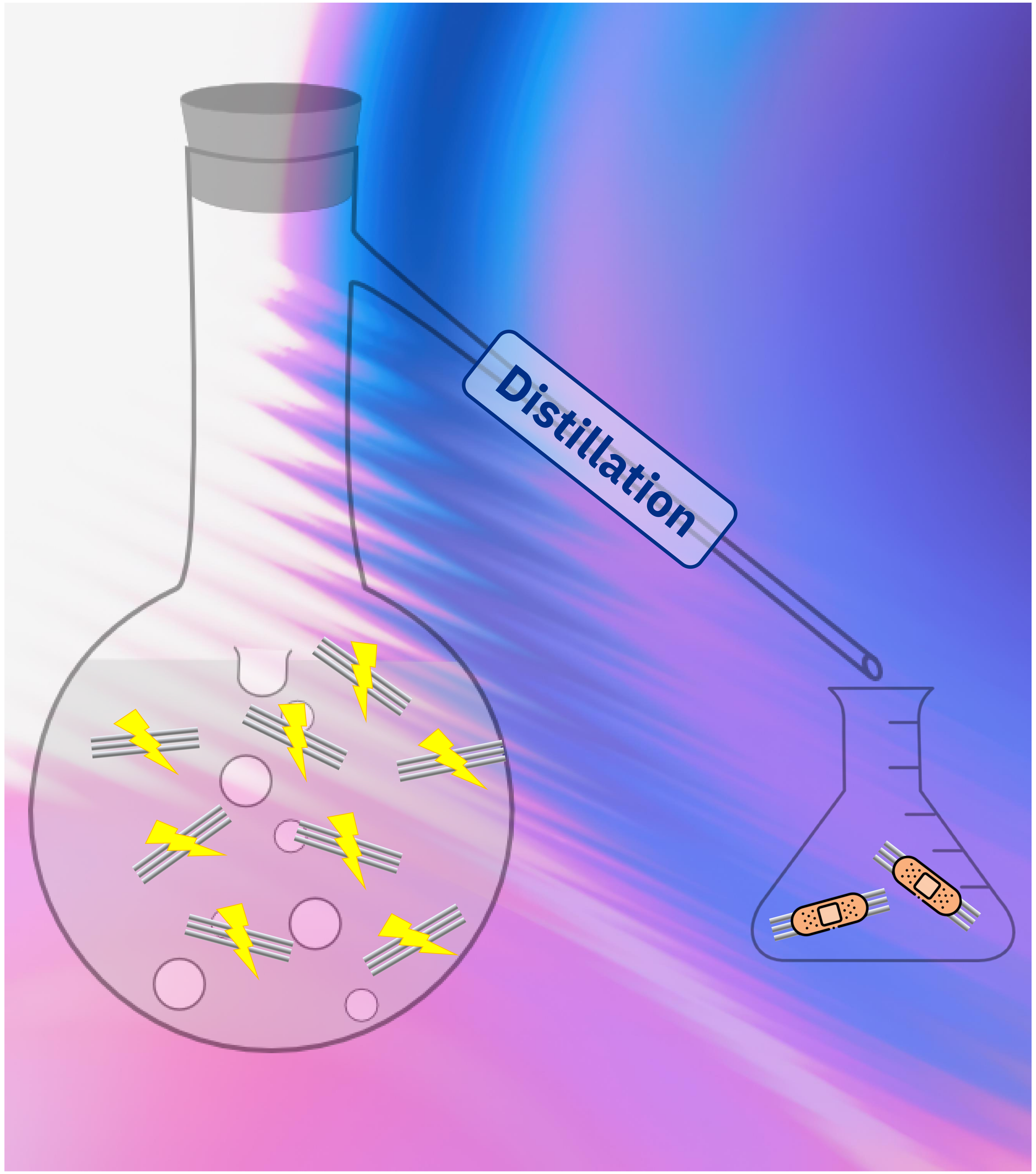
Distillation for early fault-tolerant devices

Elisa Bäumer Marty
Research Scientist
IBM Quantum

- Motivation
- IBM Roadmap:
  - Utility
  - Quantum Advantage
  - Fault-tolerance
- Entanglement distillation



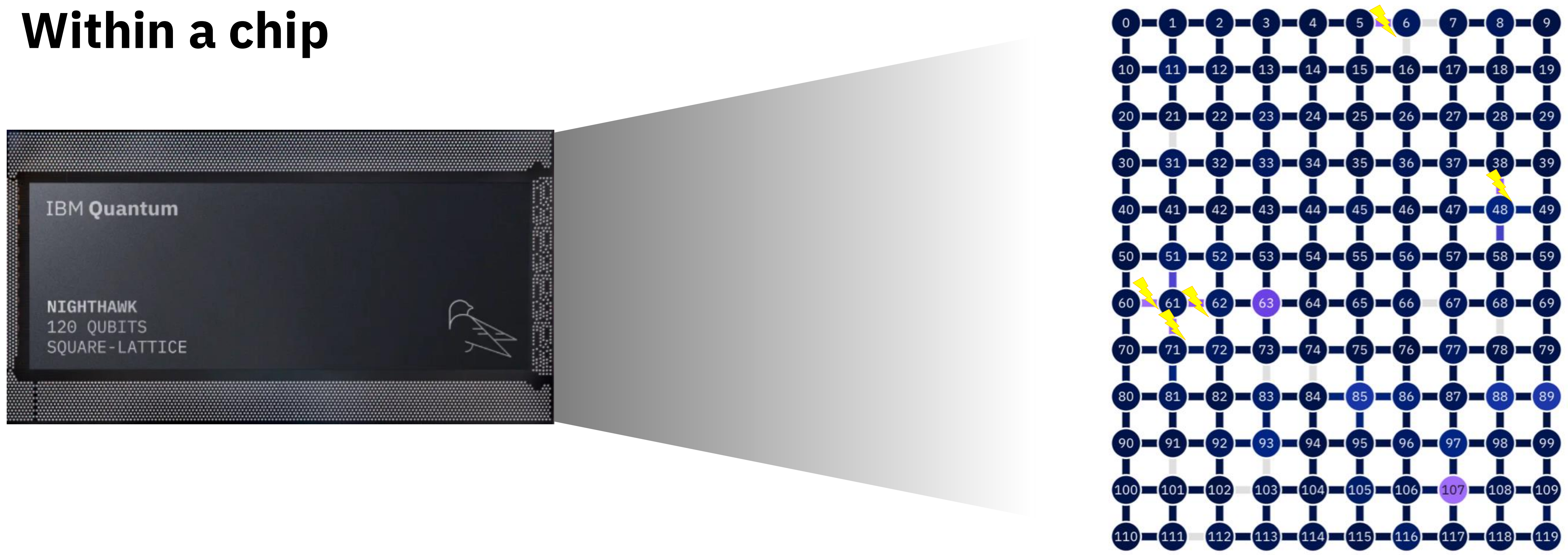
IBM

Motivation

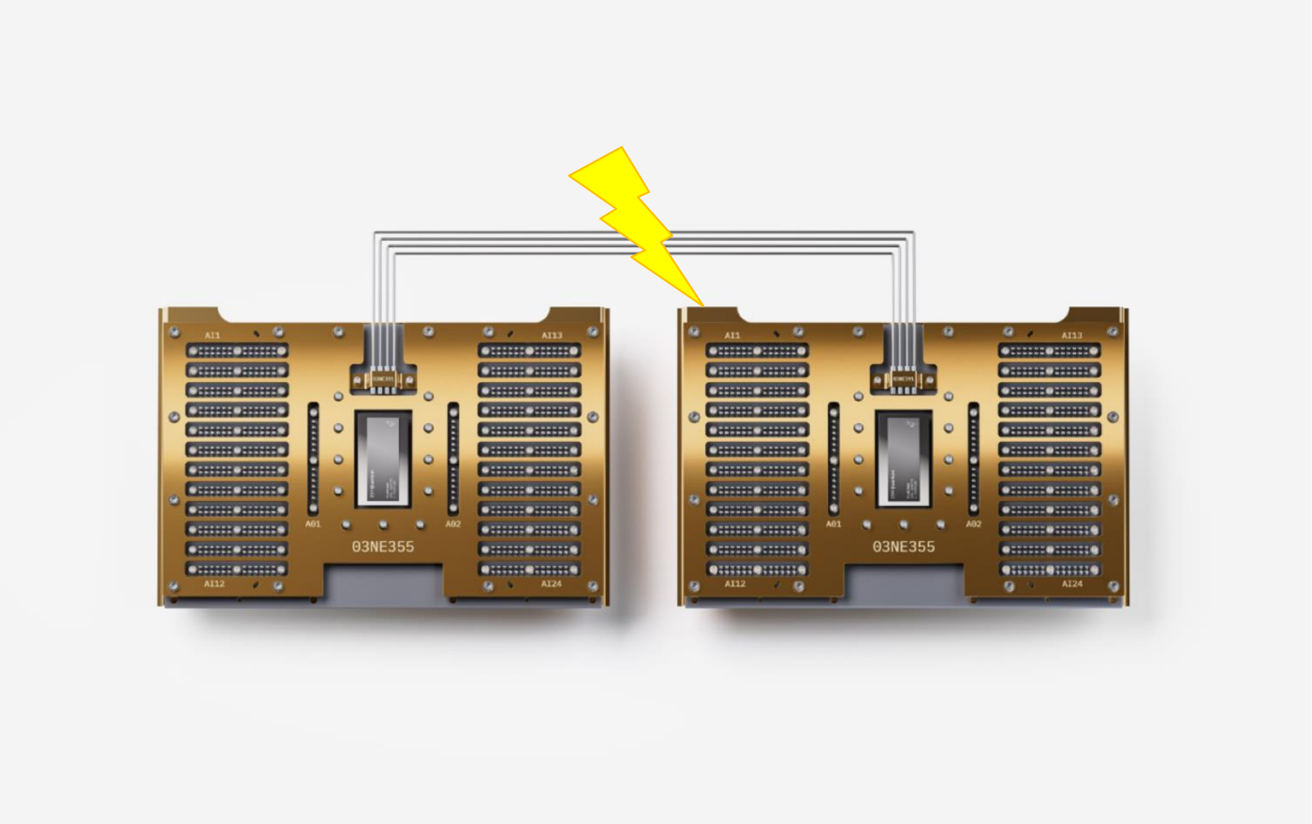
Early fault-tolerant quantum computing architectures are expected to exhibit highly non-uniform error rates:

