Research Spotlight Forum

6.4.19 Quantum Information Sciences

Architecture and Runtime for Reliable Quantum Computers

PRESENTED BY:

Moinuddin Qureshi,

Professor, ECE, Georgia Institute of Technology

(moin@gatech.edu)

Sandia National Laboratories | Georgia Institute of Technology | ILLINOIS | PURDUE UNIVERSITY | THE UNIVERSITY OF TEXAS AT AUSTIN | THE UNIVERSITY OF NEW MEXICO | NM STATE UNIVERSITY | NEW MEXICO TECH SCIENCE • ENGINEERING • RESEARCH UNIVERSITY

SAND2018-14303 O

ABOUT YOURSELF

Brief Bio: Moinuddin Qureshi is a Professor of ECE at Georgia Tech. Previously, he was a research staff member (2007-2011) at IBM T.J. Watson Research Center, where he developed the caching algorithms for Power-7 processors. He is a member of the Hall of Fame for ISCA, MICRO, and HPCA. His research has been recognized with the best paper award at MICRO 2018, best paper award at CF 2019, and two awards (and three honorable mentions) at IEEE MICRO Top Picks. His ISCA 2009 paper on Phase Change Memory was awarded the 2019 Persistent Impact Prize in recognition of "exceptional impact on the fields of study related to non-volatile memories". He was the Program Chair of MICRO 2015 and Selection Committee Co-Chair of Top Picks 2017. He received his Ph.D. (2007) and M.S. (2003) from the University of Texas at Austin.

Research Areas: Computer architecture, specifically:

(1) MEMORY SYSTEMS: 3D memories, non-volatile memories, hybrid memories

(2) QUANTUM COMPUTING: error mitigation, control processor arch, CRYO

(3) HARDWARE SECURITY: cache side channel, secure memory, speculation attacks

Research Group: "Memory Systems Lab" of 5 PhD students

Keywords: Quantum computing, error mitigation, variation-aware qubit mapping, CryoDRAM, Control-processor architecture, cache side channels, integrity trees, mitigating speculation based attacks, stacked DRAM, NVM, Alloy cache, CEASER

Homepage:

http://moin.ece.gatech.edu/

Group:

http://memlab.ece.gatech.edu/

Google-Scholar:

https://tinyurl.com/qureshi-GS

CURRENT WORK IN QUANTUM COMPUTING

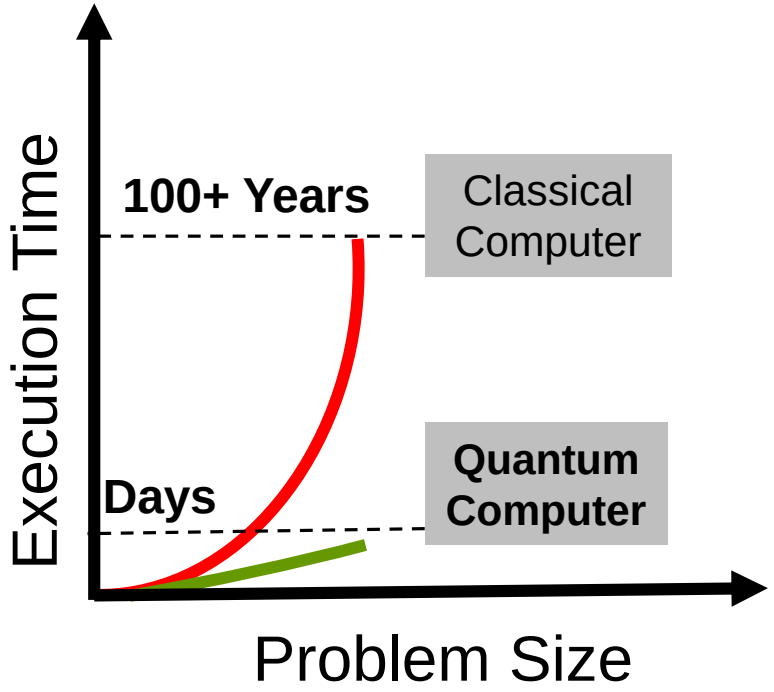
OUTLINE

1. **Variability-Aware Qubit Mapping for NISQ Computers [ASPLOS 2019]**
2. **Control Processor Architecture for Fault-Tolerant Quantum Computers [MICRO 2017]**
3. **Feasibility of Cryogenic DRAM for Quantum Computing Systems [MEMSYS 2017]**

[Tannu and Qureshi]
http://memlab.ece.gatech.edu/papers/ASPLOS_2019_1.pdf

Quantum Computers are Here!

Quantum computers can speedup hard problems



Recent demonstrations

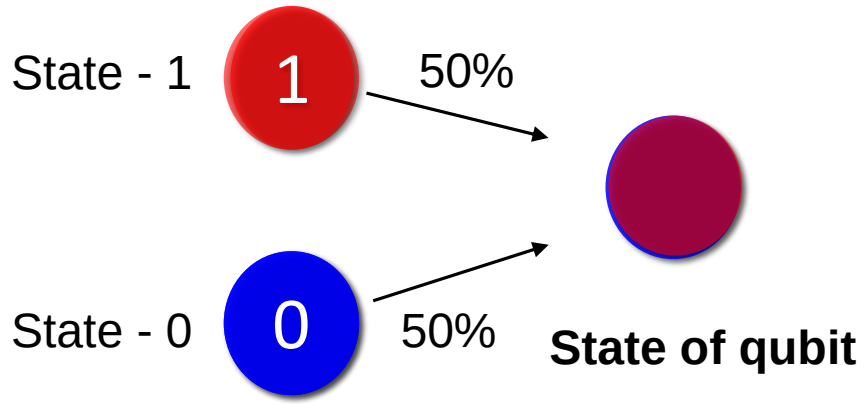
Quantum Machine	Number of Qubits*
Google	72
IBM	50
Intel	49
Rigetti	19
IonQ	11

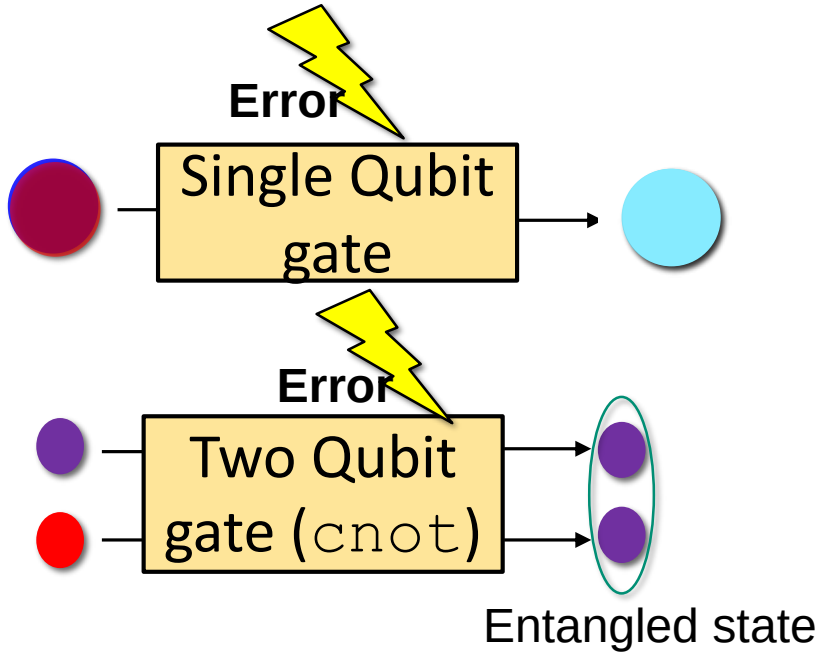
* Under test , fabricated, or announced

QC with 10+ qubits are here, QC with 100+ qubits expected soon

Quantum Computing: Background

QC operates on principles of entanglement and superposition





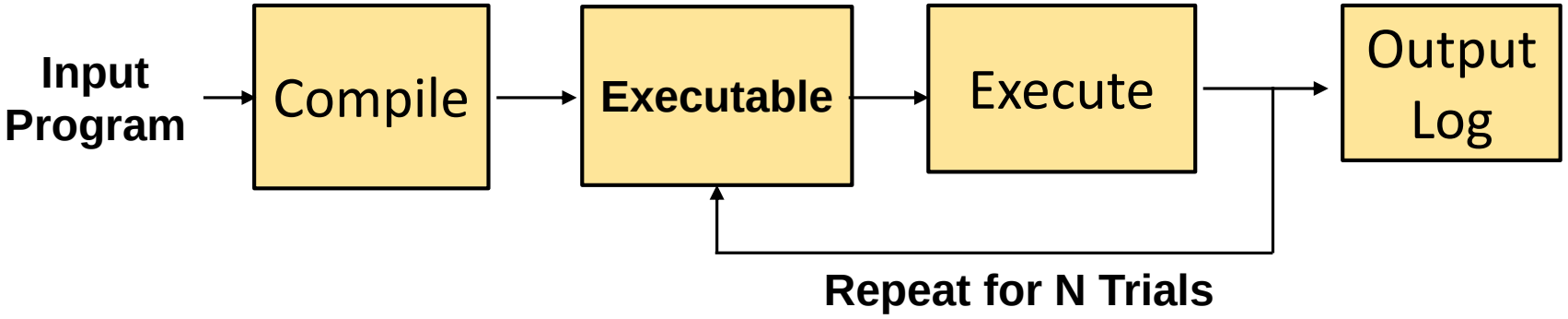
State of qubit is a superposition of state “0” and state “1”

Gate operations modulate state of the qubit

NISQ Programming Model

- Quantum Error Correction is expensive (20x-50x qubits)

- Noisy Intermediate Scale Quantum Computer (NISQ) [Preskill'18]
  - Run program without any error correction



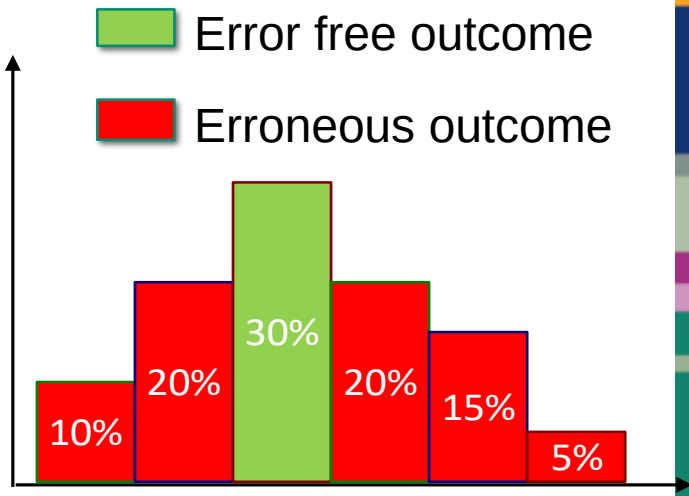


Figure of Merit: Probability of Successful Trial (PST)

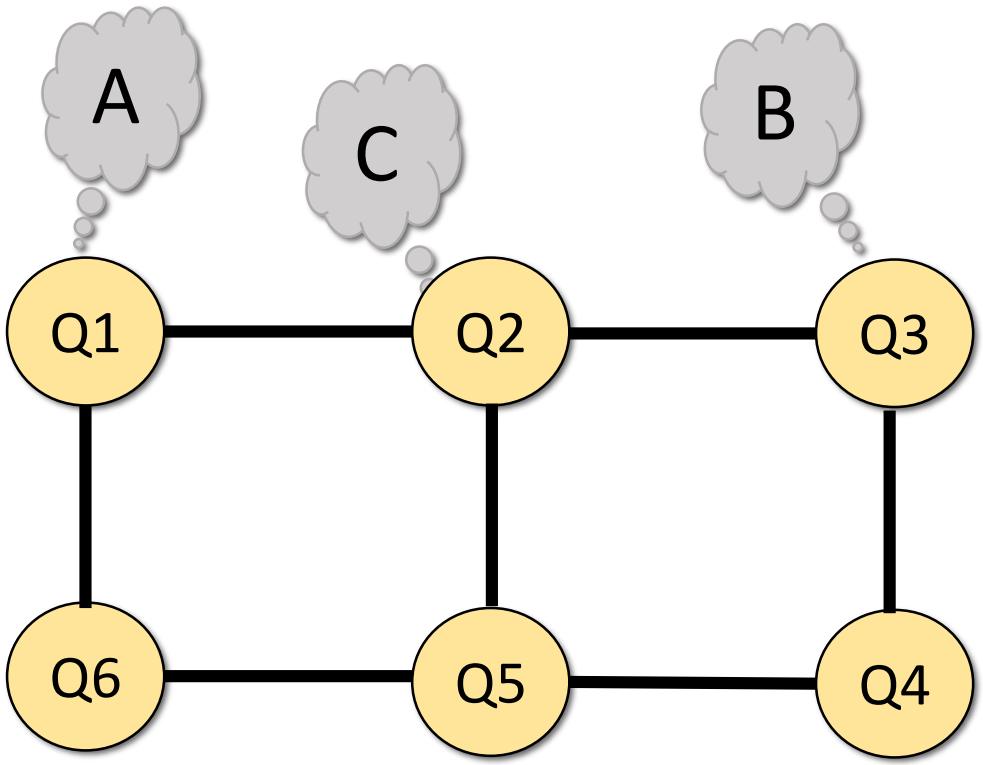
The Problem of Limited Connectivity

```
CNOT A,B
```

