Quantum Memory Hierarchies

Efficient Designs that Match Available Parallelism in Quantum Processors

Darshan D. Thaker
Tzvetan S. Metodi

UC Davis

Andrew W. Cross
Issac L. Chuang

MIT

Frederic T. Chong

UC Santa Barbara

Motivation

Study tradeoffs between area - reliability - performance.

Goals of this research: Reduce overall area of the design.

Leverage conventional architectural techniques to improve performance.

Provide abstractions for further research.

Outline

Background and prior work.

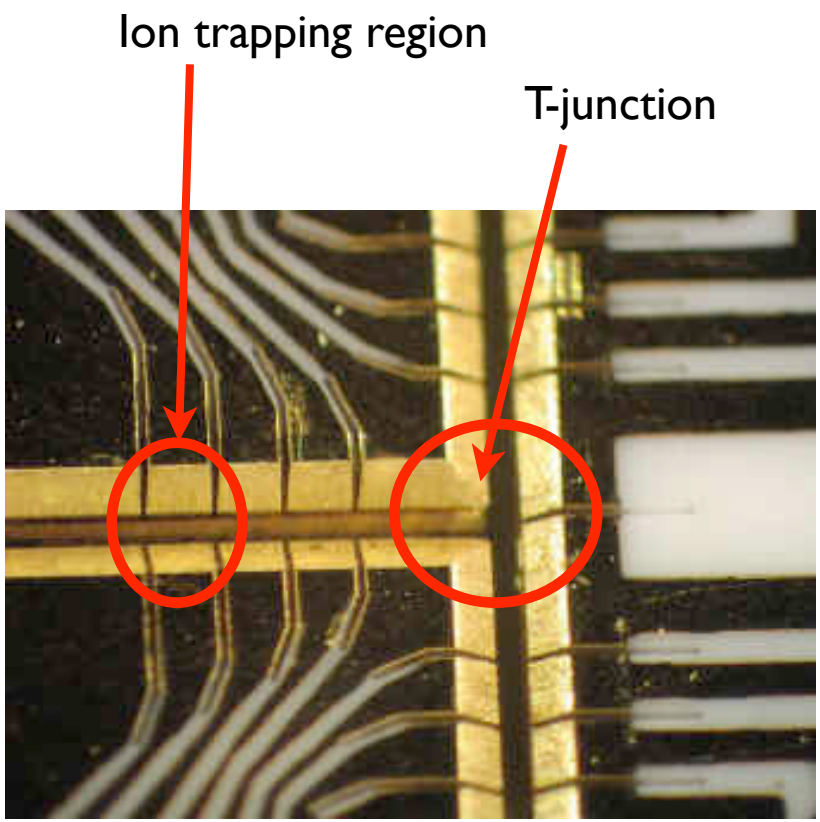
Overview of quantum error correction codes.

Specialization into memory and compute regions.

Improving performance.

Results and discussion.

Ion-Traps



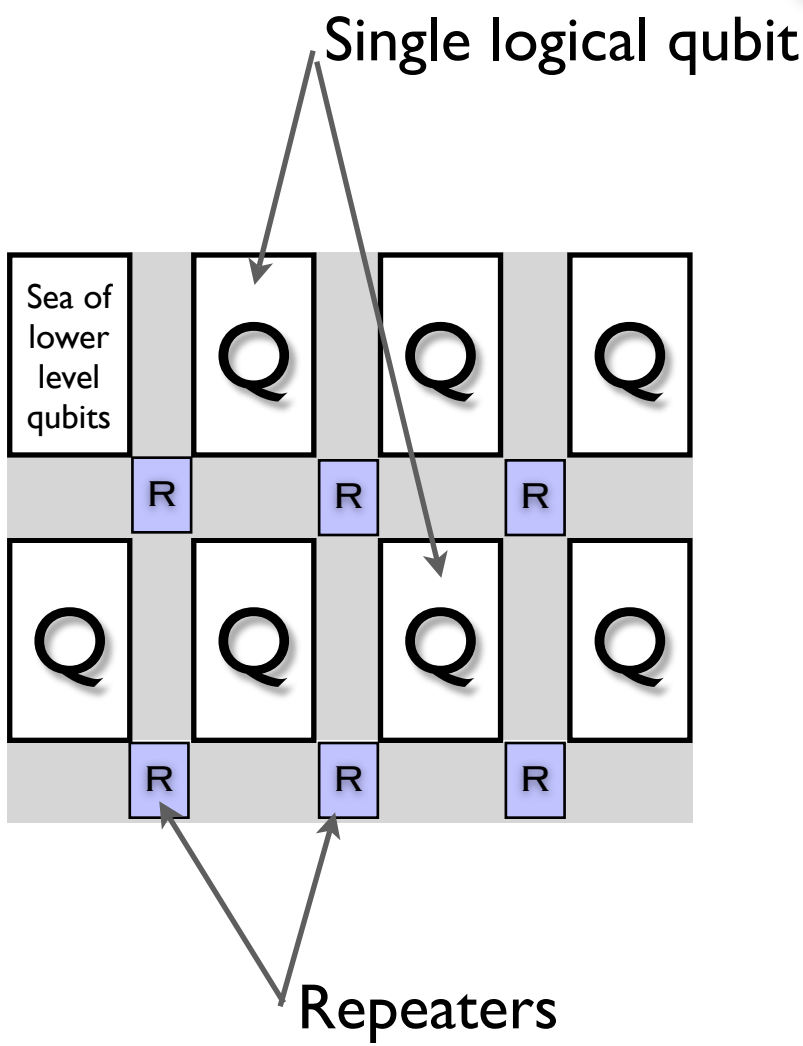
Courtesy: C.Monroe at U.Michigan

Use ions trapped in electromagnetic fields.

Lasers acting on ions induce quantum gates.

Newer traps are micromachined.

Quantum Logic Array



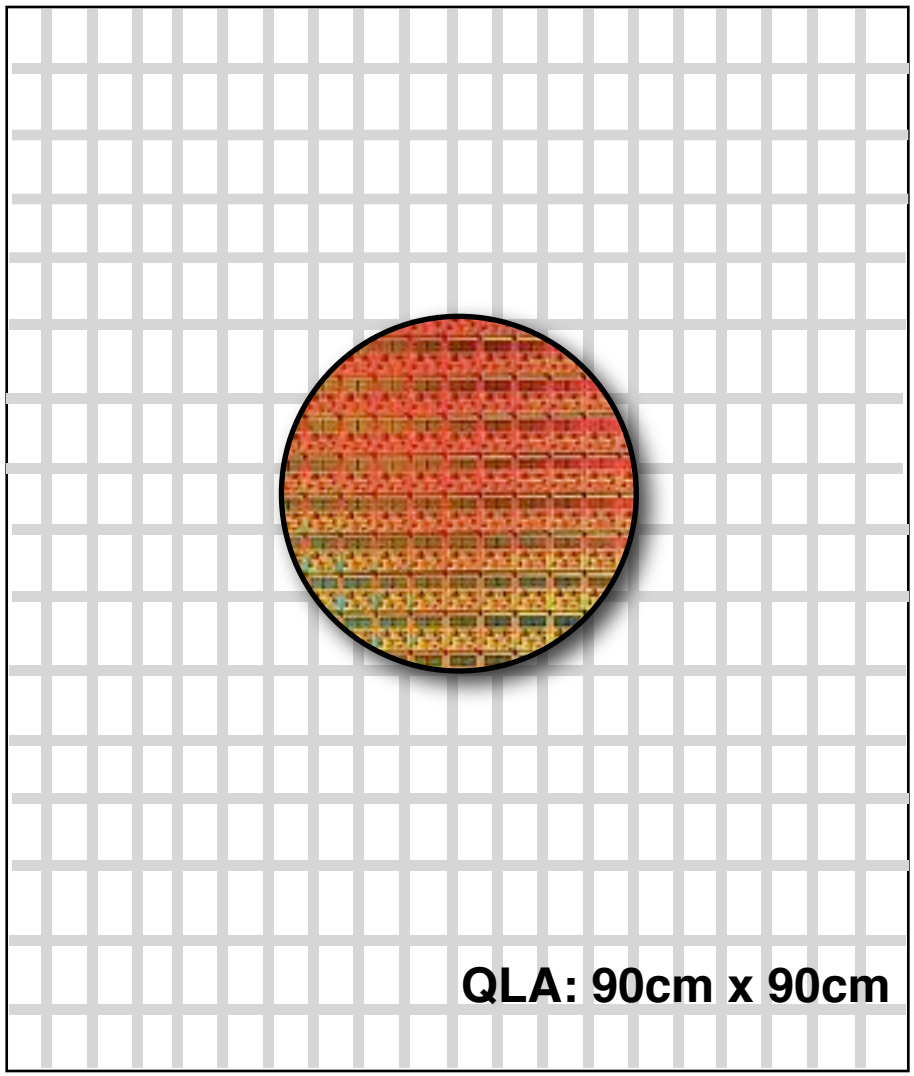
Compute-anywhere design.

Teleportation based long-distance communication.