Noisy operations

Within a chip

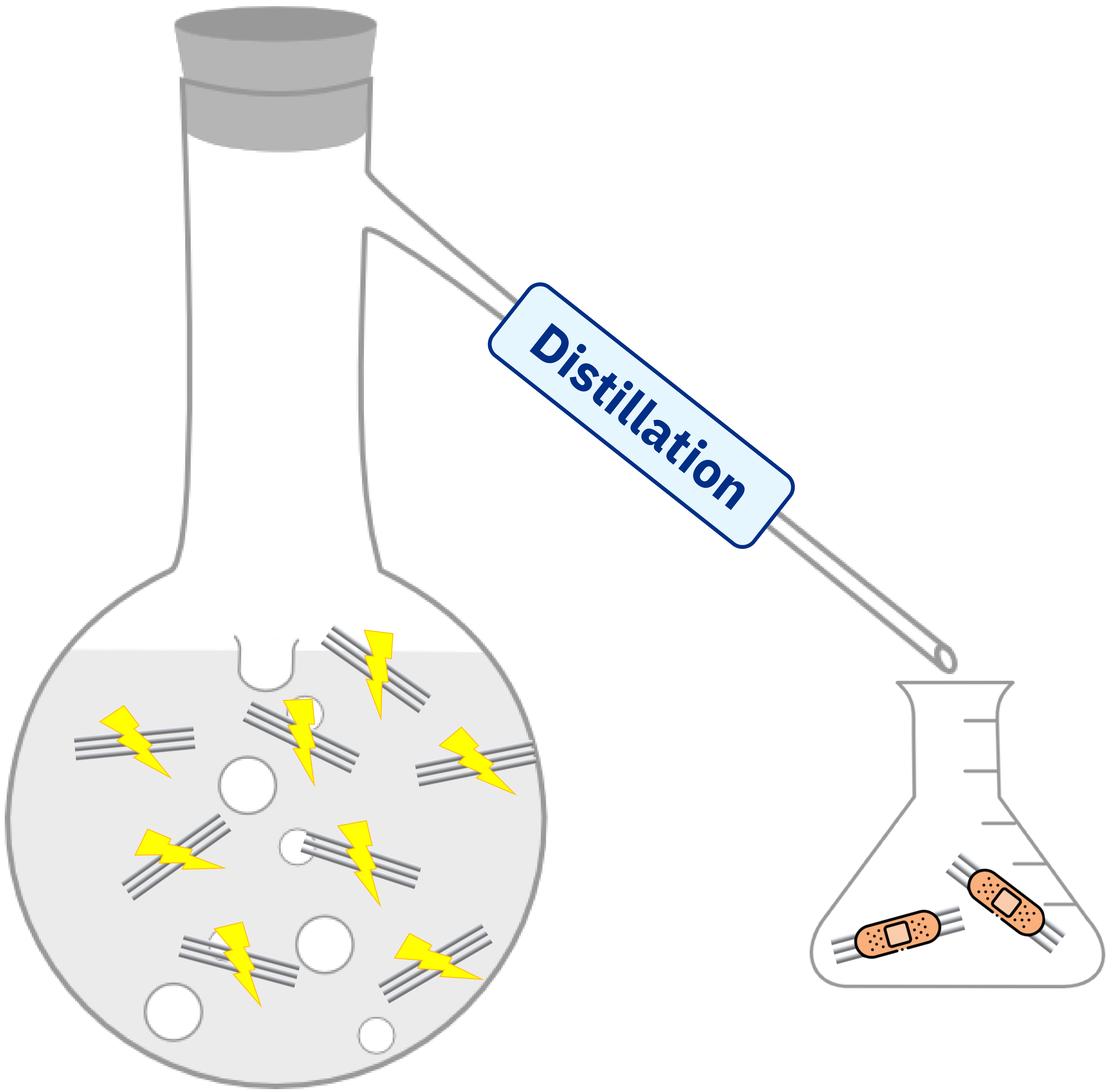


Connecting chips



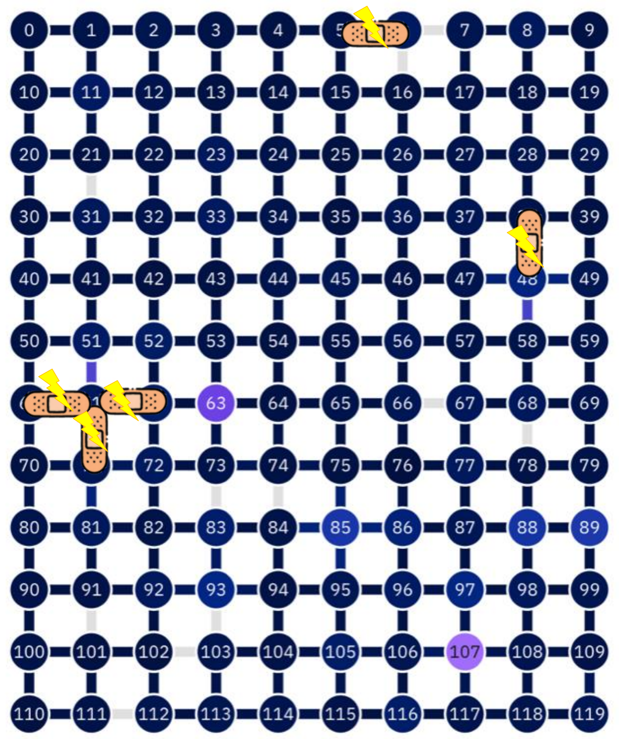
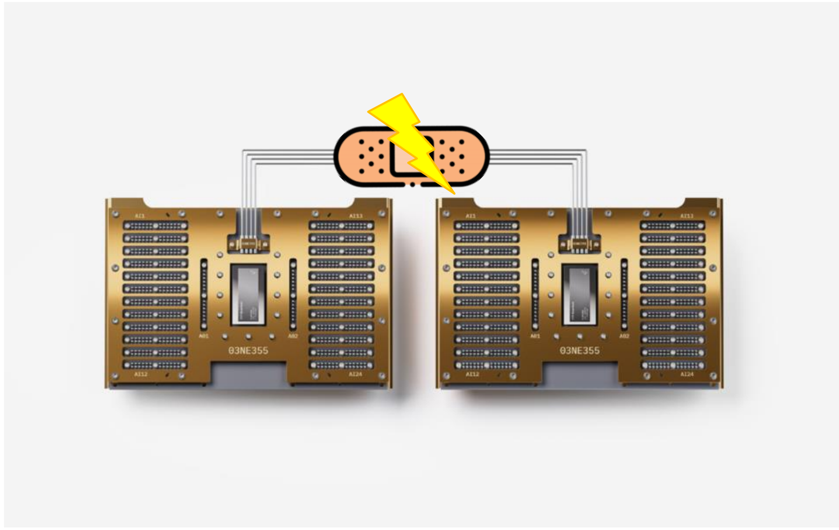
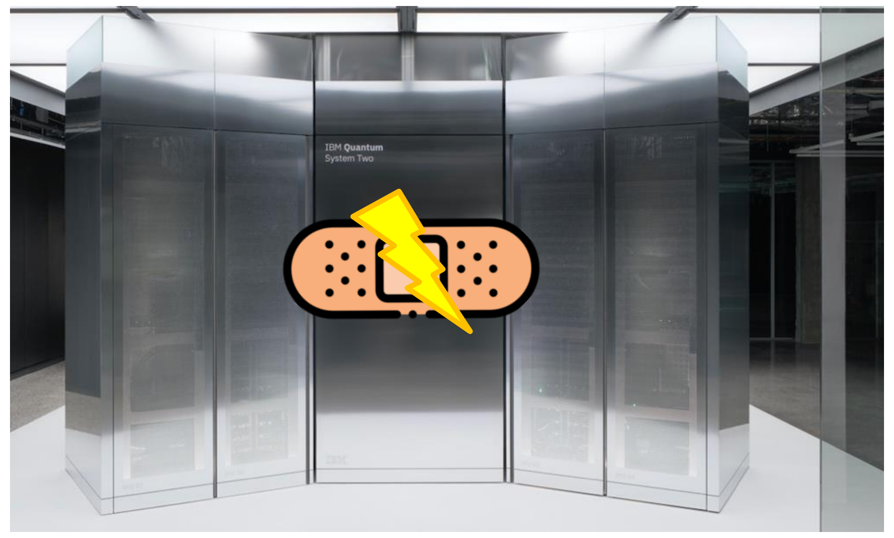
Connecting fridges

Distillation



➢ Distillation provides a natural way to trade additional operations and qubits for higher-fidelity ones:

Distilled operations 🩹

- Entanglement distillation
- Magic state distillation
- Synthillation: Distillation of larger, synthesized gates

➢ Development of suitable distillation and synthillation protocols can drastically improve the scalability of algorithms!

		Hardware
2016–2019 (completed)	Ran quantum circuits on IBM Quantum Platform	Early (completed): Canary 5 qubits; Albatross 16 qubits; Penguin 20 qubits; Prototype 53 qubits
2020 (completed)	Released multi-dimensional roadmap publicly with initial focus on scaling	Falcon (completed): Benchmarking; 27 qubits
2021 (completed)	Enhanced quantum execution speed by 100x with Qiskit Runtime	
2022 (completed)	Brought dynamic circuits to unlock more computations	Eagle (completed): Benchmarking; 127 qubits
2023 (completed)	Enhanced quantum execution speed by 5x with Quantum Serverless and execution modes	
2024 (completed)	Demonstrated accurate execution of a quantum circuit at a scale beyond exact classical simulation (5K gates on 156 qubits)	Heron (5K) (completed): Error mitigation; 5K gates; 133 qubits
2025 (completed)	Deliver quantum + HPC tools that will leverage Nighthawk, a new higher-connectivity quantum processor able to execute more complex circuits	Nighthawk (5K): Error mitigation; 5K gates; 120 qubits
2026	Enable the first examples of quantum advantage using a quantum computer with HPC	Nighthawk (7.5K) (on target): Error mitigation; 7.5K gates; 120 qubits; Up to 120x3 = 360 qubits
2027	Improve quantum circuit quality to allow 10K gates	Nighthawk (10K): Error mitigation; 10K gates; 120 qubits; Up to 120x9 = 1080 qubits
2028	Improve quantum circuit quality to allow 15K gates	Nighthawk (15K): Error mitigation; 15K gates; 120 qubits; Up to 120x9 = 1080 qubits
2029	Deliver a fault-tolerant quantum computer with the ability to run 100M gates on 200 qubits	Starling (100M): Fault-tolerant; 100M gates; 200 qubits
2033+	Beyond 2033, quantum computers will run circuits comprising a billion gates on up to 2000 qubits, unlocking the full power of quantum computing	Blue Jay (1B): Fault-tolerant; 1B gates; 2000 qubits

Code assistant (2024 onward, completed)

Applying algorithms to applications
- Functions (2024–2025, completed)
- Use case benchmarking toolkit (2026, on target)
- Computation libraries (2027 onward)

Discovering new algorithms for advantage
- Advanced classical transpilation tools (2024, completed)
- Advanced classical mitigation tools (2025, completed)
- Utility mapping tools (2026–2028, on target)
- Circuit libraries (2029 onward)

Orchestrating workloads for quantum + HPC
- Resource Management (2023 onward)
- Qiskit Serverless (2023, completed)
- Plugins for HPC (2024, completed)
- C API (2025, completed)
- Execution modes (2023–2025, completed)
- Profiling tools (2026–2027, on target)
- Workflow accelerators (2028 onward)

Accurately and efficiently executing on quantum computers
- IBM Quantum Experience (2016–2020, completed)
- Qiskit Runtime (2021 onward)
- OpenQASM 3 (2021, completed)
- Dynamic Circuits (2022, completed)
- Error mitigation (2023, completed)
- 200K CLOPS (2024, completed)
- Utility-scale dynamic circuits (2025, completed)
- Fault-tolerant ISA (2029 onward)

Hardware →

Completed
On target

Quantum computing

Key milestones in bringing useful quantum computing to the world

2023

Establish quantum utility

2026

Demonstrate quantum advantage

2029

Deliver the first large-scale, fault-tolerant quantum computer

Quantum computing

Key milestones in bringing useful quantum computing to the world

2023

Establish quantum utility

2026

Demonstrate quantum advantage

2029

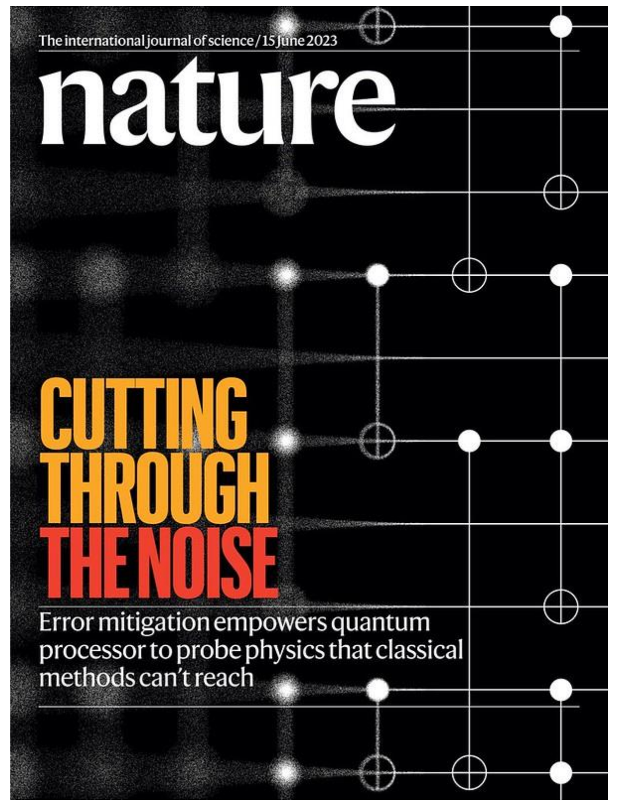
Deliver the first large-scale, fault-tolerant quantum computer

Theory | Utility | Advantage

Quantum Utility (2023)

Demonstration that a quantum computer can run quantum circuits beyond the ability of a classical computer simulating a quantum computer

Confirmation via research, papers, & theory



IBM's 2023 research paper ("Evidence for the utility of quantum computing before fault tolerance") provided evidence and methods to move the industry into the Utility era

https://www.nature.com/articles/s41586-023-06096-3

Quantum Advantage (TBD)

Demonstration that a quantum computer can solve a problem more accurately, cheaper, or more efficiently than classical computing alone

Confirmation via real-world usage

Advantage will come at different times in different domains and depends on the continued advancement of quantum algorithm implementations across industries

"Evidence of utility before fault tolerance" (2023)

A noisy quantum computer produces accurate expectation values on 127 qubits and 2880 gates, outside of brute force classical computation.

Y. Kim, A. Eddins, et al, Nature. 618, 500–505 (2023)

Quantum computing

Key milestones in bringing useful quantum computing to the world

2023

Establish quantum utility

2026

Demonstrate quantum advantage

2029

Deliver the first large-scale, fault-tolerant quantum computer

Performing an information processing task more efficiently, cost-effectively, or accurately using quantum computers than is known to be possible with classical computers alone

There are two important nuances to this →

1

We must establish trust in the outputs of a real and noisy quantum device.

A scientific experiment is only as good as confidence in methods.

2

Quantum advantage is not something that will happen at a singular moment in time.

It is a hypothesis that is subject to falsification.

We predict that quantum advantage will be found by the end of 2026...

... but only if we work together.

Application use cases

Working groups

IBM Quantum is not in this alone. Working groups bring together the best industry pioneers and scientists in their field to accelerate our path to achieving Quantum Advantage across several domain areas:

Optimization

July 19–20, Zurich, 2023

UKRI:STFC
E.On
Fraunhofer
Los Alamos

Materials and HPC

April 17–18, Chicago, 2023

Riken
Univ. Tokyo
Univ. Chicago
Oak Ridge Nat. Labs

High energy physics

November 2–3, Geneva, 2023

CERN
DESY
PSNC
Univ. Tokyo

Healthcare and life sciences

April 13–14, Cleveland, 2023

Cleveland Clinic
Univ. Toronto
Univ. Chicago
Yonsei Univ

Application use cases

White papers

A first outcome of the working groups is the white papers in their respective domains.

Optimization

Quantum Optimization: Potential, Challenges, and the Path Forward
https://arxiv.org/abs/2312.02279

Materials and HPC

Quantum-centric Supercomputing for Materials Science: A Perspective on Challenges and Future Directions
https://arxiv.org/abs/2312.09733

High energy physics

Quantum Computing for High-Energy Physics: State of the Art and Challenges. Summary of the QC4HEP Working Group
[arXiv:2307.03236](https://arxiv.org/abs/2307.03236)

Healthcare and life sciences

Towards quantum-enabled cell-centric therapeutics
[arXiv:2307.05734](https://arxiv.org/abs/2307.05734)

Quantum Optimization: Potential, Challenges, and the Path Forward*

arXiv:2312.02279v1 [quant-ph] 4 Dec 2023

Quantum-centric Supercomputing for Materials Science:
A Perspective on Challenges and Future Directions

arXiv:2312.09733v1 [quant-ph] 14 Dec 2023

Towards quantum-enabled cell-centric therapeutics

arXiv:2307.05734v2 [quant-ph] 1 Aug 2023

Quantum Computing for High-Energy Physics
State of the Art and Challenges
Summary of the QC4HEP Working Group

arXiv:2307.03236v1 [quant-ph] 6 Jul 2023

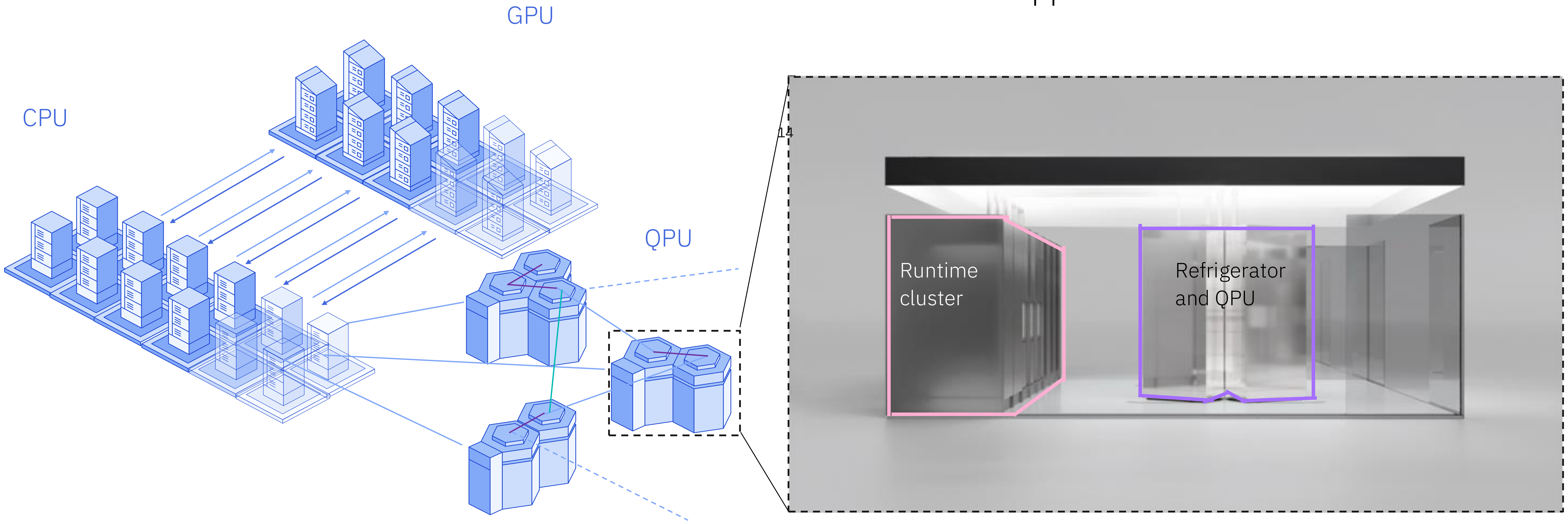
Alberto Di Meglio,[1,*] Karl Jansen,[2,3,†] Ivano Tavernelli,[4,‡] Constantia Alexandrou,[5,3] Srinivasan Arunachalam,[6] Christian W. Bauer,[7] Kerstin Borras,[8,9] Stefano Carrazza,[10,1] Arianna Crippa,[2,11] Vincent Croft,[12] Roland de Putter,[6] Andrea Delgado,[13] Vedran Dunjko,[12] Daniel J. Egger,[4] Elias Fernández-Combarro,[14] Elina Fuchs,[1,15,16] Lena Funcke,[17] Daniel González-Cuadra,[18,19] Michele Grossi,[1] Jad C. Halimeh,[20,21] Zoë Holmes,[22] Stefan Kühn,[2] Denis Lacroix,[23] Randy Lewis,[24] Donatella Lucchesi,[25,26,1] Miriam Lucio Martinez,[27,28] Federico Meloni,[8] Antonio Mezzacapo,[6] Simone Montangero,[25,26] Lento Nagano,[29] Voica Radescu,[30] Enrique Rico Ortega,[31,32,33,34] Alessandro Roggero,[35,36] Julian Schuhmacher,[4] Joao Seixas,[37,38,39] Pietro Silvi,[25,26] Panagiotis Spentzouris,[40] Francesco Tacchino,[4] Kristan Temme,[6] Koji Terashi,[29] Jordi Tura,[12,41] Cenk Tüysüz,[2,11] Sofia Vallecorsa,[1] Uwe-Jens Wiese,[42] Shinjae Yoo,[43] and Jinglei Zhang[44,45]