Not Possible no link between A and B

```
SWAP B,C
CNOT A,B
```

link between A and B, CNOT can be performed



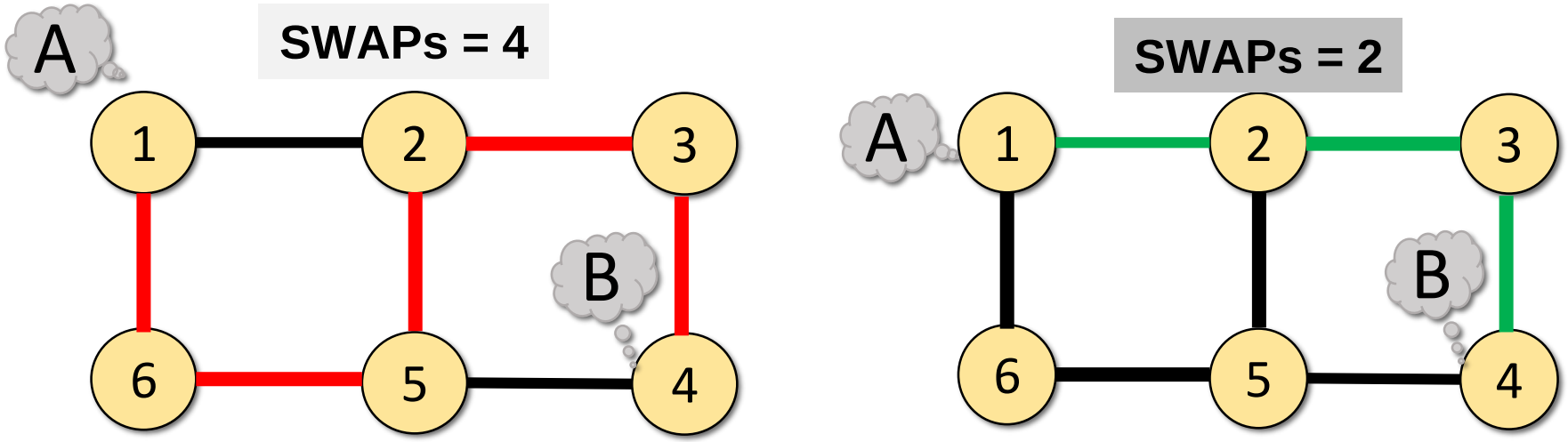
SWAP facilitate data movement

Compiler insert SWAPs → SWAPs are extra instructions (can also fail)

NISQ Compiler Policies

[1] Zulehner+, (DATE'18)
[2] Siraichi+, (CGO'18)
[3] Li+, (ASPLOS'19)

Compiler responsible for qubit allocation and movement

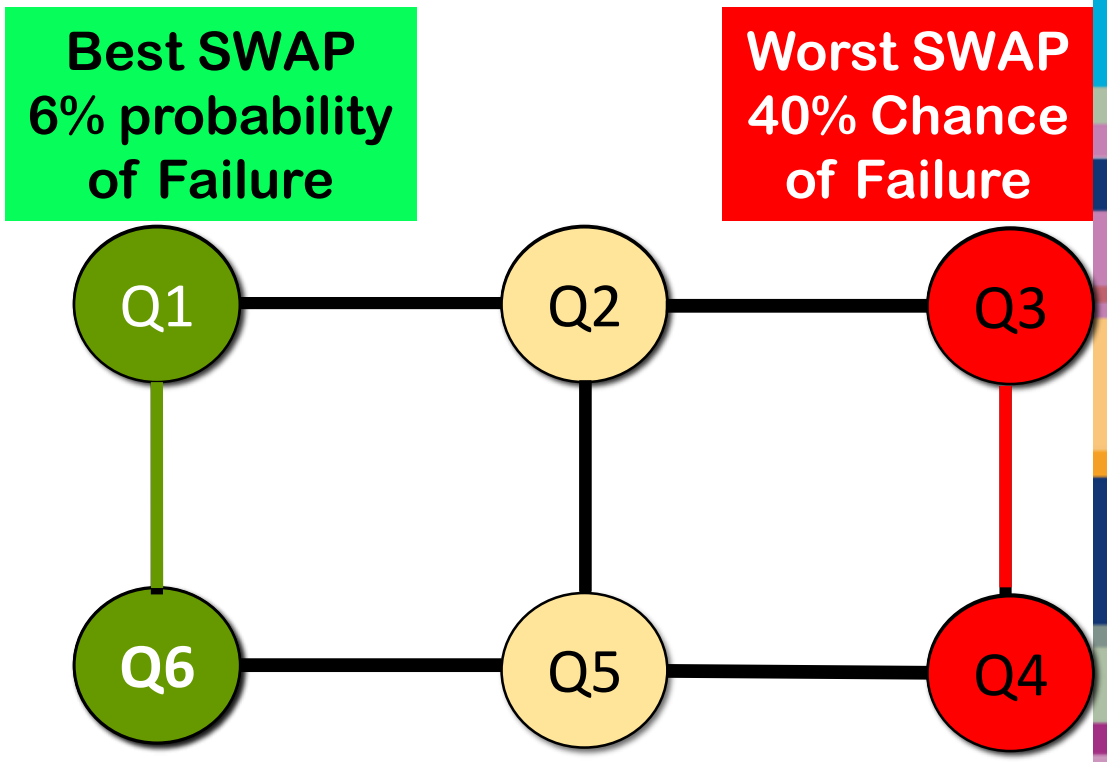


Qubit movement policy minimizing SWAPs

Existing compiler policies solely focus on minimizing SWAPs

Not All Qubits Are Created Equal!

- **Variability:** Some qubits and links fail with higher probability than others
- Avoiding certain links can improve reliability significantly



Goal: Exploit variation in error rates to improve reliability (assign more operations on reliable qubits/links)

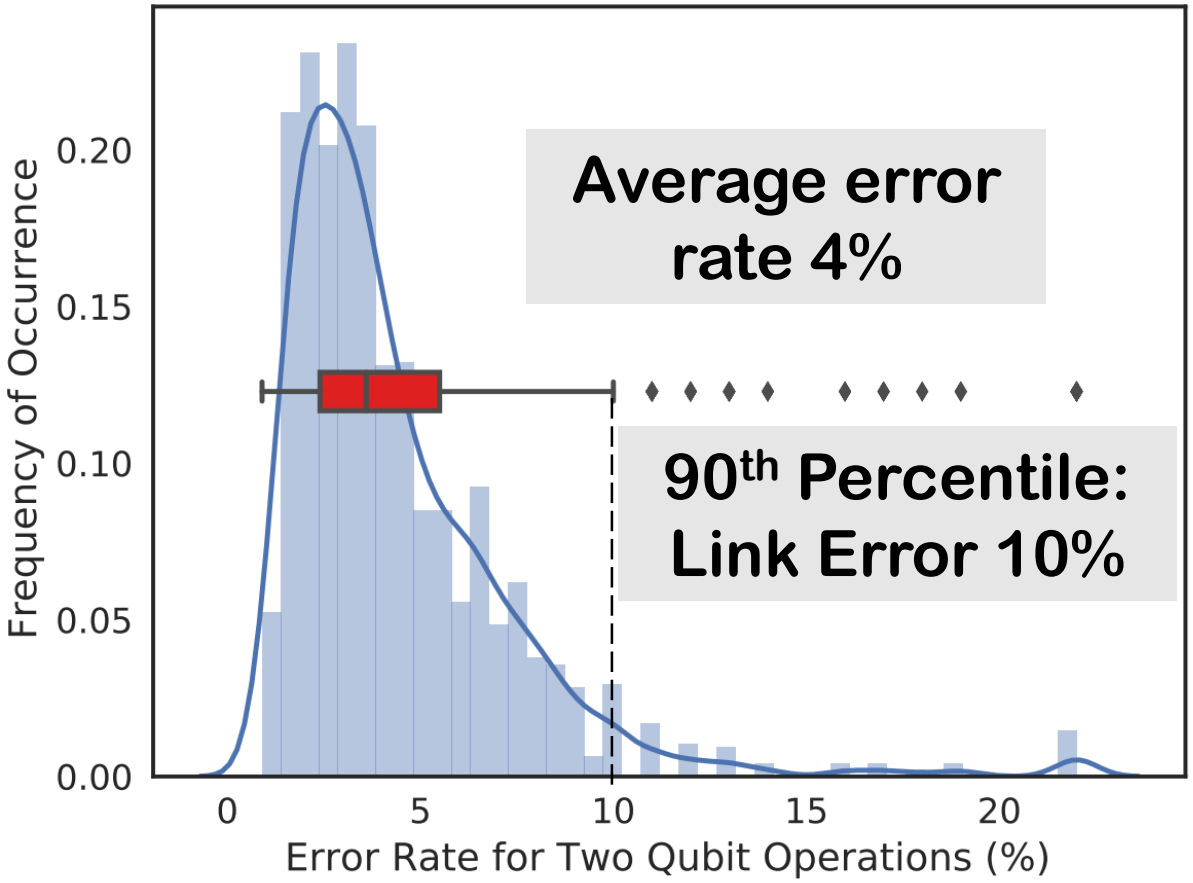
Characterization Methodology

- Gate operations require analog pulses
- Quantum computers require frequent calibration to get the best pulse
- IBM calibrates machines **multiple times a day and generate the report** contains error rates for each qubit and link
- We analyzed calibration 100+ reports over 52 days for IBMQ-20

IBM Q 20 Tokyo [ibmq_20_tokyo]

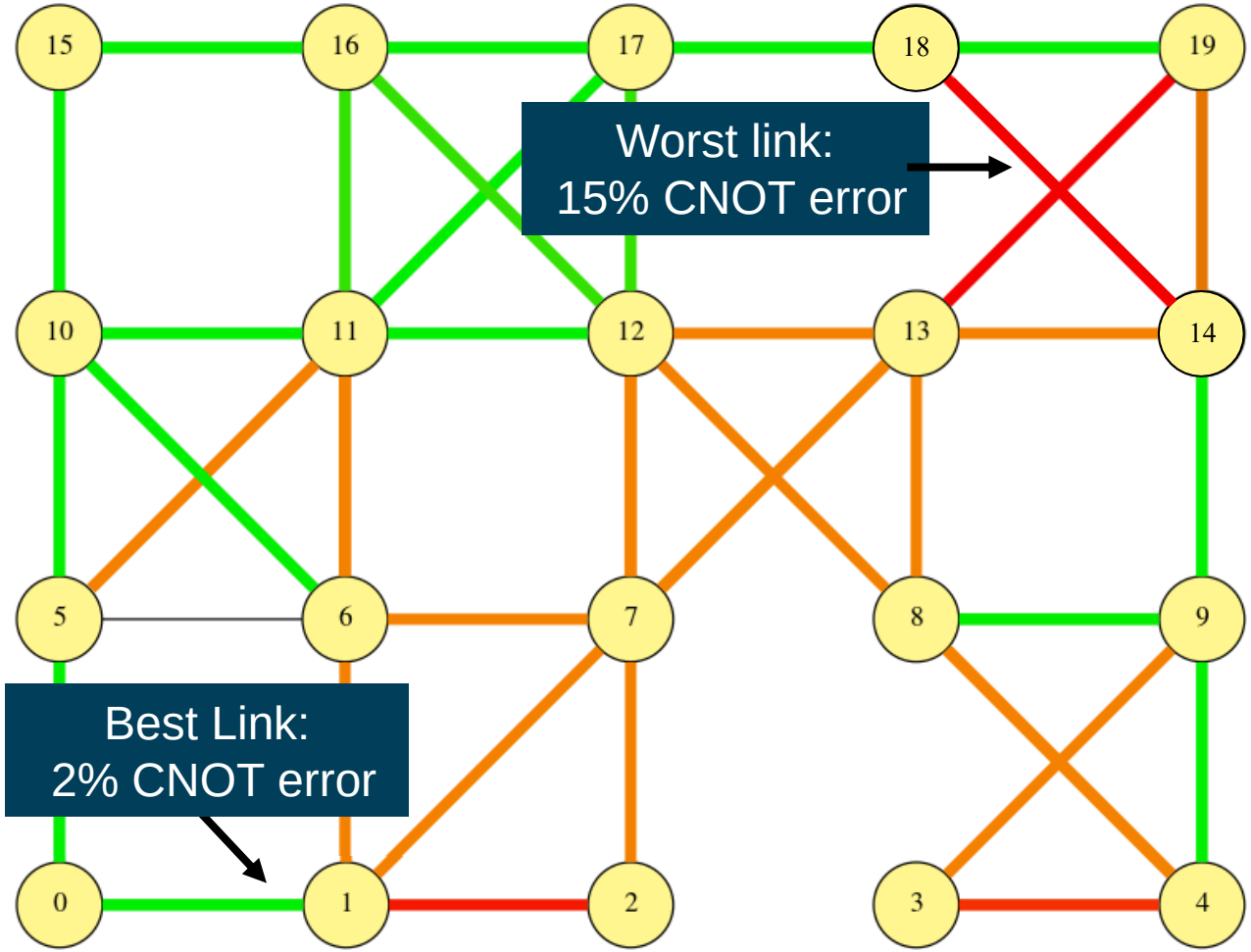
	Average
Frequency (GHz)	4.97
T1 (μs)	76.91
T2 (μs)	54.53
Gate error (10^{-3})	1.73
Readout error (10^{-2})	9.27
	AVG
MultiQubit gate error (10^{-2})	2.88

