

Integrating quantum computing into high-performance computing systems

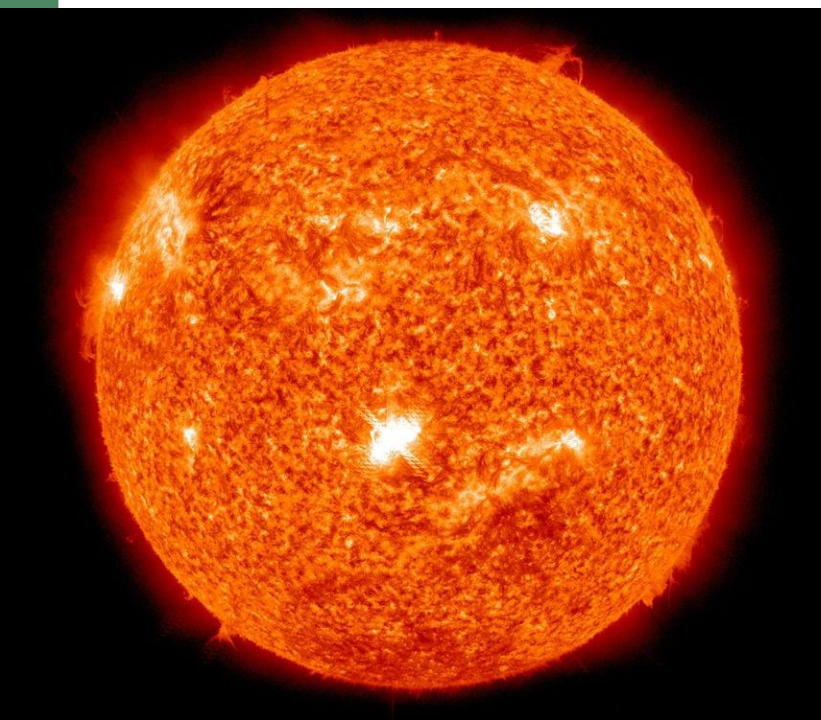
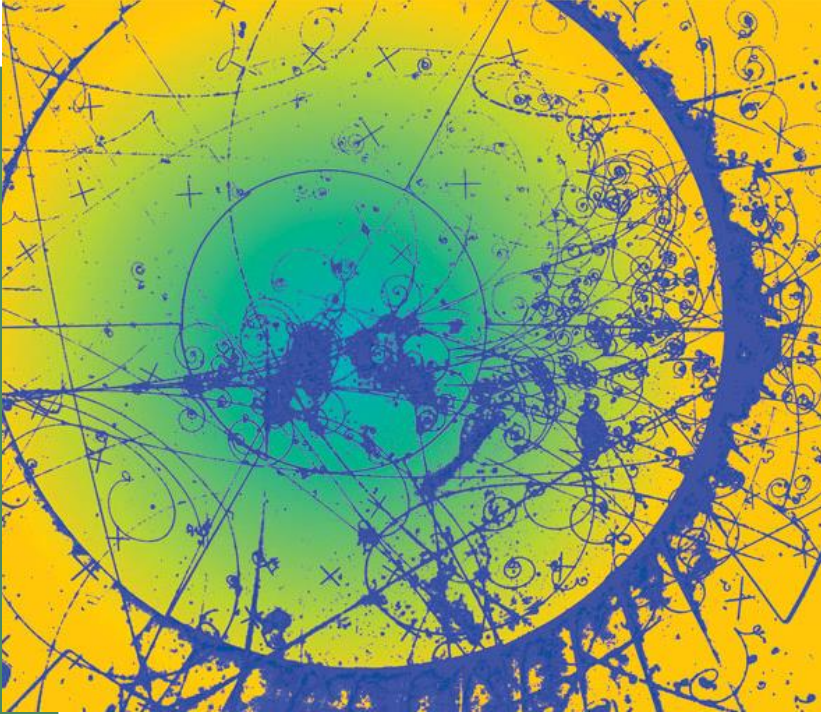
Travis Humble
Quantum Science Center
Oak Ridge National Laboratory
humblets@ornl.gov

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

This work is also supported by the Quantum Science Center, a National Quantum Information Science Research Center of the U.S. Department of Energy (DOE).

What you should learn from this presentation

- Motivation for integrating quantum computing with high-performance computing, aka, QHPC
- Terminology and techniques for evaluating QHPC
- Priority research areas to advance QHPC design and development



Oak Ridge National Laboratory



Oak Ridge National Laboratory

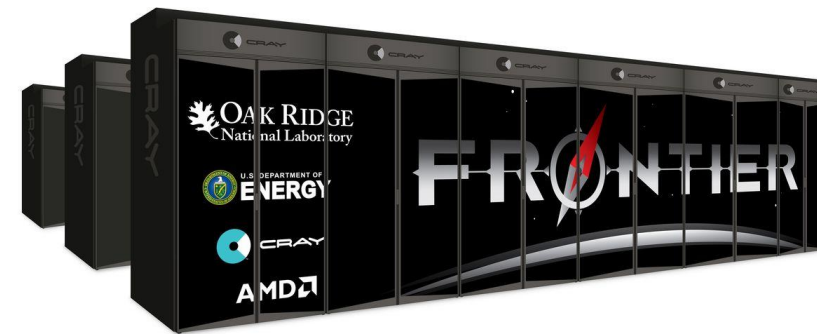




Cray XK7, 18,688 Nodes
16-core AMD Interlagos + K20X
17 PFLOPS, 8.2 MW,
#1 TOP500 (2012)



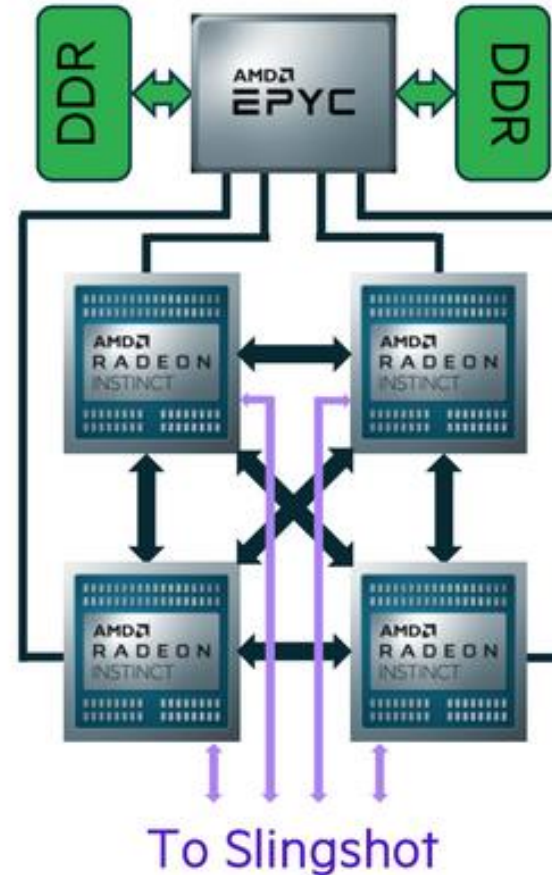
IBM, 4,600 Nodes
2 Power9 + 6 NVidia Volta
144 PFLOPS, 9.7 MW,
#1 TOP500 (2018)



CRAY EX, 9,408 Nodes
1 AMD EPYC + 4 Radeon Instinct
1.1 EXAFLOPS, 21.1 MW
#1 TOP500 (2022)

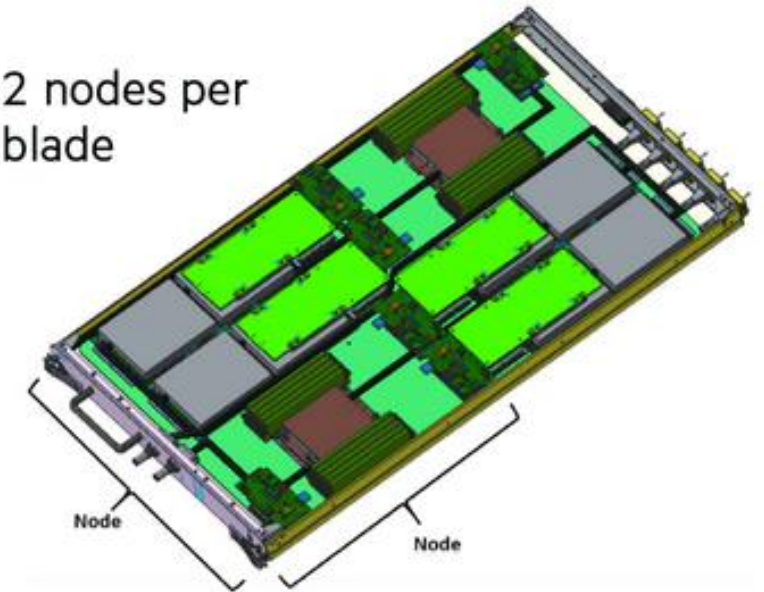
Frontier Node Spec

- 9,472 AMD Epyc 7453s "Trento" 64 core 2 GHz CPUs (606,208 cores)
- 37,888 Radeon Instinct MI250X GPUs (8,335,360 cores).
- Performs double precision operations at the same speed as single precision.
- 62.68 gigaflops/watt.



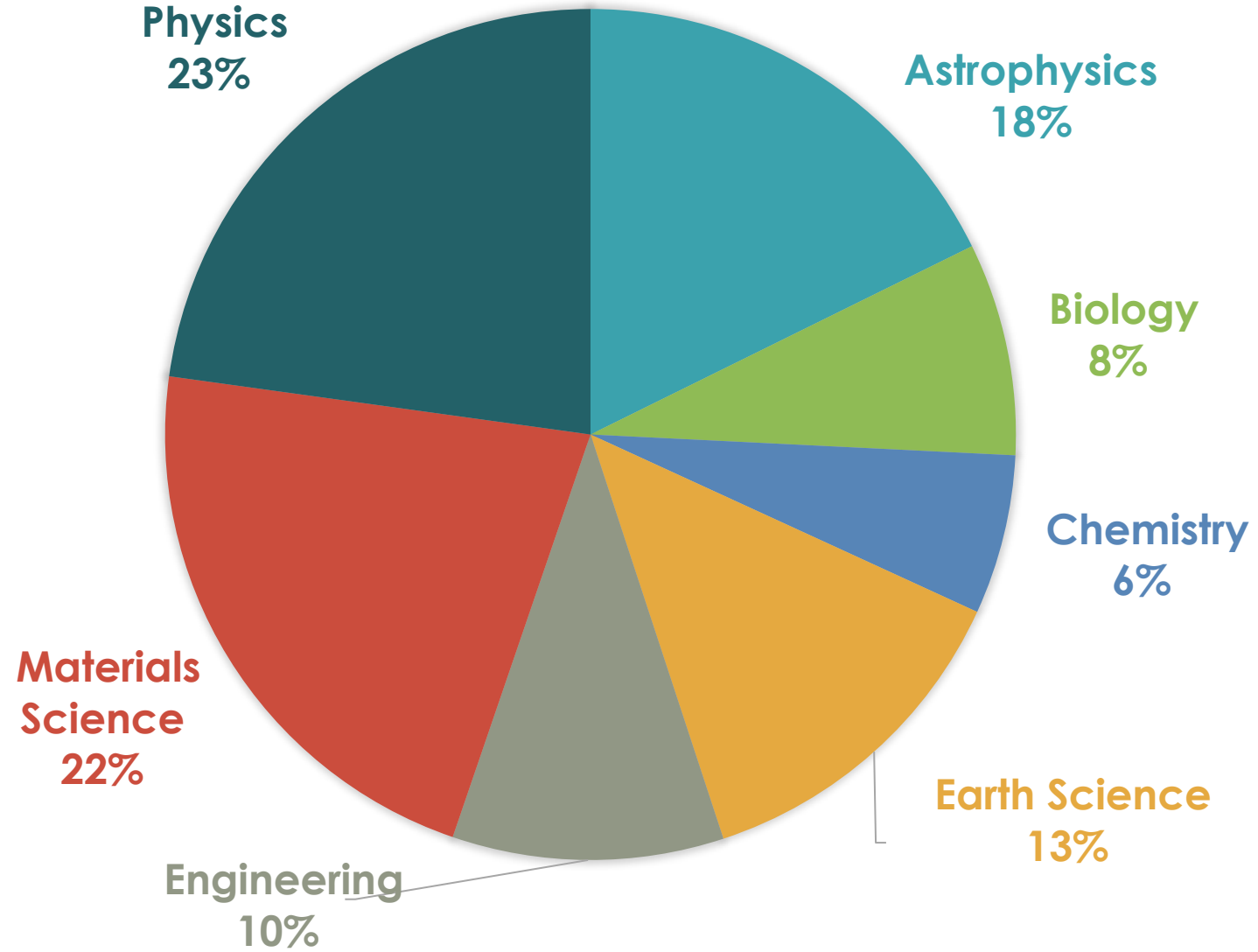
AMD GPU
(ORNL)

2 nodes per
blade



COPYRIGHT 2020 HPE





N.B. There are limits on the speedups for parallel processing.