Exponential speedup when factoring large numbers.

Unresolved Issue: Size

Quantum Logic Array



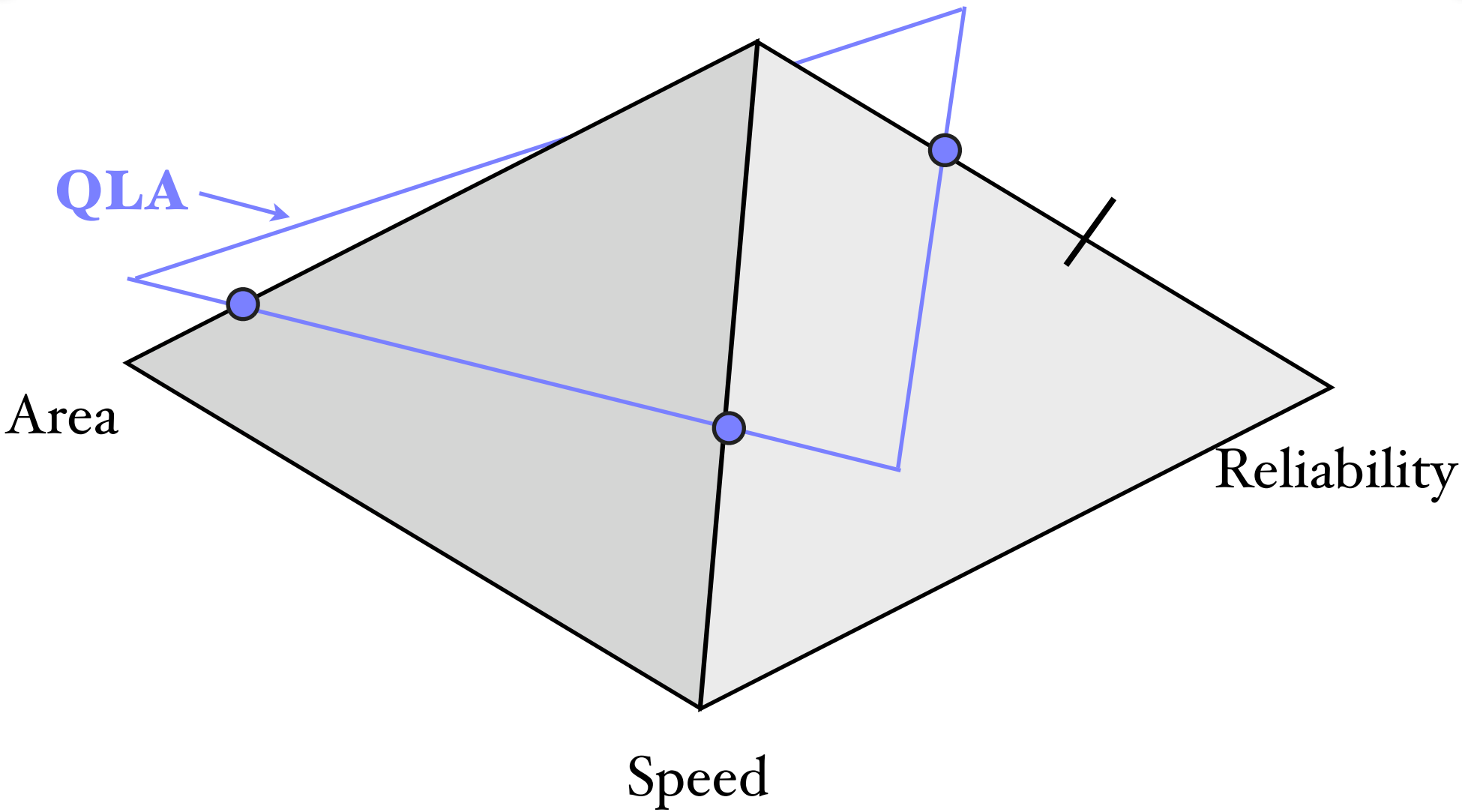
Sea of qubits design.

Teleportation based long-distance communication.

Exponential speedup when factoring large numbers.

Unresolved Issue: Size

Design Pyramid



Outline

Background and prior work.

Overview of quantum error correction codes.

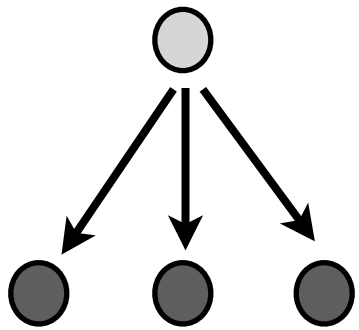
Specialization into memory and compute regions.

Improving performance.

Results and discussion.

Comparison with Classical Codes

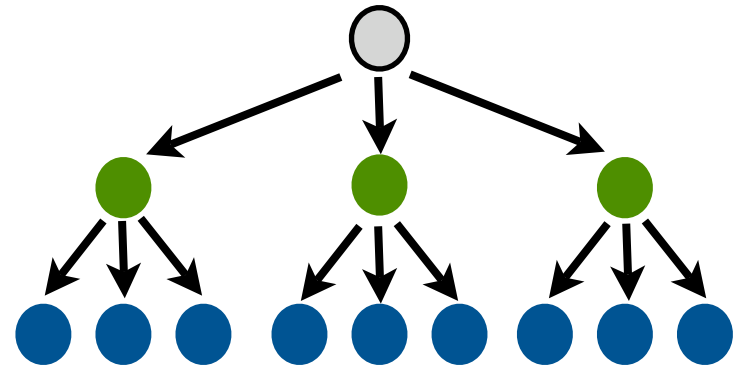
Classical three bit code



Single bit encoded as three bits.

Majority Voting.

Equivalent quantum code

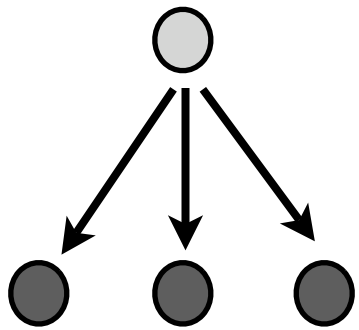


Nine qubit Shor code

Protects against *bit*-flips and *phase*-flips.

Comparison with Classical Codes

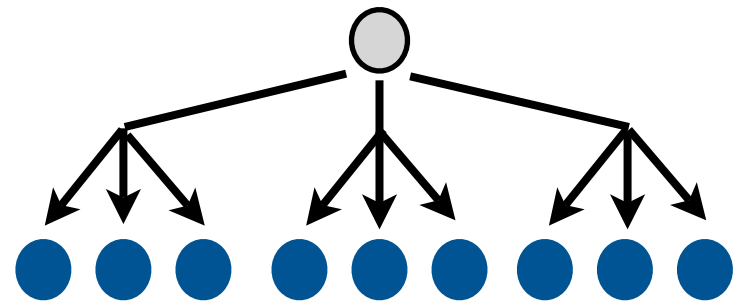
Classical three bit code



Single bit encoded as three bits.

Majority Voting.

Equivalent quantum code



Nine qubit Shor code

Protects against *bit*-flips and *phase*-flips.

Greater Reliability

Need greater reliability than provided by encoding a single time.

The No cloning theorem and restrictions on measurement require greater reliability.

Cannot use methods like checkpointing or make duplicates.

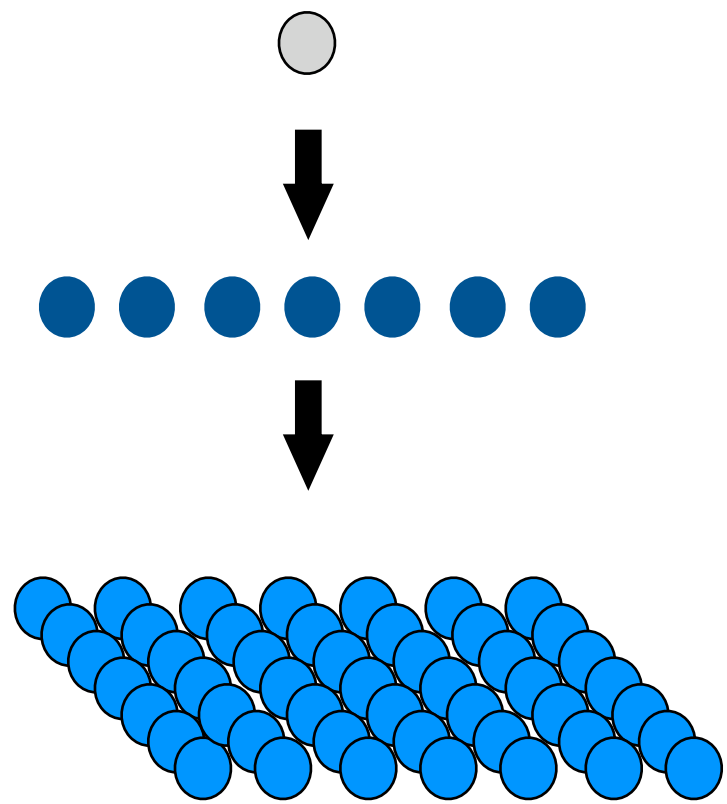
Solution: Use concatenated codes.