Two Qubit Gate Error Rate



Two qubit error rate is high and show significant variability

Spatial Variation in Two Qubit (Link) Error Rates



Link Error Rate > 7%

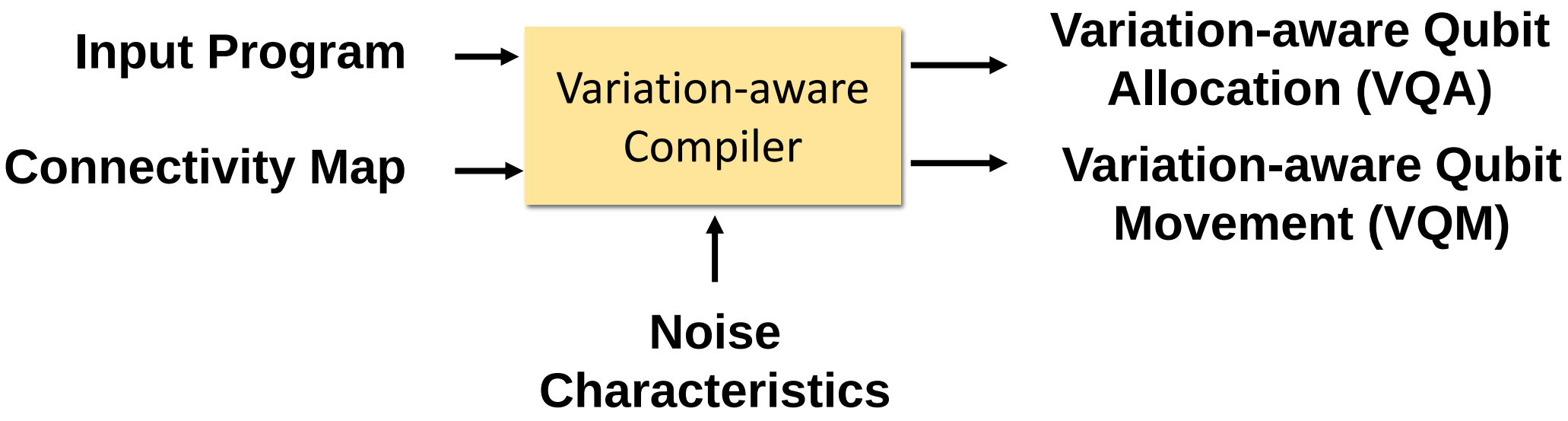
3% < Link Error Rate < 7%

Link Error Rate < 3%

Average link error rate for 76 links in IBM-Q20 machine

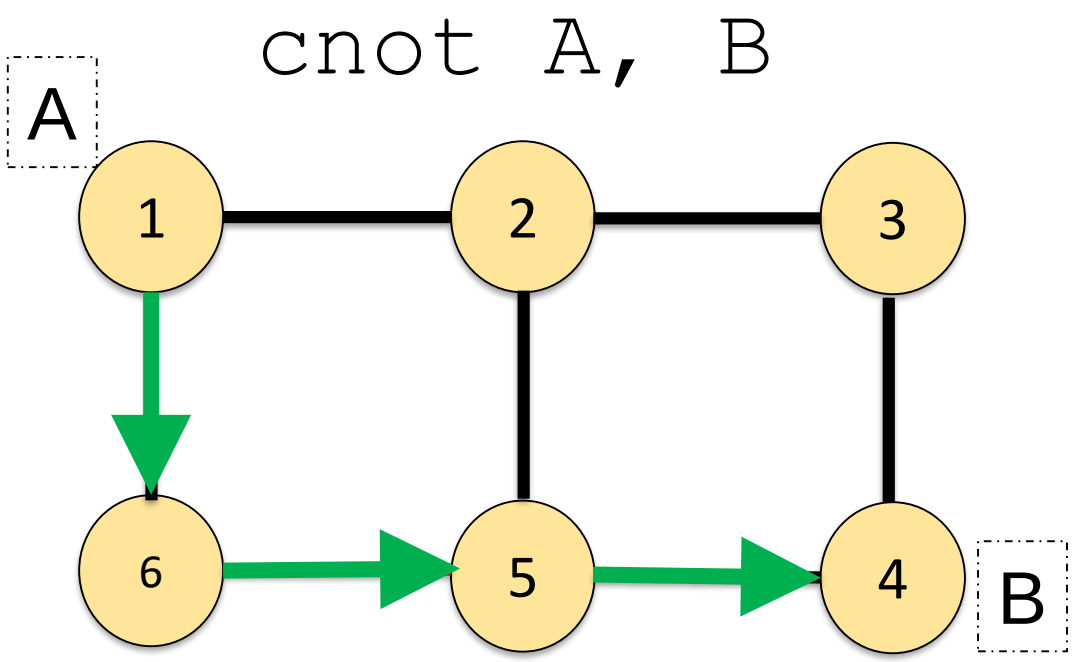
Some links are consistently more error prone than others

Variation-Aware Policy

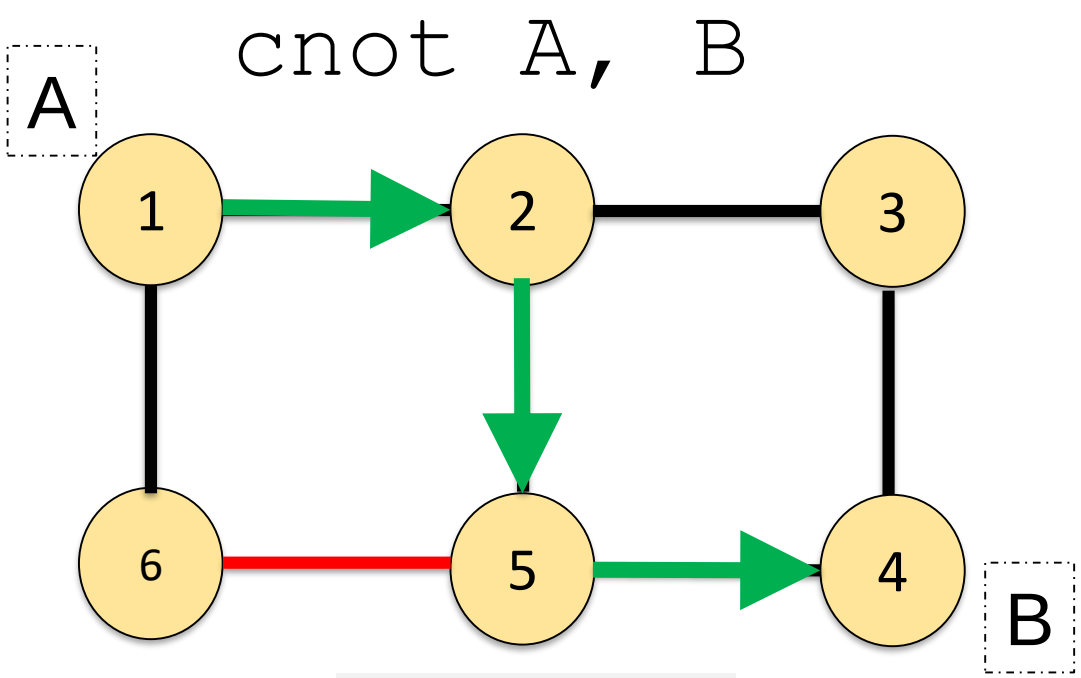


We propose variation-aware policy , to generate initial assignment and operation schedule that maximize the reliability, not just SWAP count

Qubit Movement



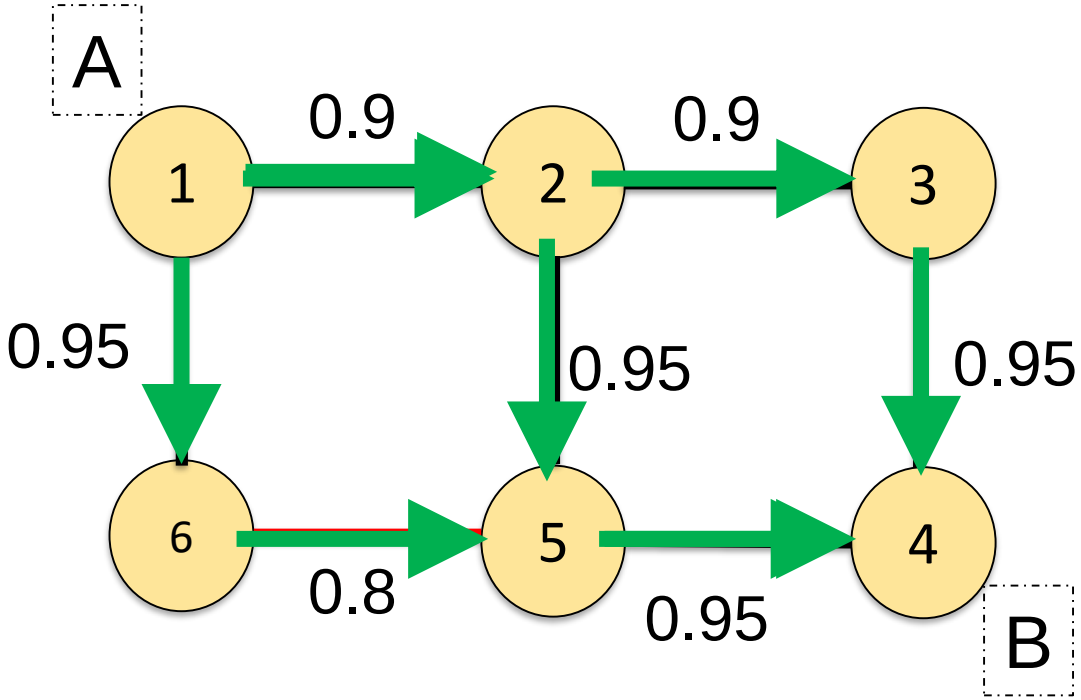
SWAPs = 2



SWAPs = 2

Multiple possible paths to entangle two qubits with identical SWAP cost

Variation-aware Qubit Movement (VQM)



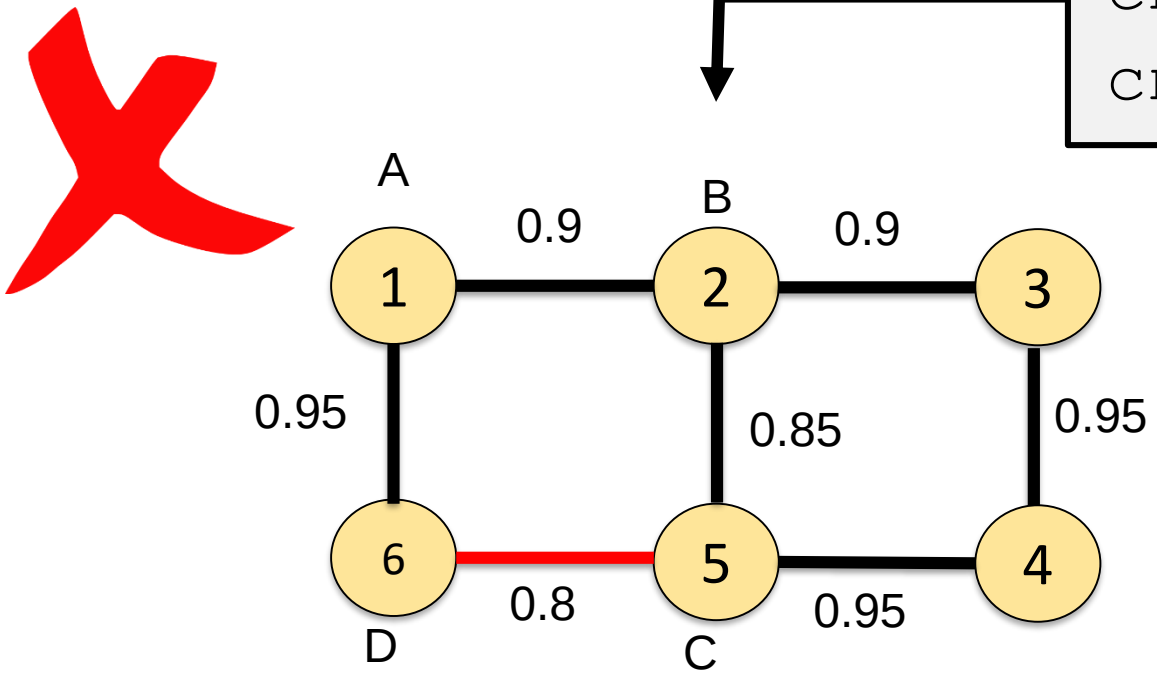

```
cnot A, B
```

Movement Path	Probability of Success
1-6-5	40%

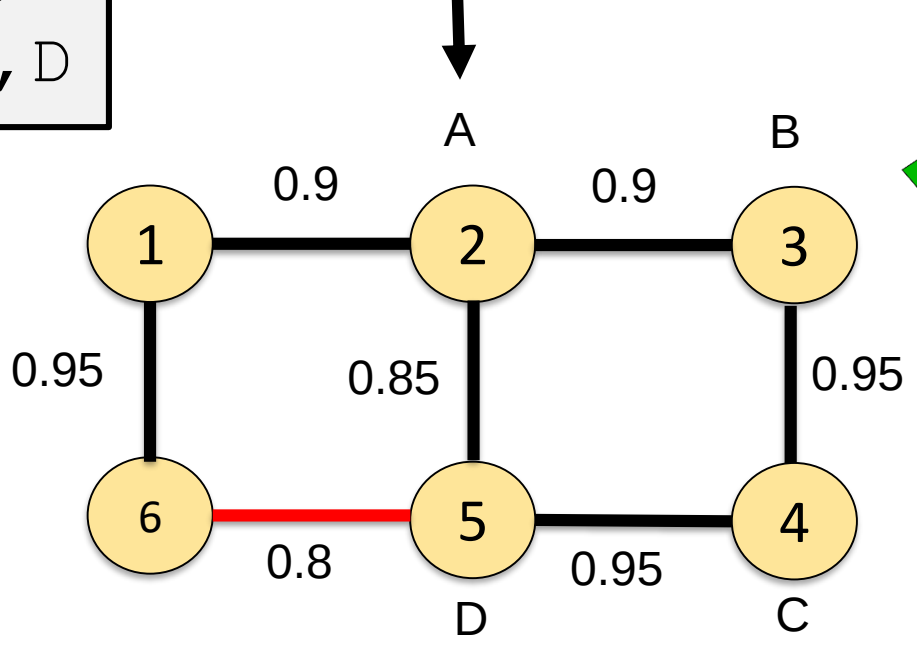
Chose a sequence of swaps that maximizes the reliability