[1] European Organization for Nuclear Research (CERN), CH-1211 Geneva, Switzerland
[2] CQTA, Deutsches Elektronen-Synchrotron DESY, Platanenallee 6, 15738 Zeuthen, Germany
[3] Computation-based Science and Technology Research Center, The Cyprus Institute, 20, Constantinou Kavafi str., 2121 Nicosia, Cyprus
[4] IBM Quantum, IBM Research – Zurich, 8803 Rüschlikon, Switzerland
[5] Department of Physics, University of Cyprus, PO Box 20537, 1678 Nicosia, Cyprus
[6] IBM Quantum, IBM Research - 1101 Kitchawan Rd, Yorktown Heights, NY, USA
[7] Physics Division LBNL - M/S 50A5104 1 Cyclotron Rd Berkeley, CA, USA
[8] Deutsches Elektronen-Synchrotron DESY, Notkestrasse 85, 22607 Hamburg, Germany
[9] RWTH Aachen University, Templergraben 55, 52062 Aachen, Germany
[10] TIF Lab, Dipartimento di Fisica, Università degli Studi di Milano and INFN Sezione di Milano, Milan, Italy
[11] Institut für Physik, Humboldt-Universität zu Berlin, Newtonstr. 15, 12489 Berlin, Germany
[12] $\langle aQa^L \rangle$ Applied Quantum Algorithms Leiden, The Netherlands
[13] Physics Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA
[14] Department of Computer Science, Facultad de Ciencias, University of Oviedo, 33007, Asturias, Spain
[15] Institute of Theoretical Phyiscs, Leibniz University Hannover, 30167 Hannover, Germany
[16] Physikalisch-Technische Bundesanstalt, 38116 Braunschweig, Germany
[17] Transdisciplinary Research Area "Building Blocks of Matter and Fundamental Interactions" (TRA Matter) and Helmholtz Institute for Radiation and Nuclear Physics (HISKP), University of Bonn, Nußallee 14-16, 53115 Bonn, Germany
[18] Institute for Theoretical Physics, University of Innsbruck, 6020 Innsbruck, Austria
[19] Institute for Quantum Optics and Quantum Information of the Austrian Academy of Sciences, 6020 Innsbruck, Austria
[20] Department of Physics and Arnold Sommerfeld Center for Theoretical Physics, Ludwig-Maximilians-Universität München,Germany
[21] Munich Center for Quantum Science and Technology, Germany
[22] Institute of Physics, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
[23] Paris-Saclay University, CNRS/IN2P3, IJCLab, 91405 Orsay, France
[24] Department of Physics and Astronomy, York University, Toronto, Ontario, Canada, M3J 1P3
[25] Department of Physics and Astronomy, Università di Padova,V. Marzolo 8, I-35131 Padova, Italy
[26] INFN - Sezione di Padova, Via Marzolo 8, 35131 Padova, Italy
[27] Nikhef – National Institute for Subatomic Physics, Science Park 105, 1098 XG Amsterdam, The Netherlands
[28] Department of Gravitational Waves and Fundamental Physics, Maastricht University, 6200 MD Maastricht, The Netherlands
[29] International Center for Elementary Particle Physics (ICEPP), The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[30] IBM Quantum, IBM Deutschland Research & Development GmbH - Schoenaicher Str. 220, 71032 Boeblingen, Germany
[31] Department of Physical Chemistry, University of the Basque Country UPV/EHU, Box 644, 48080 Bilbao, Spain
[32] Donostia International Physics Center, 20018 Donostia-San Sebastián, Spain
[33] EHU Quantum Center, University of the Basque Country UPV/EHU, P.O. Box 644, 48080 Bilbao, Spain
[34] IKERBASQUE, Basque Foundation for Science, Plaza Euskadi 5, 48009 Bilbao, Spain
[35] Department of Physics, University of Trento, via Sommarive 14, I–38123, Povo, Trento, Italy
[36] INFN-TIFPA Trento Institute of Fundamental Physics and Applications, via Sommarive 14, I–38123, Trento, Italy
[37] Instituto Superior Técnico, Dep. Física, Lisboa, Portugal
[38] Center of Physics and Engineering of Advanced Materials (CeFEMA), Instituto Superior Técnico, Lisboa, Portugal,
[39] Laboratory of Physics for Materials and Emergent Technologies (LaPMET), Portugal
[40] Fermi National Accelerator Laboratory, Kirk and, Pine St, Batavia, IL 60510, USA
[41] Instituut-Lorentz, Universiteit Leiden, P.O. Box 9506, 2300 RA Leiden, The Netherlands

Quantum-centric supercomputing

Delivering impactful quantum computing requires the interplay of quantum and classical resources at scale

Quantum-centric supercomputing is the path toward industrial-scale applications



Quantum computing

Key milestones in bringing useful quantum computing to the world

2023

Establish quantum utility

2026

Demonstrate quantum advantage

2029

Deliver the first large-scale, fault-tolerant quantum computer

Errors in Quantum Systems

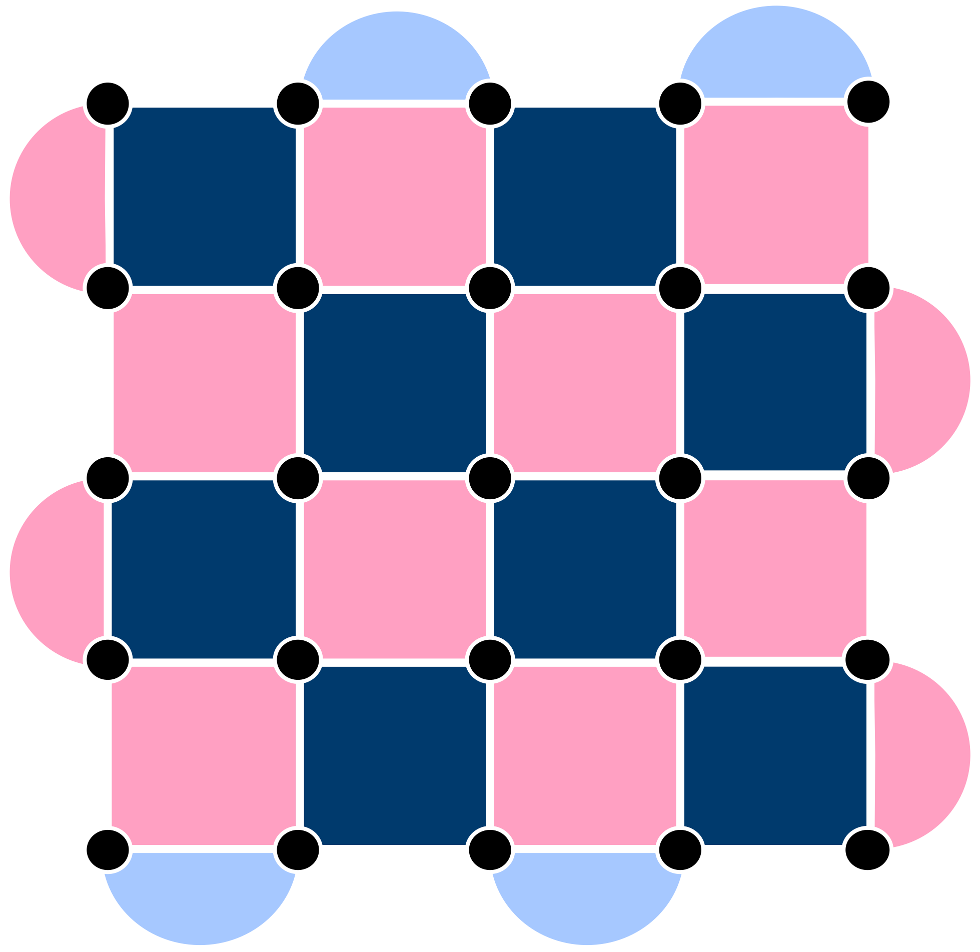
Classical error correction: repetition code 0 → 000, 1 → 111

Quantum error correction: No-Cloning Theorem

→ Quantum Information **cannot** be copied

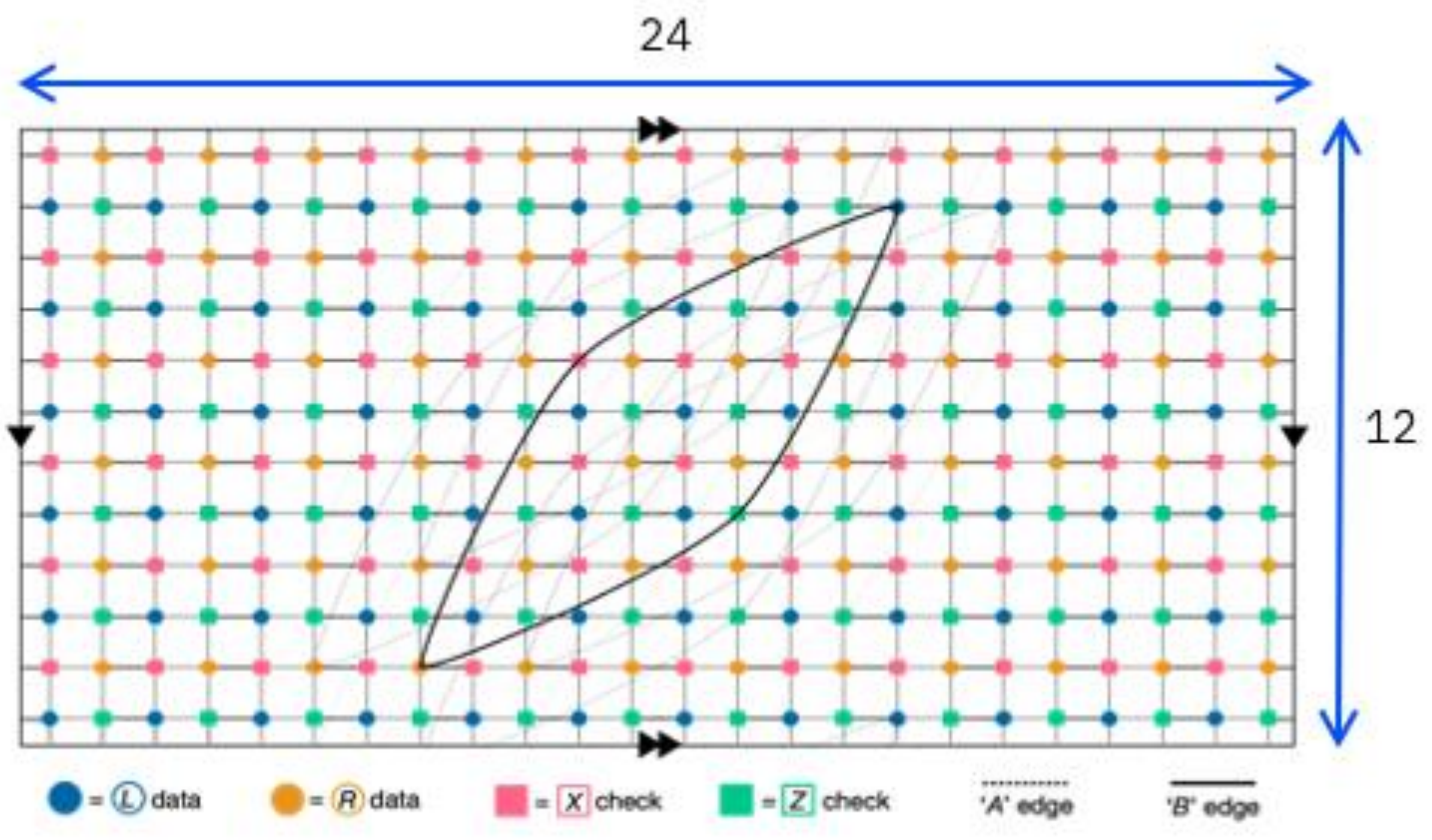
➔ **We cannot copy arbitrary qubits,** but need to **protect bits and phases**

➔ **Huge overhead needed**, e.g. surface code requires physical qubits per logical qubit with distance d $\approx d^2$



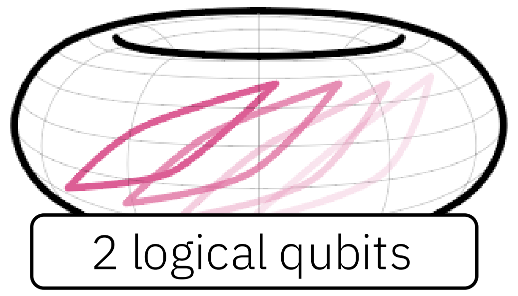
Scaling beyond the surface code

The gross code is **10x more efficient** than the surface code



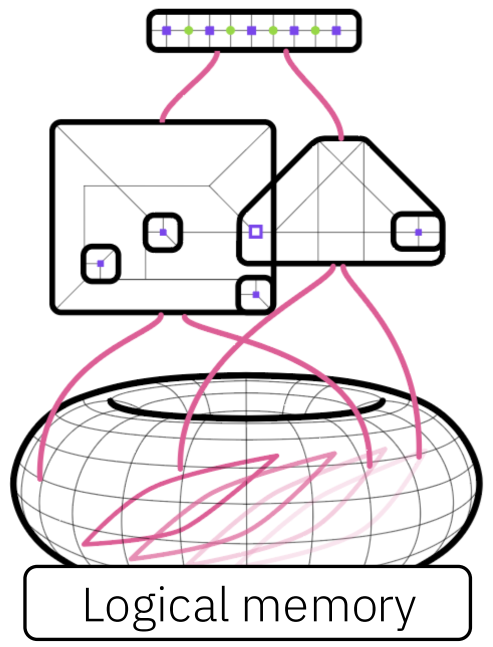
IBM fault-tolerant quantum computing roadmap

Loon (2025)



