

Realization of Thread Level Parallelism on Quantum Devices

Keren Li¹, Shiyao Hou², and Zipeng Wu³

¹*Shenzhen University, nanshan, Shenzhen, China*

²*Sichuan Normal University, Chengdu, China*

³*Shenzhen SpinQ Technology Co., Shenzhen, China*

Scaling up quantum devices is a central challenge for realizing practical quantum computation. Modular quantum architectures promise scalability, yet experiments to date have relied on either $\lesssim 10^3$ -qubit monolithic chips or fragile interconnects with high loss. Here, we introduce a classical linkage scheme that merges multiple independent quantum processing units (QPUs) into a single logical device, enabling thread-level parallelism (TLP). Theoretically, we show that quantum routines with product-state inputs and low-rank entangling layers can be re-expressed in an efficient parallelizable form. Experimentally, we validate this architecture on clusters comprising up to sixteen benchtop nuclear magnetic resonance (NMR) quantum nodes. A four-qubit Greenberger-Horne-Zeilinger (GHZ) state is partitioned into parallel two-qubit subcircuits, achieving a fidelity of 93.8 % with respect to the ideal state. A non-Hermitian evolution, implemented via a truncated Cauchy integral on Hermitian Hamiltonians, reproduces exact observables with high accuracy. Our results demonstrate that classical links suffice to scale up the logical size of quantum computations and realize general, non-unitary channels on today's hardware, opening an experimentally accessible route toward software-defined, clustered quantum accelerators.

Acknowledgement: National Natural Science Foundation of China, Guangdong Basic and Applied Basic Research Foundation, the Guangdong Provincial Quantum Science Strategic Initiative

- [1] Realization of Thread Level Parallelism on Quantum Devices, arXiv preprint arXiv:2511.05436