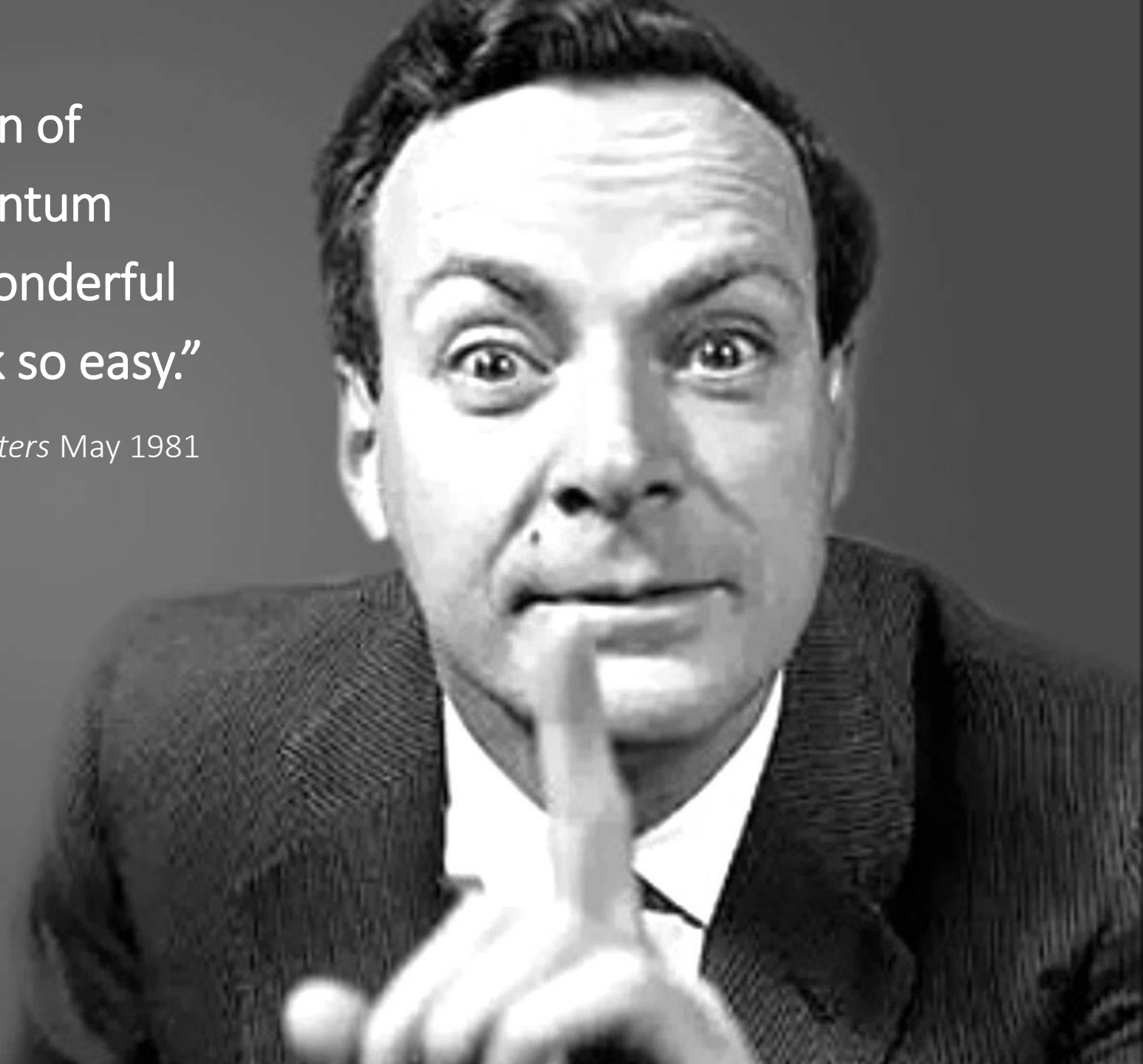
- **Amdahl's law** says that the benefits from parallel processing are limited by the portion of the program that can be parallelized.
 - Let p be the fraction of the processing time that can be parallelized.
 - Let s be the speedup from parallelizing the process, eg, number of parallel operations.
 - $S(s)$ is then the relative improvement in performance (speed).
- What happens as s goes to infinity?

Amdahl's law

$$S(s) = \frac{1}{(1 - p) + \frac{p}{s}}$$

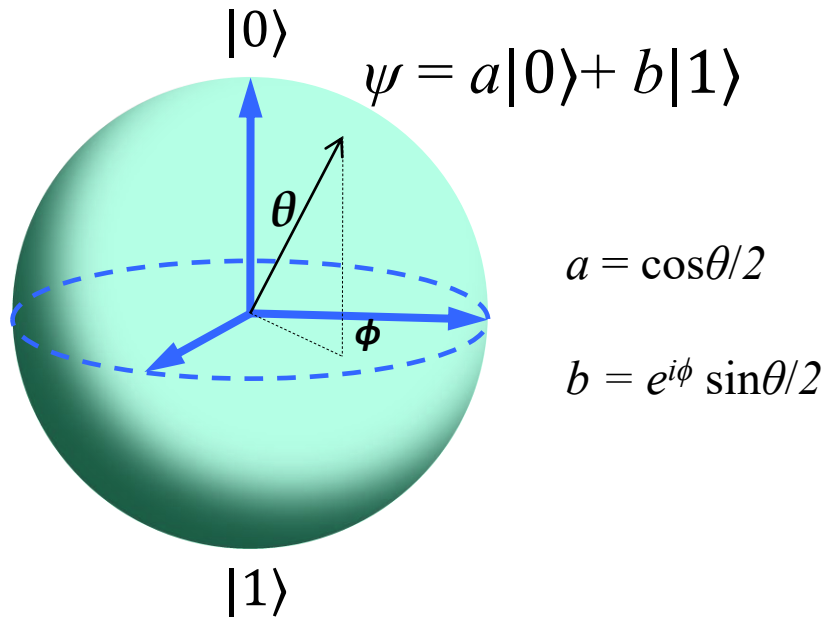
“If you want to make a simulation of nature, you'd better make it quantum mechanical, and by golly it's a wonderful problem, because it doesn't look so easy.”

Richard Feynman, *Simulating Physics with Computers* May 1981



Basic Requirements of a Quantum Computer

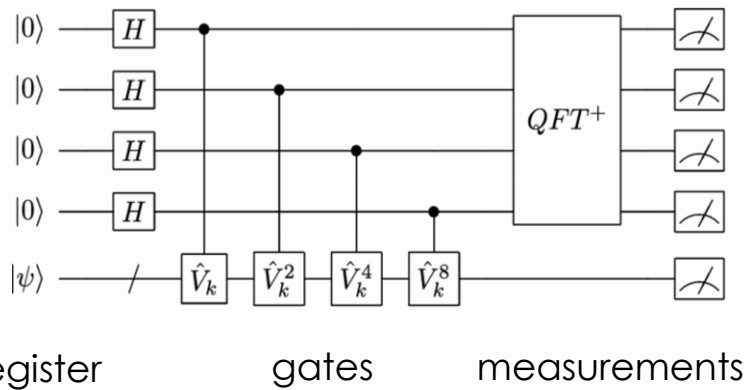
A qubit



- A scalable system of qubits
- The ability to initialize qubits in fiducial states
- A universal set of quantum gates
- Decoherence times longer than gate operation times
- A qubit-specific measurement capability

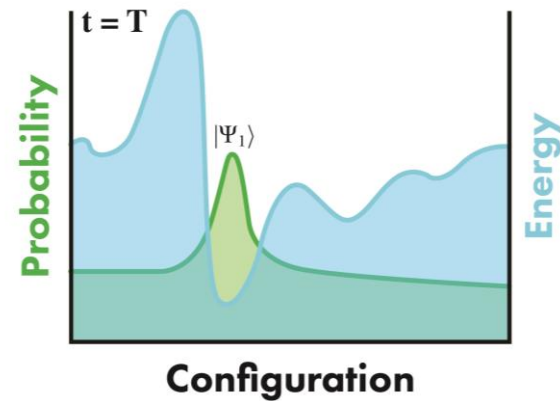
Digital Computation

- Fast, discrete transformations of the computational state
- Easily translated to Boolean logic
- Universality defined by discrete gate set



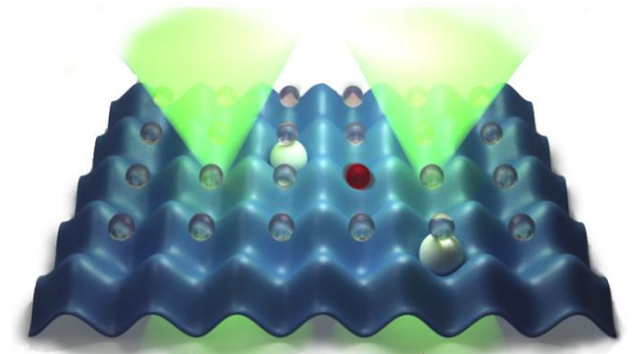
Adiabatic Computation

- Slow, continuous transformations of the computational state
- Easily translated to optimization
- Universality defined by Hamiltonian control



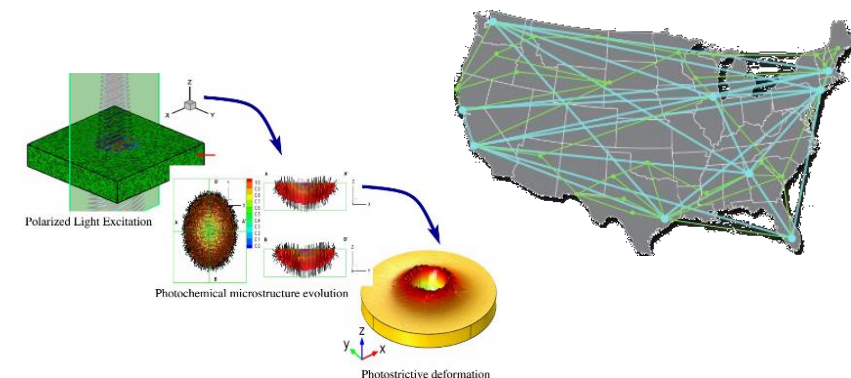
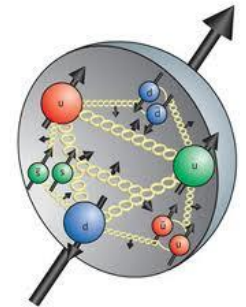
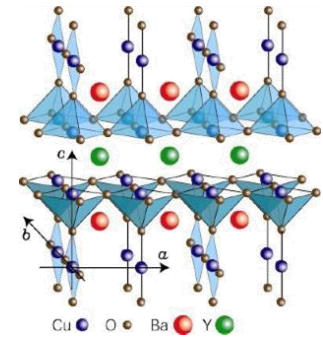
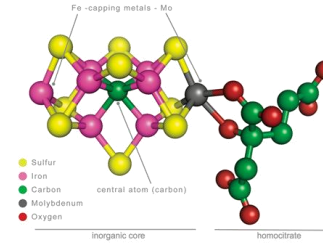
Analog Computation

- Fast, continuous transformation of the computational state
- Easily translated to quantum simulation
- Universality defined by Hamiltonian control



Scientific Computing with Quantum Computers

- Algorithms in the quantum computing model have been found to take fewer steps to solve problems
 - Quantum Simulation
 - Partition Functions
 - Discrete Optimization
 - Machine Learning
 - Factoring
 - Unstructured Search
 - Eigensystems
 - Linear Systems
- Several physical domains motivate quantum computing as a paradigm for scientific computing
 - High-energy Physics
 - Materials Science
 - Chemistry
 - Biological Systems
 - Artificial Intelligence
 - Data Analytics
 - Planning and Routing
 - Verification and Validation

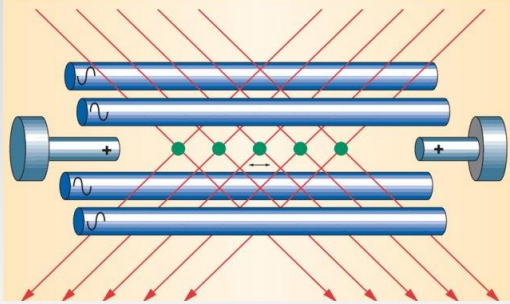


Working Terminology



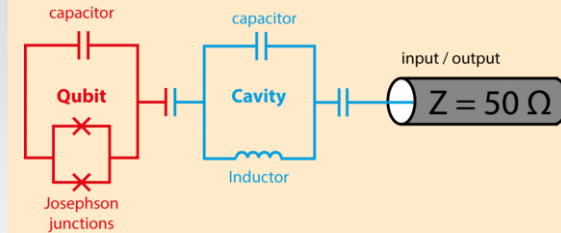
- **classical algorithm:** a defined sequence of transformations to solve a specific problem by manipulating a classical logic state.
- **quantum algorithm:** a defined sequence of transformations to solve a specific problem by manipulating a quantum state.
- **hybrid algorithm:** a combination of a quantum and classical algorithm.
- **quantum circuit:** a visual schematic to represent a sequence of transformations acting on a quantum state.
- **quantum program:** a sequence of ordered instructions for a quantum computer.
- **quantum application:** a quantum program to solve a specific problem.
- **quantum advantage:** an improvement in performance of a quantum application relative to a classical baseline.

