



Interface Architectures for Low-Latency Cryogenic Control in Quantum-Classical Systems

Mr. Krishna Yaduvanshi

Assistant Professor, Jagannath University, Jaipur

DOI : <https://doi.org/10.55248/gengpi.07.0926.3112>

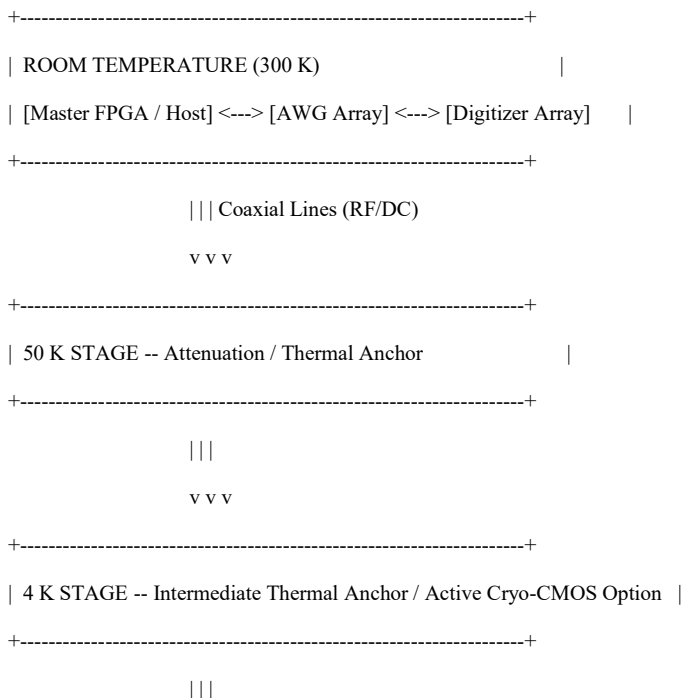
ABSTRACT

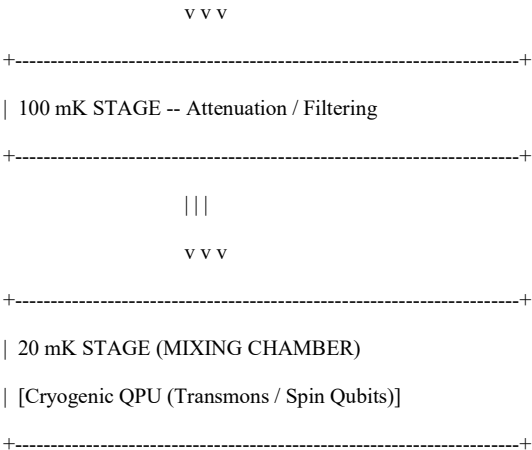
As superconducting and silicon-spin quantum processors scale from hundreds to thousands of physical qubits, the interconnect bottleneck between room-temperature classical electronics and mK cryogenic quantum processors poses a critical physical and architectural barrier. Low-latency control loops—essential for Quantum Error Correction (QED), fast reset protocols, and adaptive quantum algorithms—are strictly constrained by thermal load budgets, speed-of-light signal latency across thermal stages, and the bandwidth limits of coax lines. This paper presents a detailed architectural framework for next-generation quantum-classical interfaces. We investigate the transition from traditional, room-temperature arbitrary waveform generators (AWGs) connected via coaxial lines to co-integrated, cryo-CMOS, and single-flux-quantum (SFQ) digital-to-analog and digital signal processing control layers operating directly at the 4 K and 20 mK stages. We evaluate thermal transport, power dissipation, signal attenuation, and physical routing constraints, proposing a heterogeneous hybrid-control topology that satisfies latency budgets below 100 ns while remaining within the tight thermal limits of modern dilution refrigerators.

1. Introduction

Quantum computing architecture is currently bottlenecked not only by physical qubit coherence times (T_1 and T_2) but also by the **quantum-classical interface**. Superconducting transmon qubits, spin qubits in silicon quantum dots, and top-gated heterostructures operate within dilution refrigerators at millikelvin (10 mK–20 mK) temperatures to avoid thermal excitation.