Concatenated Codes

1 logical qubit

Level 1:
7 physical qubits

Level 2:
49 physical qubits



Reliability increases doubly exponentially.

Exponentially slower.

Exponentially greater resources.

Concatenated Steane Code

Outline

Background and prior work.

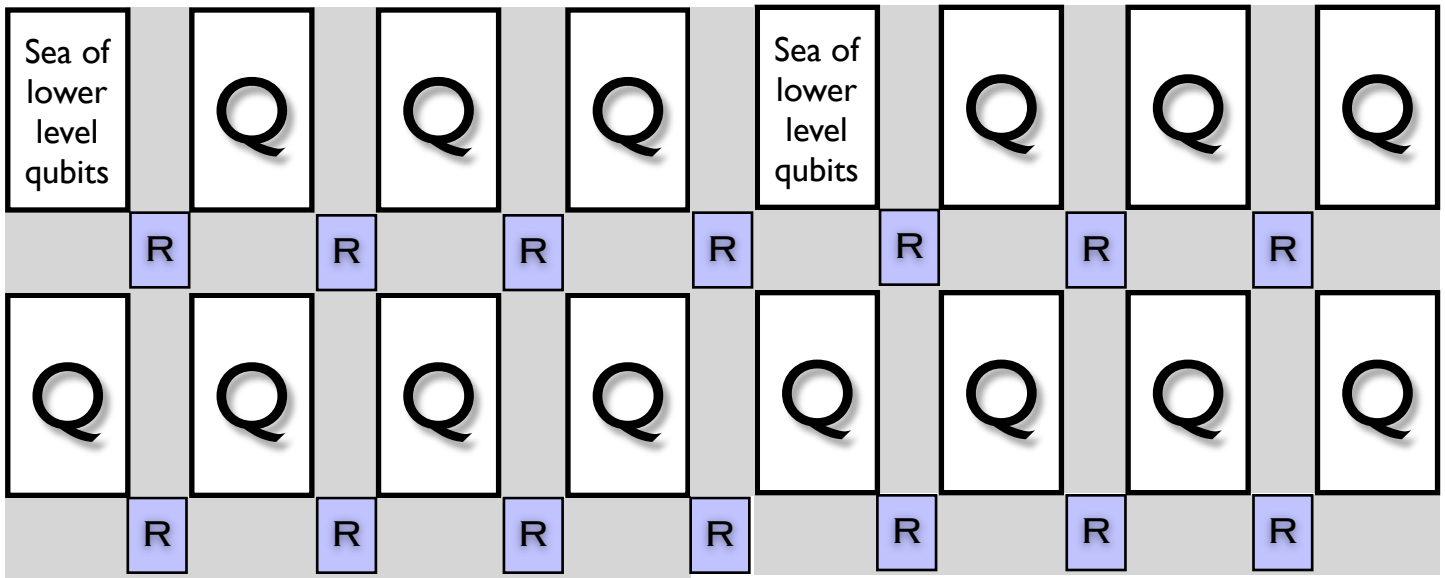
Overview of quantum error correction codes.

Specialization into memory and compute regions.

Improving performance.

Results and discussion.

Quantum Logic Array



Conventional wisdom: Max. parallelism necessary to minimize computation time and reduce prob. of failure.

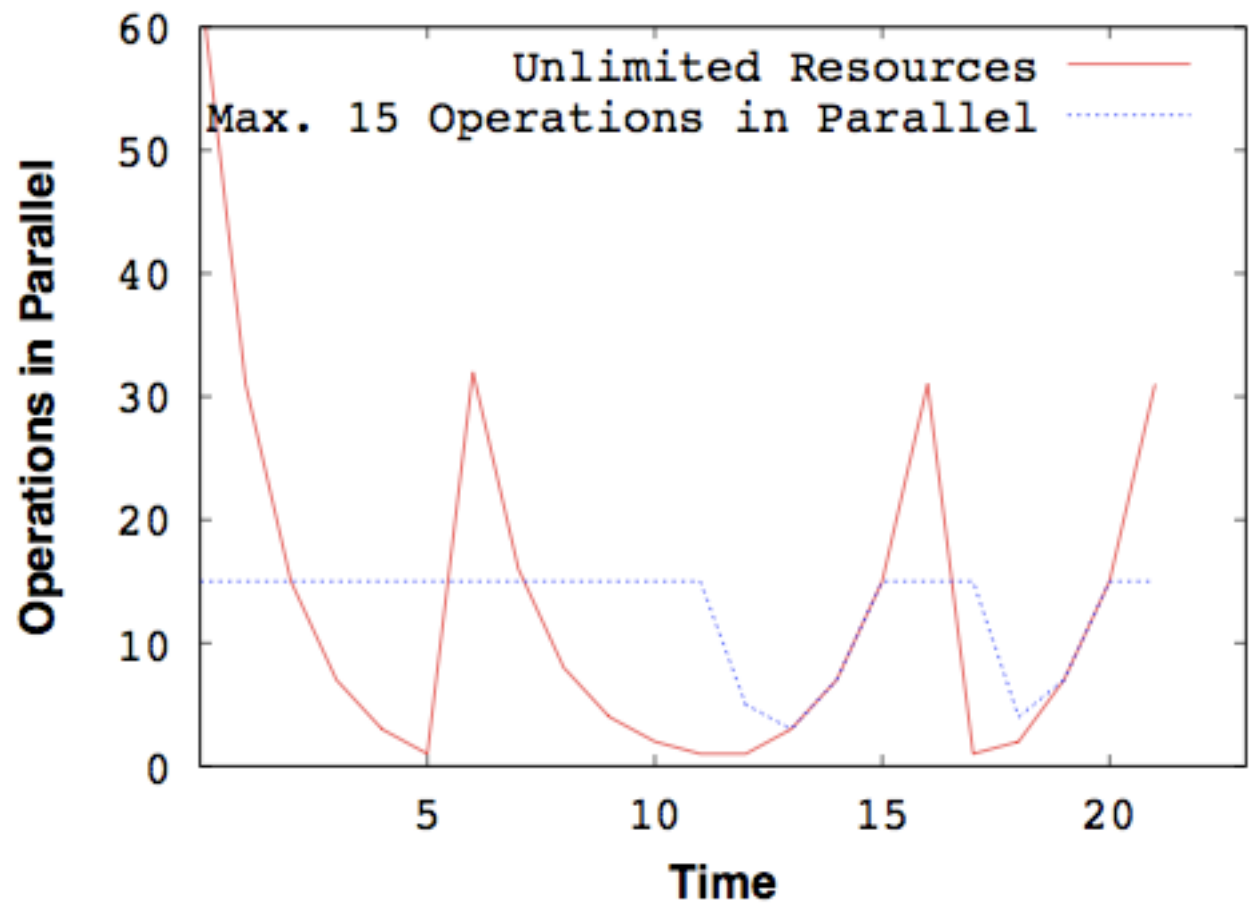
Modular Exponentiation

Shor's quantum algorithm to find factors of very large numbers yields exponential speedup over classical algorithms.

Modular exponentiation is the most compute intensive part of Shor's factoring algorithm.

Primary component: Draper carry-lookahead adder (quantum version of the classical adder).

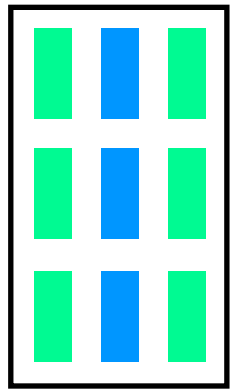
App. Constrained Parallelism



Create slower but denser memory region and faster but sparse compute region.