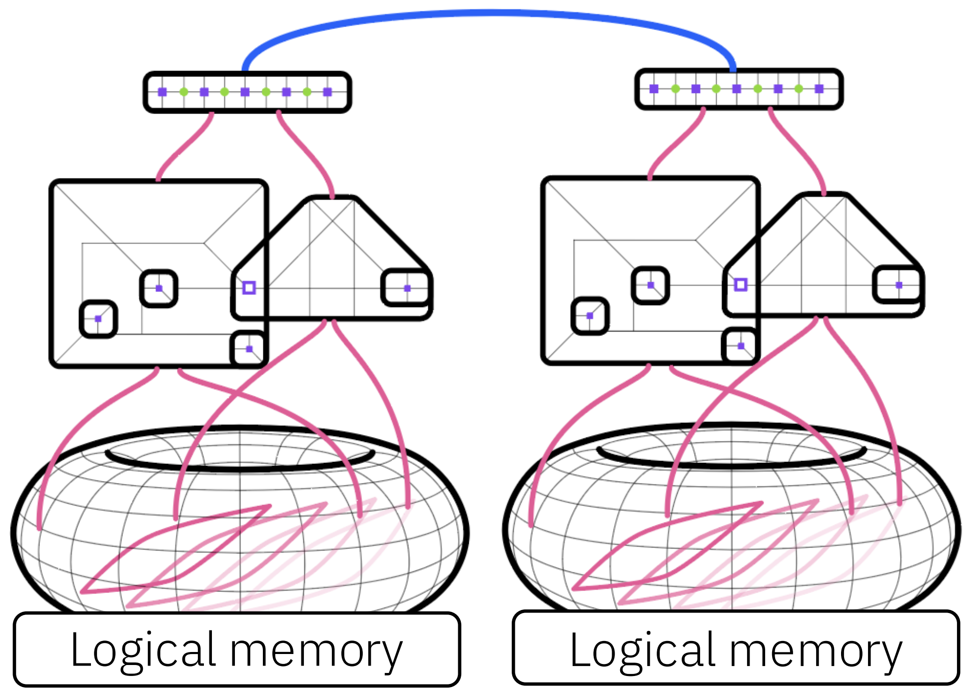
- 6-way connectivity
- c-coupler demo
- Automated design

Kookaburra (2026)



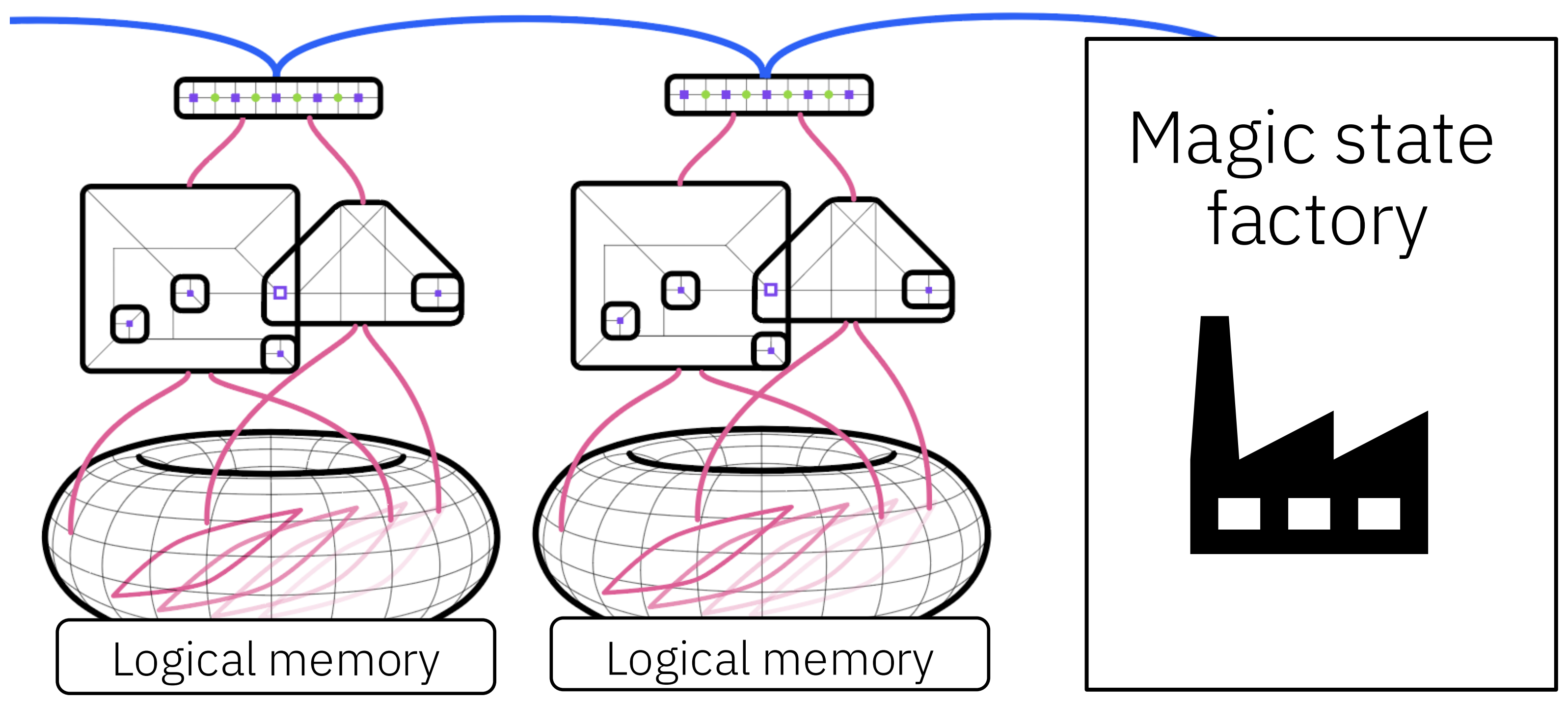
- “Long” c-couplers
- Real-time decoding
- LPU (+~100 qubits) for logical operations

Cockatoo (2027)



- Two blocks of gross code + LPU
- Module-to-module logical communication over l-couplers

Starling (2028)

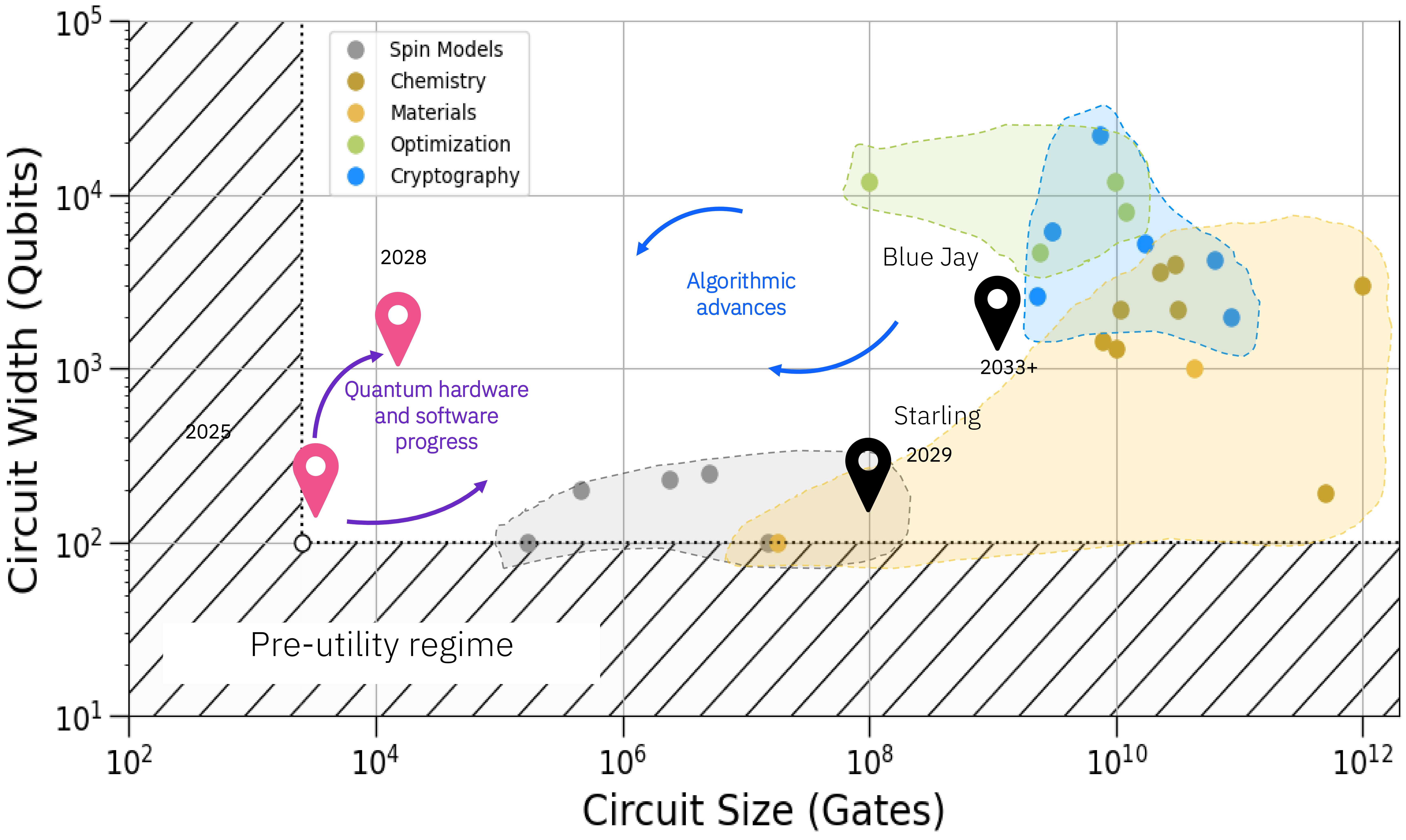


- Multiple blocks of gross code + LPUs
- Universal computation with magic state distillation

What can you do with a large-scale, fault-tolerant quantum computer?

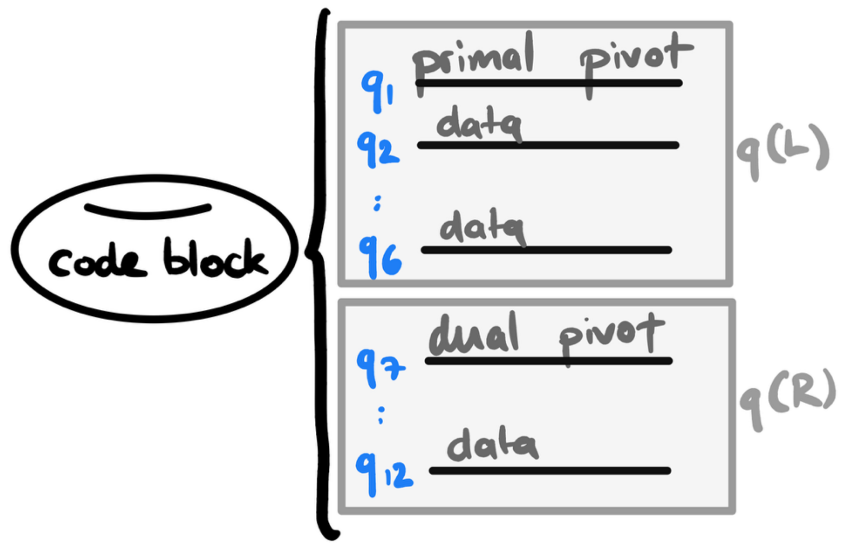
Adoption of quantum computing systems enables new applications or use cases to be discovered, potentially with current or near-future systems.

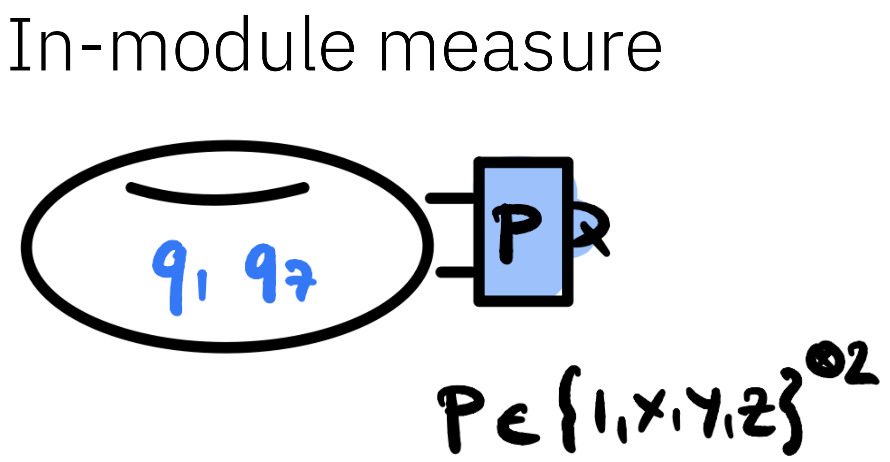
Investments in algorithm research should enable known applications to be realized sooner.

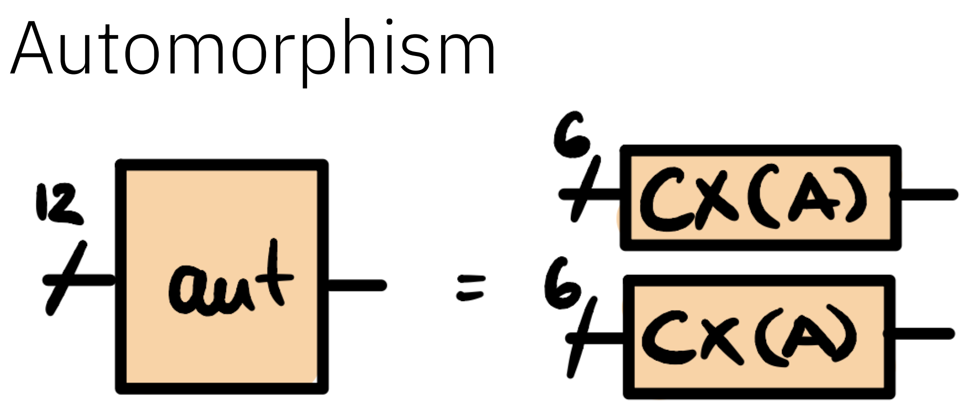


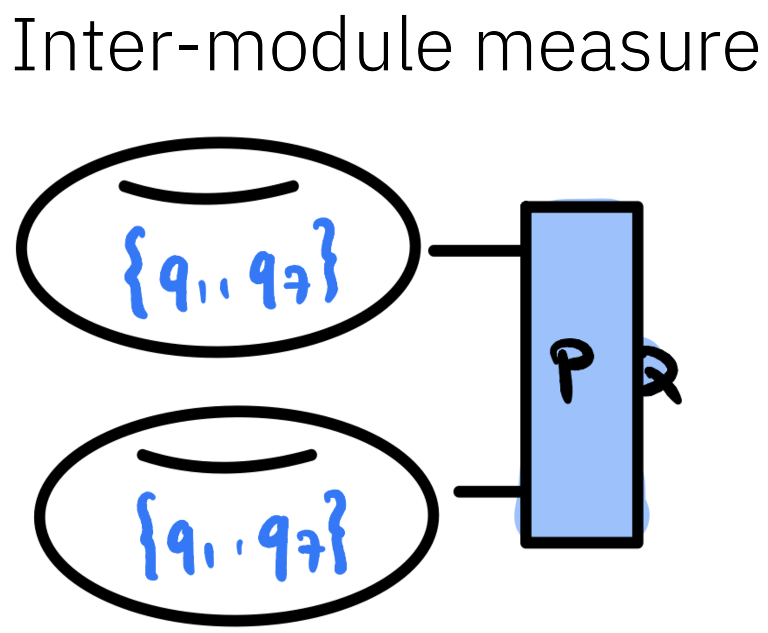
How to compute on the Gross code

- The **basic instruction set** is given by the bicycle instruction set architecture (ISA) – very far from what we currently use on near-term devices





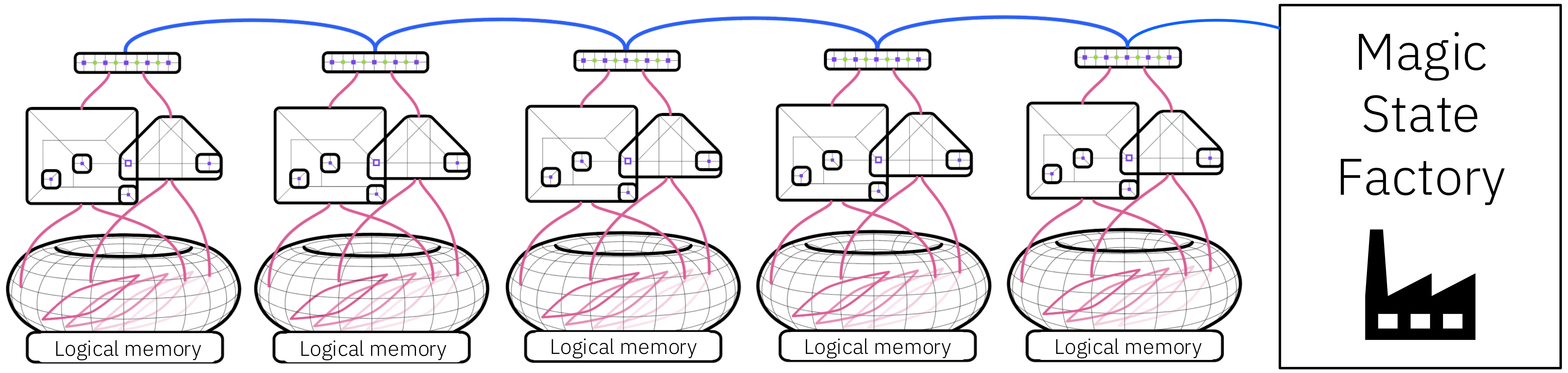




➢ This enables fault-tolerant implementation of the full Clifford group

➢ To reach a universal gate set, we need so-called «magic states», which are then injected via teleportation using Cliffords and measurements

➢ General gates are typically decomposed into Cliffords and T-gates (magic)



So now we can do any quantum computation?