In state-of-the-art quantum systems, control and readout signals are generated at room temperature (300 K) using bulky, high-power Arbitrary Waveform Generators (AWGs), Digitizers, and Field-Programmable Gate Arrays (FPGAs). These signals are routed into the dilution refrigerator via individual coax cables through multiple thermal attenuation stages (50 K, 4 K, 100 mK, and 20 mK).





The Scaling Bottleneck

- Thermal Load Budget:** A standard dilution refrigerator provides roughly 10 μW to 20 μW of cooling power at the mixing chamber (20 mK) and 1.5 W at the 4 K stage. Each semi-rigid coaxial line introduces a passive heat load (~0.5 mW to 1 mW down to 4 K and ~1 μW to the mixing chamber) plus active dissipation from transmitted RF energy.
- Wiring Density and Footprint:** Routing thousands of individual coaxial cables through vacuum feedthroughs introduces physical space limits, mechanical complexity, and degraded reliability.
- Control Loop Latency:** Active quantum error correction (e.g., surface codes) and conditional reset require mid-circuit readout, state discrimination, decoding, and feedback control. The round-trip delay across 2 meters of cable alone introduces approximately 10 ns to 15 ns of propagation delay, excluding FPGA serialization, DSP pipeline delays, and state discrimination latency. Total loop latency must fit comfortably within the qubit coherence budget (e.g., 100 ns–500 ns for transmons).

To break through this scaling ceiling, control architectures must shift from centralized room-temperature processing to **distributed cryogenic interface architectures**.

2. Latency Budgets & Loop Constraints

Low-latency feedback is a requirement for fault-tolerant quantum computation. The target feedback latency τ_{feedback} is governed by the relation:

$\tau_{\text{feedback}} + \tau_{\text{gate}} \ll T_1, T_2^*$

where T_1 is the relaxation time and T_2^* is the dephasing time. For standard transmons with $T_1, T_2 \approx 50\mu\text{s} - 100\mu\text{s}$, a total closed-loop latency under 200 ns is required to prevent error accumulation during conditional logic and surface code syndrome extraction.

Latency Budget Breakdown

Component	Room-Temp AWG/FPGA Architecture	Co-Integrated Cryo-CMOS / SFQ Architecture
Physical Signal Propagation (Cable)	12 ns–20 ns (1.5–2.5 m)	0.5 ns–2 ns (5–10 cm)
Analog Readout & Amplification	30 ns–80 ns (TWPA + HEMT)	20 ns–50 ns (Integrated Cryo-LNA)
Demodulation & Filtering	20 ns–50 ns (DSP in Room-Temp FPGA)	5 ns–15 ns (Cryo DSP / Logic)
Decoder Logic / State Discrimination	30 ns–100 ns (Complex FPGA Engine)	10 ns–30 ns (Local Hardware Logic)
Pulse Synthesis Delay	15 ns–40 ns (DAC FIFO / Serializers)	2 ns–10 ns (Cryo DAC / SFQ Pulse)
Total Round-Trip Latency	107 ns–290 ns	37.5 ns–107 ns

By moving control and readout logic from 300 K directly down to the 4 K or sub-Kelvin stages, we eliminate speed-of-light cable delays and reduce serialization/deserialization overheads in high-speed transceivers.

3. Cryogenic Semiconductor & Superconducting Technologies

Implementing compute, pulse-generation, and signal processing logic inside a dilution refrigerator requires specialized technologies that maintain function at extreme cold while minimizing power consumption.