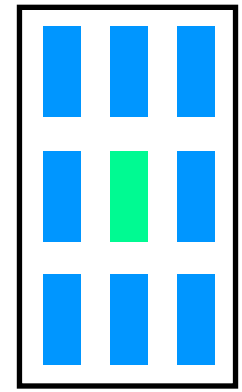
Specialization

Compute Region



Ancilla : Data
2 : 1

Memory Region



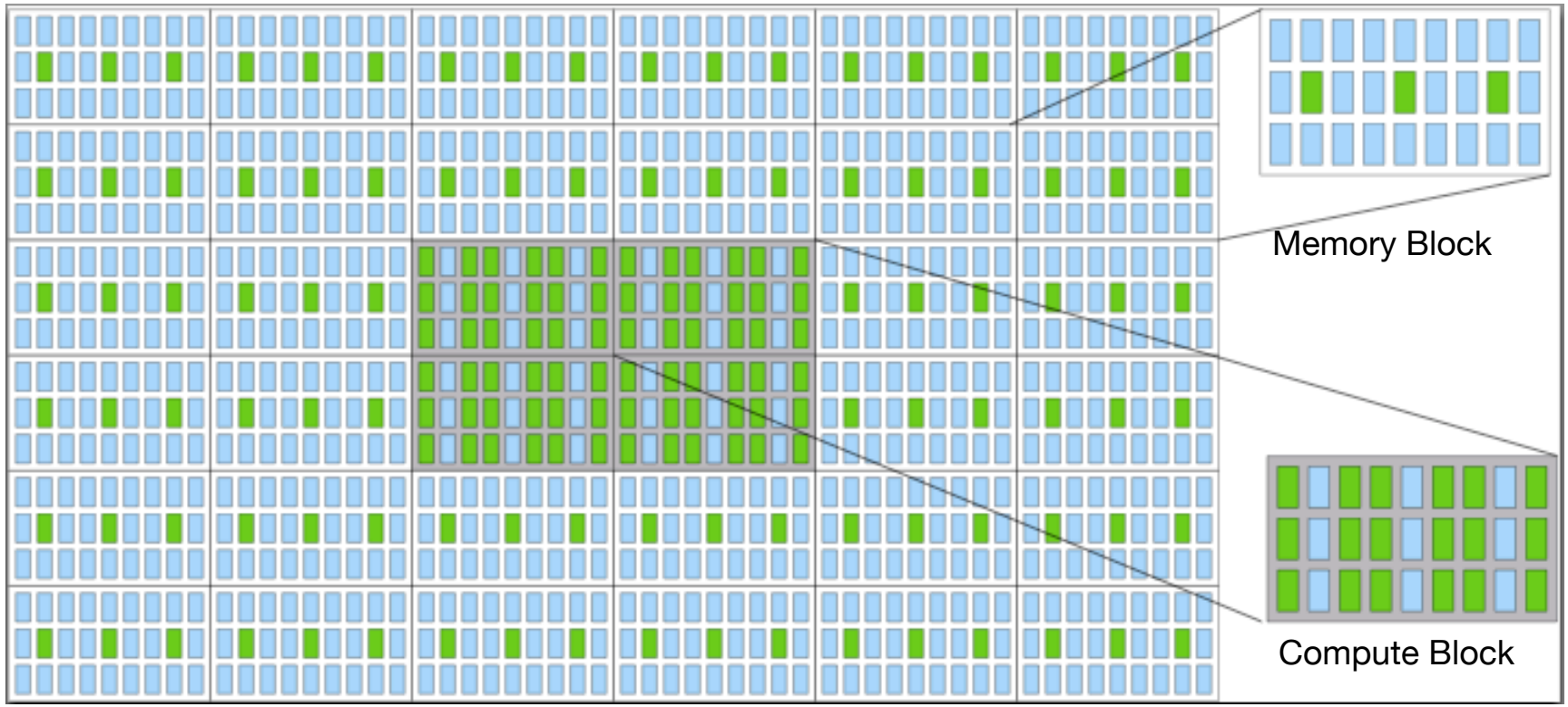
Ancilla : Data
1 : 8

Logical data qubits

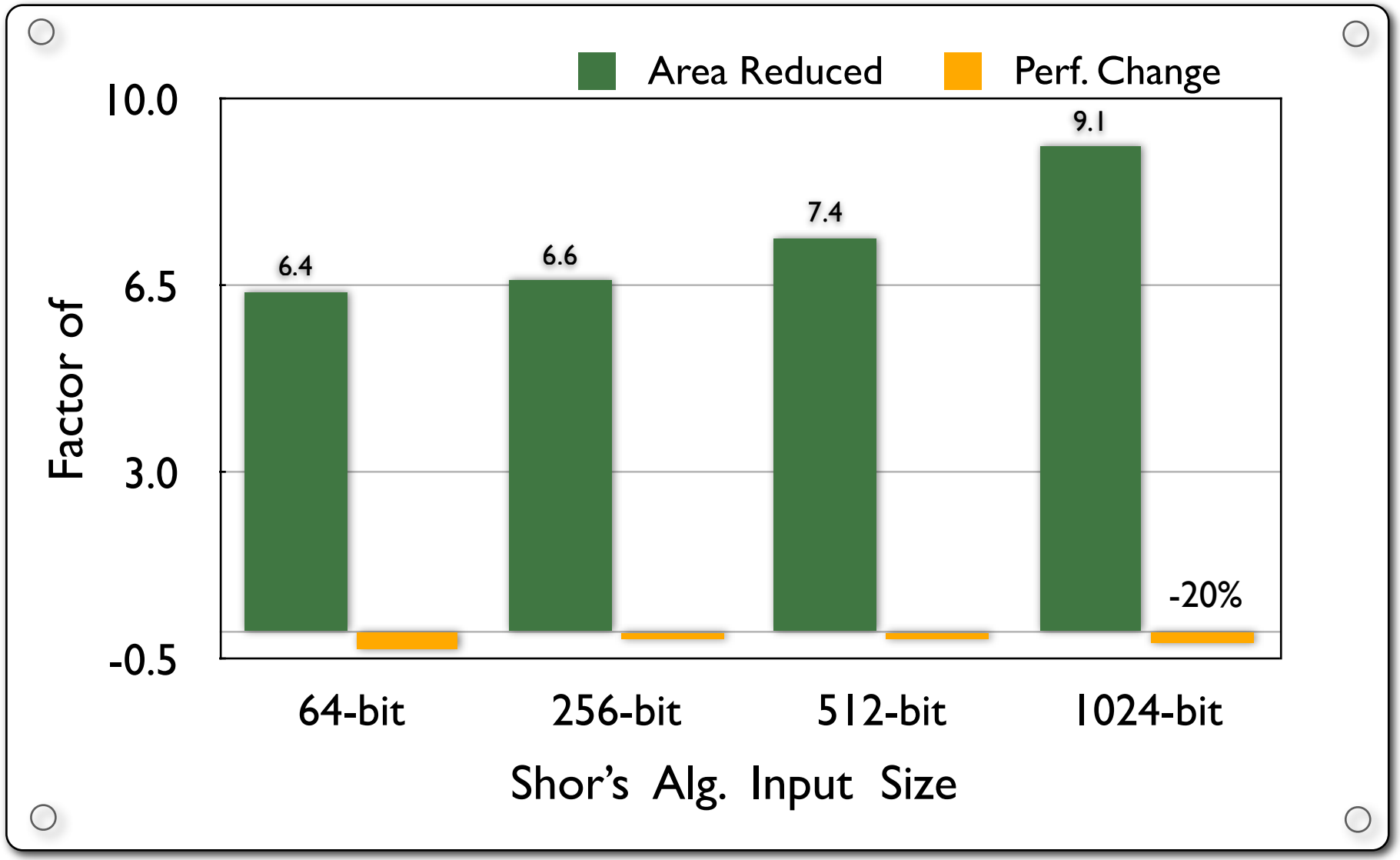
Logical ancilla qubits

An ion when idle has a lifetime of ~10 sec

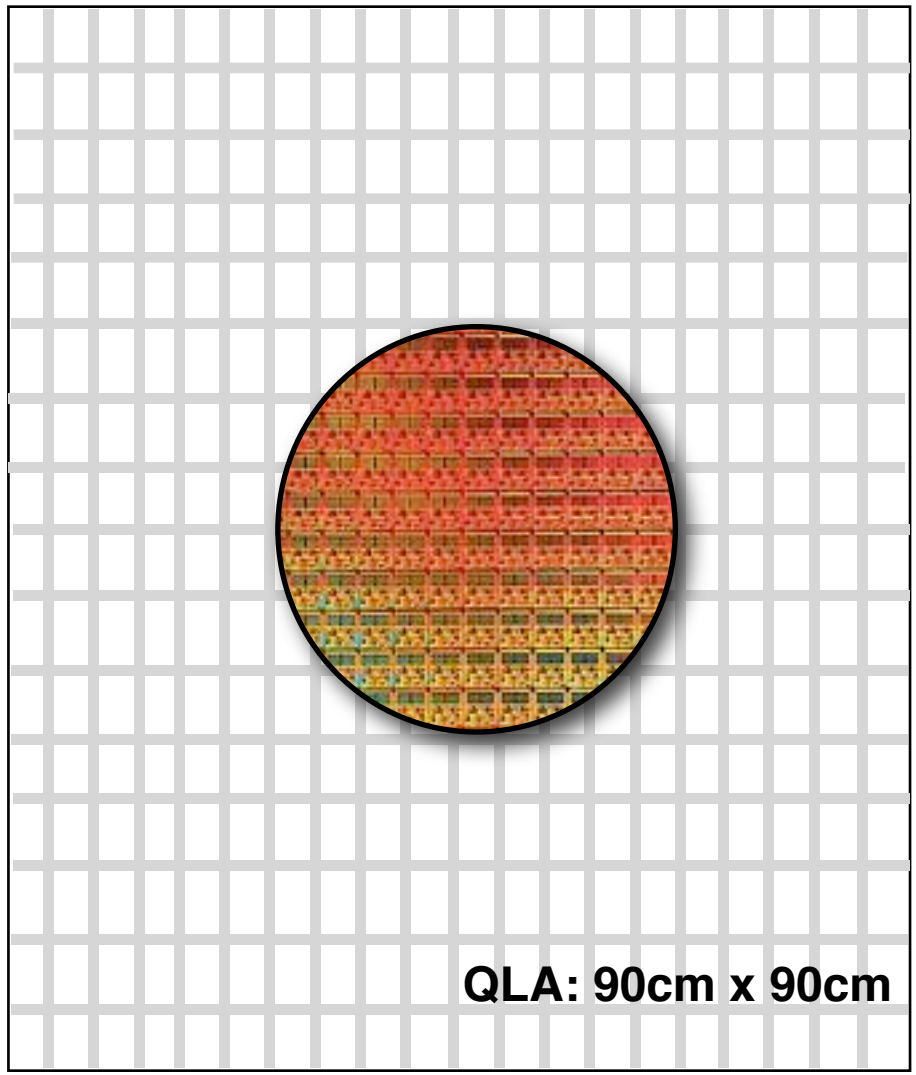
CQLA: Compressed QLA