Atoms & Ions



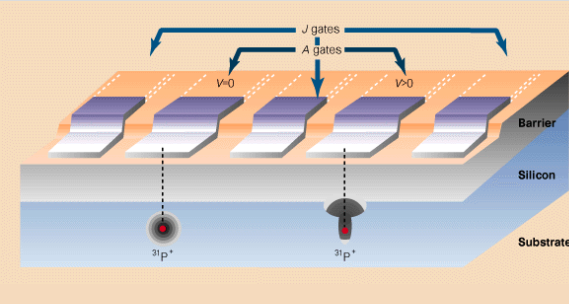
- Pure, reproducible atoms
- Synergy with quantum clocks and optical engineering
- MHZ energy scales

Superconductors



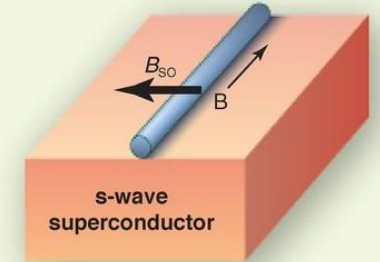
- Synthesis of purified superconducting materials
- Known materials and fabrications methods
- Integrated μm circuit
- GHZ energy scales

Semiconductors



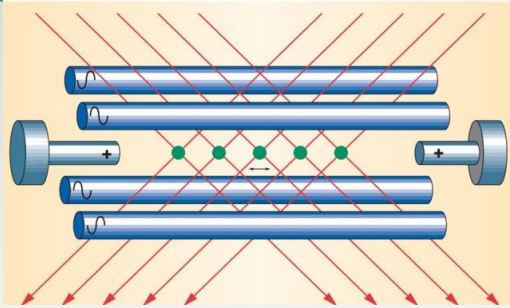
- Synthesis of purified semiconducting materials, dopants
- Known materials and fabrications methods
- Integrated nm circuit
- GHZ energy scales

Topological Materials

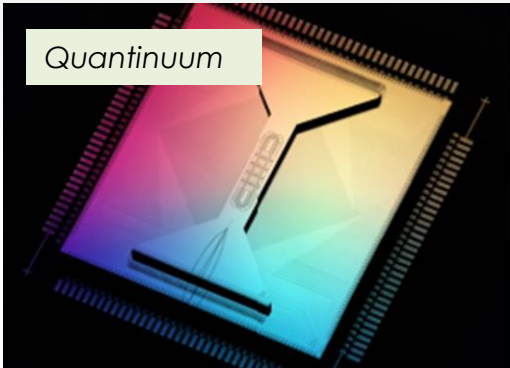


- Synthesis of topological quantum materials
- New fabrication techniques
- Integrated nm circuit
- GHZ energy scales

Atoms & Ions



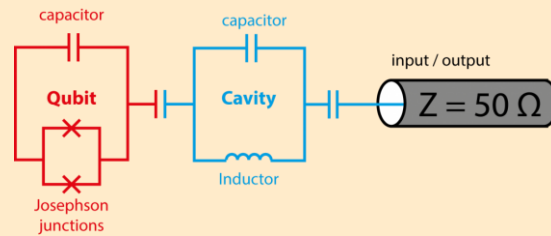
Quantinuum



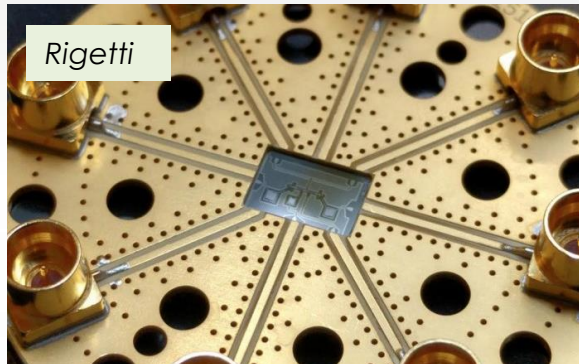
IonQ



Superconductors



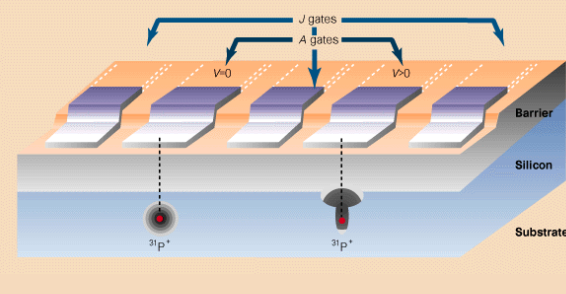
Rigetti



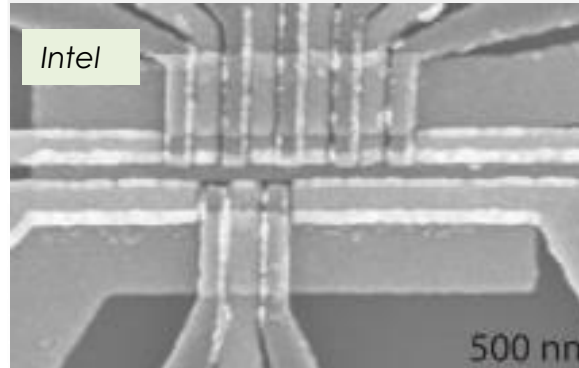
IBM



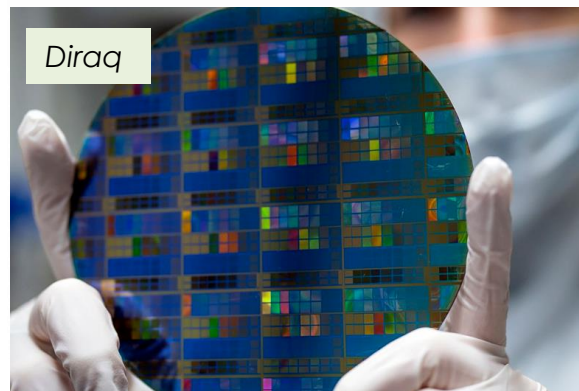
Semiconductors



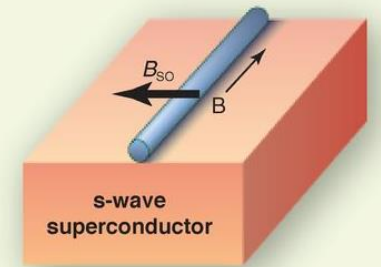
Intel



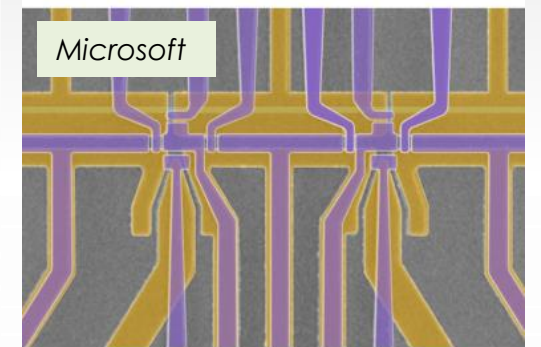
Dirac



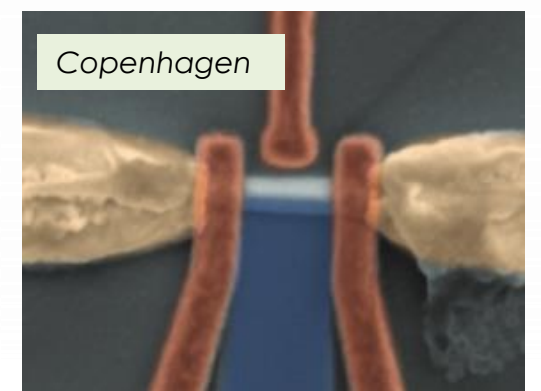
Topological Materials



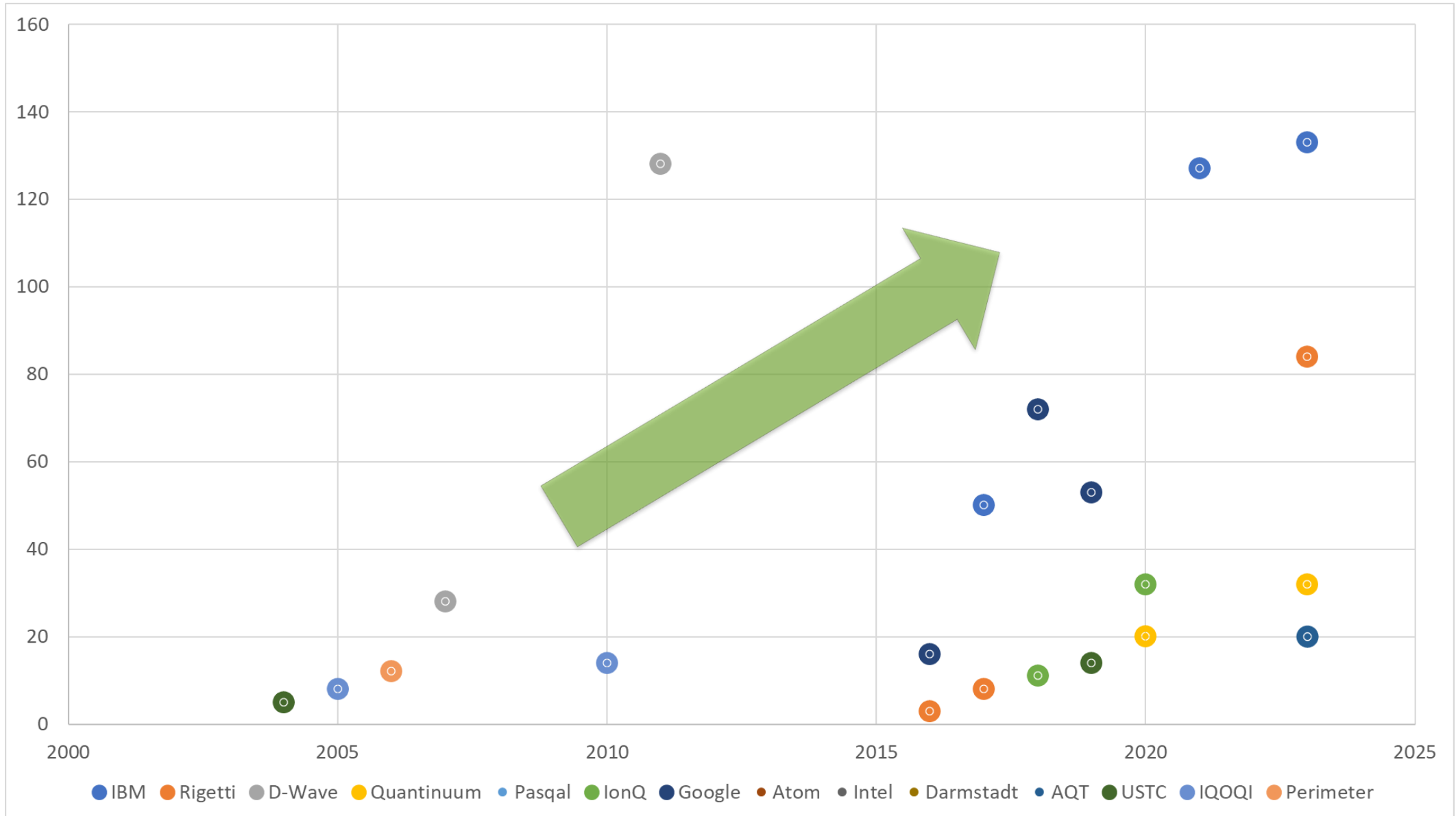
Microsoft



Copenhagen



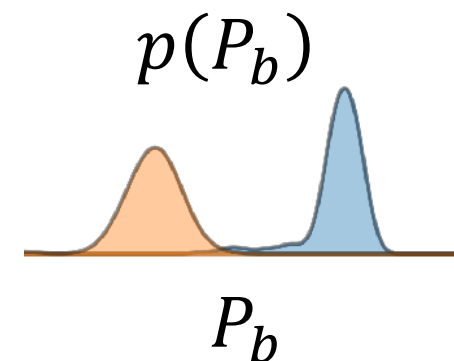
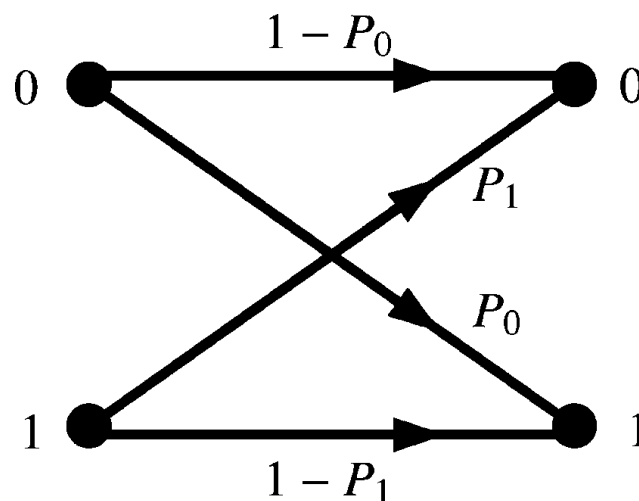
A scalable system of qubits



The ability to initialize qubits in fiducial states

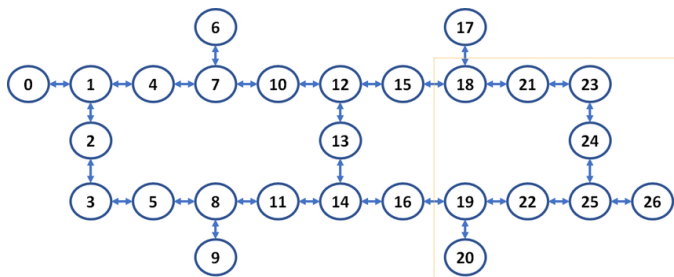
- Noise corrupts information encoded in the quantum register
 - This is caused by coupling of the register to the environment which leads to decoherence, decay, and leakage.

A binary asymmetric model for readout error, where outcome b has probability P_b to flip.



