with a tunable coupler. *Physical Review X 11, 021058 (2021).*

**[9]** **Kim, Y., Eddins, A., Anand, S., Wei, K. X., van den Berg, E., Rosenblatt, S., Nayfeh, H., Wu, Y., Zaletel, M., Temme, K., & Kandala, A.** Evidence for the utility of quantum computing before fault tolerance. *Nature 618, 500 (2023).*

**[10]** **Acharya, R., et al. (Google Quantum AI)** Suppressing quantum errors by scaling a surface code logical qubit. *Nature 614, 676 (2023).*

**[11]** **Higgott, O., & Gidney, C.** Sparse blossom: correcting a million errors per core second with minimum-weight matching. *arXiv:2303.15933 (2023).*

**[12]** **Skoric, L., Browne, D. E., Barnes, K. M., Gillespie, N. I., & Campbell, E. T.** Parallel window decoding enables scalable fault tolerant quantum computation. *Nature Communications 14, 7040 (2023).*

**[13]** **Litinski, D.** A game of surface codes: Large-scale quantum computing with lattice surgery. *Quantum 3, 128 (2019).*

**[14]** **Bravyi, S., & Kitaev, A.** Universal quantum computation with ideal Clifford gates and noisy ancillas. *Physical Review A 71, 022316 (2005).*

**[15]** **Gidney, C., & Ekerå, M.** How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits. *Quantum 5, 433 (2021).*

**[16]** **Eisert, J., et al.** Quantum certification and benchmarking. *Nature Reviews Physics 2, 382 (2020).*

**[17]** **Magesan, E., Gambetta, J. M., & Emerson, J.** Scalable and robust randomized benchmarking of quantum processes. *Physical Review Letters 106, 180504 (2011).*

**[18]** **Boixo, S., et al.** Characterizing quantum supremacy in near-term devices. *Nature Physics 14, 595 (2018).*

**[19]** **Proctor, T., Carignan-Dugas, A., Rudinger, K., Nielsen, E., Blume-Kohout, R., & Young, K.** Direct randomized benchmarking for multiqubit devices. *Physical Review Letters 123, 030503 (2019).*

**[20]** **Cabrillo, C., Cirac, J. I., García-Fernández, P., & Zoller, P.** Creation of entangled states of distant atoms by interference. *Physical Review A 59, 1025 (1999).*

**[21]** **Monroe, C., Raussendorf, R., Ruthven, A., Brown, K. R., Maunz, P., Duan, L.-M., & Kim, J.** Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects. *Physical Review A 89, 022317 (2014).*

**[22]** **Mirhosseini, M., Sipahigil, A., Kalaee, M., & Painter, O.** Superconducting qubit to optical photon transduction. *Nature 588, 599 (2020).*

**[23]** **Bluvstein, D., et al.** Logical quantum processor based on reconfigurable atom arrays. *Nature 626, 58 (2024).*

**[24]** **Moses, S. A., et al. (Quantinuum)** A race-track trapped-ion quantum processor. *Physical Review X 13, 041052 (2023).*

**[25]** **Zhao, H., Chen, W. D., Kejriwal, A., et al.** Quantum-enabled microwave-to-optical transduction via silicon nanomechanics. *Nature Nanotechnology 20, 602-608 (2025).*

**[26]** **van Thiel, T. C., Weaver, M. J., Berto, F., et al.** Optical readout of a superconducting qubit using a piezo-optomechanical transducer. *Nature Physics 21, 401-405 (2025).*

**[27]** **PsiQuantum team.** A manufacturable platform for photonic quantum computing. *Nature 641, 876-883 (2025).*

**[28]** **Aghaee Rad, H., Ainsworth, T., Alexander, R. N., et al. (Xanadu).** Scaling and networking a modular photonic quantum computer. *Nature 638, 912-919 (2025).*

**[29]** **Gidney, C.** How to factor 2048 bit RSA integers with less than a million noisy qubits. *arXiv:2505.15917 (2025).*

**[30]** **Bennett, C. H., Brassard, G., Popescu, S., Schumacher, B., Smolin, J. A., & Wootters, W. K.** Purification of noisy entanglement and faithful teleportation via noisy channels. *Physical Review Letters 76, 722 (1996).*