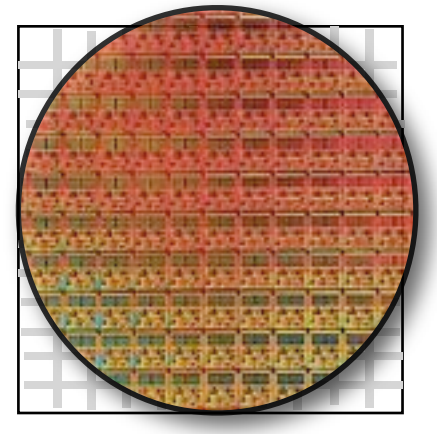


Area Reduction



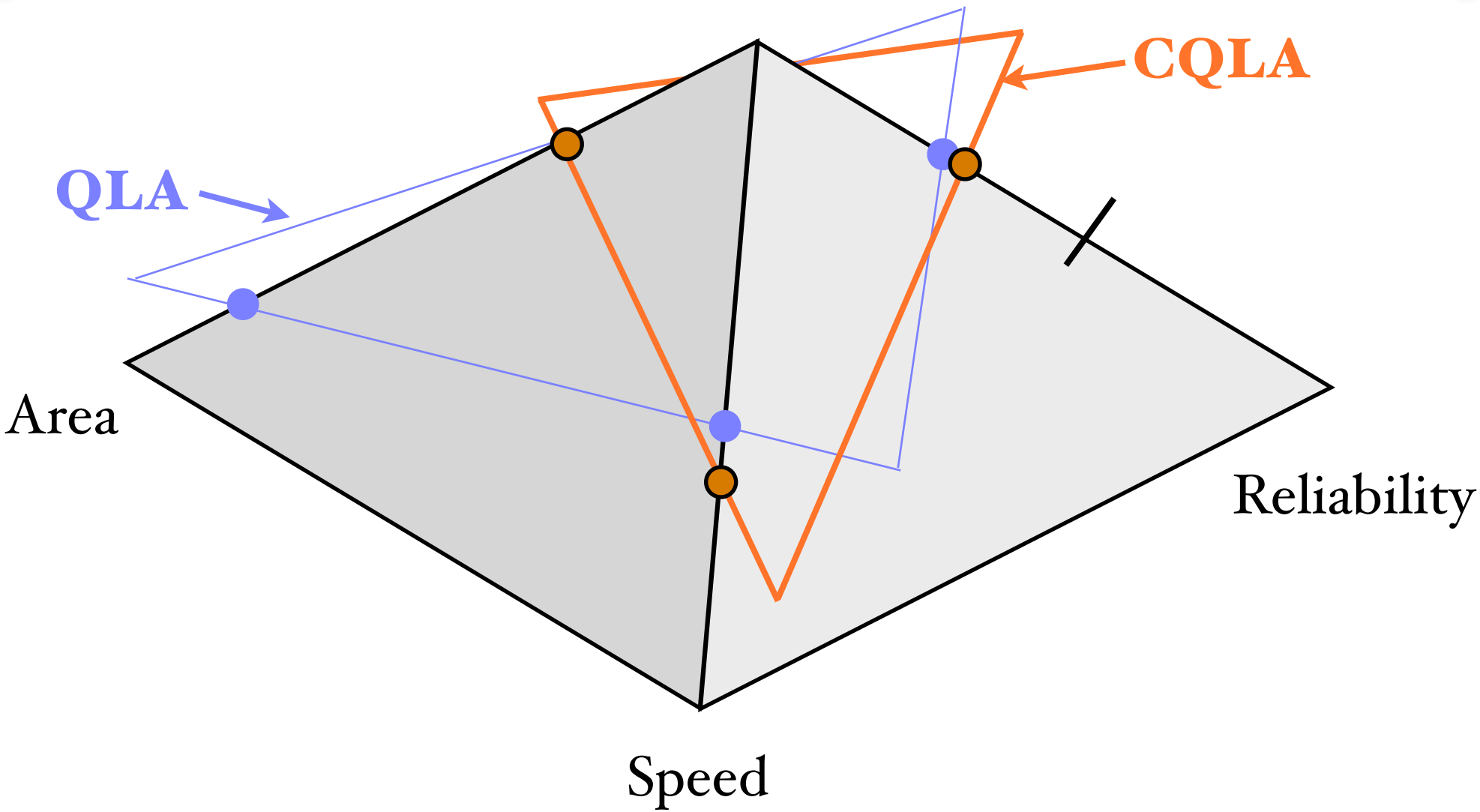
CQLA: Reduced Size





CQLA: 28cm x 28cm

Design Pyramid: CQLA



Outline

Background and prior work.

Overview of quantum error correction codes.

Specialization into memory and compute regions.

Improving performance.

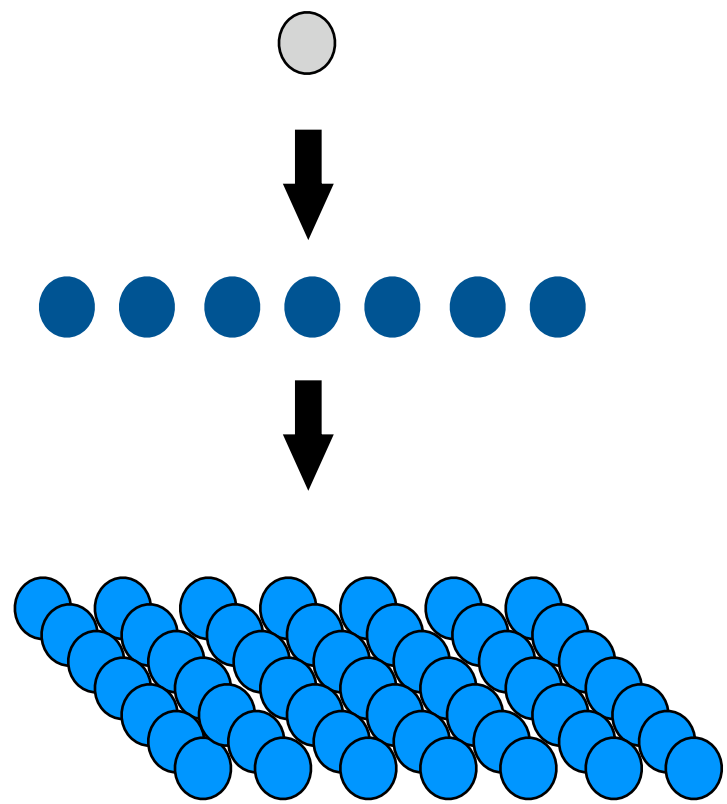
Results and discussion.