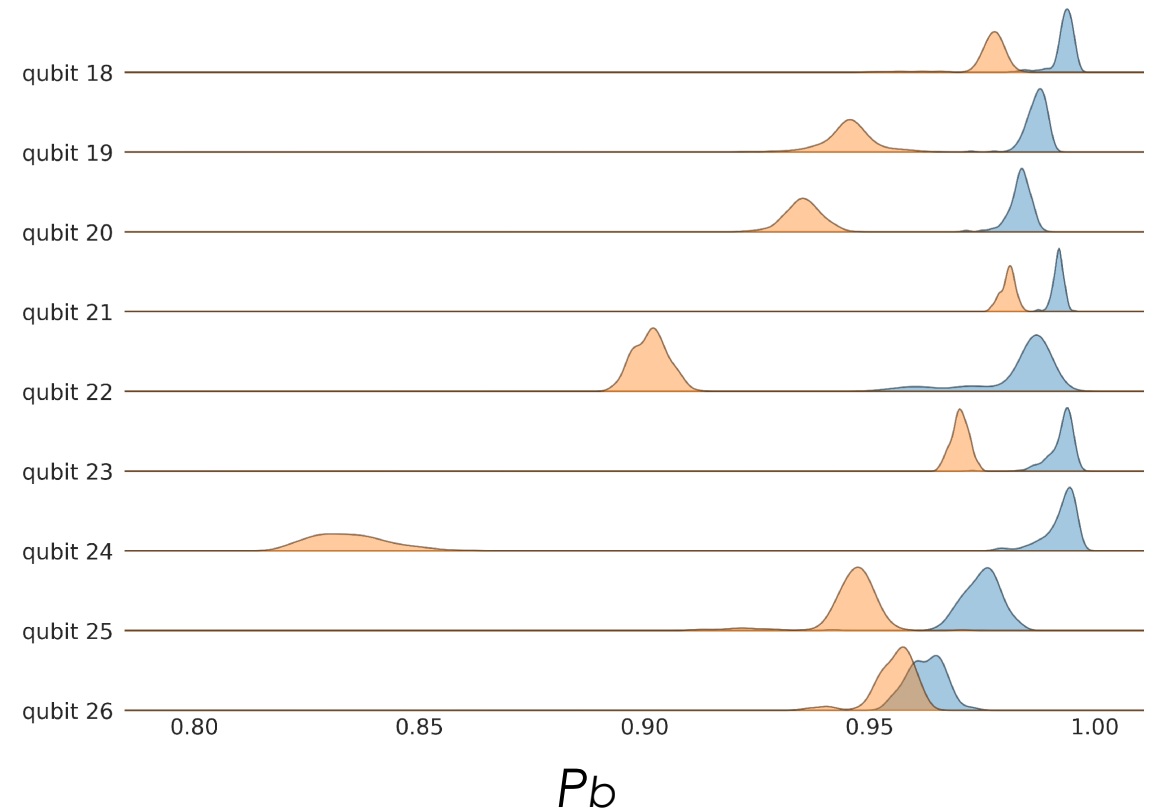
The ability to initialize qubits in fiducial states

- Noise corrupts information encoded in the quantum register
 - This is caused by coupling of the register to the environment which leads to decoherence, decay, and leakage.



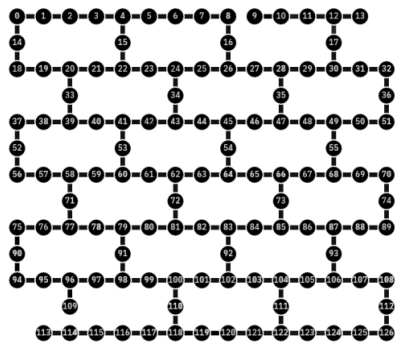
ibmq_toronto

Distribution of readout error by register element



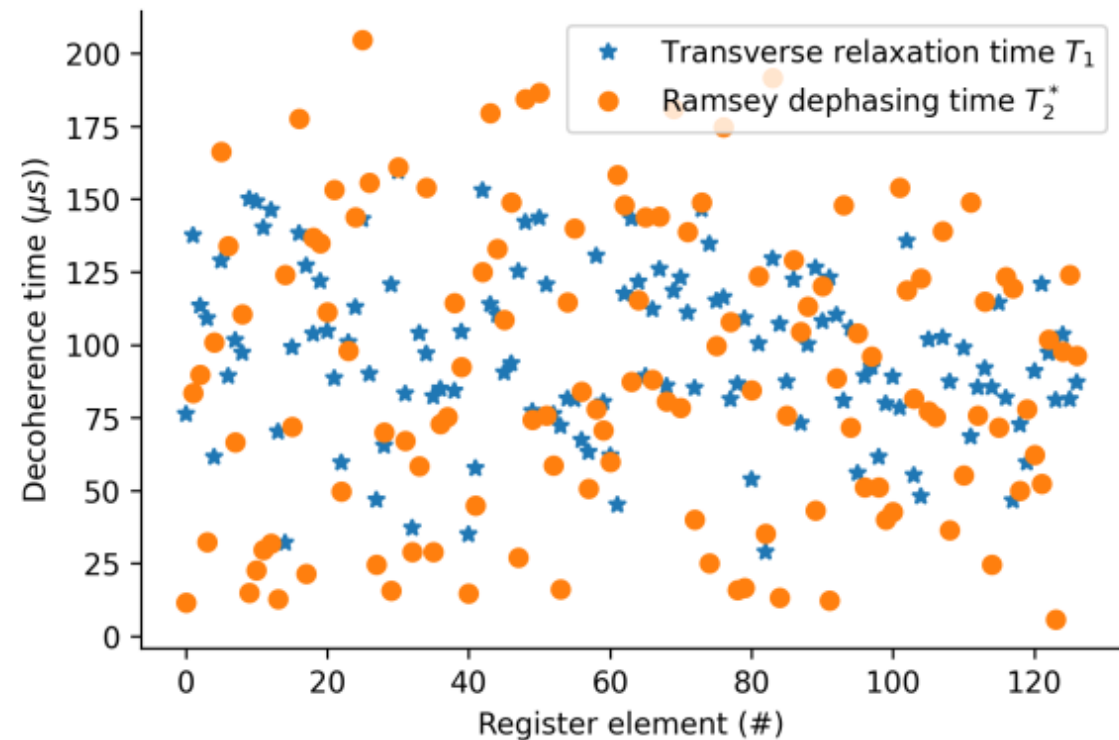
Decoherence times longer than gate operation times

- Programs must be completed before information is lost
 - A high ratio of information lifetime (decoherence) to gate duration is essential for useful calculations.



lbm_washington

Estimated T_1 , T_2 by register element



N. B. What are the T_1 and T_2 times?

- T_1 and T_2 are the time scales on which a quantum system loses energy and coherence, respectively.
 - T_1 is the **relaxation time**
 - T_2 is the **dephasing time**
- These time scales are important for controlling states of a quantum register.
 - A key concern in NMR techniques for probing magnetic spin interactions.
 - Both processes contribute to the more general process of **decoherence**

$$T_1 > T_2$$

Let a single qubit undergo dephasing and relaxation

$$\begin{pmatrix} |a_0|^2 & a_0 b_0^* \\ a_0^* b_0 & |b_0|^2 \end{pmatrix} \rightarrow \begin{pmatrix} |a_t|^2 & a_0 b_t^* e^{-t/T_2} \\ a_t^* b_t e^{-t/T_2} & |b_t|^2 \end{pmatrix}$$

$$|a_0|^2 + |b_0|^2 = |a_t|^2 + |b_t|^2 = 1$$

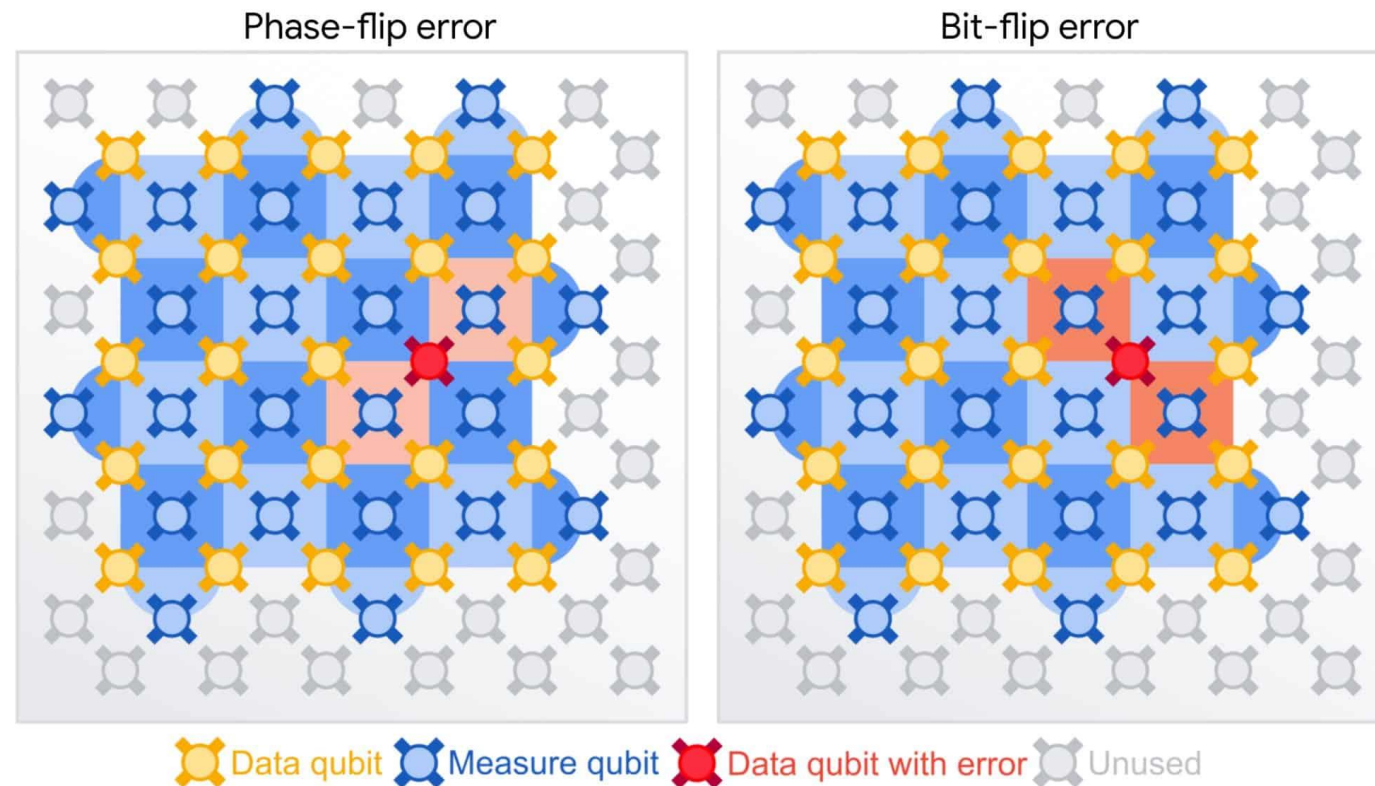
$$|a_t|^2 = |a_0|^2 (1 - e^{-t/T_1})$$

$$|b_t|^2 = 1 - |a_t|^2$$

Sustained growth requires fault tolerance

System engineering requires minimum qubit capacity and fidelity to reach fault tolerance.

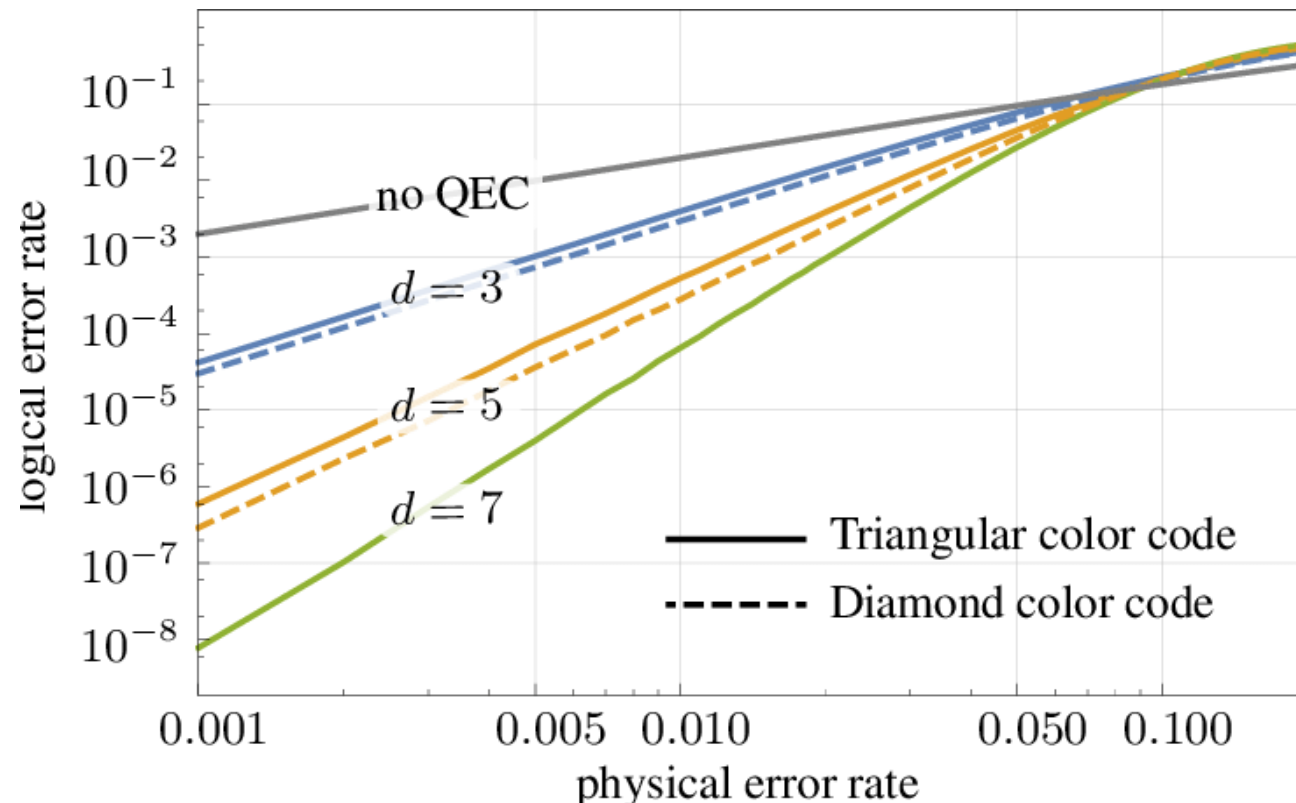
Redundant encoding of information adds complexity in development and use.