Variation Aware Qubit Allocation (VQA)

```
cnot A,B
cnot B,C
cnot C,D
```



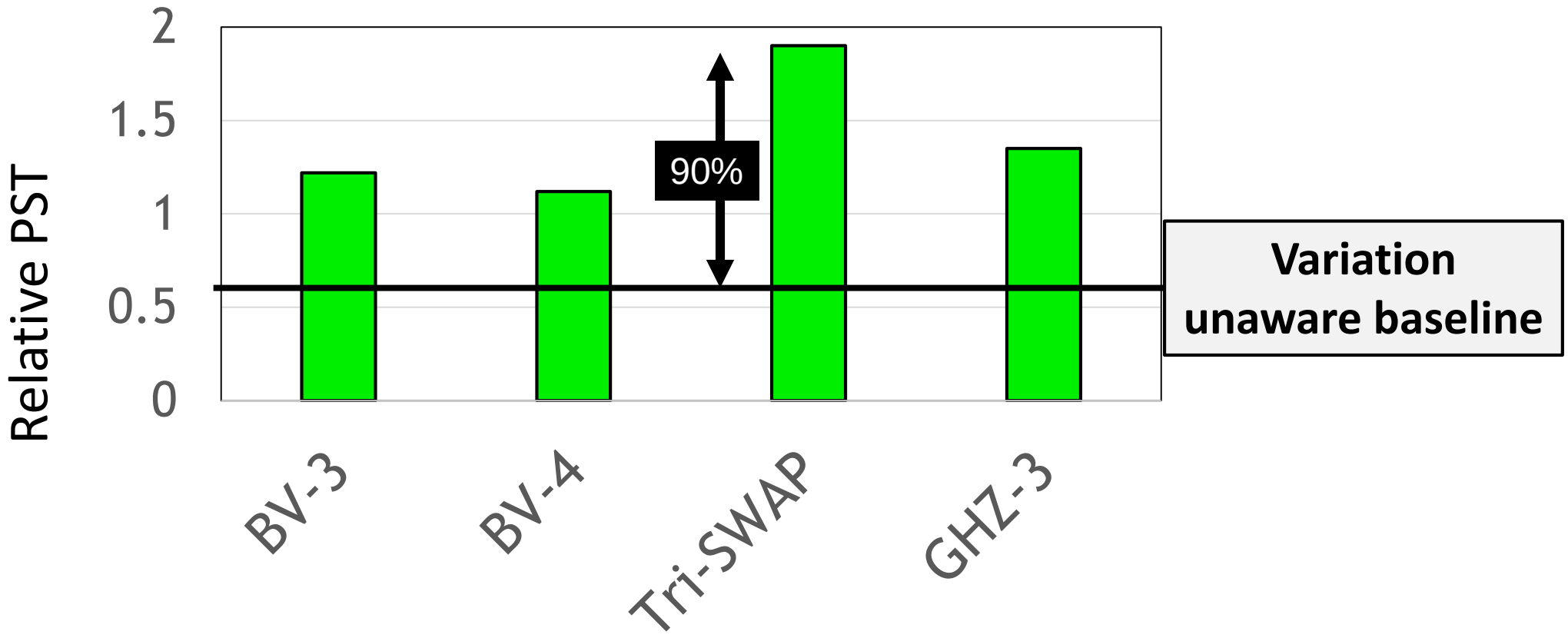
SWAPs = 0 **PST = 0.61**



SWAPs = 0 **PST = 0.77**

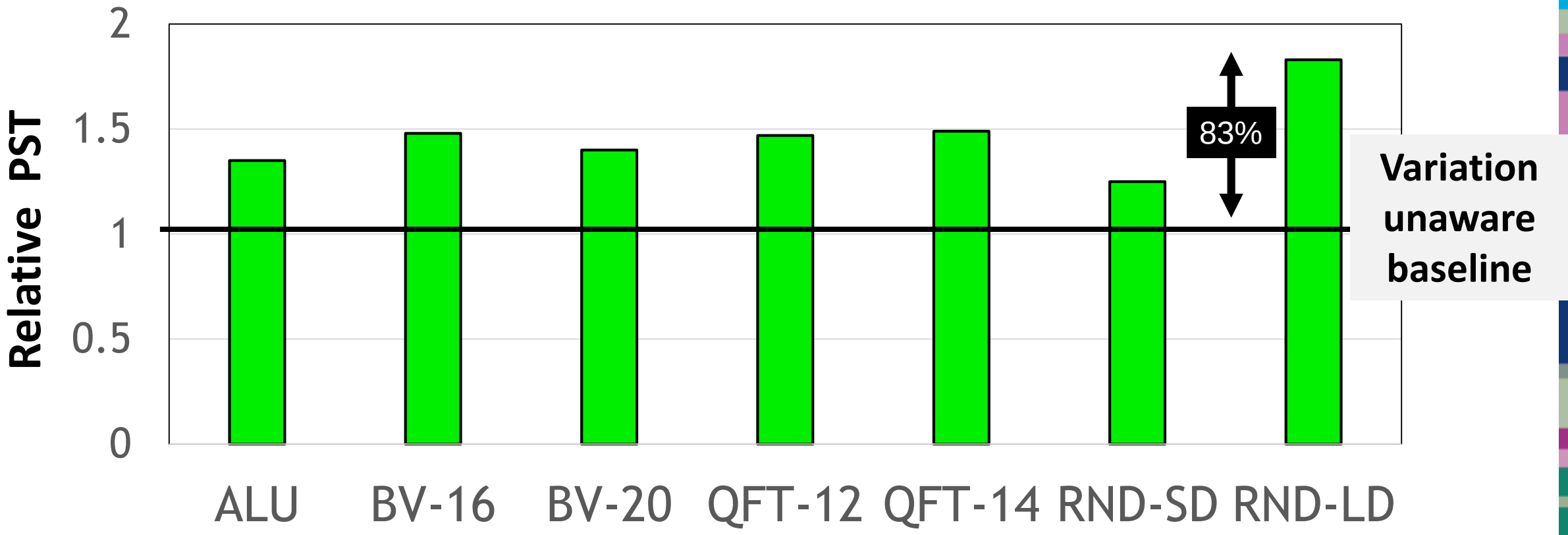
Choose qubits that maximizes the reliability

Evaluations on Real System: IBM-Q5



On IBM-Q5, VQA+VQM improves the reliability up to 1.9x

Evaluations for IBM-Q20 Simulator



VQA+VQM improves PST up to 1.83x over variation-unaware baseline

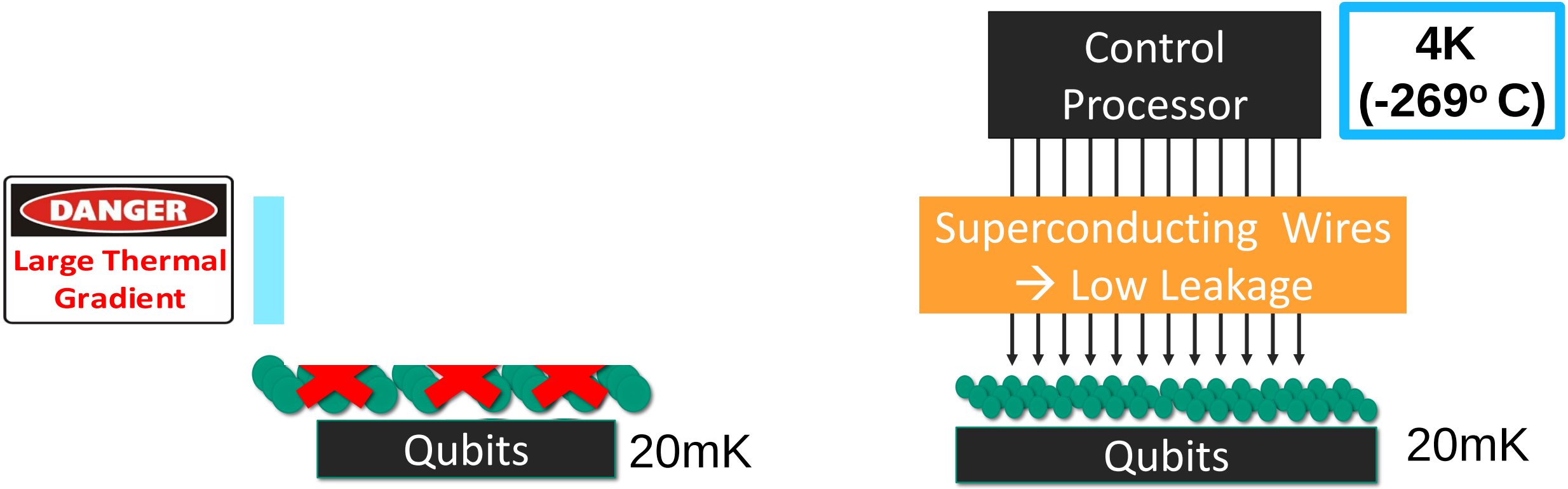
CURRENT WORK IN QUANTUM COMPUTING

OUTLINE

1. **Variability-Aware Qubit Mapping for NISQ Computers [ASPLOS 2019]**
2. **Control Processor Architecture for Fault-Tolerant Quantum Computers [MICRO 2017]**
3. **Feasibility of Cryogenic DRAM for Quantum Computing Systems [MEMSYS 2017]**

[Tannu, Myers, Nair, Carmean, and Qureshi]
http://memlab.ece.gatech.edu/papers/MICRO_2017_1.pdf

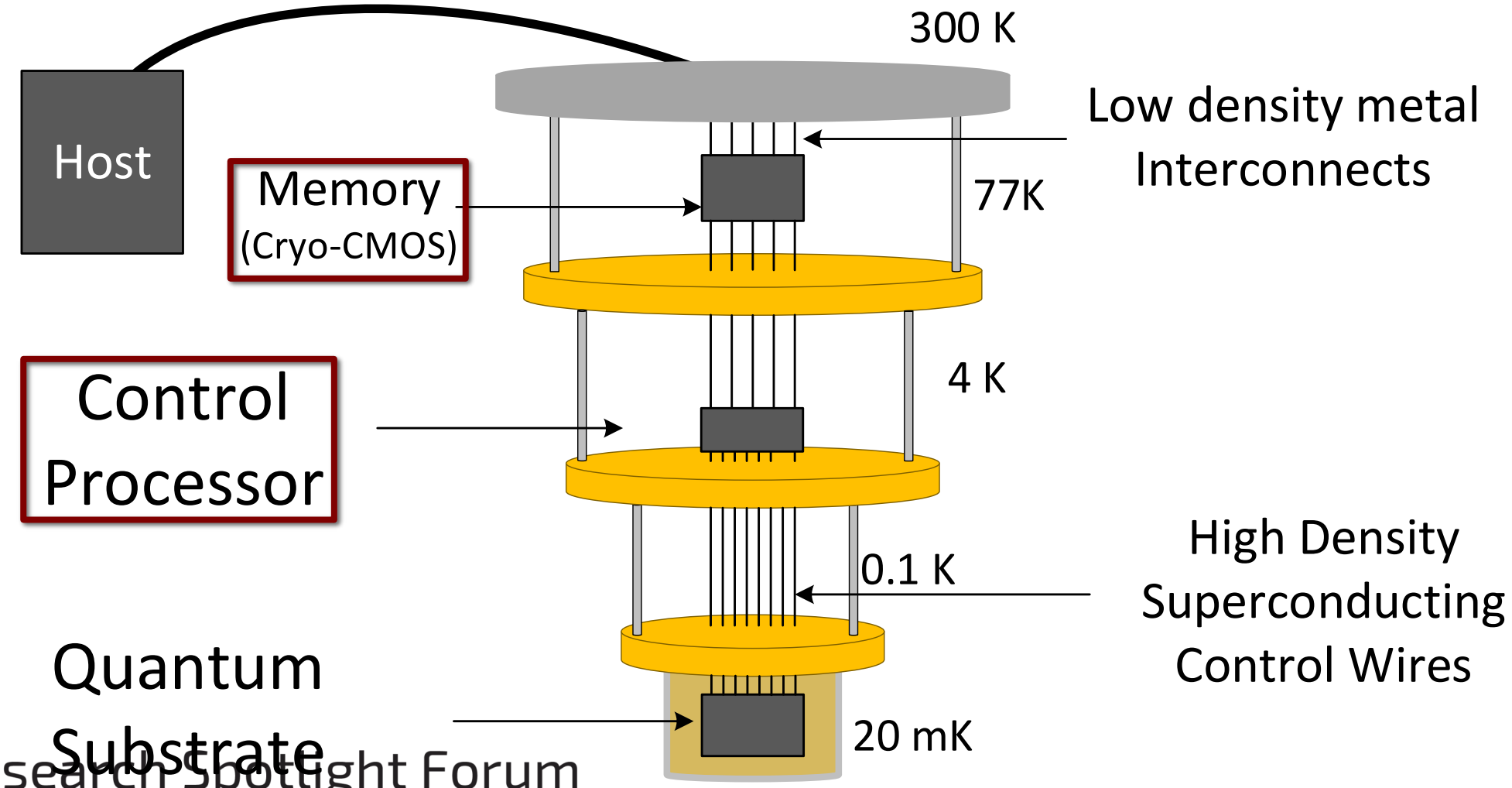
Cryogenic Control Processor



Cryogenic Control Processor is essential for scalable Quantum Computer

(Ref: Cryogenic Control Architecture for Large-Scale Quantum Computing by Hornibrook et. al)