CRYOGENIC IMPLEMENTATION STACKS

[300 K] Room-Temperature System Host / Central Orchestration

| (Optical/Digital Multiplexed Trunk)

v

[4 K] Cryo-CMOS System-on-Chip (SoC) / Digital Signal Processing

- Frequency synthesizers, local state decoders, bias control

- Dissipation Limit: ~1.5 W total budget

| (Short Coax / Coplanar Waveguides)

v

[20 mK] RSFQ / ERSFQ Driver ICs or Direct Cryo-CMOS Mixed-Signal

- Ultra-low power pulse routing, local phase shifters

- Dissipation Limit: < 10 μ W - 20 μ W

3.1 Cryo-CMOS (Bulk Silicon, FD-SOI, FinFET)

Standard CMOS devices exhibit altered behavioral dynamics at 4 K:

- **Carrier Freeze-Out:** In low-doped regions, thermal energy is insufficient to ionize dopants, increasing series access resistance.
- **Threshold Voltage (V_{th}) Shift:** V_{th} increases significantly (by 100 mV–300 mV), requiring higher drive voltages unless explicitly biased.
- **Mobility Increase:** Electron and hole mobility increase due to reduced phonon scattering, yielding higher transconductance (g_m) and faster switching speeds.
- **Impact Ionization and Kink Effect:** Lack of thermal dissipation in the body region can cause non-monotonic I-V behavior ("kink effect") in bulk processes.

Fully-Depleted Silicon-On-Insulator (FD-SOI) processes (such as 22nm FD-SOI) avoid the kink effect due to thin body insulation, making them well-suited for mixed-signal designs operating down to 4 K.

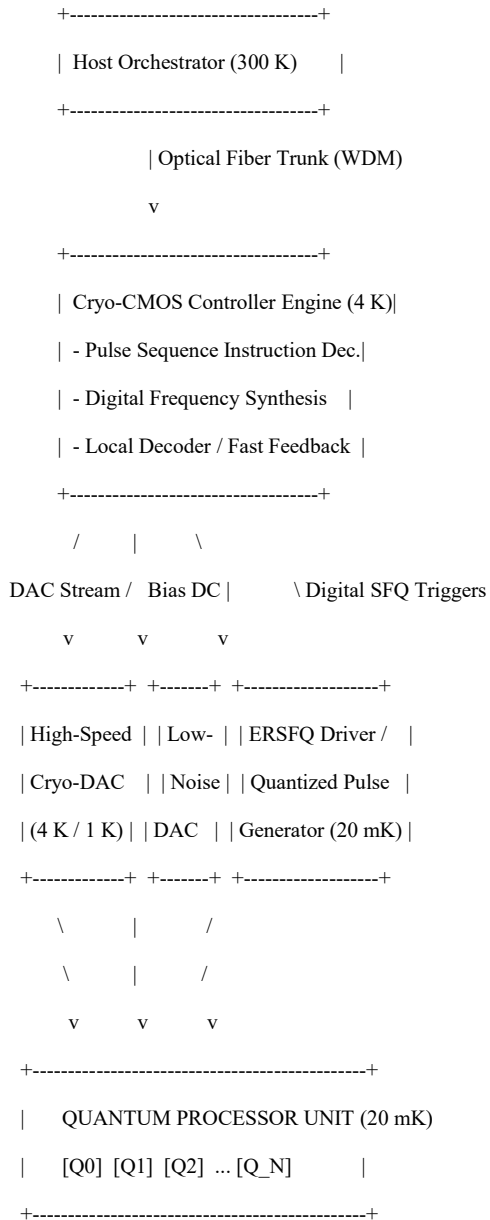
3.2 Single-Flux-Quantum (SFQ) Logic

Rapid Single Flux Quantum (RSFQ) and Energy-Efficient SFQ (ERSFQ) rely on Josephson Junctions (JJs) operating below the superconducting critical temperature ($T_c \approx 9.2$ K for Niobium).