Sustained growth requires fault tolerance

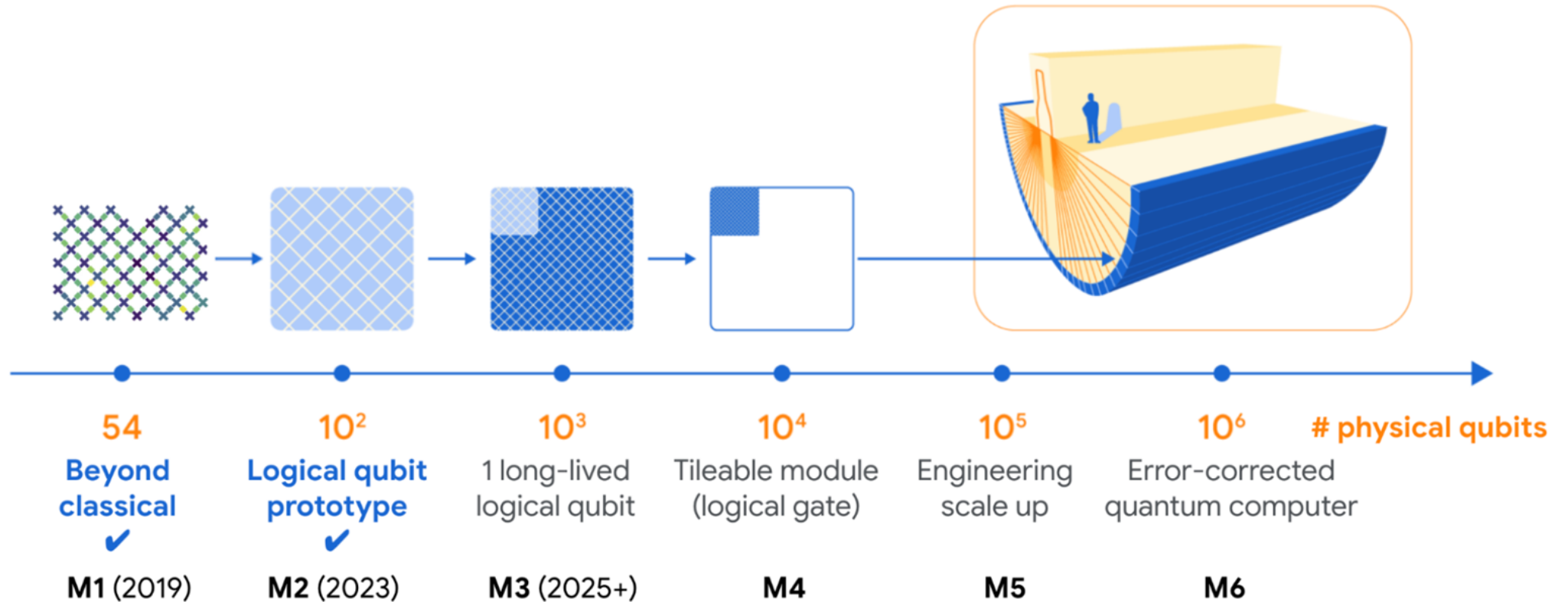
System engineering requires minimum qubit capacity and fidelity to reach fault tolerance.

Redundant encoding of information adds complexity in development and use.



Sustained growth requires fault tolerance

Quantum error correction	–	Enabled	At scale
# Physical qubits	10 – 100	100 – 1000	$10^4 – 10^6$
# Logical qubits	–	1	10 – 1000+
Logical error	10^{-3}	$10^{-2} – 10^{-6}$	$10^{-6} – 10^{-12}$



N.B. Accumulating errors with a binomial distribution

- If each gate has a probability of error (**error rate**) p , what is the cumulative circuit error after n gates?
 - A sequence of n independent random events with probabilities p_i is modeled by the **binomial distribution**.
 - Let $p_i = p$. Then the probability to observe exactly k errors is

$$f(k, n, p) = \binom{n}{k} p^k (1 - p)^{n-k}$$

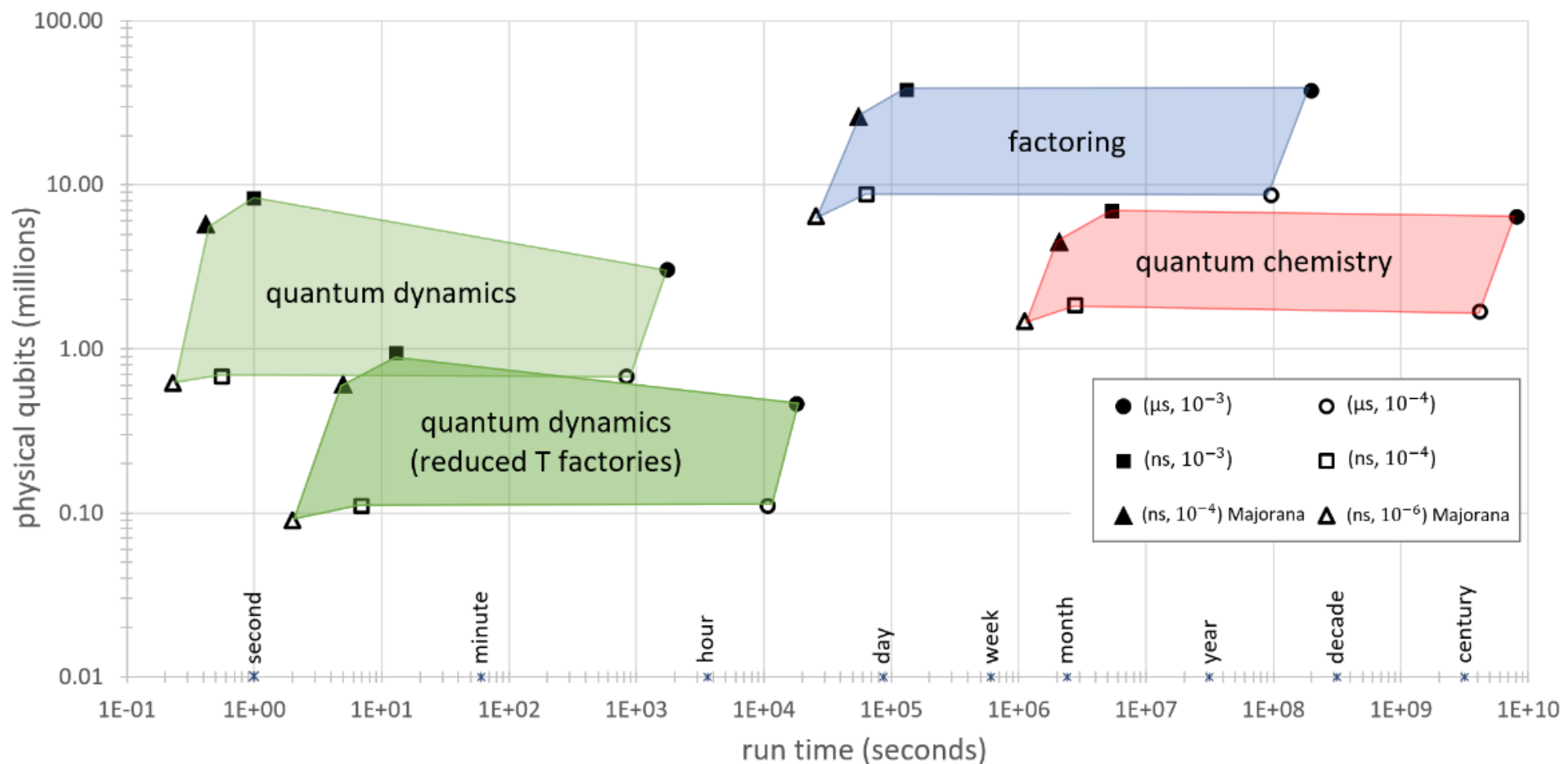
with the binomial coefficient “n choose k” defined as

$$\binom{n}{k} = \frac{n!}{k! (n - k)!}$$

- The cumulative error for n events is then

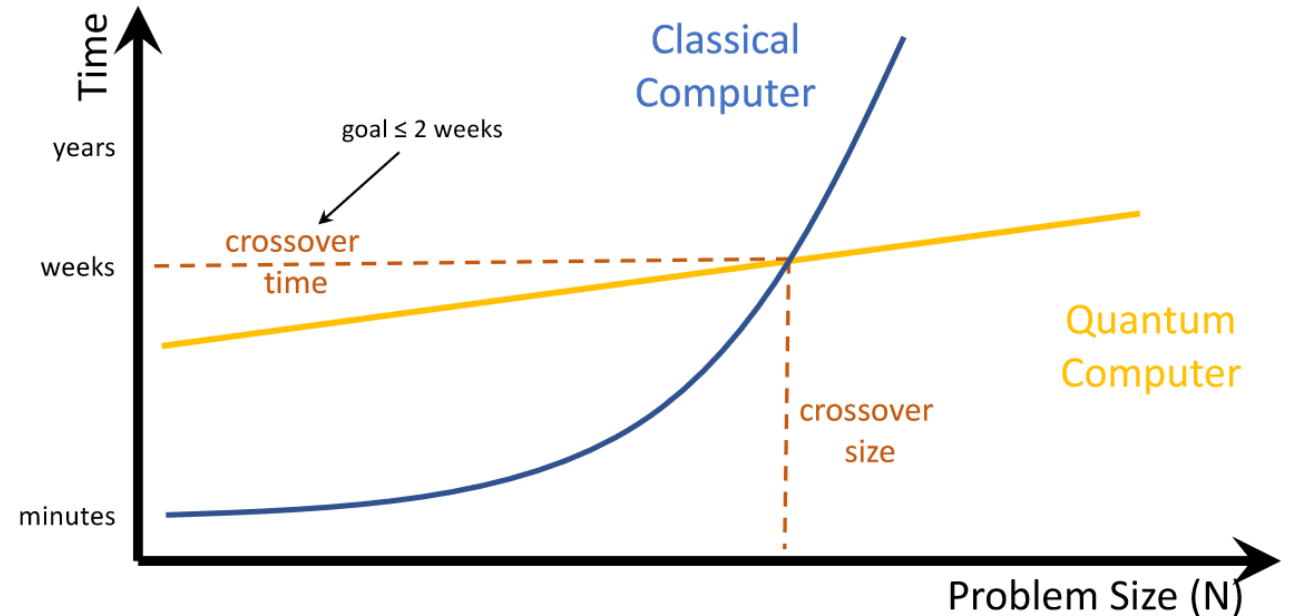
$$\text{Prob(error)} = \sum_{k=1}^n f(k, n, p)$$

Estimates for resource requirements to reach quantum advantage



What is quantum advantage?

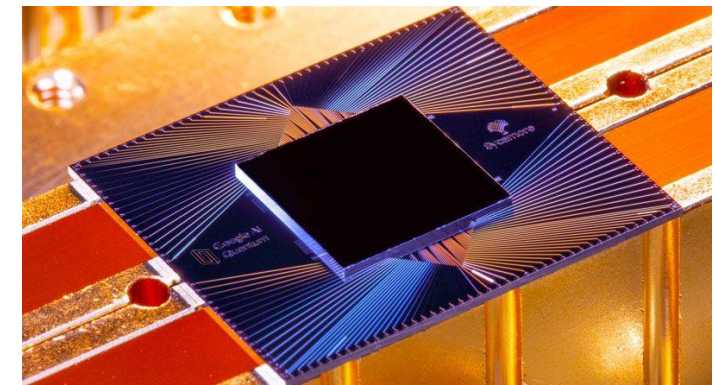
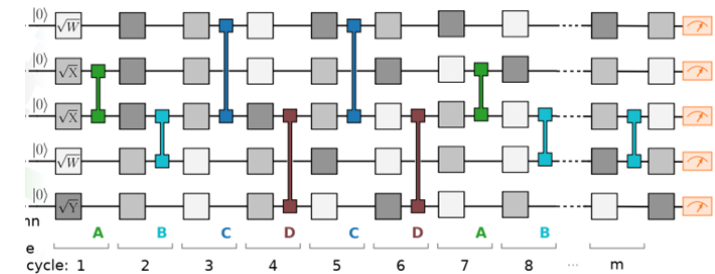
- Quantum advantage usually means improvements in solving a problem using quantum computing relative to best-in-class conventional methods.
 - But there are other interpretations and nuanced definitions.
- A practical concern is the smallest problem size for which a quantum computer would show a quantum advantage?
 - The “answer” is a function of scientific advancement.



A first test of quantum advantage over conventional HPC

- The first demonstration of a quantum computer outperforming every other computer was made in 2019.
 - The Google Sycamore processor outperformed the DOE Summit supercomputer on solving a synthetic benchmark problem called **random circuit sampling**.
- This was the first evidence that quantum computing had begun to surpass conventional computer capabilities.
 - The result drive questions about defining quantum advantage and conventional methods for random circuit sampling.
 - [Latest results from Quantinuum](#) extend these ideas.

Quantum circuits with m “layers” of 1- and 2-qubit gates were randomly generated and executed on the Google sycamore process.