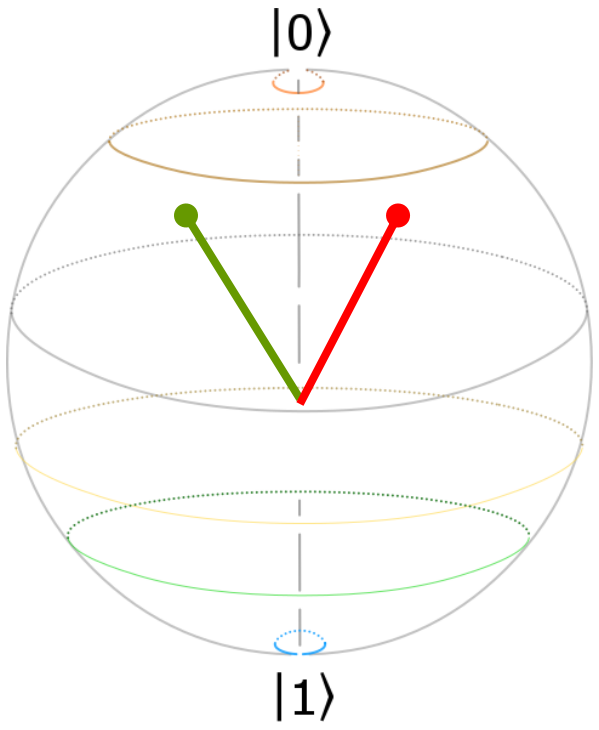
Scalable Organization: Thermal Hierarchy



Quantum bits are fickle

Even at *20mK,* qubits can lose state

Classical Bit



Quantum Bit

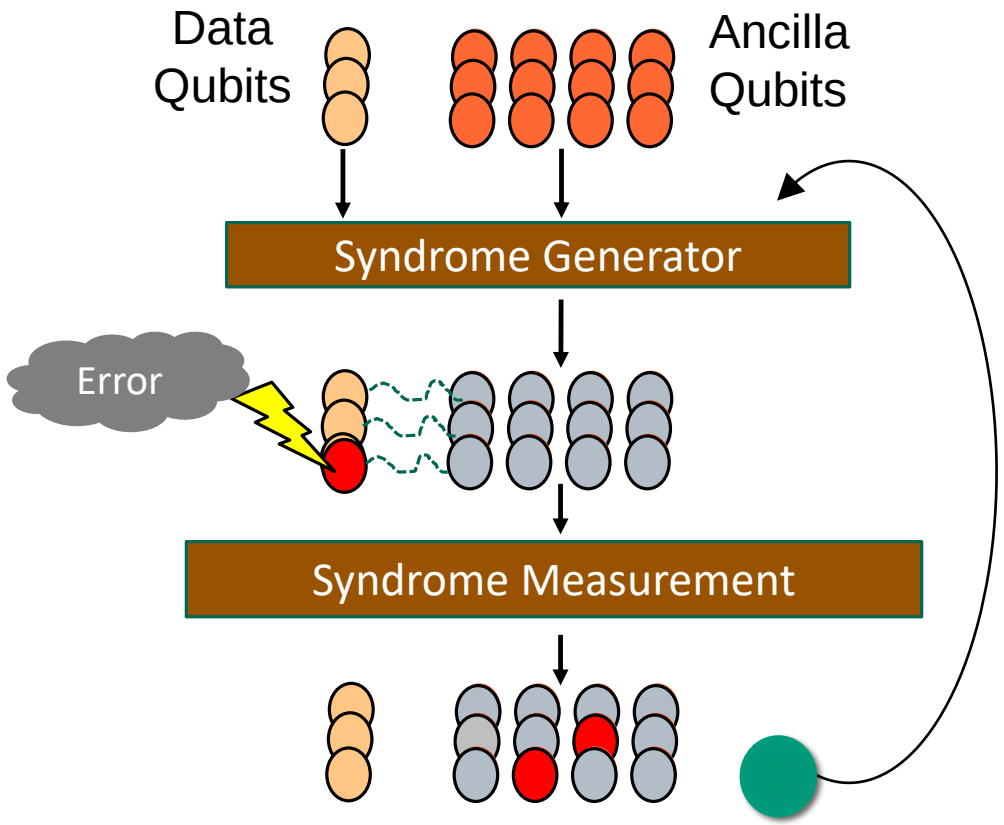
Need Error Correction to protect Quantum bits

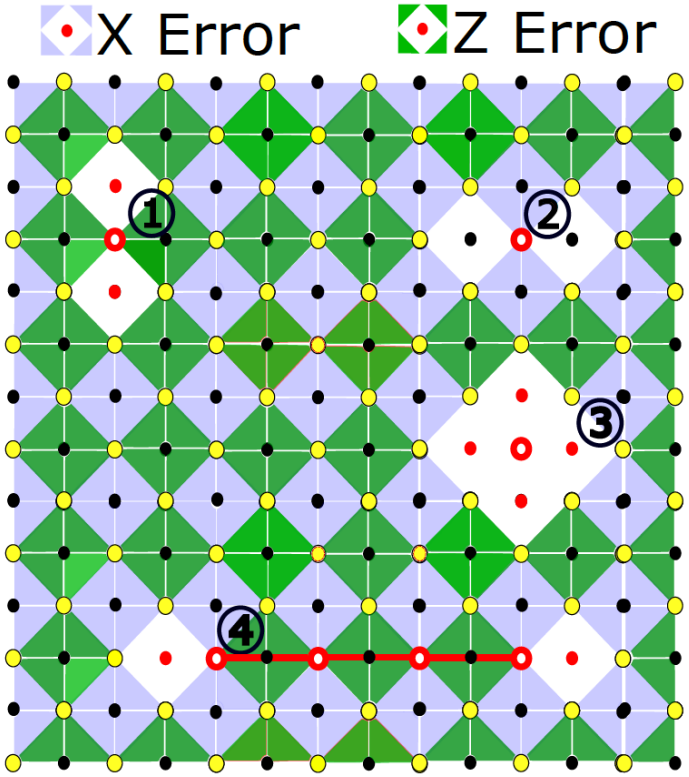
Quantum Error Correction

- Copying qubits is not allowed
- Measurement destroys the qubit state
- Quantum Error Correction can protect qubits

Continuous Quantum Error is essential to protect the state

Quantum Error Correction





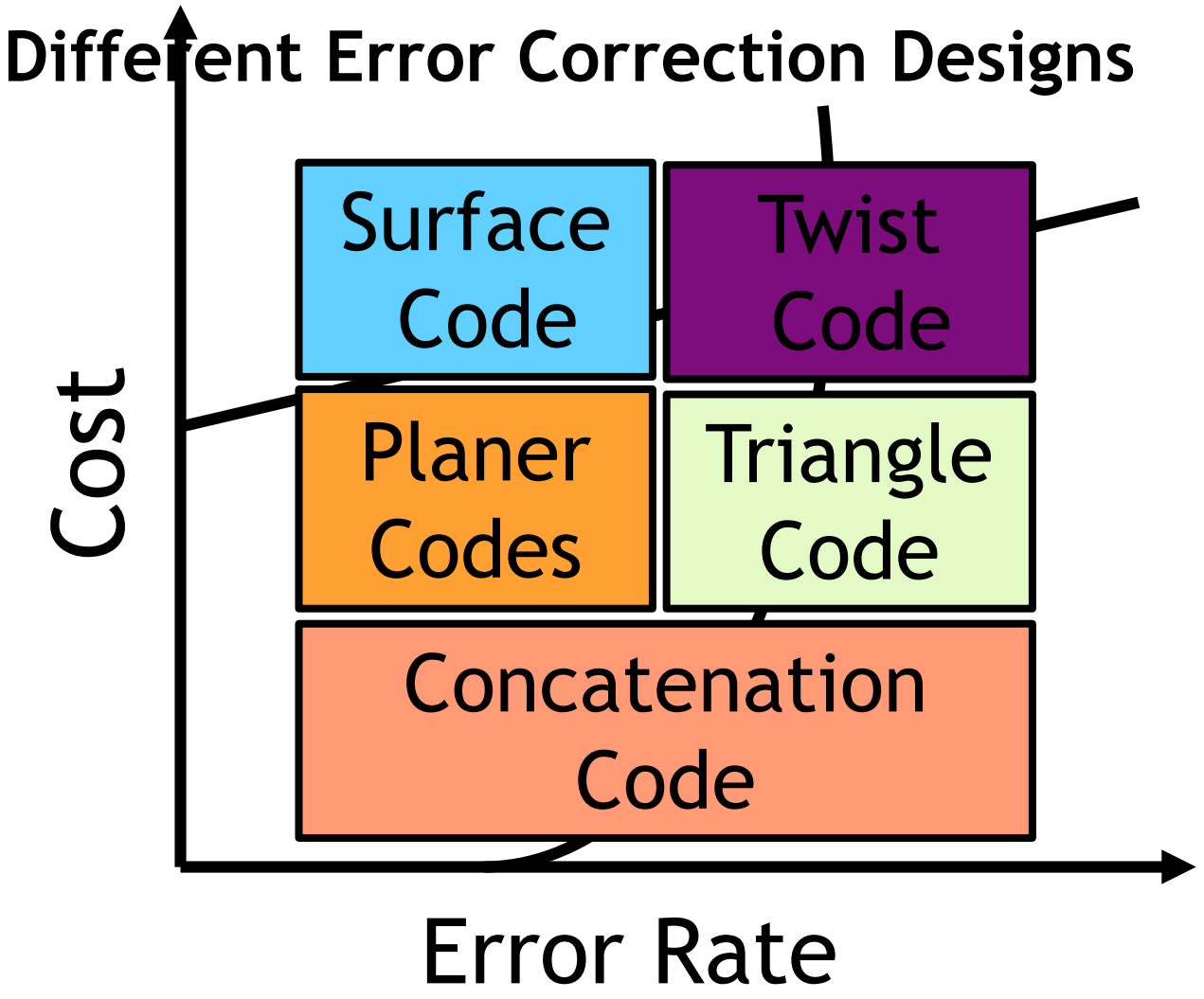
Surface Code

Programmable Quantum Error Correction

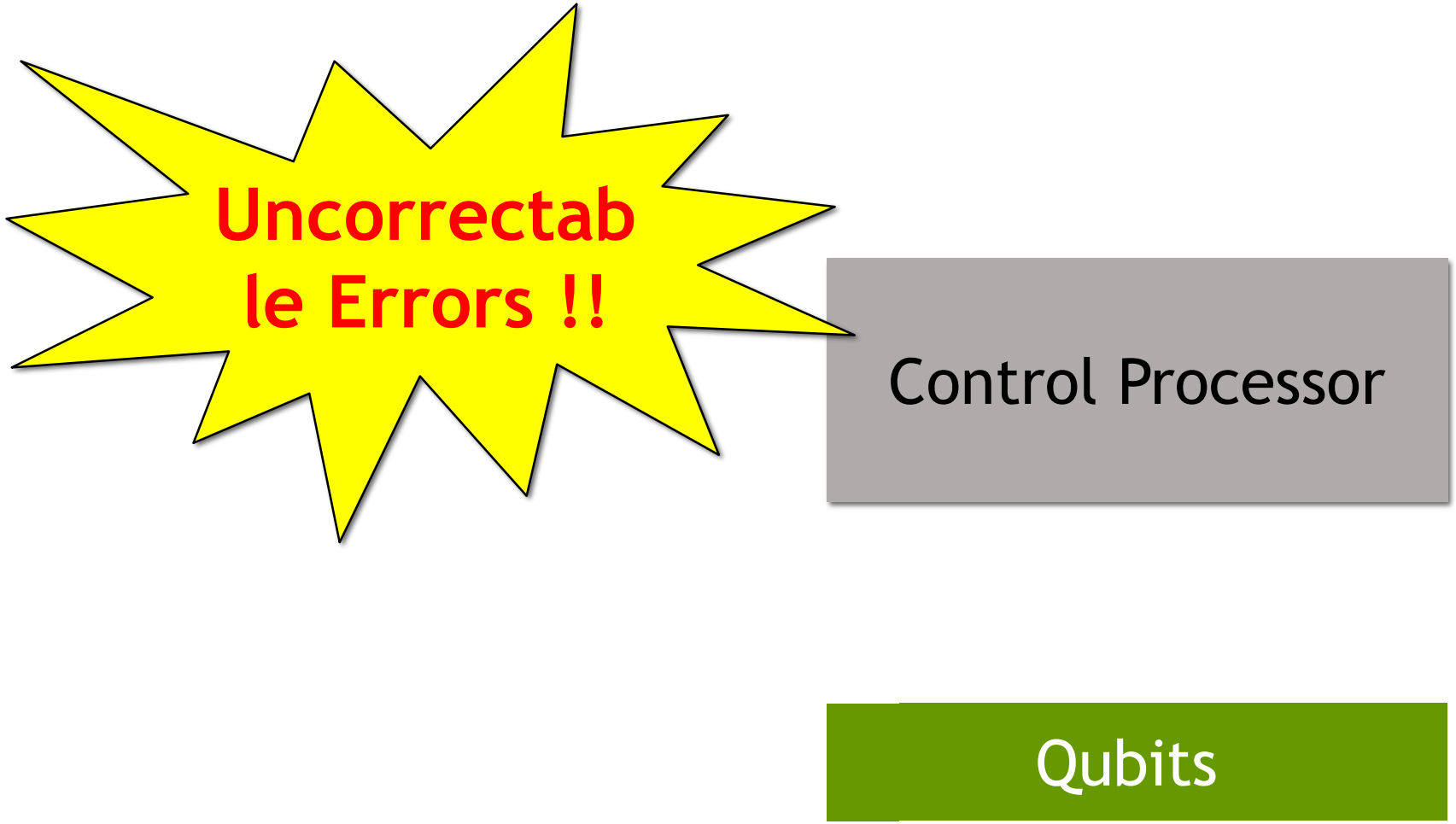
- Different QECC Designs

- New error correction codes → QECC needs to be **programmable**

- Software managed QECC → Compiler inserts QECC instructions in regular instruction stream