Concatenated Codes

1 logical qubit

Level 1:
7 physical qubits

Level 2:
49 physical qubits



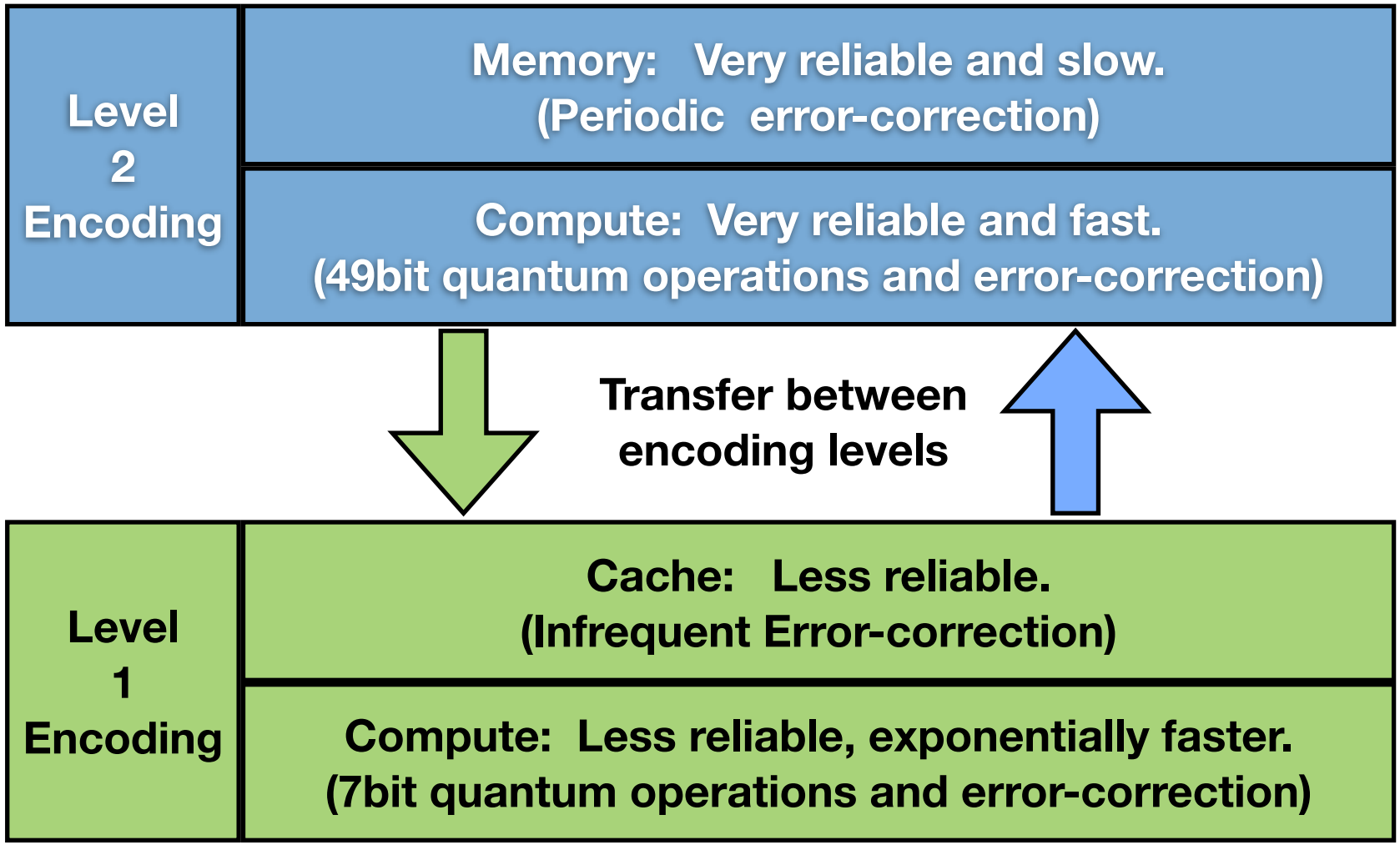
Reliability increases doubly exponentially.

Exponentially slower.

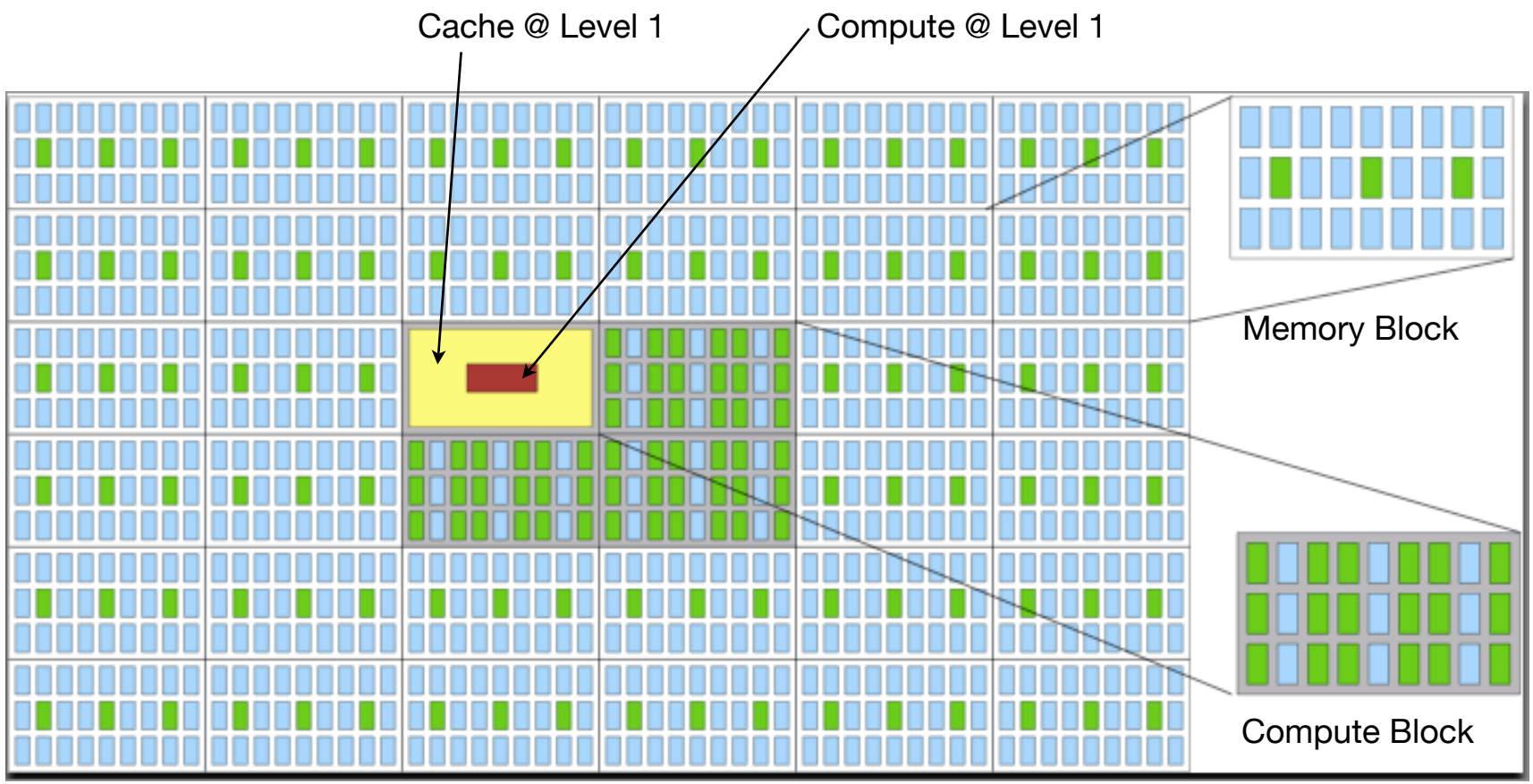
Exponentially greater resources.