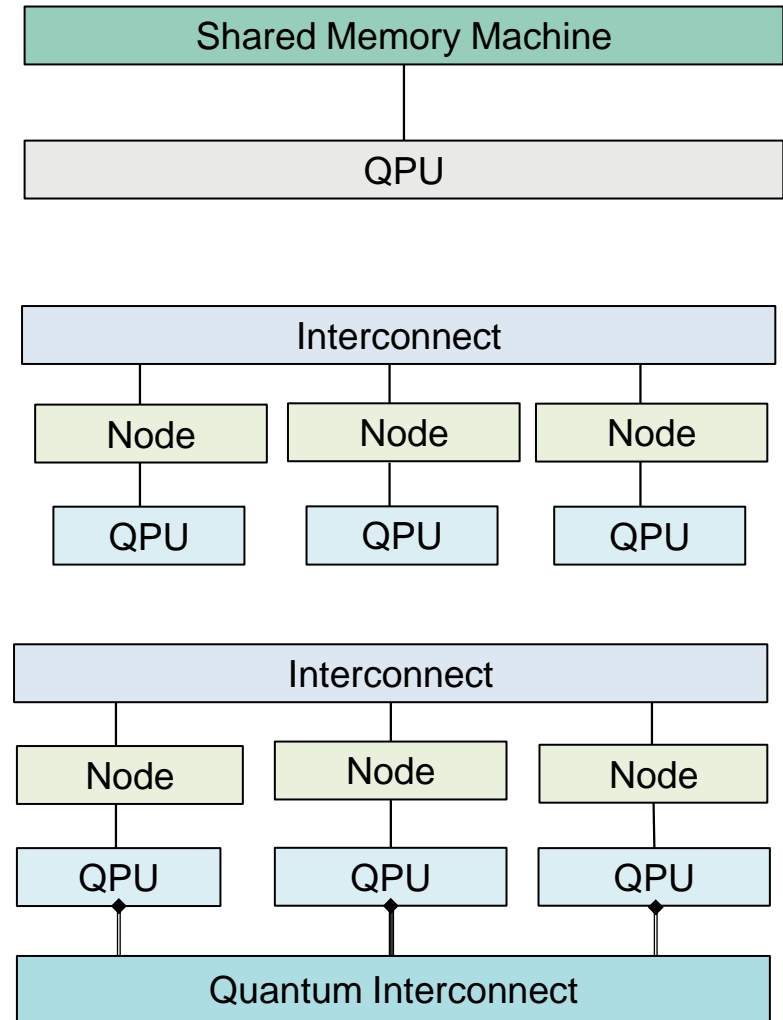
Quantum advantage for value creation

- Quantum computing technology has the potential to affect the 17 United Nations Sustainable Development Goals
 - **Zero hunger**; More efficient nitrogen fixation to enhance food supplies
 - **Good health** and well-being; Faster and cheaper drug development
 - **Clean water** and sanitation; Enhanced water treatment capabilities
 - Affordable, **clean energy**; Energy system optimization
 - **Climate action**; Improved weather modeling and analysis

	Applications	Value creation potential ¹ (\$B)	
		Low	High
Cryptography (\$40-\$80B)	Encryption/decryption	\$40	\$80
Optimization (\$100-\$220B)	Aerospace: Flight route optimization	\$20	\$50
	Finance: Portfolio optimization	\$20	\$50
	Finance: Risk management	\$10	\$20
	Logistics: Vehicle routing/network optimization	\$50	\$100
Machine learning (\$150-\$220B)	Automotive: Automated vehicle, AI algorithms	\$0	\$10
	Finance: Fraud and money-laundering prevention	\$20	\$30
	High tech: Search and ads optimization	\$50	\$100
	Other: Varied AI applications	\$80+	\$80+
Simulation (\$160-\$330B)	Aerospace: Computational fluid dynamics	\$10	\$20
	Aerospace: Materials development	\$10	\$20
	Automotive: Computational fluid dynamics	\$0	\$10
	Automotive: Materials and structural design	\$10	\$15
	Chemistry: Catalyst and enzyme design	\$20	\$50
	Energy: Solar conversion	\$10	\$30
	Finance: Market simulation (e.g. derivatives pricing)	\$20	\$35
	High tech: Battery design	\$20	\$40
	Manufacturing: Materials design	\$20	\$30
	Pharma: Drug discovery and development	\$40	\$80

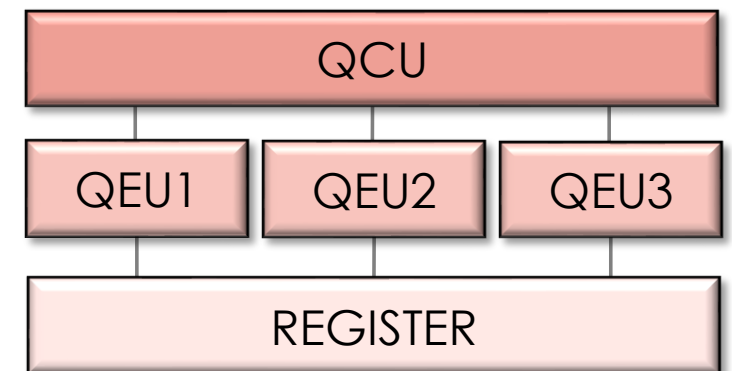
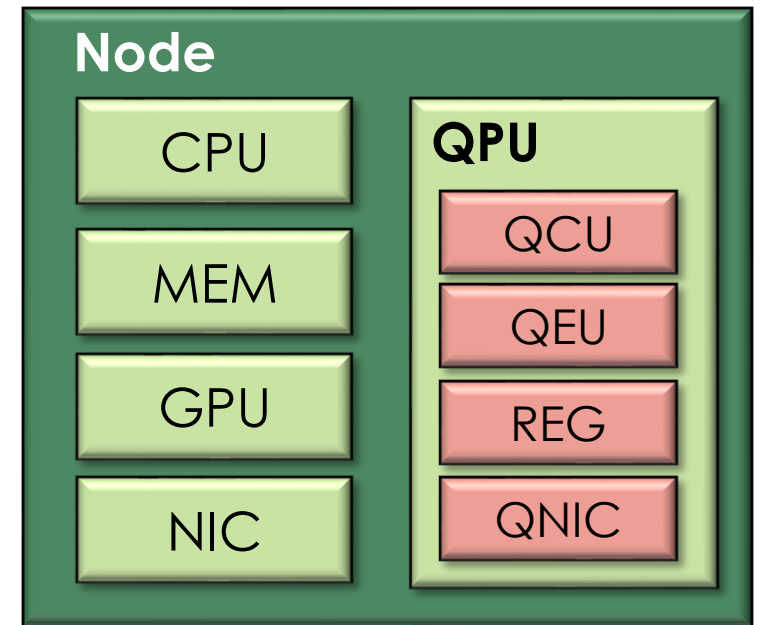
Quantum High-Performance Computing

- Are QPUs compatible with modern scientific computing?
 - When do QPU's accelerate applications relative to state-of-the-art HPC?
 - What are the behavioral and functional requirements placed on the processor?
- How do we integrate conventional workflows with emerging quantum methods?
 - What are the programming and execution models?
 - What are the methods for performance and resource management



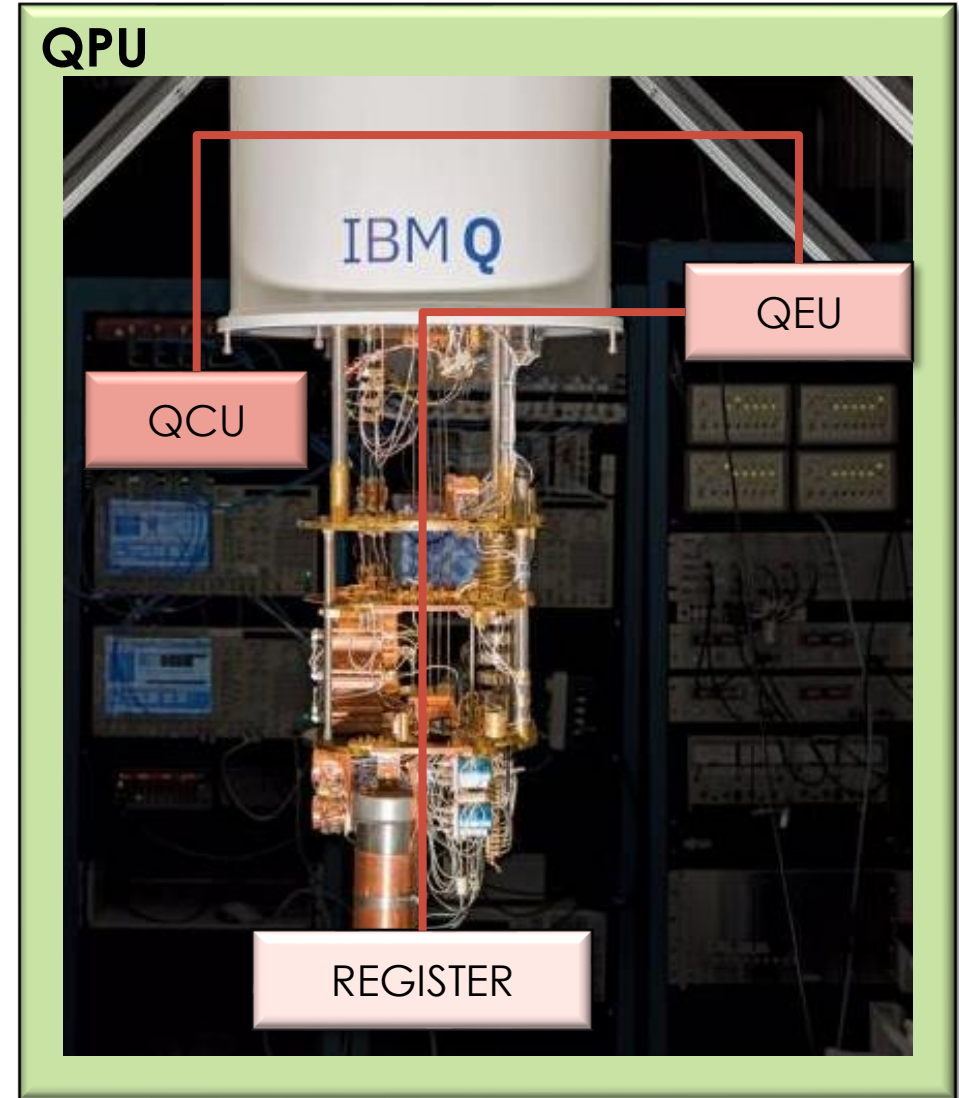
Quantum Node Terminology

- A **node** is part of a computer system that may be composed from CPUs GPUs, and memory hierarchies as well as QPUs.
- A **quantum processing unit** (QPU) encompasses methods for parsing and executing quantum programs.
- The **quantum control unit** (QCU) parses instruction sent by the CPU to the QPU.
- A **quantum execution unit** (QEU) applies fields to initiate gates. There may be multiple QEU's.
- Applied fields drive changes in the **quantum register**. The register state stores the value of the computation.
- I/O is based on fields to prepare and measure the register in computational basis states.
- **Network interfaces** for the conventional (NIC) and quantum (QNIC) interconnects support communication



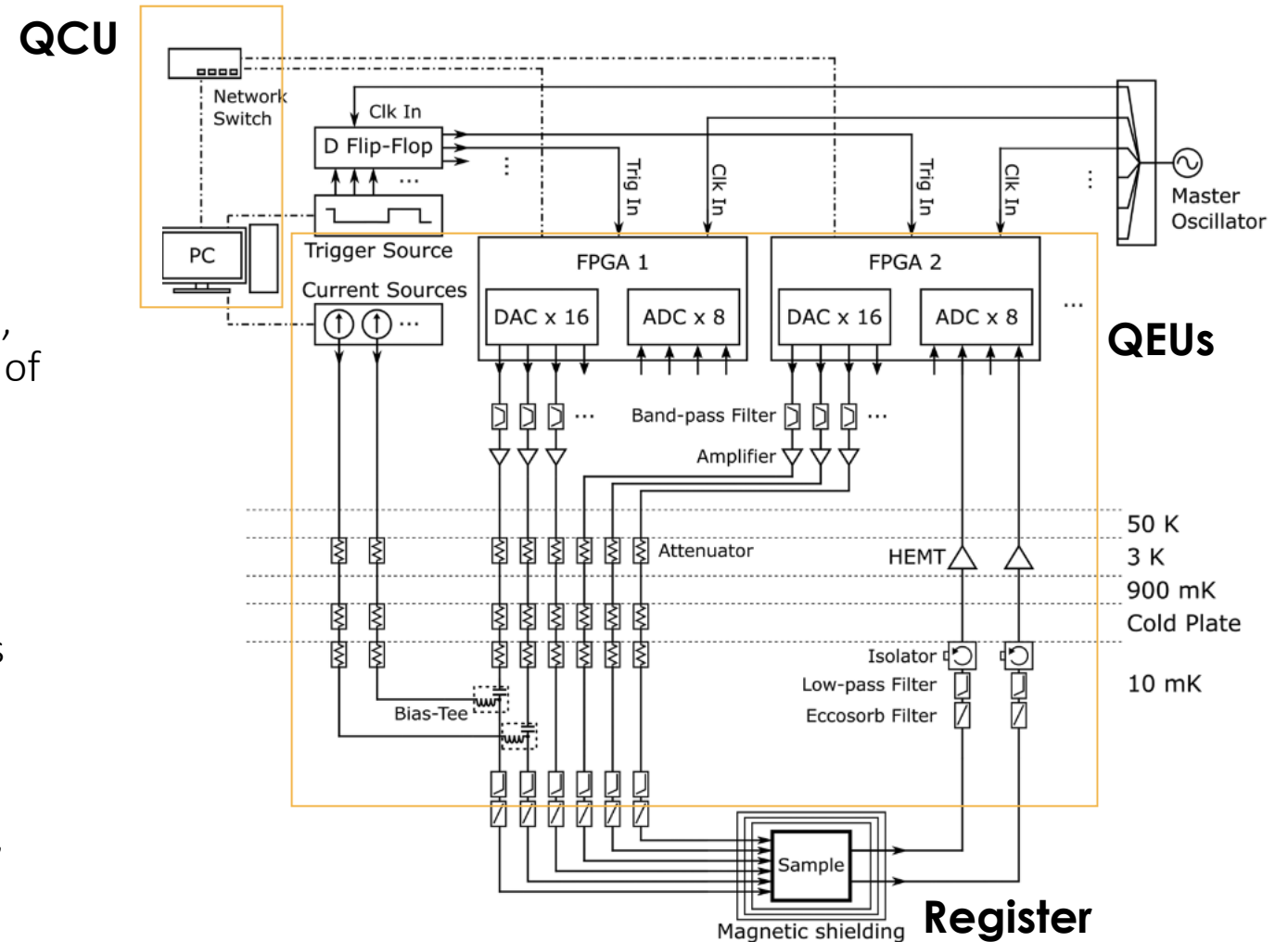
Quantum Node Model

- Digitized waveforms drive analog signal generation to interface with register
- Arbitrary waveform generators (AWGs), filters, and amplifiers transform EM fields in and out of on-chip resonators.
- Filtered signals drive digital signal processing workflows to recover classical information.
- Classical information includes register state as well as diagnostics.
- Thermodynamic controls expressed by electromagnetic shielding, ultra-high vacuum, cryogenic cooling