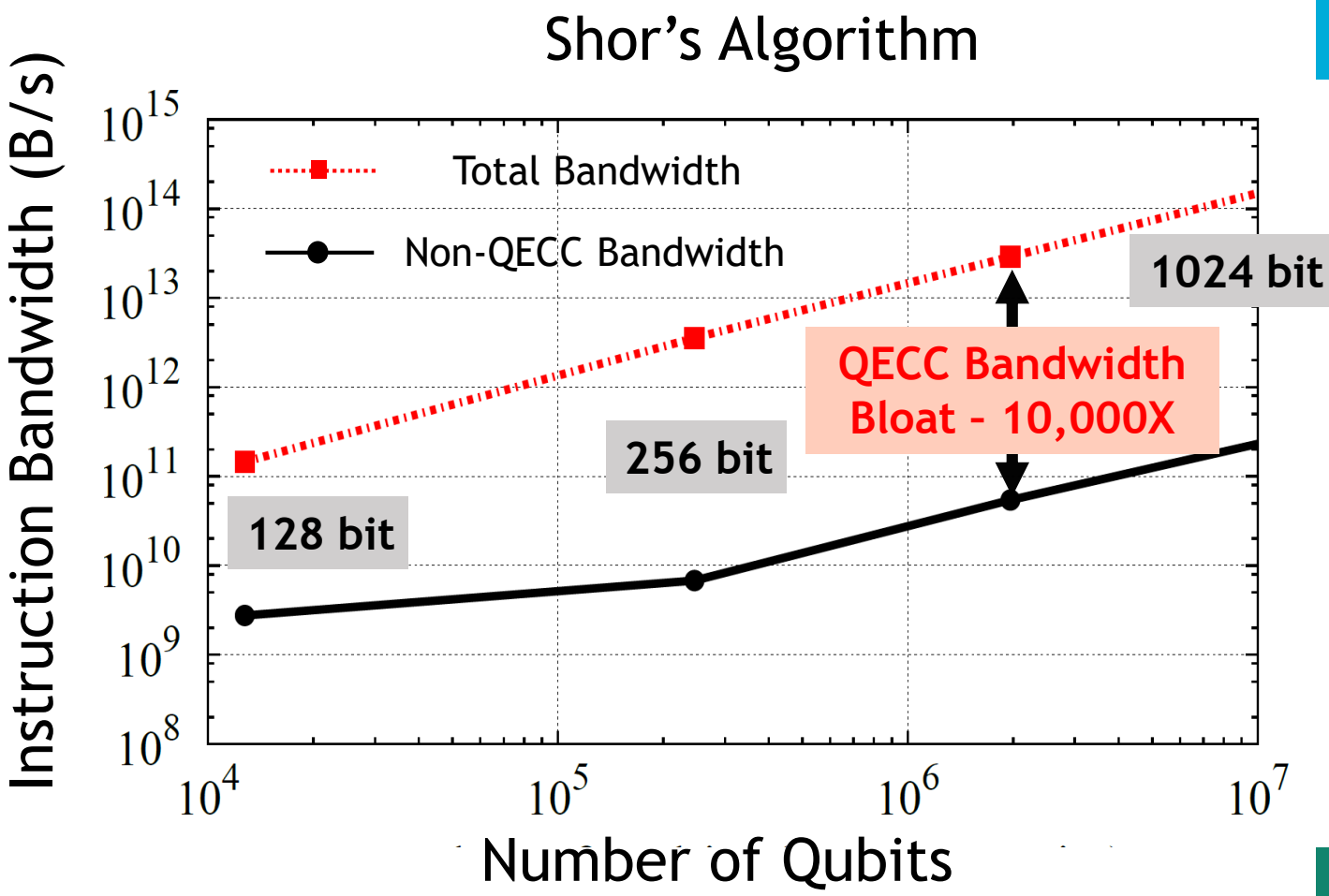
Baseline Architecture



Problem: Instruction Bandwidth Bottleneck

- Can't use i-cache -- delay in delivery of QECC instructions results in error

- SW-QECC + no i-cache→ instruction bandwidth **must** scale linearly with number of qubits

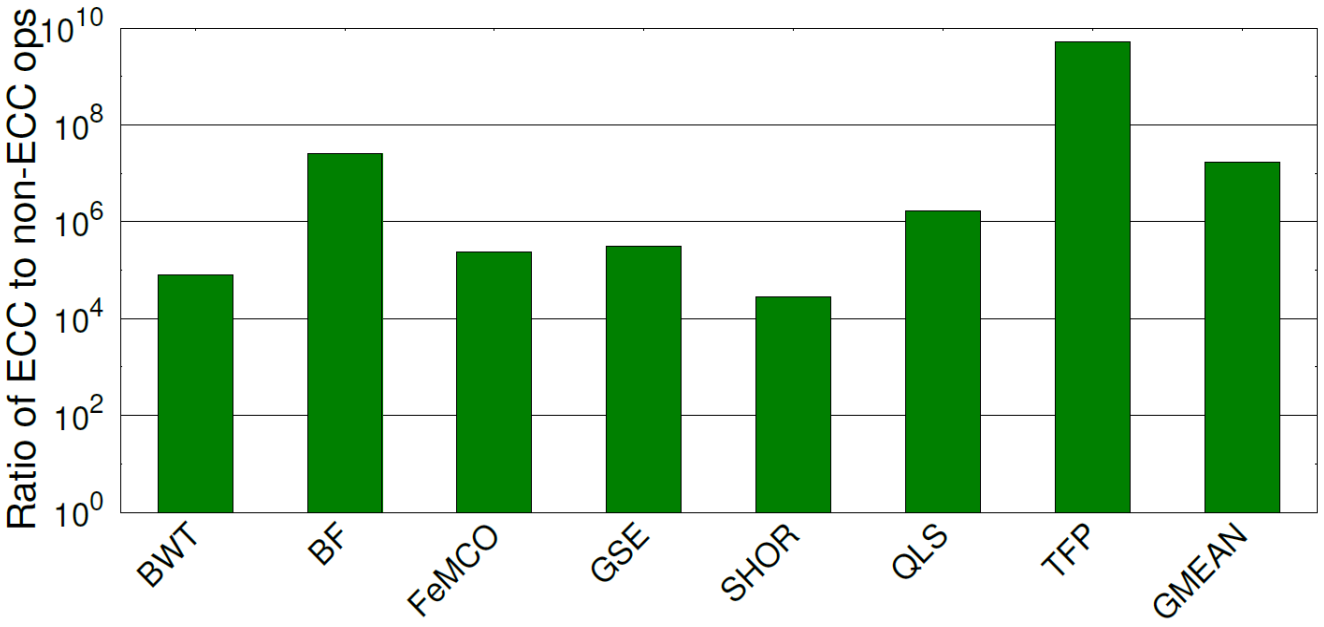


Instruction Bandwidth bottleneck ➔ 99.99% of instructions are QECC

QECC Instruction Bandwidth Bloat

- Realistic Quantum workloads require substantially large number of qubits

- Large number of qubits → must support large instruction bandwidth



QECC instruction bloat is dominant in all large scale quantum workloads

	Programmability	BW Efficiency
Software-Managed QECC	✓	✗
Hardcoded QECC	✗	✓

Goal: To enable programmable Quantum Error Correction without bandwidth bloat

Insight

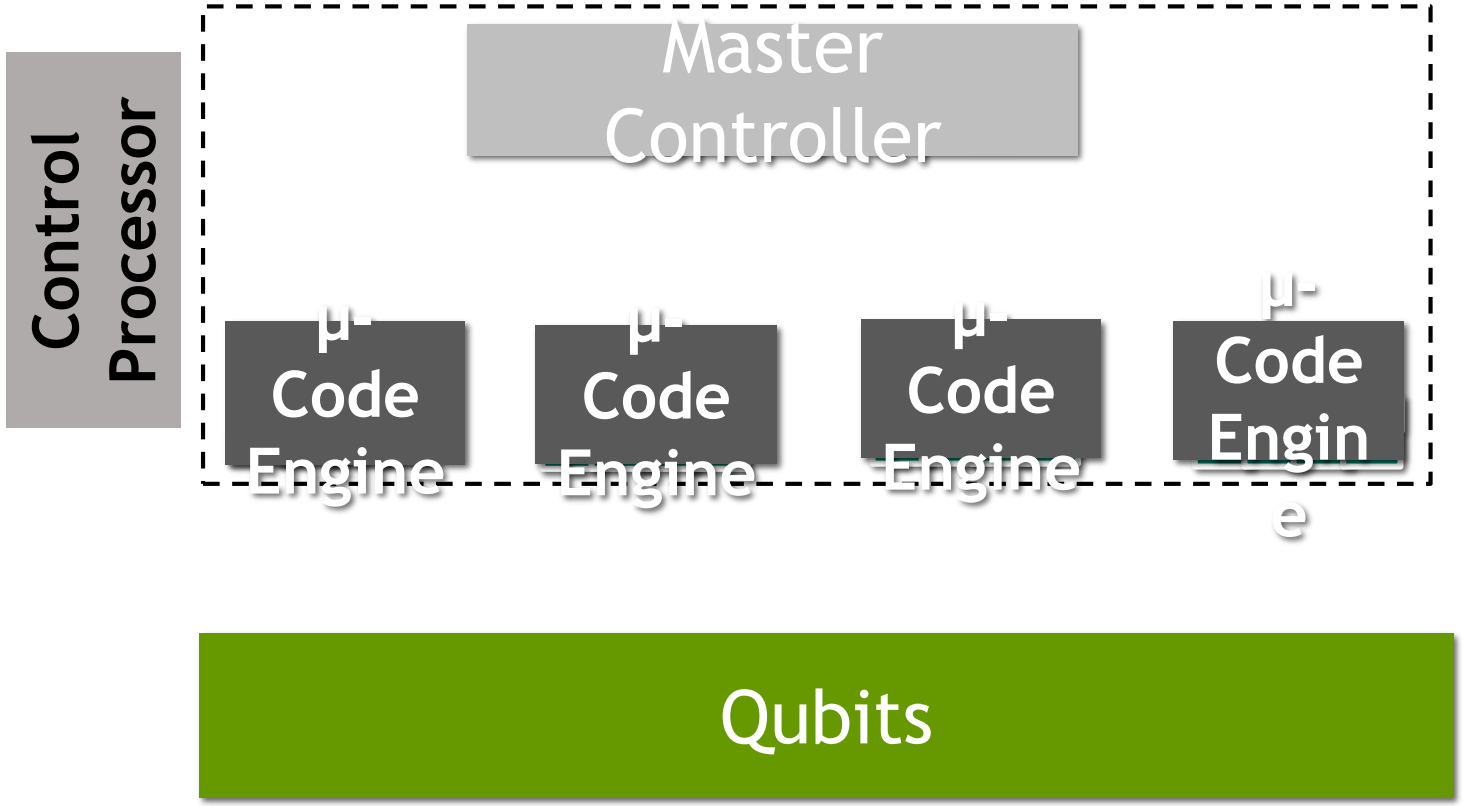
QECC can be executed independently without any global synchronization

↓

QECC is simple enough to manage in hardware using programmable microcode

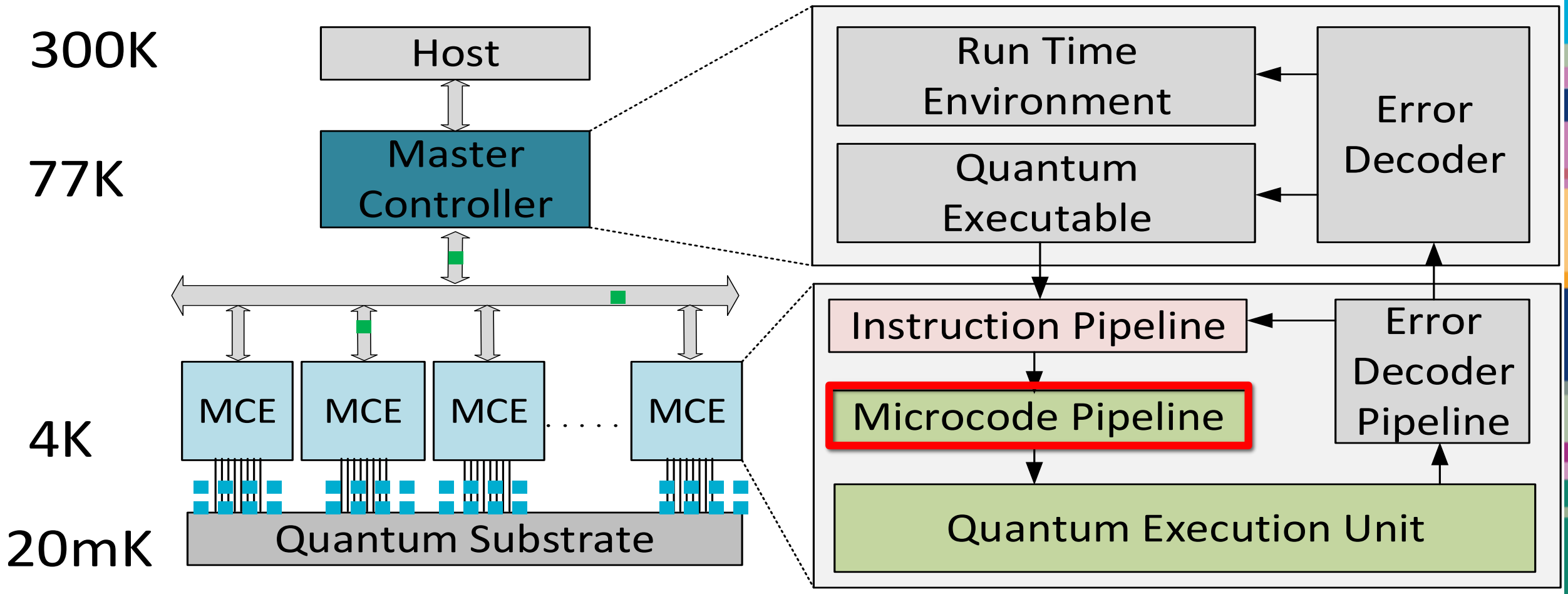
Quantum Error Correction Substrate (QuEST)