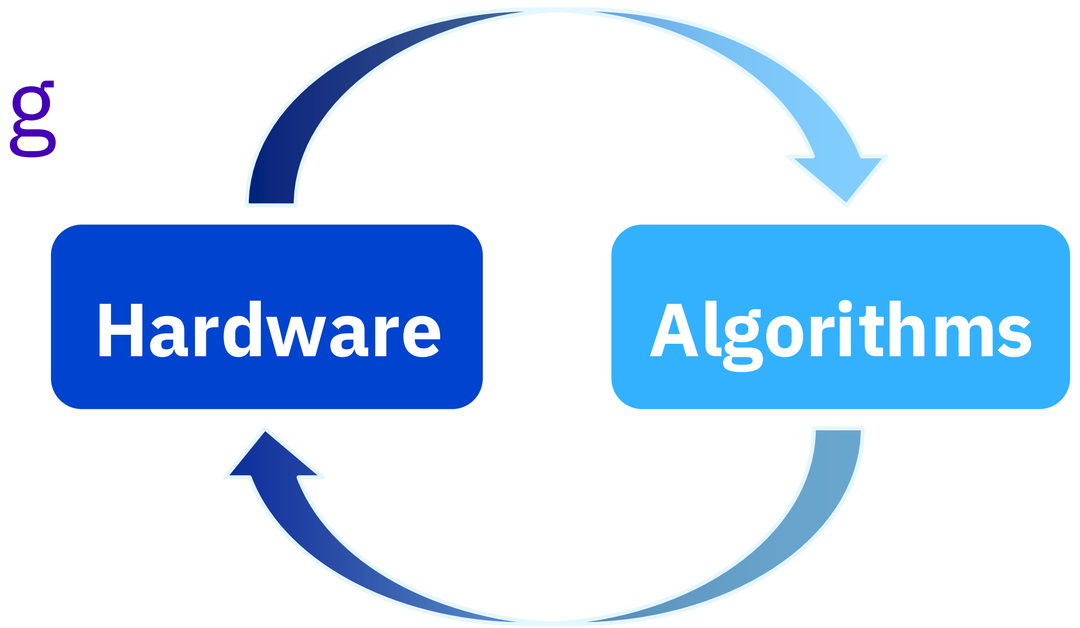
Resource estimation on the Gross code

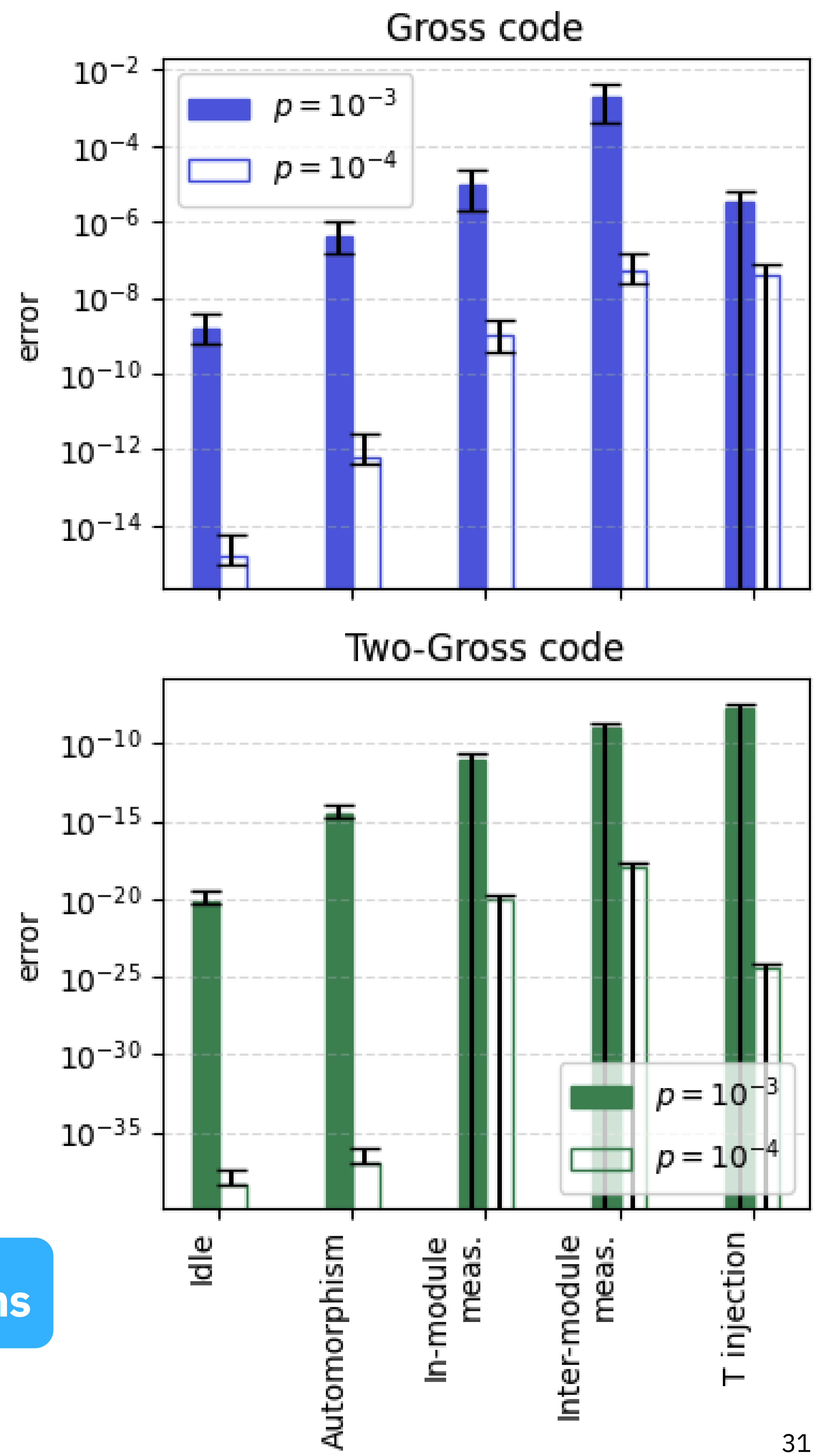
- Even on fault-tolerant devices we will have a **fixed error budget** and thus need to optimize our resources
- Costs for different operations drastically change in the fault-tolerant regime, e.g. an arbitrary phase gate with precision $\varepsilon = 10^{-12}$ requires ~200 T-gates
  → Phase gates that used to be "for free" are now very expensive
- **T-injection and inter-module measurements dominate the costs**
- Layout and transpilation matter a lot, e.g. teleportation of gates vs qubits

➢ **Goals:**

i. Resource estimates to check feasibility and identify bottlenecks

ii. Re-design and optimize quantum algorithms according to the resource costs in the fault-tolerant regime





Distillation factories

- **Magic state distillation:** Using many copies of a noisy magic state and (fault-tolerant) Clifford operations plus measurements we can distill a few higher-fidelity magic states
- **Entanglement distillation:** Using many noisy inter-module gates we can distill a few higher-fidelity links

➢ We can trade **time** and **number of qubits** for **precision**

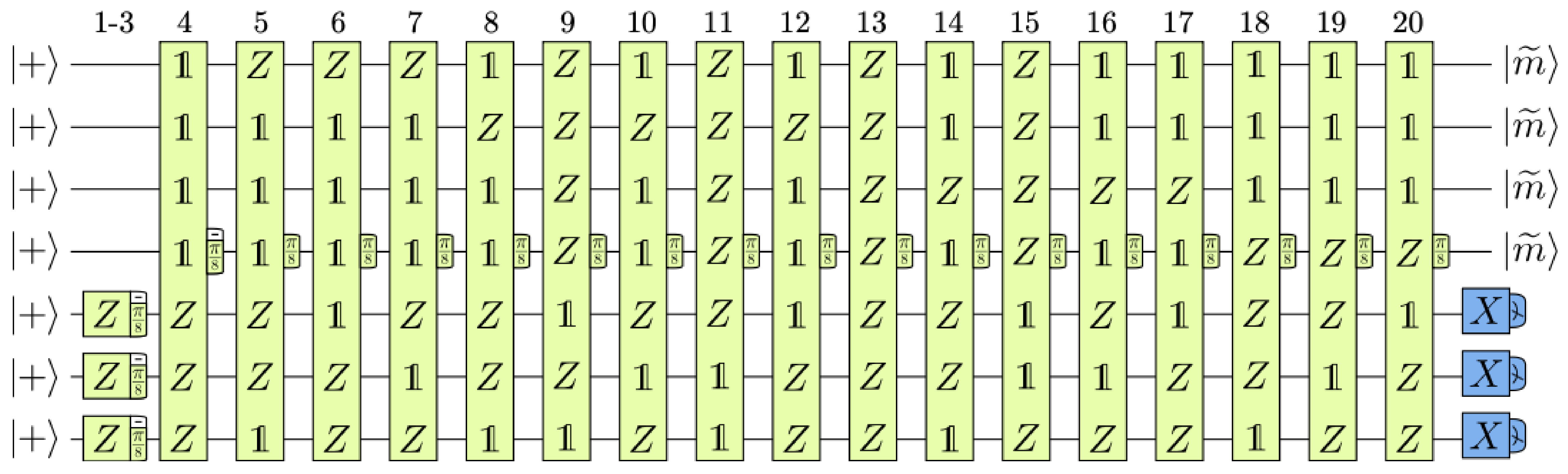


Figure 4: 20-to-4 distillation circuit.

D. Litinski, Magic State Distillation: Not as Costly as You Think, Quantum 3, 205 (2019)

➢ **Goals:**

i. Find suitable distillation protocols for our architecture

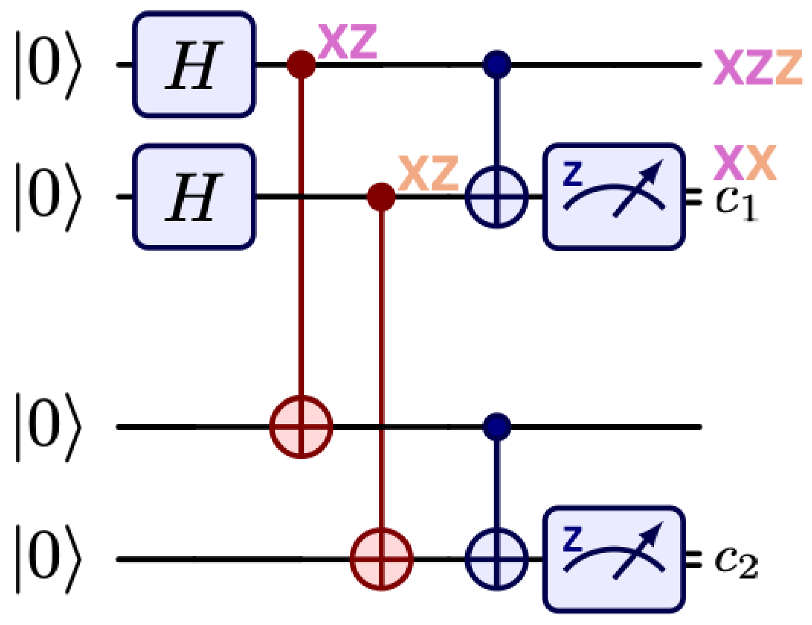
ii. Find a good balance of distillation depending on the algorithm

Standard entanglement distillation protocols

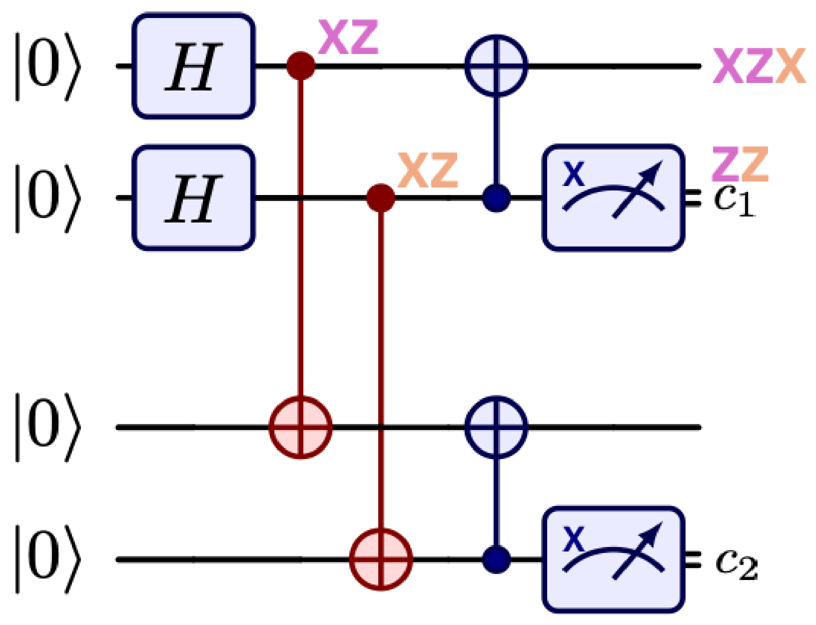
1) Generate several noisy Bell pairs.
2) Use only local operations and classical communication.
3) Postselect to obtain fewer, higher-fidelity Bell pairs.