Concatenated Steane Code

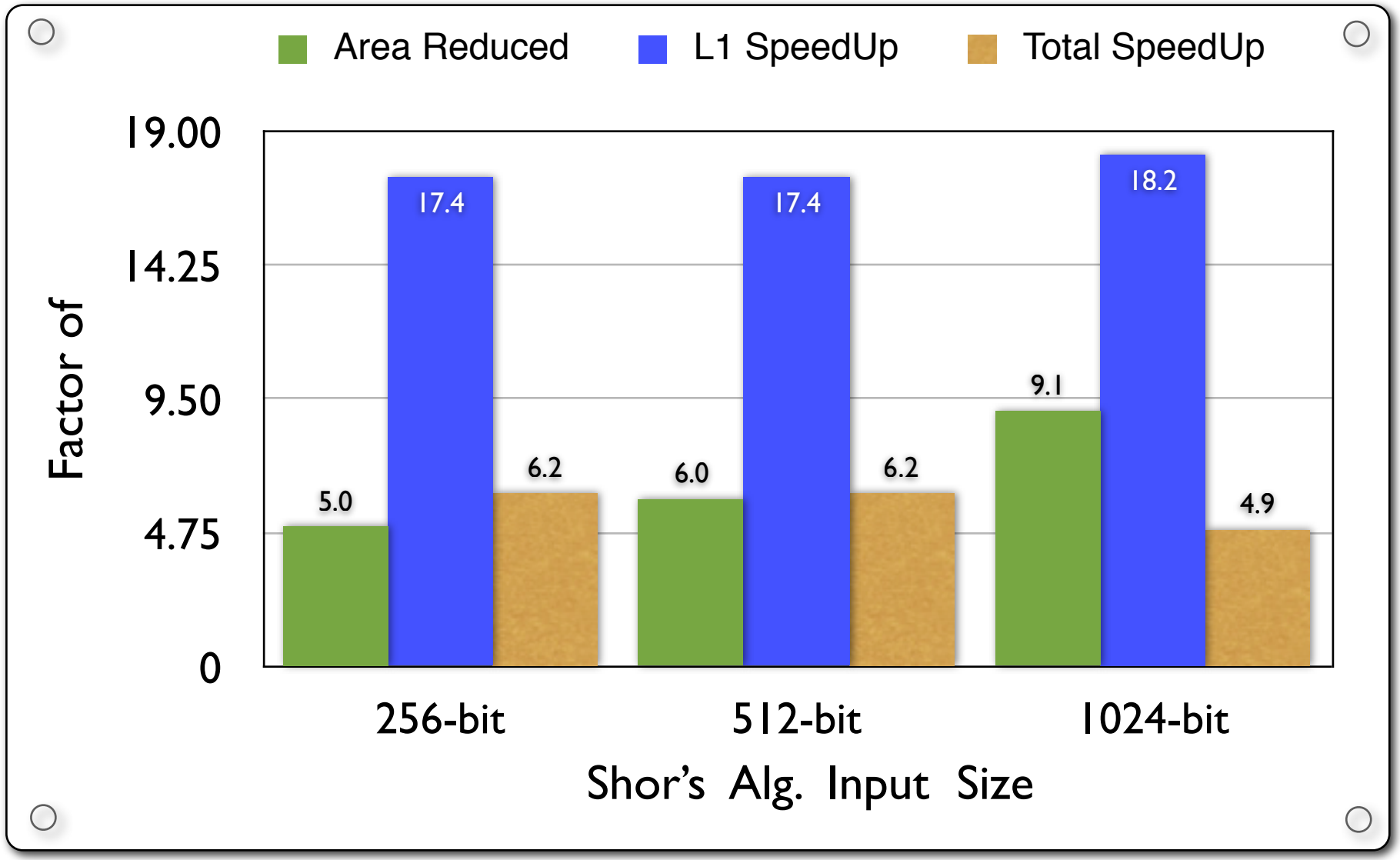
Compute @ Level 1



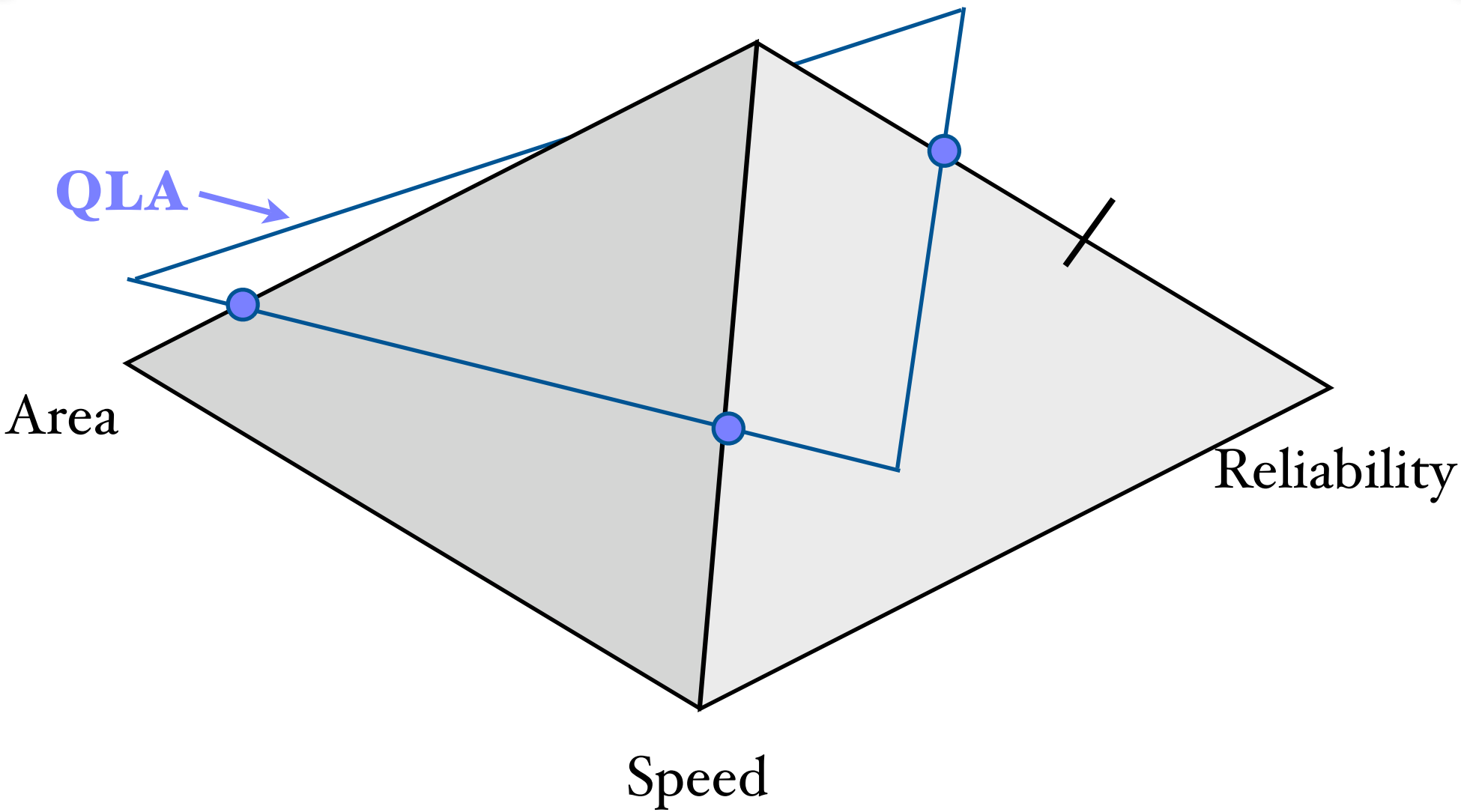
Faster CQLA



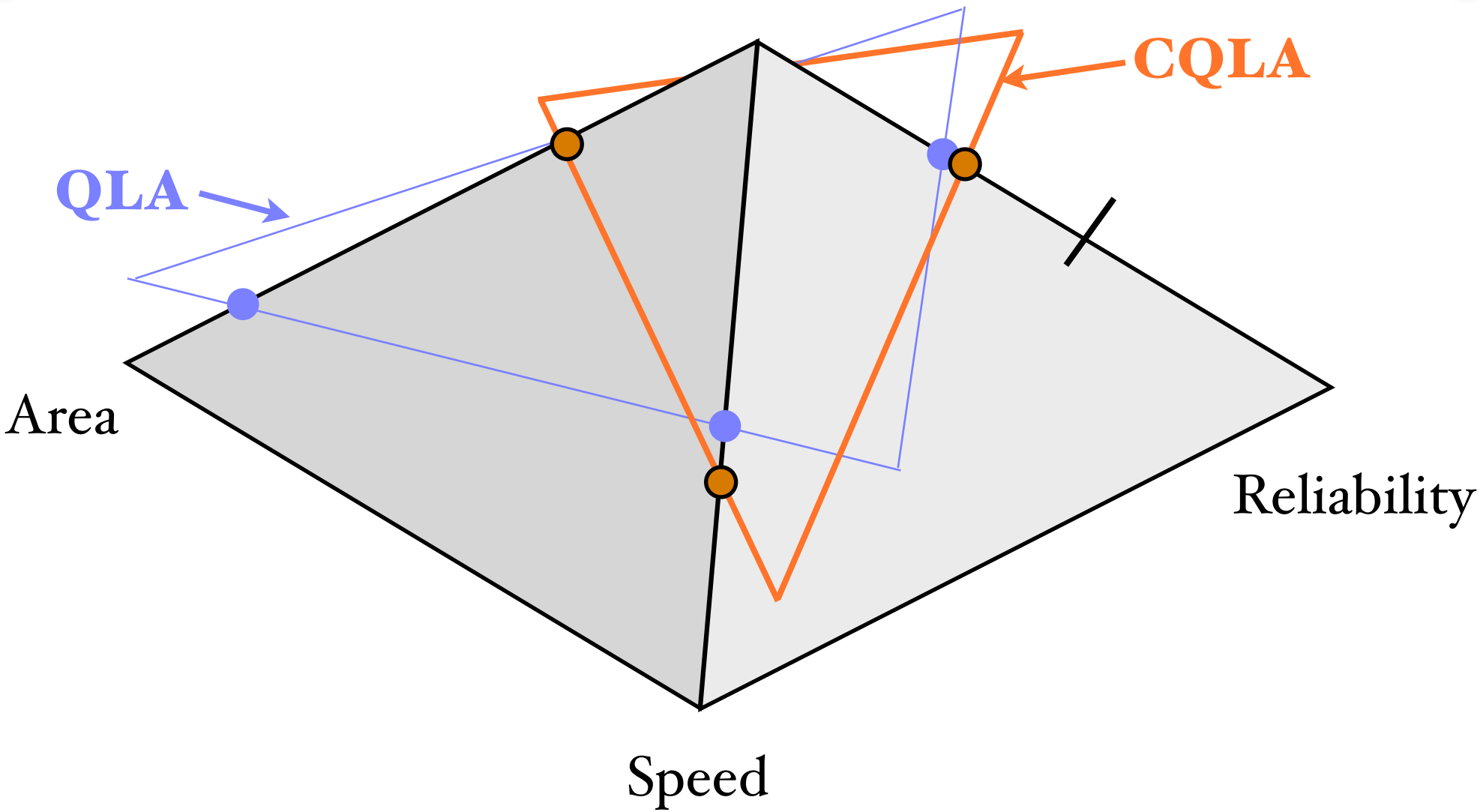
Overall Results



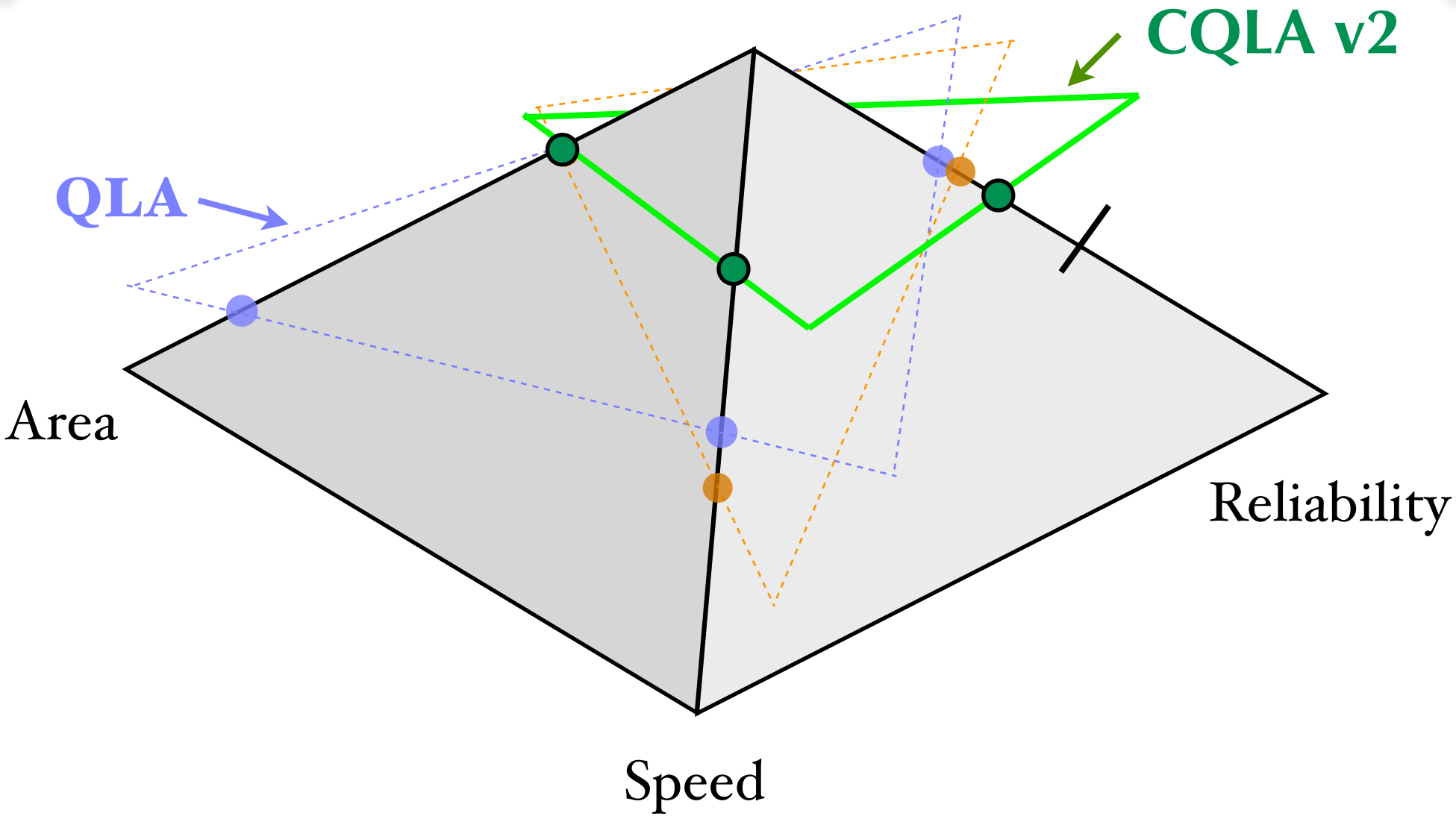
Design Pyramid: QLA



Design Pyramid: CQLA



Design Pyramid: CQLA v2



Discussion

Parallelism in quantum computing constrained by applications.

Different scheduling mechanisms of quantum operations.

Introduced a memory hierarchy for quantum computers.

Area reduced factor of 9 and speedup of factor of 4.

Discussion - 2

Even better results using the Bacon-Shor quantum error-correction code.

- Area reduced by a factor of 13.
- Speedup of factor of 8.

Details of transfer networks to enable change in encodings.

Future Work

Limited control signals: Incorporate studies of laser resources and laser power.

Incorporating fault tolerance into compiler optimization: Compiler techniques to reduce error-correction costs.

¿ Questions ?

Project webpage: http://aar.cs.ucdavis.edu/qarc

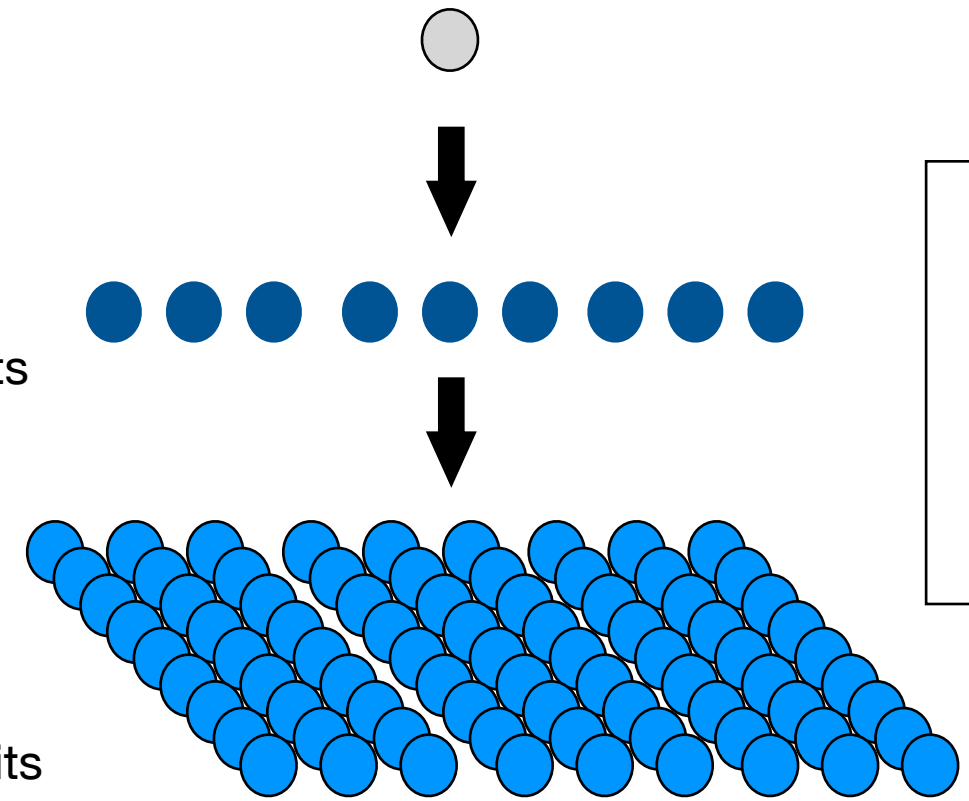
Your questions...

Overall Results

Par Xfer	Adder Size	L1 SpeedUp	L2 SpeedUp	Adder SpeedUp	Area Reduced
Steane [[7, 1, 3]] Code					