Quantum Node Model

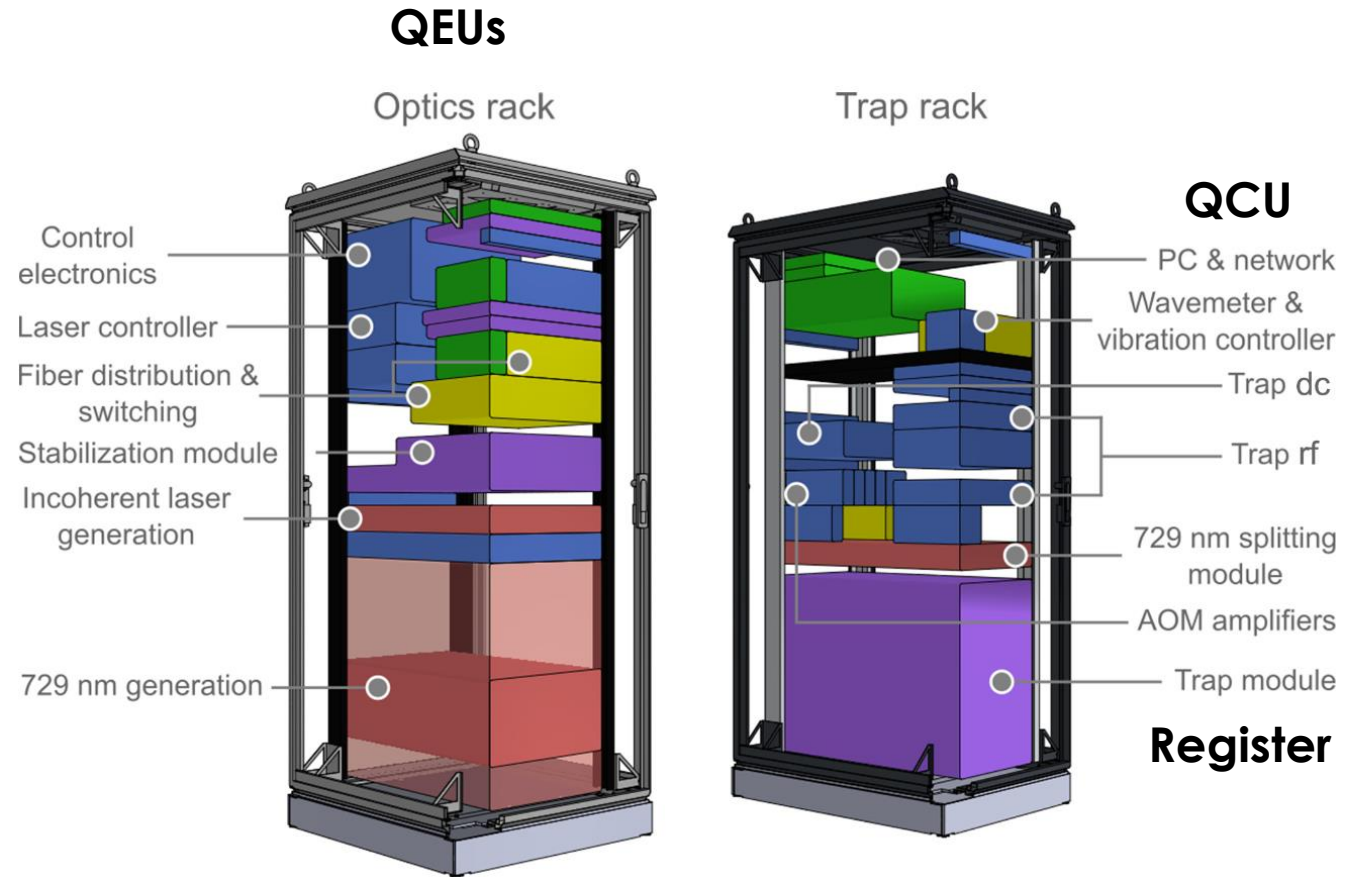
- Digitized waveforms drive analog signal generation to interface with register
- Arbitrary waveform generators (AWGs), filters, and amplifiers transform EM fields in and out of on-chip resonators.
- Filtered signals drive digital signal processing workflows to recover classical information.
- Classical information includes register state as well as diagnostics.
- Thermodynamic controls expressed by electromagnetic shielding, ultra-high vacuum, cryogenic cooling



K. H. Park et al., "ICARUS-Q: Integrated control and readout unit for scalable quantum processors," Rev. Sci. Instrum. 93, 104704 (2022).

Quantum Node Model

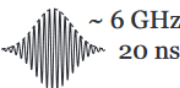
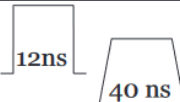

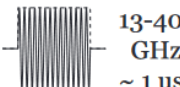
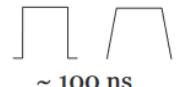

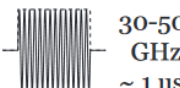





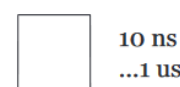

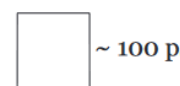

- Digitized waveforms drive analog signal generation to interface with register
- Arbitrary waveform generators (AWGs), filters, and amplifiers transform EM fields in and out of on-chip resonators.
- Filtered signals drive digital signal processing workflows to recover classical information.
- Classical information includes register state as well as diagnostics.
- Thermodynamic controls expressed by electromagnetic shielding, ultra-high vacuum, cryogenic cooling



I. Pogorelov et al., "[Compact Ion-Trap Quantum Computing Demonstrator](#)," PRX Quantum 2, 020343 (2021)

Quantum Node Model

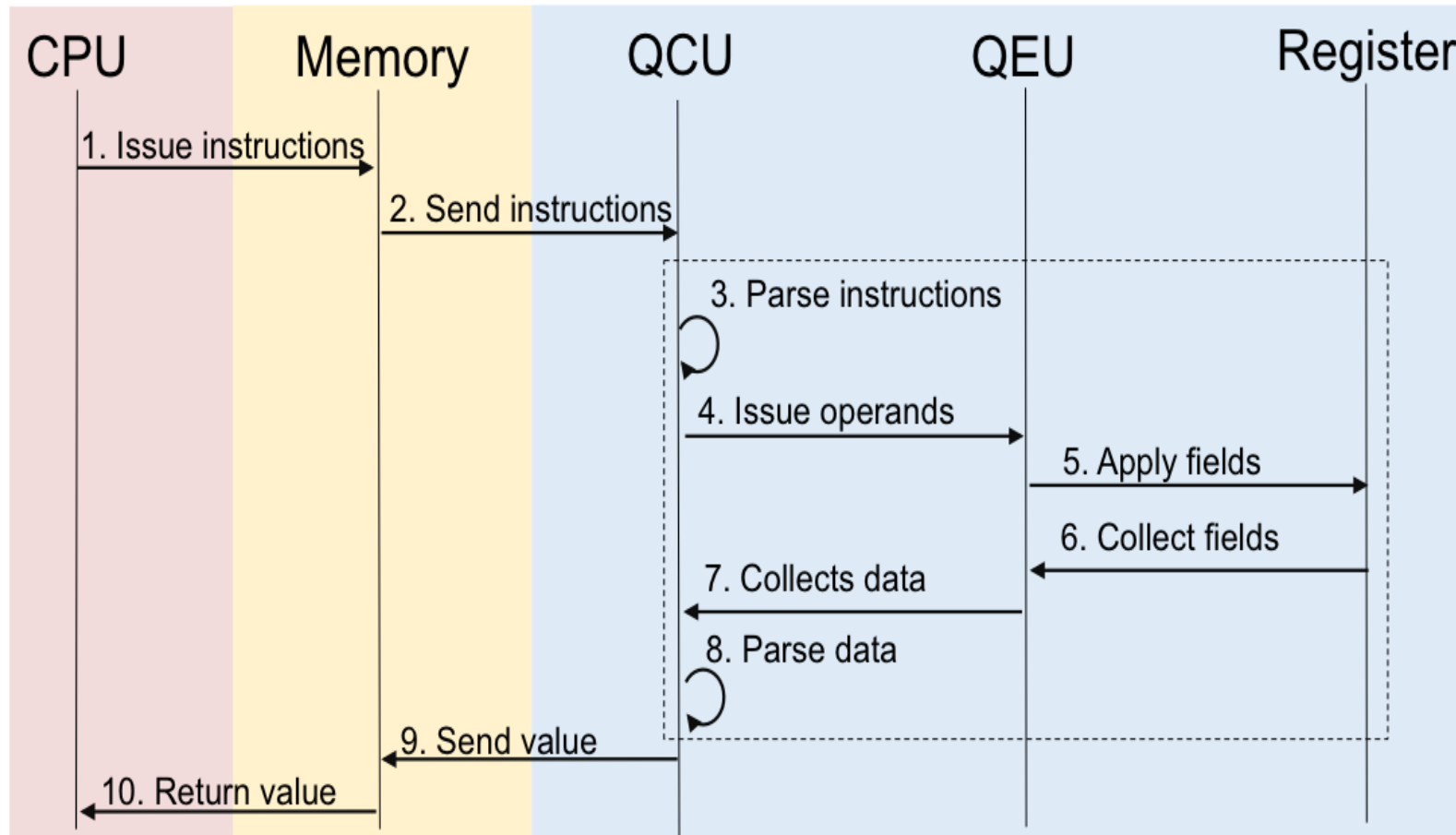
- Digitized waveforms drive analog signal generation to interface with register
- Arbitrary waveform generators (AWGs), filters, and amplifiers transform EM fields in and out of on-chip resonators.
- Filtered signals drive digital signal processing workflows to recover classical information.
- Classical information includes register state as well as diagnostics.
- Thermodynamic controls expressed by electromagnetic shielding, ultra-high vacuum, cryogenic cooling

Technology	T_2^*	1-Qubit gate	2-Qubit gate	Qubit read-out	DC-Biasing
Superconducting qubits (Transmons)	2.5 us	 ~ 6 GHz 20 ns	 12ns 40 ns	 7-8 GHz, ~ 1 us	flux-bias current
Single-electron spin qubits in a quantum dot	120 us	 13-40 GHz ~ 1 us	 ~ 100 ns		gate voltage
Single-electron spin qubits in a donor system	160 us	 30-50 GHz ~ 1 us	 ~ 100 ns		gate voltage
Singlet-triplet qubit	700 ns	 ~ 1 ns	 ~ 1 us		gate voltage
Exchange-only qubit	2.3 us	 10 ns ...1 us	Sequence of pulses between different quantum dots		gate voltage
Hybrid qubit	< 10 ns	 ~ 100 ps	Sequence of pulses between different quantum dots		gate voltage

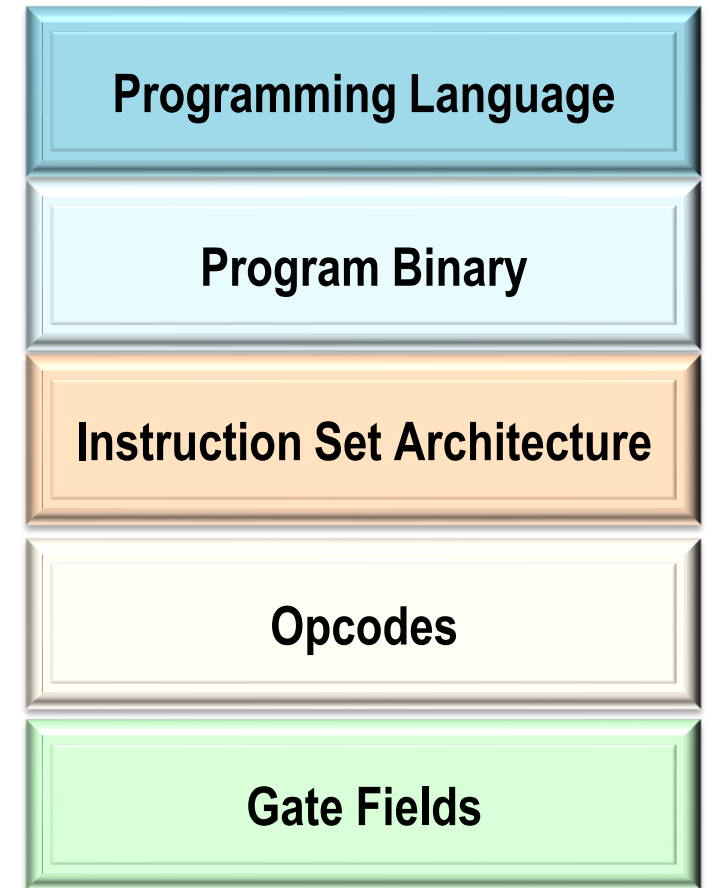
von Dijk et al., "[The electronic interface for quantum processors](#)," Microprocessors and Microsystems, 66, 90 (2019)