The protocols accept if
$c_1 = c_2 \wedge c_3 = c_4 \wedge c_5 = c_6$

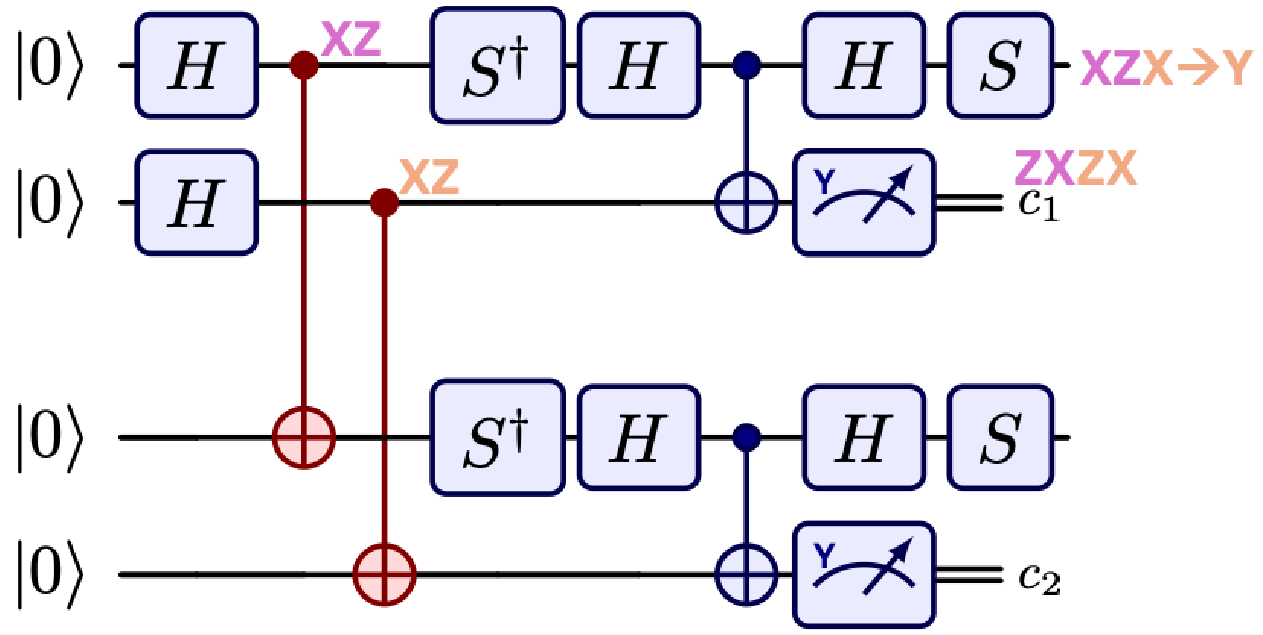
Z-recurrence:



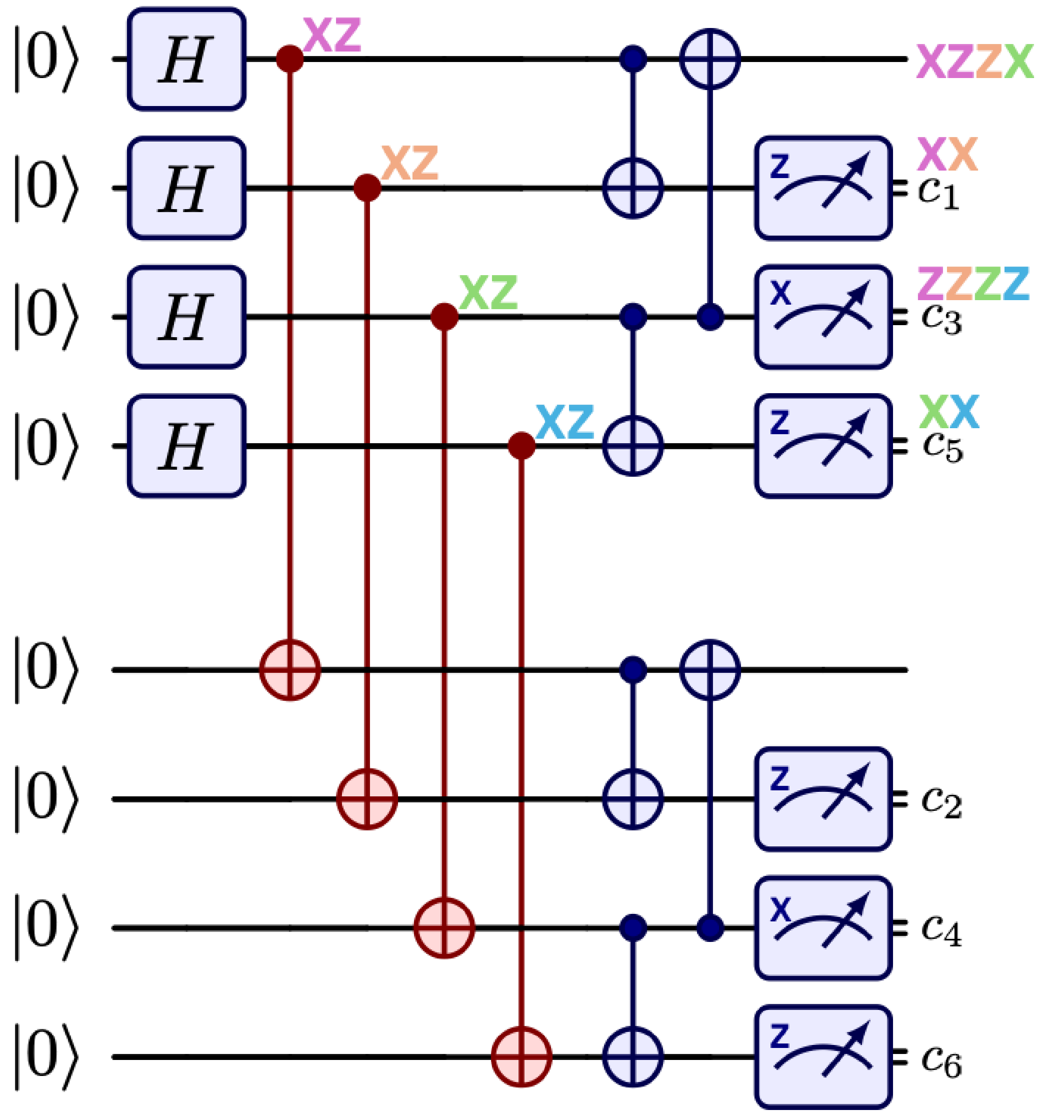
X-recurrence:



Y-recurrence:



Two-round recurrence



➢ Single recurrence cannot detect an arbitrary error

➢ Two-round recurrence can detect an arbitrary error

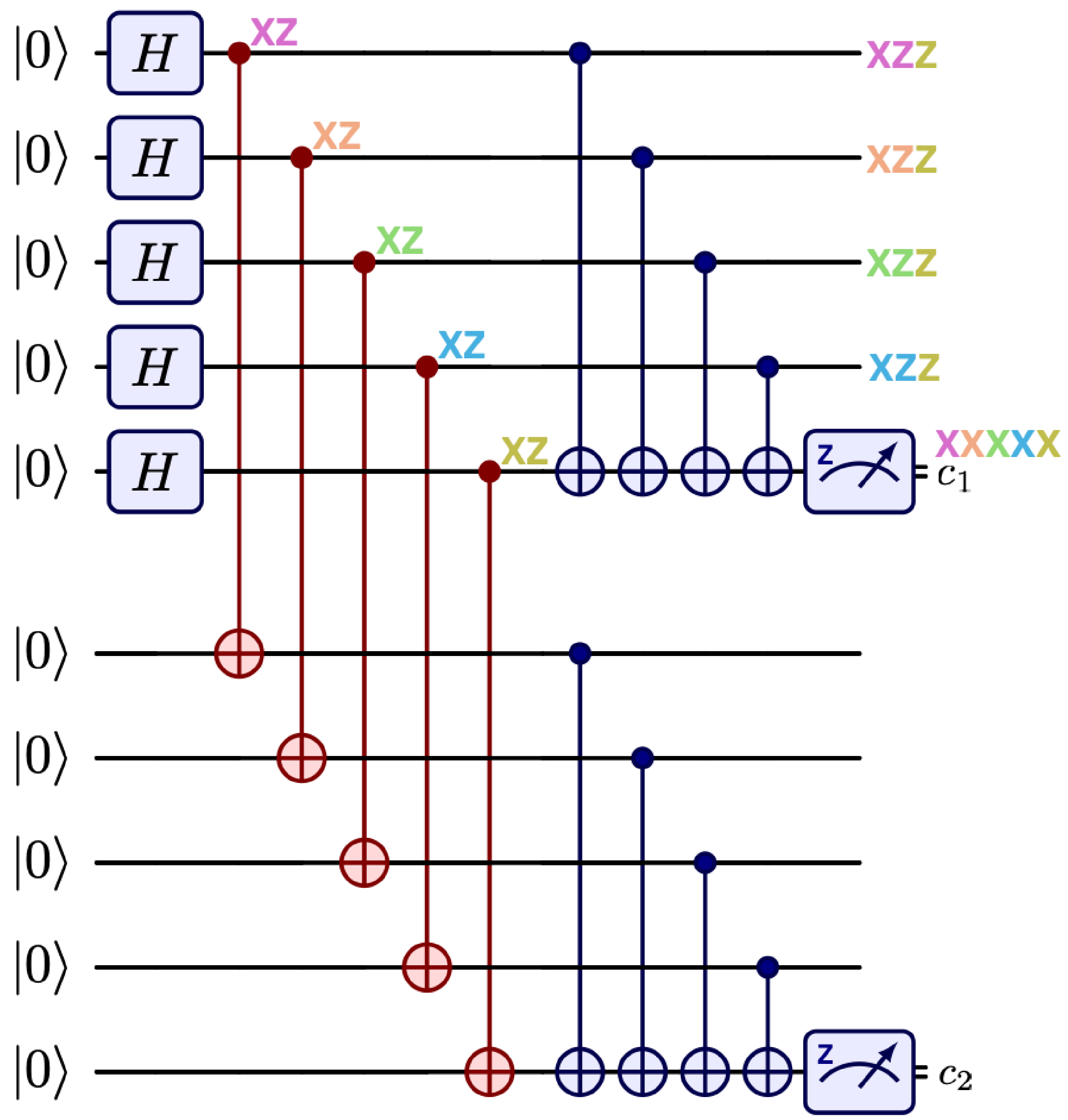
Standard entanglement distillation protocols

1) Generate several noisy Bell pairs.
2) Use only local operations and classical communication.
3) Postselect to obtain fewer, higher-fidelity Bell pairs.

The protocols accept if
$c_1 = c_2 \wedge c_3 = c_4 \wedge c_5 = c_6$

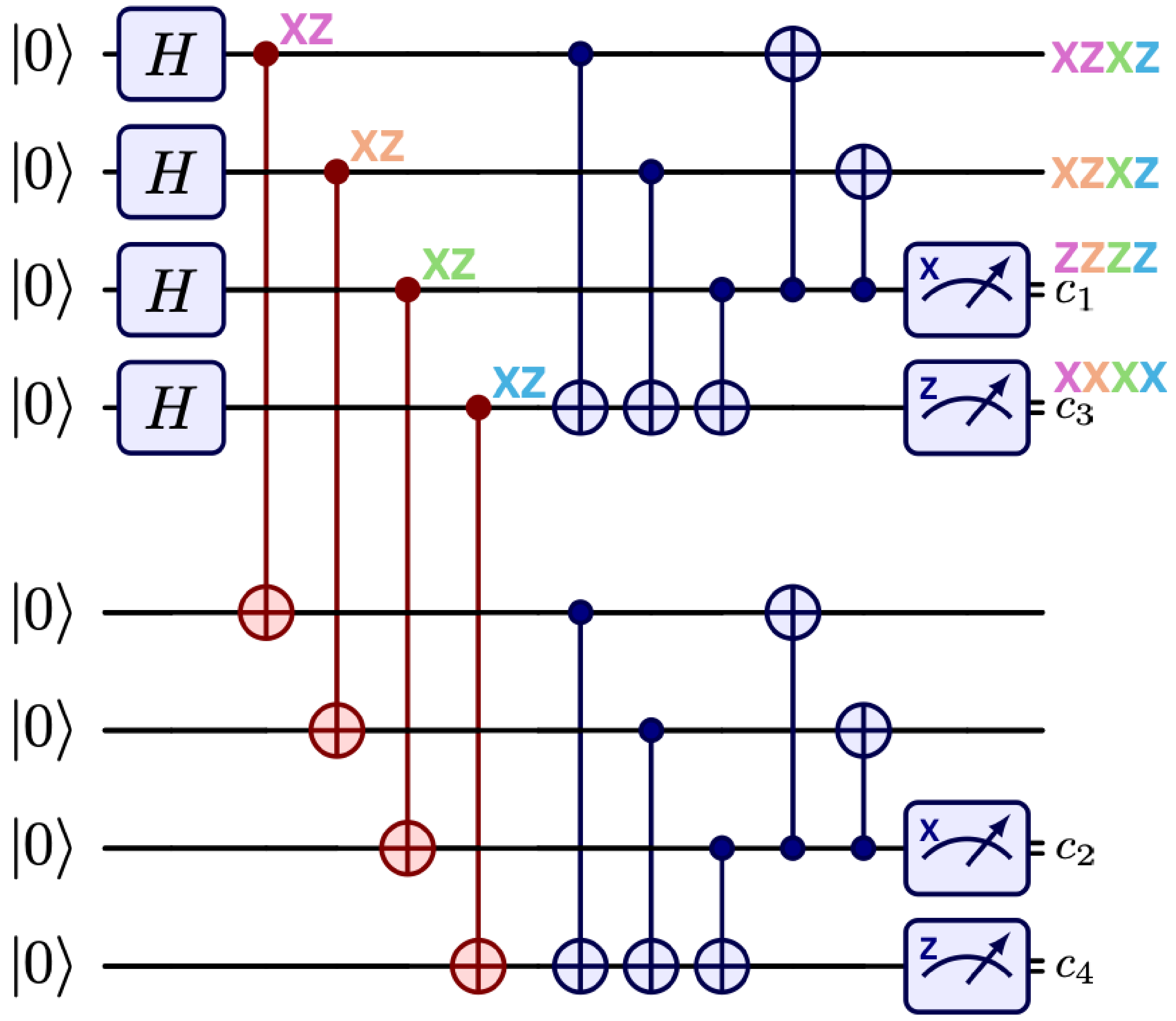
➢ Higher output protocols:

Maneva-Smolin



➢ Cannot detect an arbitrary error

Leung-Shor



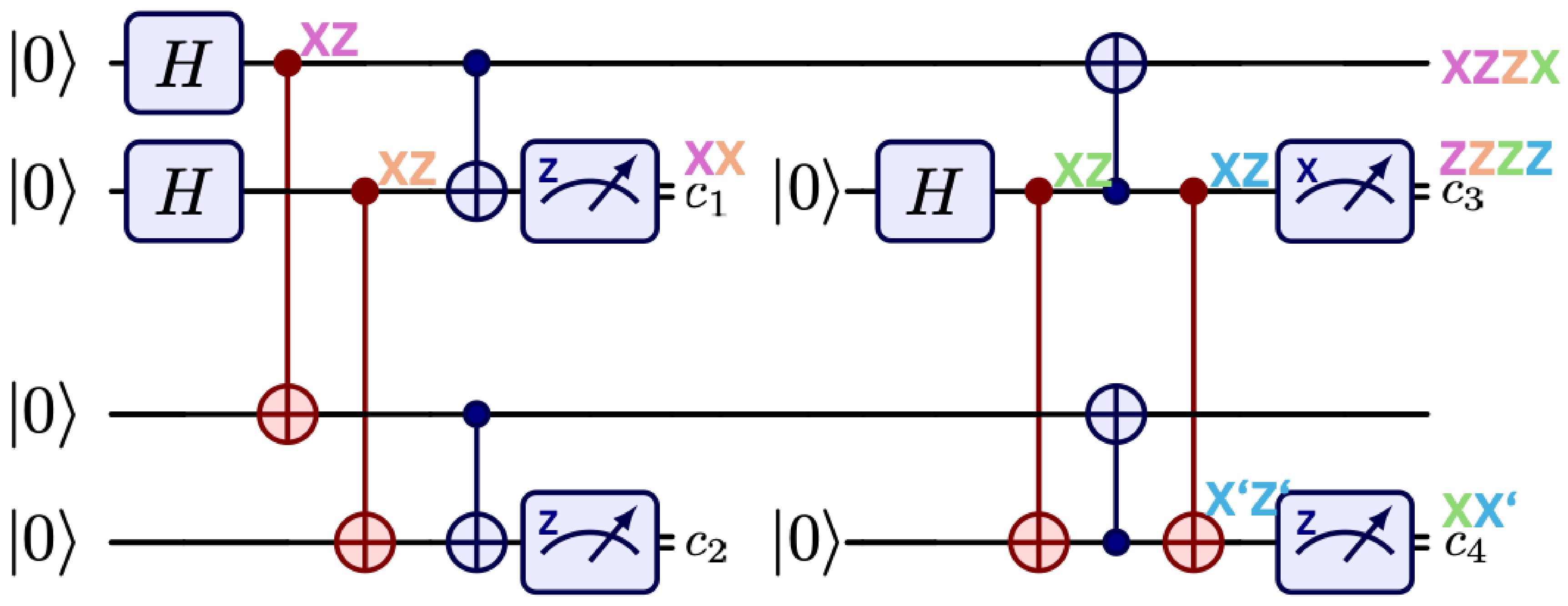
➢ Detects arbitrary error and outputs two Bell pairs

arXiv:2605.26757

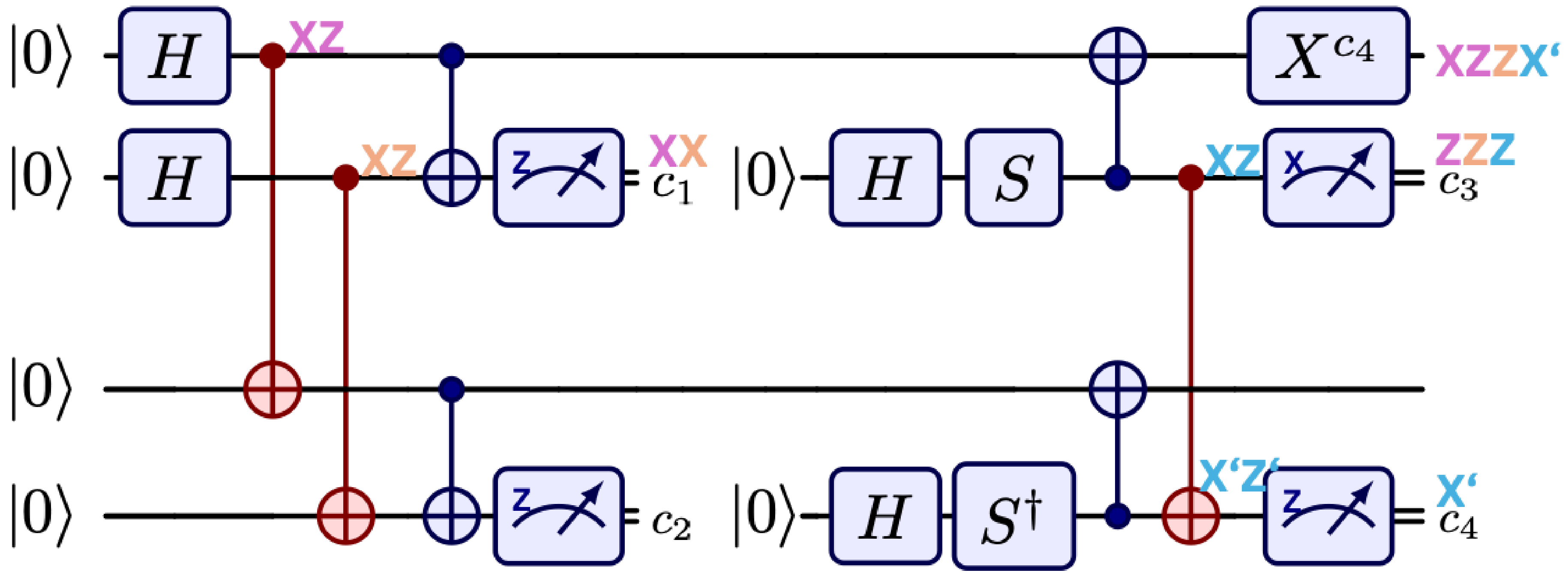
Our new protocols

Main idea: Use the noisy inter-module entangling operation also inside the distillation circuit.

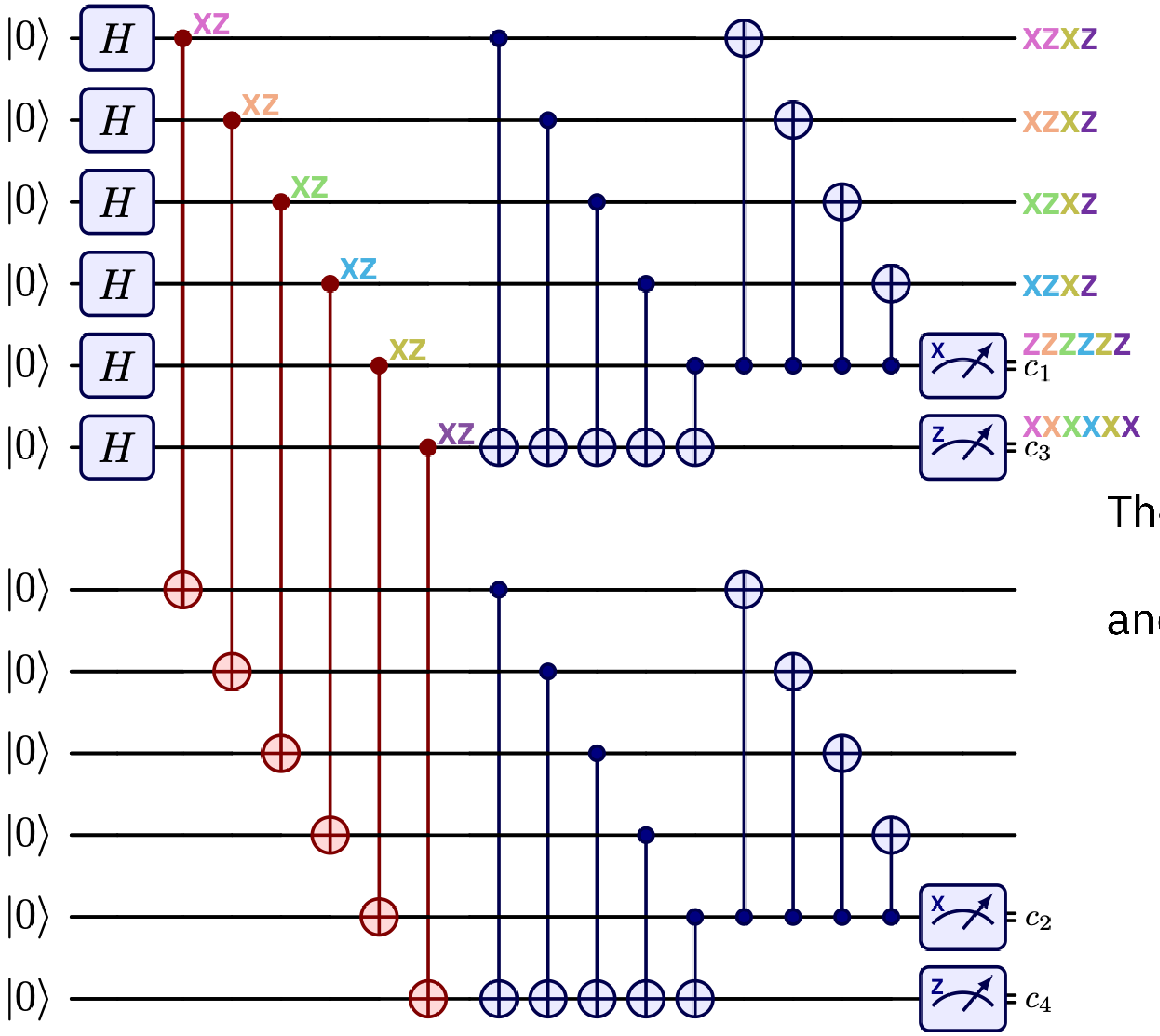
Space-optimal



Space-optimal reduced



High-rate



The protocols accept if

$$c_1 = c_2 \wedge c_3 = c_4\,,$$

and in (a) additionally requires

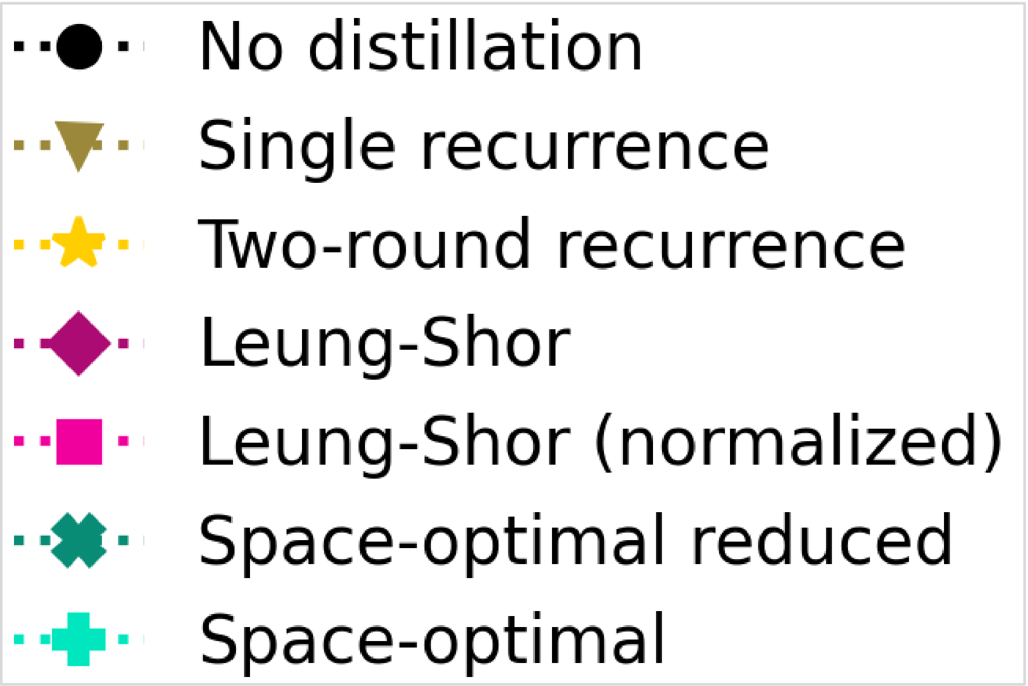
$$c_3 = c_4 = 0$$

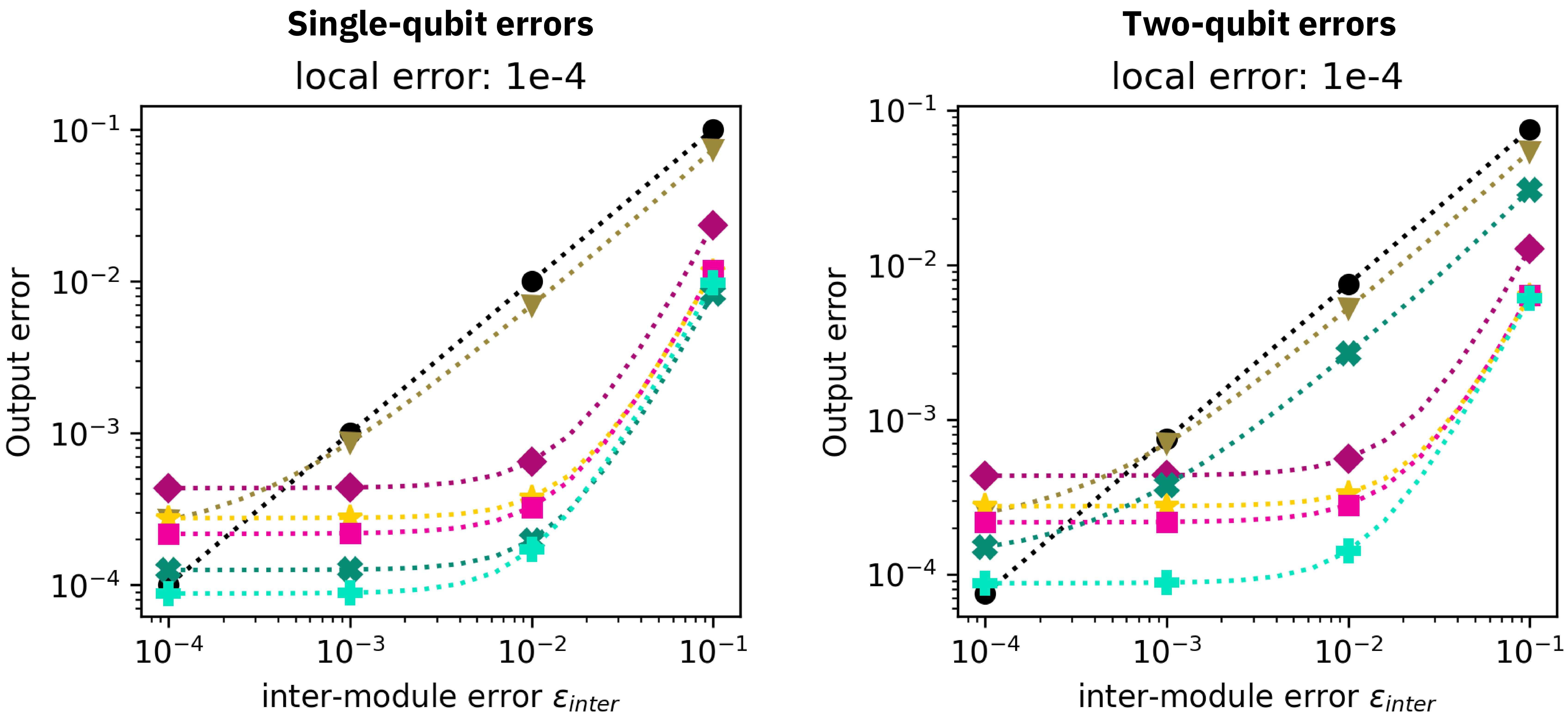
Comparison

Protocol	# qubits per module	Depth 1-D	Depth all-to-all	Duration [ns]	# output Bell pairs	# inter-module operations	Output error rate
Single recurrence [Fig. 1(a)] [6]	2	5 (+1)	2 (+1)	928 (+1320)	1	2	$\frac{2}{3}(\varepsilon_1 + \frac{2}{3}\varepsilon_2)$
Two-round recurrence [Fig. 1(d)] [15]	3	10 (+2)	5 (+2)	3296 (+1288)	1	4	$\frac{8}{9}(\varepsilon_1 + \frac{2}{3}\varepsilon_2)^2$
	4	14 (+1)	3 (+1)	2688 (+1288)			
Maneva-Smolin [Fig. 1(e)] [17]	m	$10m - 18$	m (+1)	see Fig. 4(c)	$m - 1$	m	$\frac{m}{3}(\varepsilon_1 + \frac{2}{3}\varepsilon_2)$
Leung-Shor [Fig. 1(f)] [18]	4	19 (+2)	4 (+2)	3240 (+1440)	2	4	$\frac{14}{9}(\varepsilon_1 + \frac{2}{3}\varepsilon_2)^2$
Space-optimal [Fig. 2(a)] [ours]	**2**	8 (+2)	5 (+2)	2608 (+1376)	1	4	$\frac{7}{9}\varepsilon_1^2 + \frac{4}{9}\varepsilon_2^2 + \frac{32}{27}\varepsilon_1\varepsilon_2$
	3	11 (+1)	3 (+2)	1856 (+1320)			
Space-optimal reduced [Fig. 2(b)] [ours]	2	7 (+2)	4 (+2)	2520 (+1376)	1	3	$\frac{2}{3}\varepsilon_1^2 + \frac{4}{9}\varepsilon_2$
	3	6 (+4)	3 (+2)	688 (+1648)			
High-rate [Fig. 2(c)] [ours]	m	$8m-13$ (+2)	m (+2)	see Fig. 4(c)	$m - 2$	m	$(\frac{1}{6}m^2 - \frac{7}{18}m + \frac{4}{9})(\varepsilon_1 + \frac{2}{3}\varepsilon_2)^2$

Output error rates are conditioned on successful postselection and reported to leading order in the inter-module error probabilities. For multi-output protocols, the reported error is the joint error. Linear error suppression is colored red due to its inefficiency.