10	256	17.4	0.9	6.2	5.0
	512	17.4	0.9	6.3	6.0
	1024	18.2	0.8	**4.9**	**9.1**
5	256	10.4	0.9	4.0	5.0
	512	10.4	0.9	4.0	6.0
	1024	10.9	0.8	**2.9**	**9.1**
Bacon-Shor [[9, 1, 3]] Code					
10	256	9.6	1.5	5.9	7.4
	512	9.6	2.3	8.8	8.8
	1024	10.1	2.0	**8.1**	**13.4**
5	256	5.2	1.5	3.7	7.4
	512	5.2	2.3	5.4	8.8
	1024	5.4	2.0	**4.9**	**13.4**

Concatenated Codes

1 logical qubit



Level 1:
9 physical qubits

Level 2:
81 physical qubits

Reliability increases doubly exponentially.

Exponentially slower.

Exponentially greater resources.

Improve Performance

Let memory remain at Level 2 encoding.

Compute at Level 1 encoding.

Drawbacks: Reliability degrades.
Transfer between Level 1 and Level 2 is very expensive.

Use a cache to alleviate transfer costs.

Size Reduction

Input Size	Compute Blocks	Area Reduced (Factor of)	Speed Up
64-bit	9	6.4	0.7
	16	3.7	0.9
256-bit	36	6.6	0.9
	49	5.0	0.9
512-bit	64	7.4	0.9
	81	6.0	0.9
1024-bit	**100**	**9.1**	**0.8**
	121	7.8	0.9

Overall Results

Parallel Transfers	Input Size	L1 Speed Up	L2 Speed Up	Total Speed Up	Area Reduced (Factor of)
5	256	10.4	0.9	4.0	5.0
	512	10.4	0.9	4.0	6.0
	1024	10.9	0.8	2.9	9.1
10	256	17.4	0.9	6.2	5.0
	512	17.4	0.9	6.3	6.0
	1024	**18.2**	**0.8**	**4.9**	**9.1**