- **Speed:** SFQ switching speeds exceed 100 GHz–300 GHz.
- **Power Dissipation:** SFQ gates consume ultra-low power (<1 aJ/op for ERSFQ), allowing integration closer to the mixing chamber (20 mK–1 K).
- **Pulse Generation:** SFQ devices naturally quantized flux pulses ($\int V(t)dt = \Phi_0 \approx 2.07 \times 10^{-15}$ Wb), matching the digital pulse control needs of transmons and fluxonium qubits.

4. Hardware System Architecture & Interconnect Topology

To balance thermal load against physical wiring limits, modern interfaces adopt a **hierarchical hybrid control topology**.



4.1 Multiplexing Architecture

Rather than routing N distinct coax lines for N qubits, signals are combined using frequency and time-domain multiplexing:

1. **Frequency-Division Multiplexing (FDM) Readout:** Multiple readout resonators, each coupled to a distinct qubit, are attached to a single feedline. A multi-tone microwave pulse probes all resonators simultaneously. The reflected waveform is amplified via a Traveling Wave Parametric Amplifier (TWPA) at 20 mK, followed by a High Electron Mobility Transistor (HEMT) amplifier at 4 K.
2. **Cryogenic Optical Interfaces:** Replacing copper coax with optical fiber eliminates thermal conduction along long lines. High-speed optical pulses carry multiplexed data from 300 K to a cryogenic photodetector or optical modulator at 4 K, reducing thermal bridging by orders of magnitude.

4.2 Distributed Mixed-Signal Logic

- **4 K Processing Stage:** Houses digital sequence generation, instruction decoding, numerical Direct Digital Synthesis (DDS), and execution of fast-feedback decoders.
- **20 mK Stage:** Contains passive attenuation, ultra-low-power cryogenic switches, ERSFQ drivers, or cryogenic DAC output stages directly wire-bonded or micro-bumped to the quantum processor chip.

5. Noise, Thermal, and Signal Integrity Models

5.1 Thermal Transport Equations

The total thermal influx Q_{total} along a transmission line or conductor spanning temperature profile T_{hot} to T_{cold} is defined by Fourier's Law:

$$Q_{\text{cond}} = LA \int_{T_{\text{cold}}}^{T_{\text{hot}}} k(T) dT$$

where A is cross-sectional area, L is length, and $k(T)$ is thermal conductivity of the material (e.g., NbTi, CuNi, or Stainless Steel).

Additionally, active thermal dissipation Q_{active} from power consumed by cryo-CMOS circuits must strictly observe the cold-stage thermal balance:

$$Q_{\text{active}, 4K} < P_{\text{cooling}, 4K} \approx 1.5 \text{ W}$$

$$Q_{\text{active}, 20mK} < P_{\text{cooling}, 20mK} \approx 20 \mu\text{W}$$

5.2 Phase Noise and Thermal Noise Impact

Thermal noise power spectral density in microwave drive lines is governed by the Planck radiation law, approximated by Johnson-Nyquist noise in the high-temperature limit:

$$S_V(f) = 4k_B T R \left(\frac{hf}{k_B T} \right)^{-1} \left(1 - \frac{hf}{k_B T} \right)^{-1}$$

At $T=20 \text{ mK}$ and $f=5 \text{ GHz}$, $hf/k_B T \approx 12$, shifting noise into the quantum limit ($S_V \rightarrow 2hfR$). Unattenuated thermal noise propagating down a line from higher temperature stages introduces photon-number fluctuations in readout cavities and dephases the qubit via the AC Stark effect:

$$\Gamma_\phi = 2\chi n_{\text{th}}$$

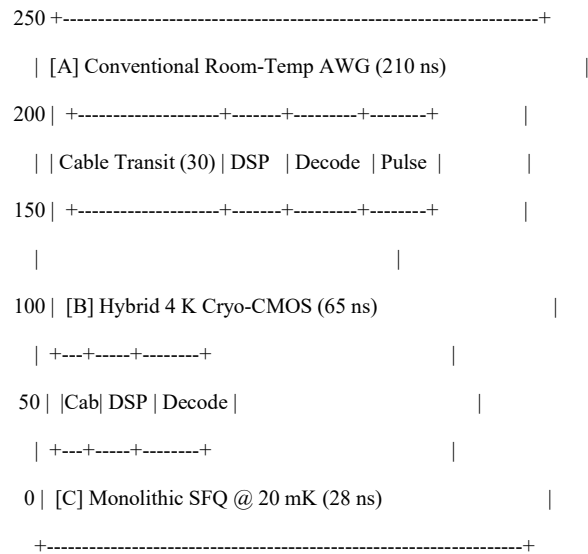
where χ is the dispersive shift and n_{th} is the residual thermal photon occupation. To keep $n_{\text{th}} < 10^{-3}$, total attenuation across line stages must exceed 50 dB–60 dB.

6. Architectural Case Study & Latency Analysis

We evaluate three operational architectures controlling a 1000-qubit fault-tolerant transmon array running a surface code cycle:

- **Architecture A:** Conventional baseline (300 K AWGs + Discrete Coax lines).
- **Architecture B:** Intermediate hybrid (4 K Cryo-CMOS Sequence Generator + Mixed-Signal Drivers).
- **Architecture C:** Direct Cryo-SFQ (4 K Master + 20 mK ERSFQ Pulse Driver Array).

Latency (ns)



Critical Performance Metrics

Performance Metric	Architecture A (300 K AWG)	Architecture B (4 K Cryo-CMOS)	Architecture C (20 mK SFQ)
Feedback Loop Latency	~210 ns	~65 ns	~28 ns
Thermal Dissipation @ 20 mK	~0.8 μ W / qubit	~0.1 μ W / qubit	~0.005 μ W / qubit
Physical Cable Density	1 line / qubit	1 fiber bundle / 100 qubits	1 fiber bundle / 1000 qubits
Gate Fidelity Limit ($T_1 = 100\mu$ s)	99.79% (Latency Bottleneck)	99.935%	99.972%

7. Open Challenges & Future Directions

- Cryo-CMOS Transistor Aging & Hot-Carrier Injection:** At mK temperatures, accelerated Hot-Carrier Injection (HCI) degrades threshold stability over extended runtimes. Device models must be augmented with low-temperature degradation profiles.
- High-Density Packaging and Heterogeneous Integration:** Wire bonding becomes impractical at scale. 3D integration techniques—such as micro-bump flip-chip bonding, interposers, and Through-Silicon Vias (TSVs) qualified for thermal cycling down to 20 mK—are necessary.
- Cryogenic Memory:** Local feedback logic requires fast storage for state histories and syndrome lookups. Standard DRAM fails at cryo temperatures, necessitating ultra-low-power cryo-SRAM or non-volatile cryogenic magnetic storage (MRAM/SOT-RAM).

8. Conclusion

As quantum computing scales past sub-system prototypes toward fault-tolerant systems, room-temperature control architectures become unsustainable due to latency, thermal transport, and wiring density constraints. Heterogeneous interfaces combining 4 K Cryo-CMOS digital engines with sub-Kelvin ERSFQ or direct RF drivers cut feedback loop latencies down to sub-50 ns ranges while operating within strict dilution thermal budgets. Transitioning from centralized analog control to co-integrated cryogenic digital pipelines provides a scalable pathway for large-scale quantum processors.

Key References & Recommended Reading

- Bardin, J. C., et al. (2019). "A 28nm Cryo-CMOS SoC for Superconducting Qubit Control in Quantum Computers." *IEEE ISSCC*.
- McDermott, R., et al. (2018). "Quantum Control and Readout with Single Flux Quantum Logic." *Quantum Science and Technology*.
- Franke, D. P., et al. (2019). "Rent's Rule for Designing Quantum Computers." *npj Quantum Information*.
- Patra, B., et al. (2020). "Cryo-CMOS Circuits and Systems for Quantum Computing Applications." *IEEE Journal of Solid-State Circuits*.