Simulations

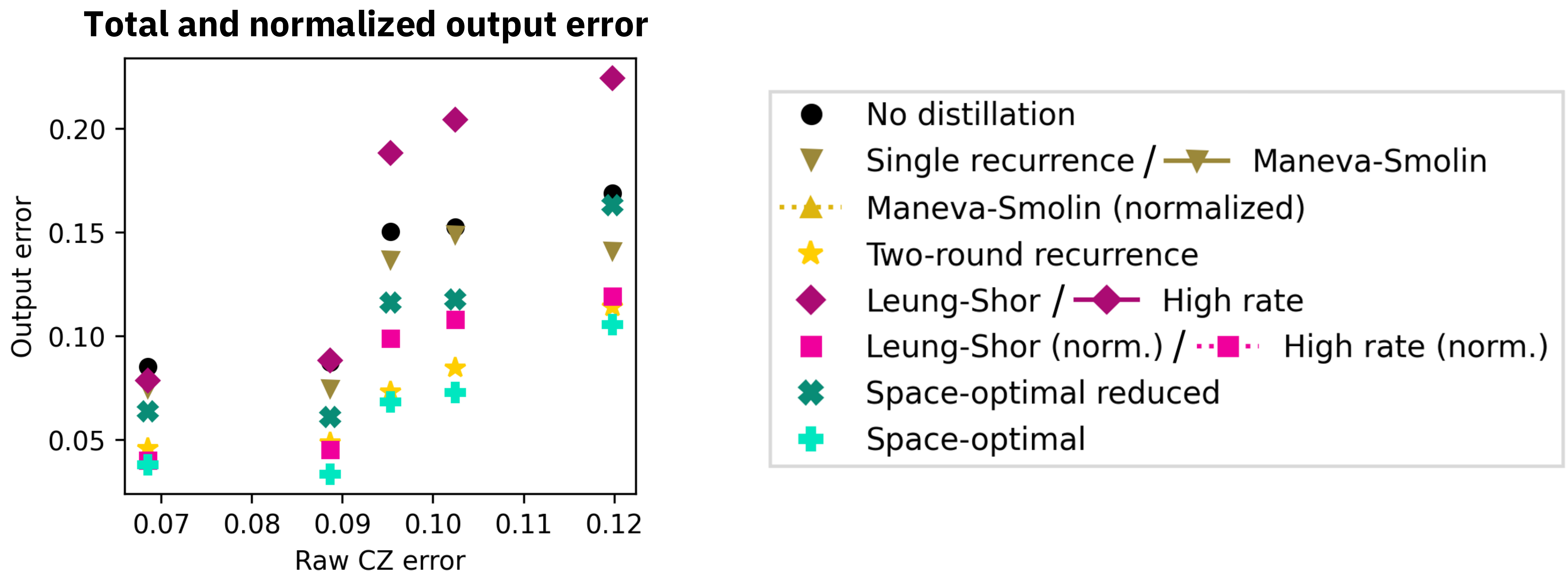


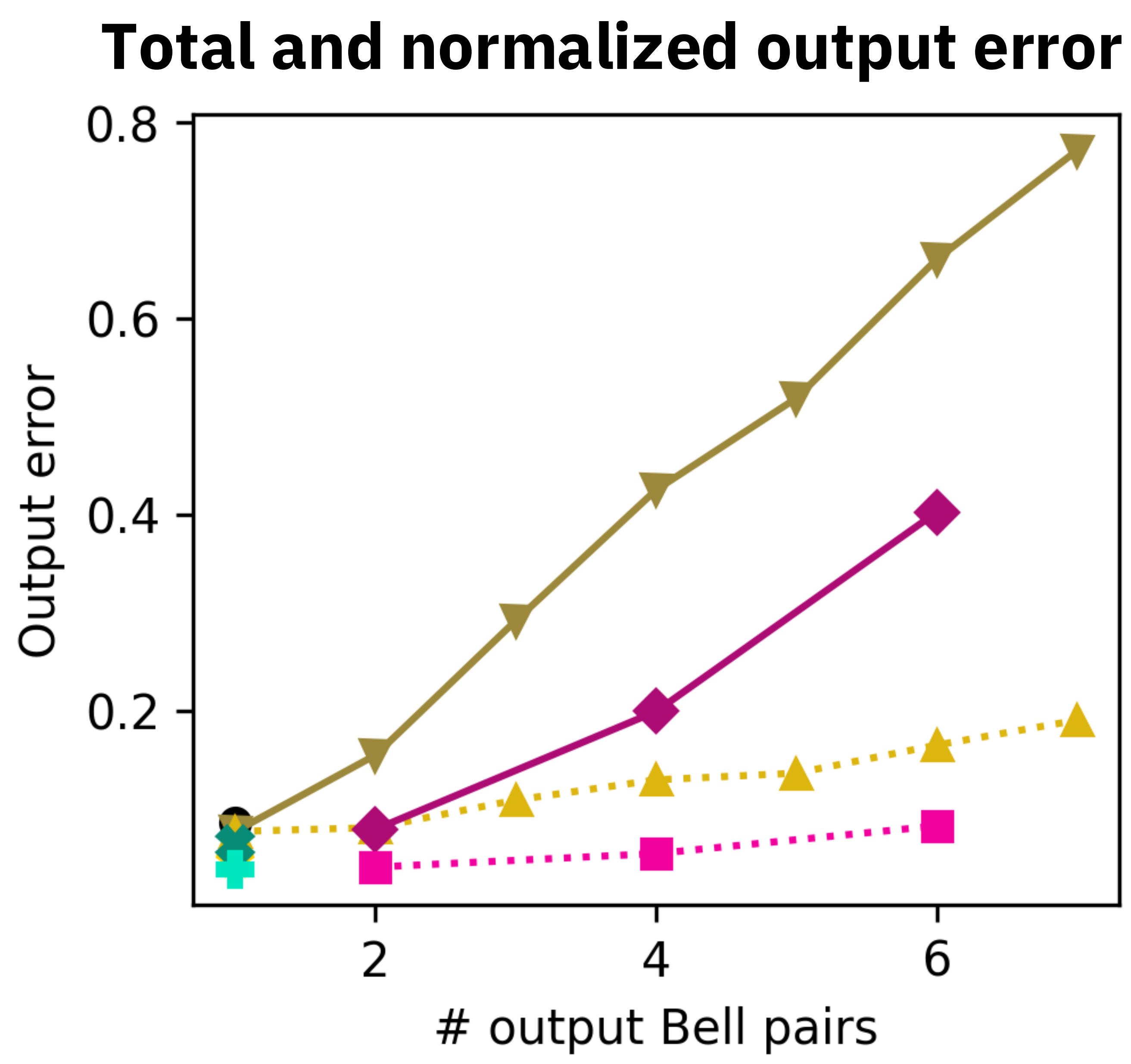


- ➤ Clear difference between the linearly scaling protocols and those with quadratic error suppression.
- ➤ Reduced protocol shows expected quadratic suppression in ε_1, but only linear suppression in ε_2, which emphasizes its limited utility that is confined to single-qubit errors.
- ➤ All other protocols show a fairly invariant behavior under the two different noise models.
- ➤ Output errors are lower bounded by the local errors.

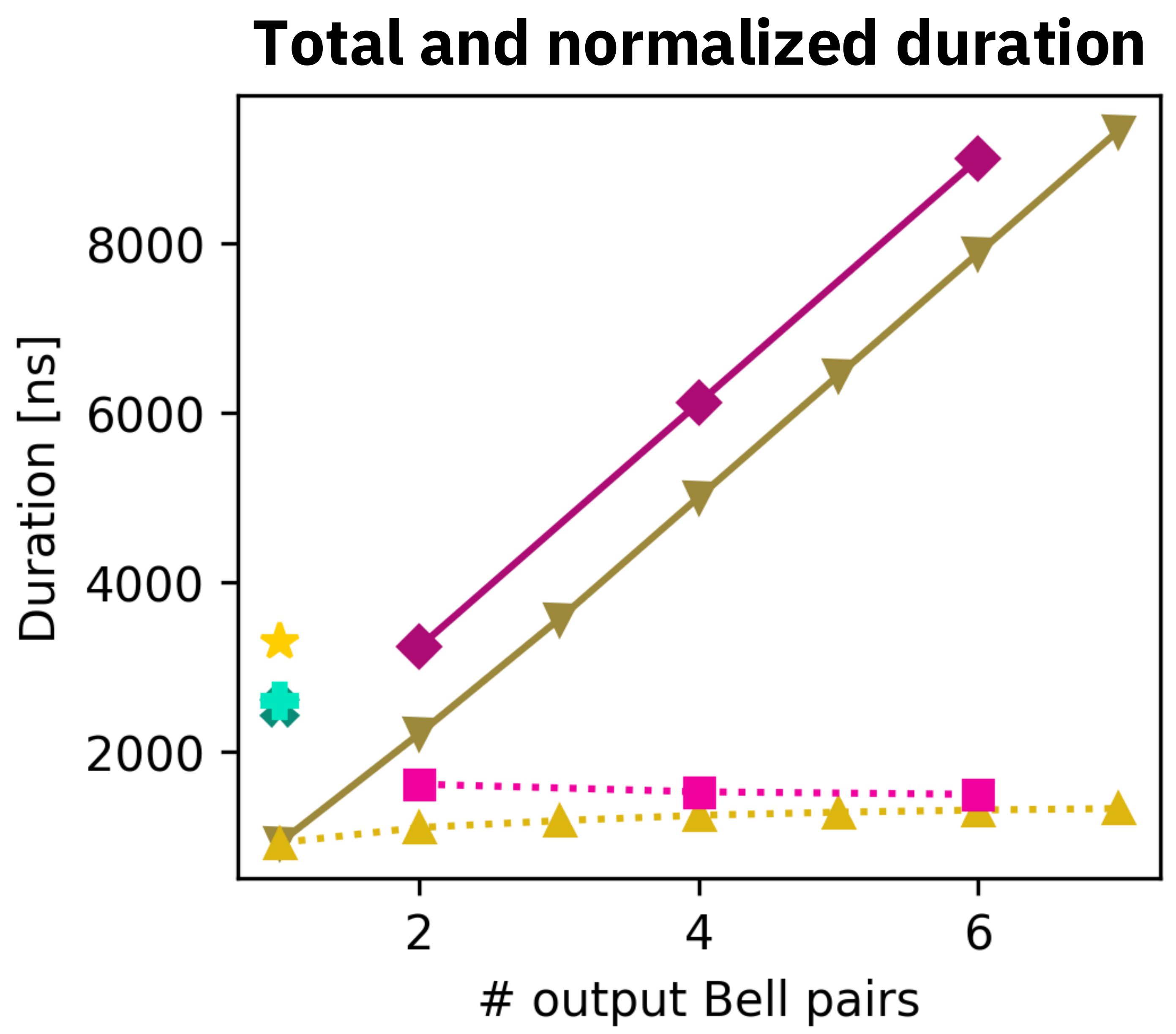
- ➤ **Across all simulations our general space-optimal protocol shows the best performance**

Experiments on near-term devices (ibm_pittsburgh)



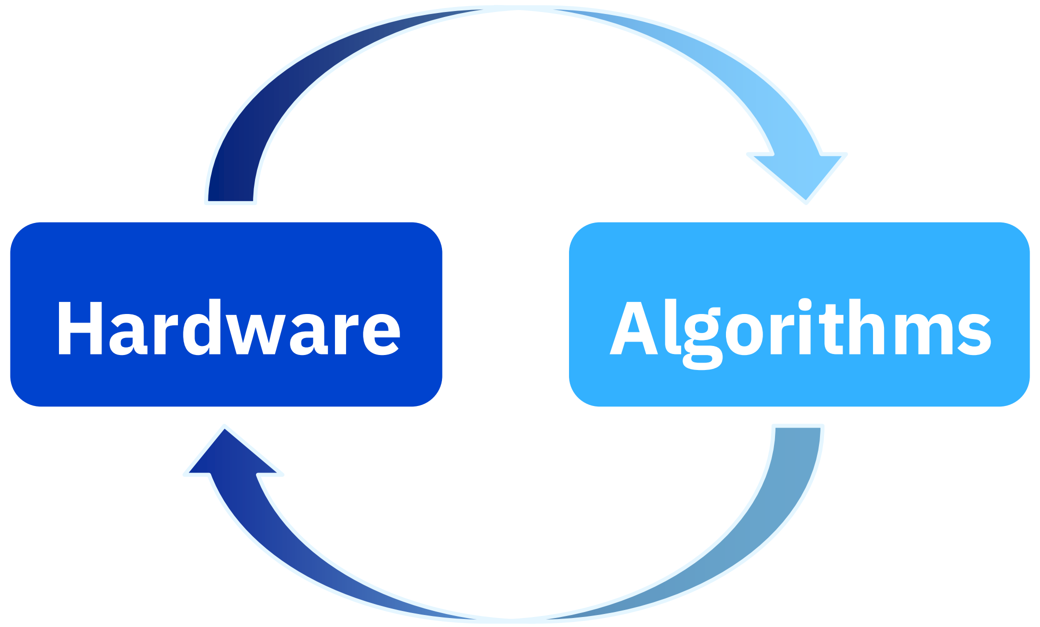


- Output error rate lower bounded by the local operations, which is here only roughly one order of magnitude better.
- Single recurrence gives only very small improvement over no distillation, but second round improves the error substantially (as expected) and achieves similar performance as normalized Leung-Shor protocol.
- High-rate protocol clearly better fidelity than Maneva-Smolin protocol.
- **The best-performing protocol is our space-optimal protocol, as it consistently yields the lowest output error and achieves the shortest overall runtime among the protocols with quadratic error suppression.**



Conclusion & Outlook

- Reusing the noisy inter-module entangling operation during distillation can reduce both space and time overhead
- ➢ Our space-optimal protocol is a practical primitive for overcoming noisy links in modular quantum computing, as it requires only two qubits per module to reach quadratic suppression of inter-module errors and also achieves the best performance in our simulations and experiments
- Potential next steps:
  - Compare distillation on physical level vs logical level
  - Look for further suitable distillation and synthillation protocols
- IBM's mission to bring useful quantum computing to the world requires that hardware, software and error correction advance together!
- ➢ **We need to re-evaluate resources for quantum algorithms in the fault-tolerant regime and optimize them accordingly**



Thank you for your attention!