- µcode engine continuously issues QECC µops
- Master controller issues regular instructions. µcode Engine decodes it to µops

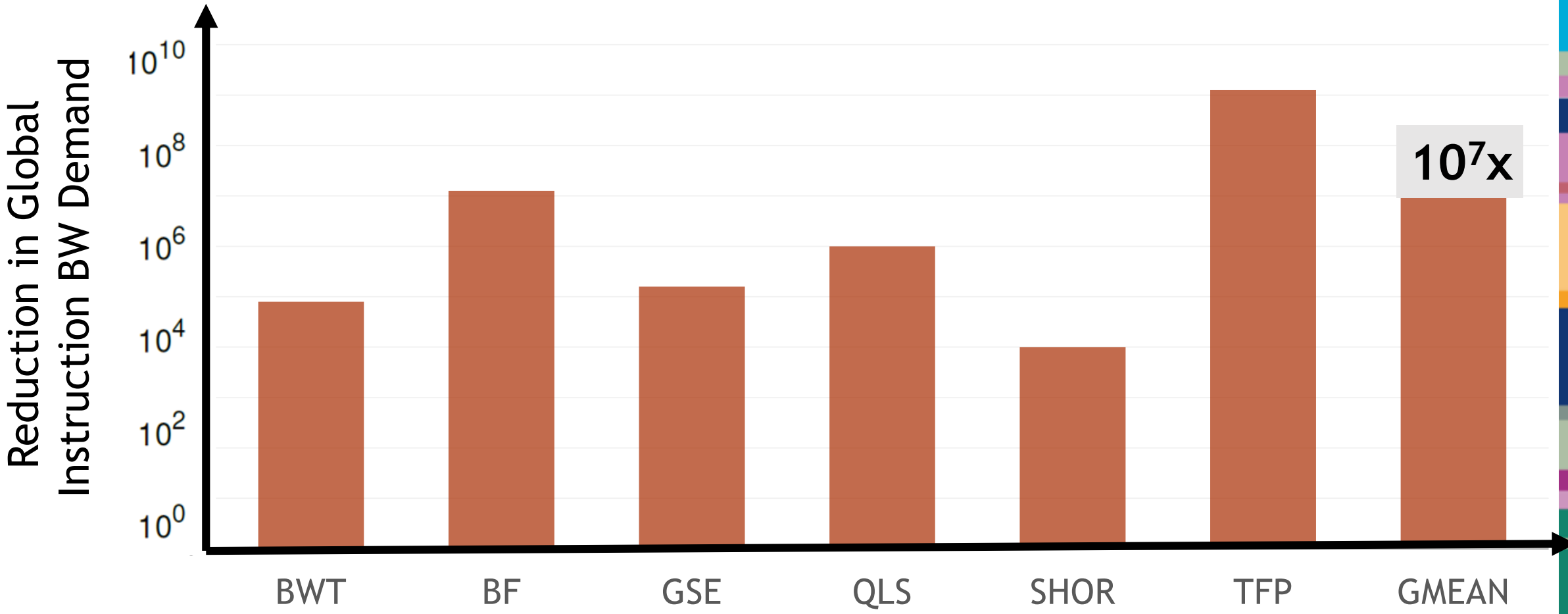


QuEST alleviates instruction bandwidth by issuing QECC ops locally

QuEST -- Organization



Evaluations



QuEST reduces global bandwidth demand by seven orders of magnitude

Key Takeaway for QUEST

	Programmability	BW Efficiency
Software-Managed QECC	✓	✗
Hardcoded QECC	✗	✓
QuEST (µcode)	✓	✓

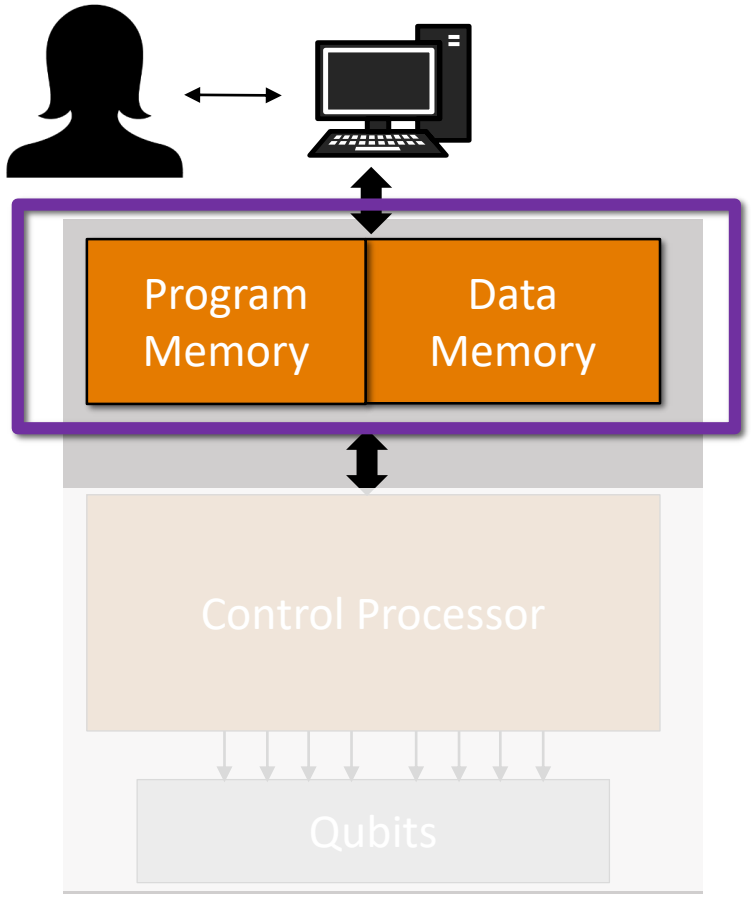
CURRENT WORK IN QUANTUM COMPUTING

OUTLINE

1. **Variability-Aware Qubit Mapping for NISQ Computers [ASPLOS 2019]**
2. **Control Processor Architecture for Fault-Tolerant Quantum Computers [MICRO 2017]**
3. **Feasibility of Cryogenic DRAM for Quantum Computing Systems [MEMSYS 2017]**

[Tannu, Carmean, and Qureshi]
http://memlab.ece.gatech.edu/papers/MEMSYS_2017_2.pdf

Memory for Quantum Computers



Quantum Computer

- ❖ Program Memory + Data Memory → Stores Quantum Executable , Data (~10s GB)
- ❖ Memory must be kept at cryo temperature to avoid large thermal gradient
- ❖ Josephson Junction technology works at 4K → Limited memory density (only few Mb)

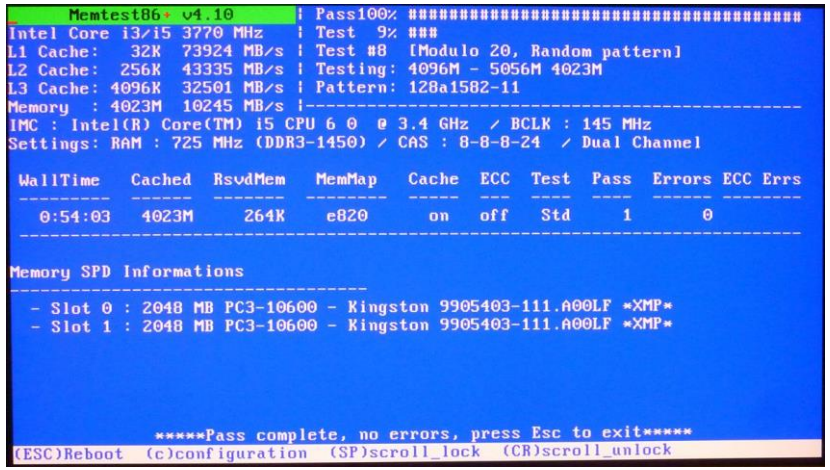
Quantum computers require substantial memory capacity at cryo temperature

Does commodity DRAM work at cryogenic temperatures ?

Goal: To characterize a DRAM at cryogenic temperature to understand the functionality and error patterns

How to Test DRAM at Cryogenic Temperature?

- Conventional memory testing → Memtest86 running on host, dedicated memory testers

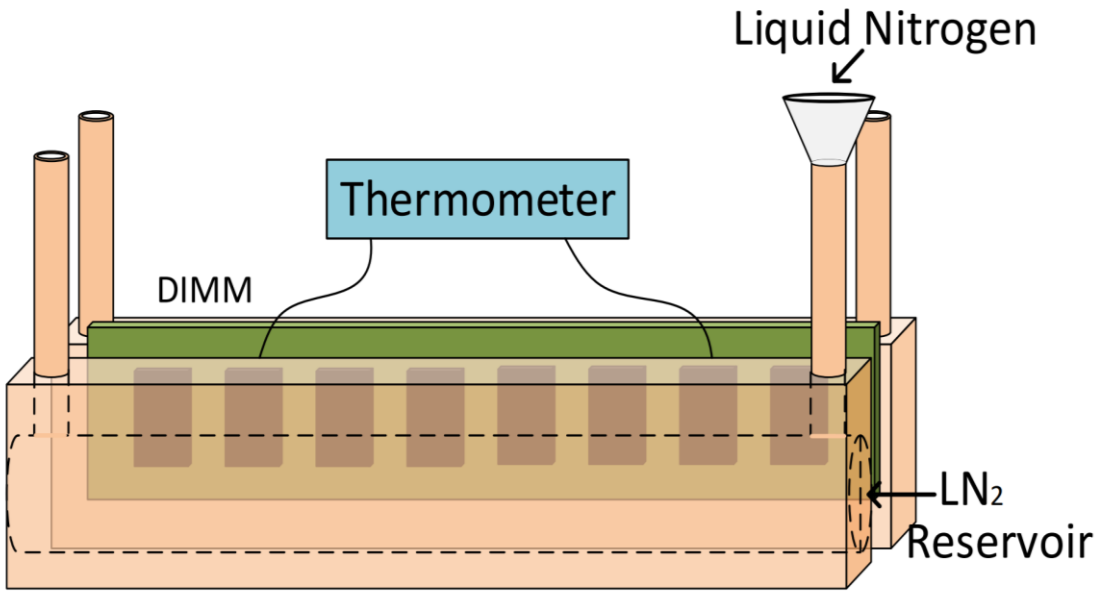


- Host machines or memory testers <u>do not work at cryogenic temperatures</u>

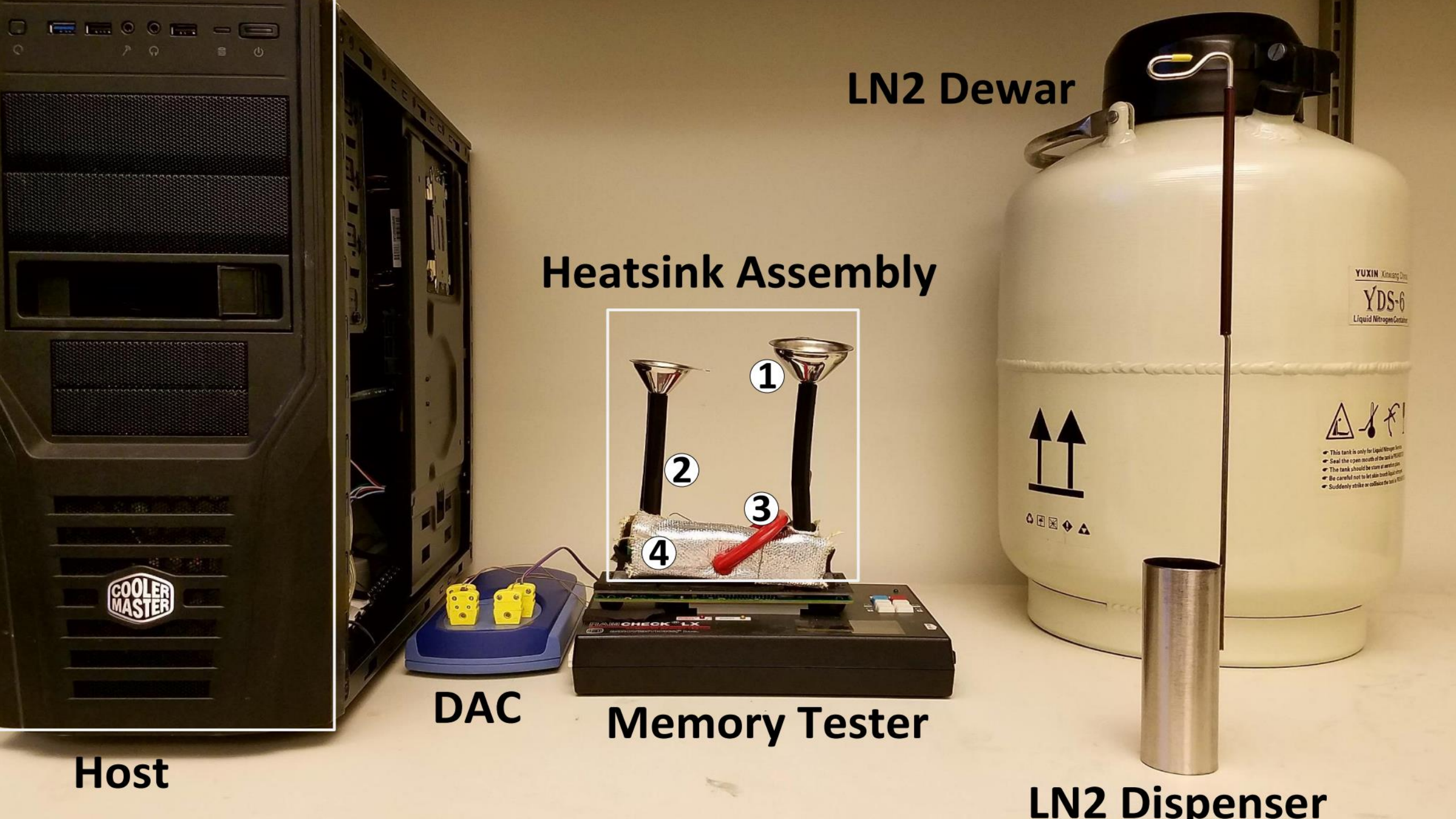
Need mechanism to reduce DIMM temperature without affecting tester

Isolated Cooling of DIMM

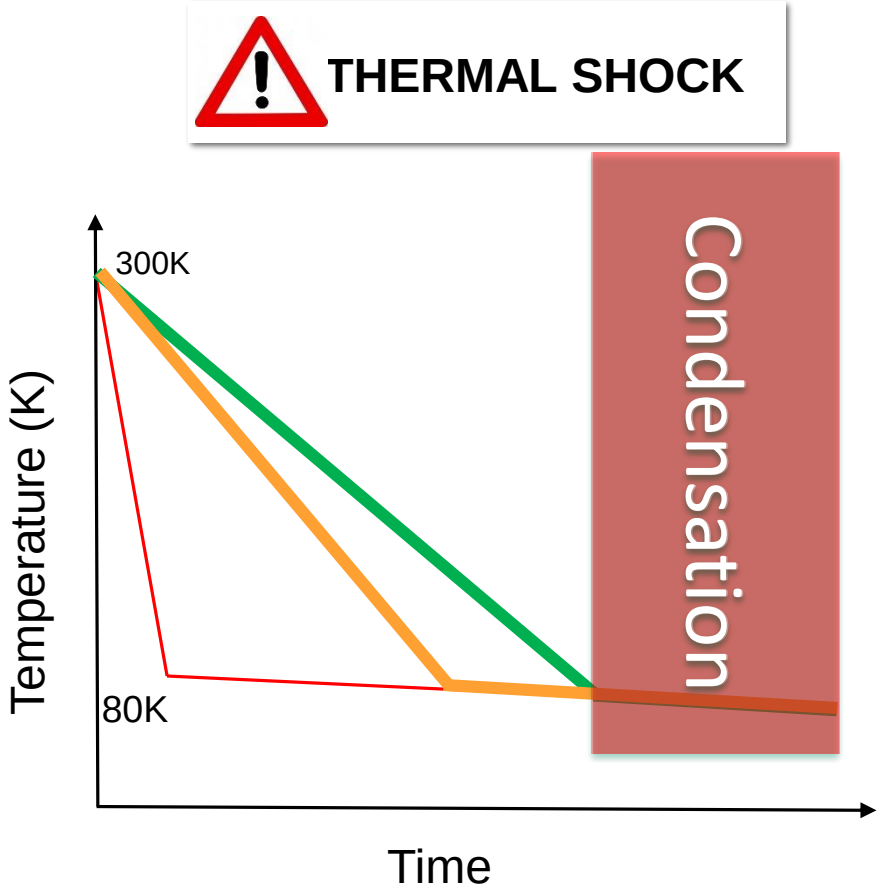
- Need cryogenic coolant → Liquid Nitrogen (boils at 77K)
- Need isolated cooling of DIMMs → Compact cryogenic heatsink
- DIMM is sandwiched between two heatsinks and can be cooled down to 80K



Compact heatsink with Liquid Nitrogen provides isolated cooling of a DIMM


LN2 Dewar
Heatsink Assembly
1
2
3
4
DAC
Memory Tester
Host
LN2 Dispenser
COOLER MASTER
YUXIN
YDS-6
Liquid Nitrogen Container

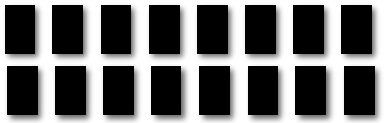
Challenges: Thermal Shock & Ice Condensation



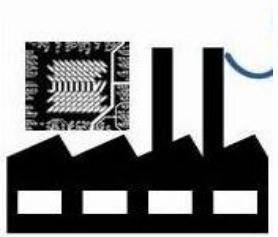
Limit rate of cooling & use isolation chamber to reduce condensation

Experimental Methodology

Number of DIMMS → **55**



Number of Chips → **750**

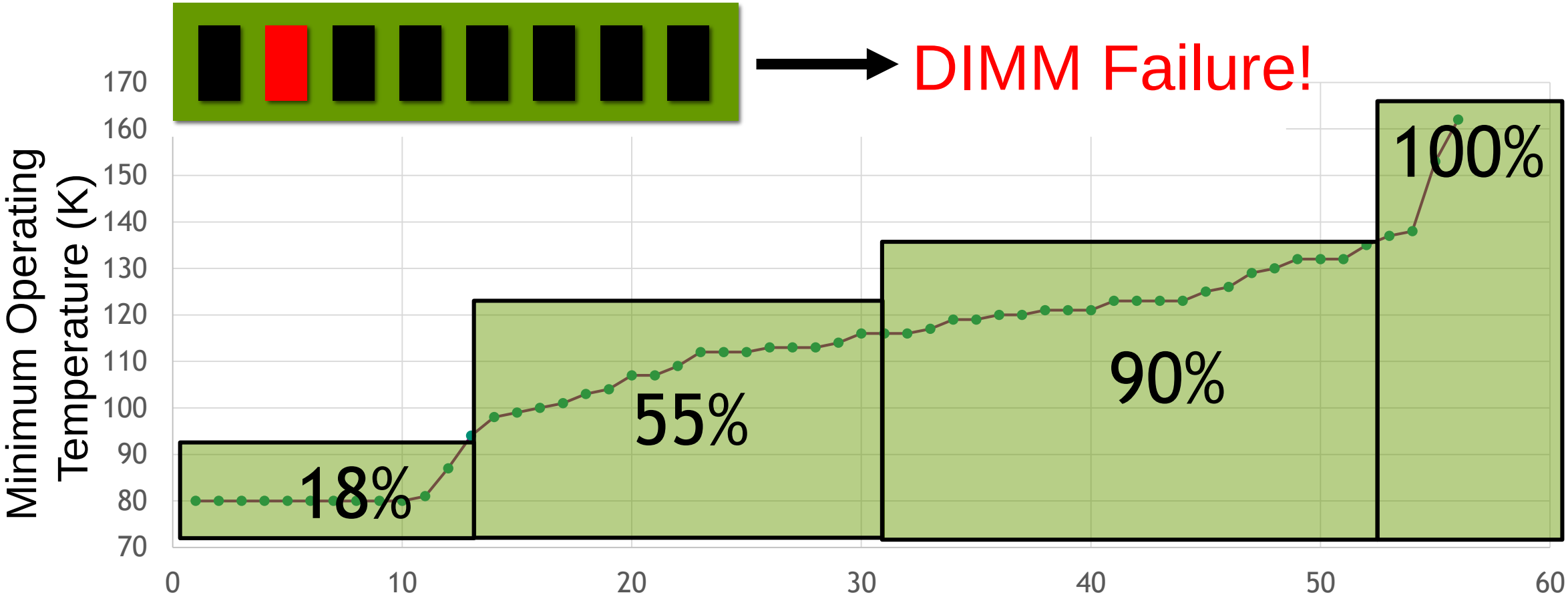


Number of Vendors → **5**

- ❖Verify memory functionality by using march-tests
- ❖Fault → single bit fault in a burst
- ❖Minimum Operational Temperature (MOT)

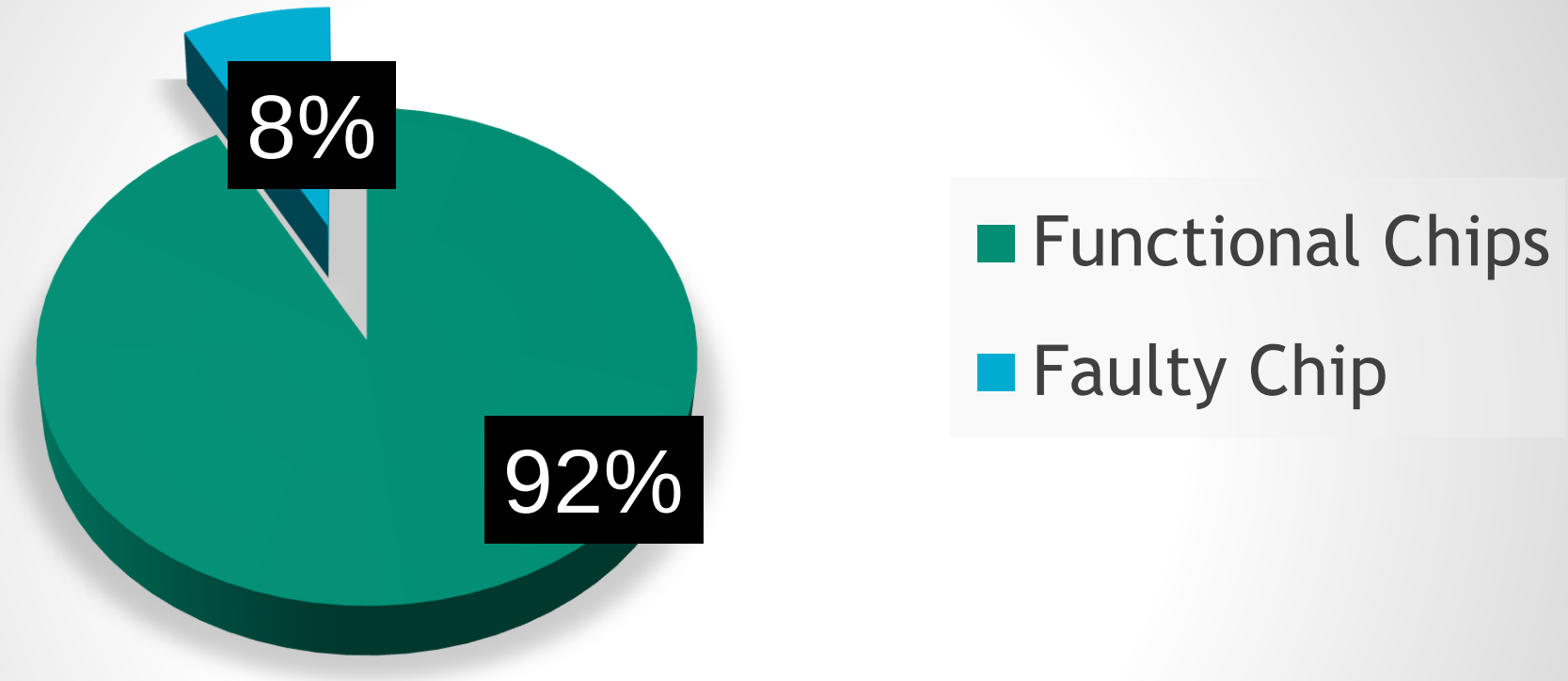
→ Minimum temperature at which no faults are observed

Minimum Operational Temperature for DIMMs



18% DIMMs are functional below 90K

Chip Failures



92% of chips worked at 90K — Pick cryogenic tolerant chips

Fault Granularity

- Single bit errors → Uncorrelated faults

- Linear codes (SECDED, BCH) are still effective

Uncorrelated faults → Conventional ECC can be effective for Cryo DRAM

Cold DRAM – Take Away

- Quantum computers need dense memory at low temperature
- Does DRAM Work at cryogenic temperature?
- Experiments show → most commodity DRAM chips work at 90K
- Error patterns are amenable to existing fault tolerance techniques

FUNDING SOURCES

List funding sources of current projects, if applicable- especially joint work with Sandia, other NM Universities, or other Sandia Academic Alliance Universities (UT Austin, Georgia Tech, UIUC, or Purdue).

- NSF and SRC (Hybrid memory research)
- Intel (3D memory research)
- Microsoft (Gift for research in quantum computing)

RESEARCH NEEDS

Describe any gaps or new directions that would benefit from collaboration.

- Access to larger-scale quantum machines and test-bed to test error mitigation algorithms
- Access to characterization data for quantum machines to enable new error mitigation algorithms
- Access to emerging NISQ algorithms and applications
- Funding opportunities (for collaboration and joint works)