Quantum Node Execution Model

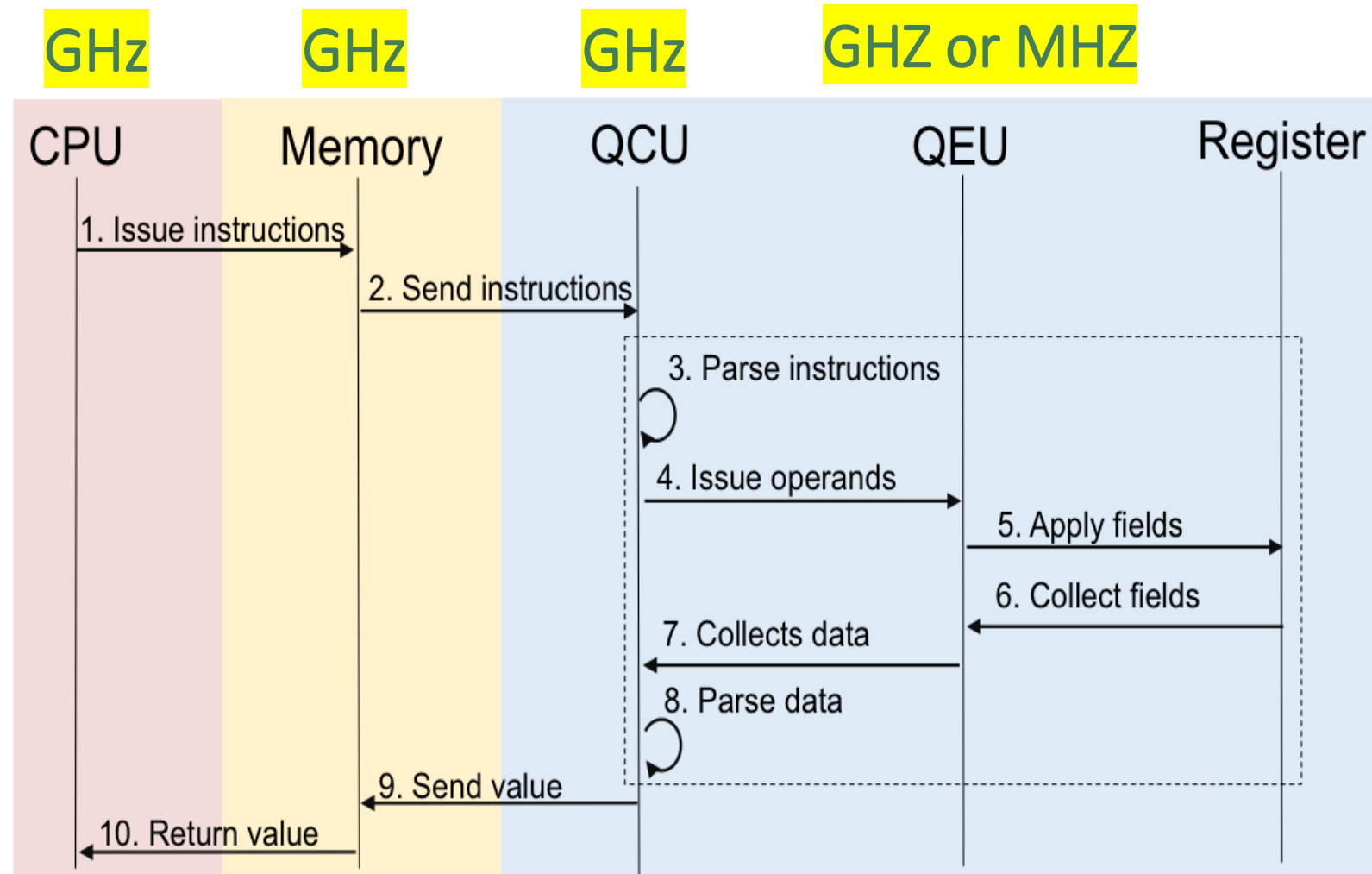
- Quantum execution models define how the QPU carries out the program instructions



Language Hierarchy



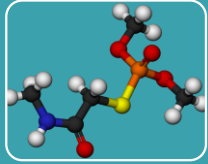
N.B. Component clocks are important for input-output interfaces.



GHz or MHz

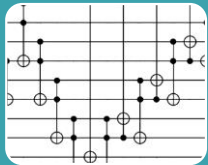
- Modern conventional control systems operate with GHz clock rates.
- Register state may evolve on GHz to MHz clock scales.
- IO rates are set by register rates and decoherence time scales.

Quantum Device Programming Stack



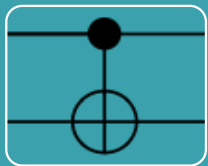
Application-level Libraries

- Domain specific data types and interfaces
- Optimize workflow and post-processing methods



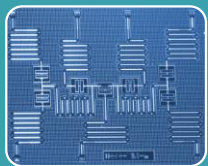
Circuit-level Libraries

- Templated designs for algorithmic primitives and basic data types



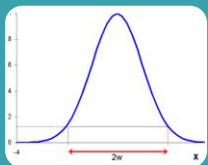
Gate-level Libraries

- Optimized sequences for concurrent operations and memory interfaces



Device-specific Libraries

- Expose tuned gate operations for device constraints



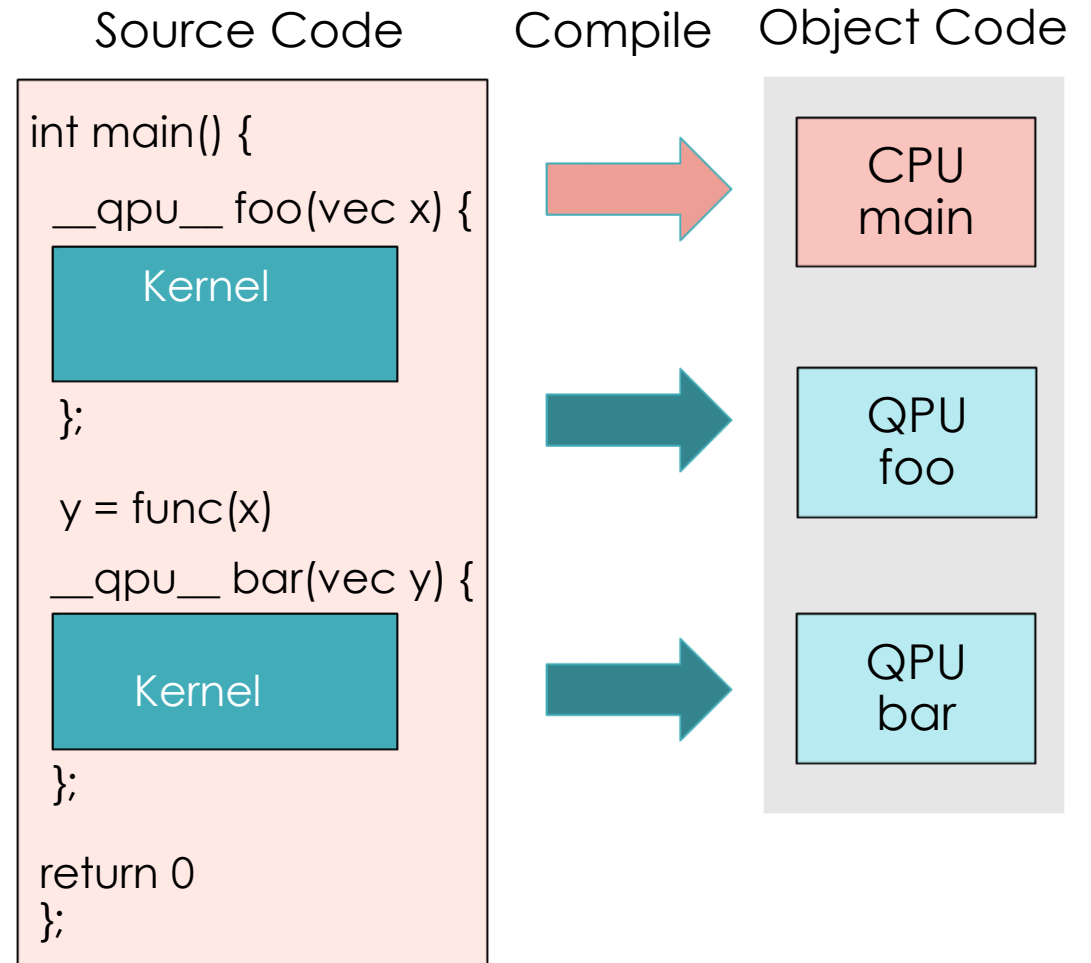
Analog Device Controls

- Expose device-specific methods for gate and pulse operations

- Many prototype languages, libraries, and interfaces for quantum computing!
 - Qiskit, Q#, pytket, cirq, pennylane
- Addresses multiple perspective for users and developers
 - Application developer, library developer, control system engineer, hardware developer
- Monolithic integration leads “full stack” development...but the diversity of independent concerns makes this approach unsustainable

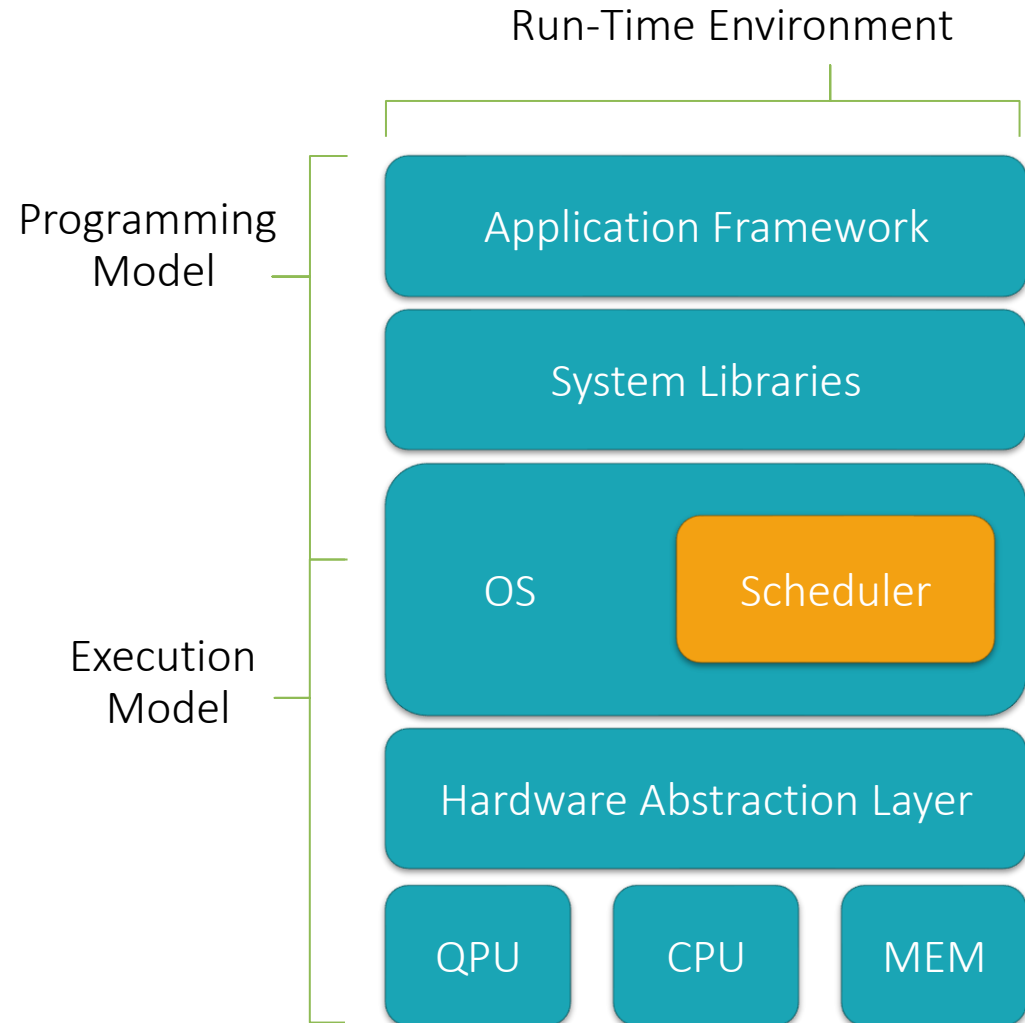
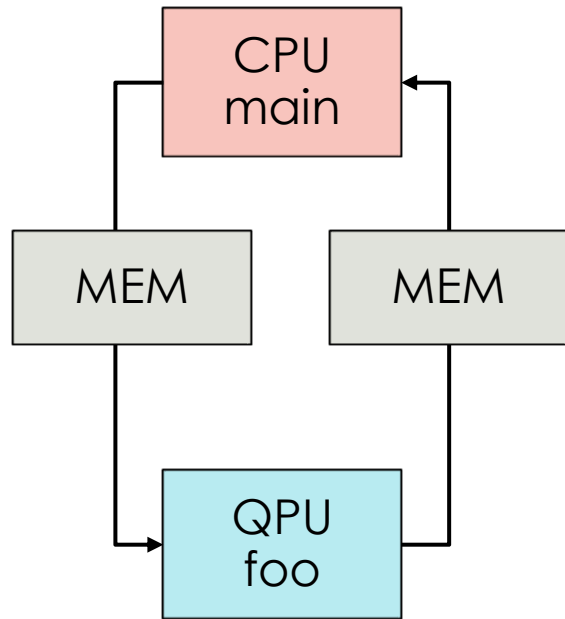
Example: XACC Programming Framework

- ORNL developed XACC, a mixed-language, directive-based programming framework
 - Language and hardware “agnostic”
 - Keywords identify quantum kernels in DSL
 - LLVM toolchain triggers backend compilers
 - Example: Host C/C++ program with OpenQASM kernel on Rigetti QPU
- QIR is an industry-supported quantum intermediate representation for expressing and optimizing programs that combine quantum and classical instructions.
 - Transformations tune classical programming methods
 - Case study: Using XACC framework to implement QIR programs

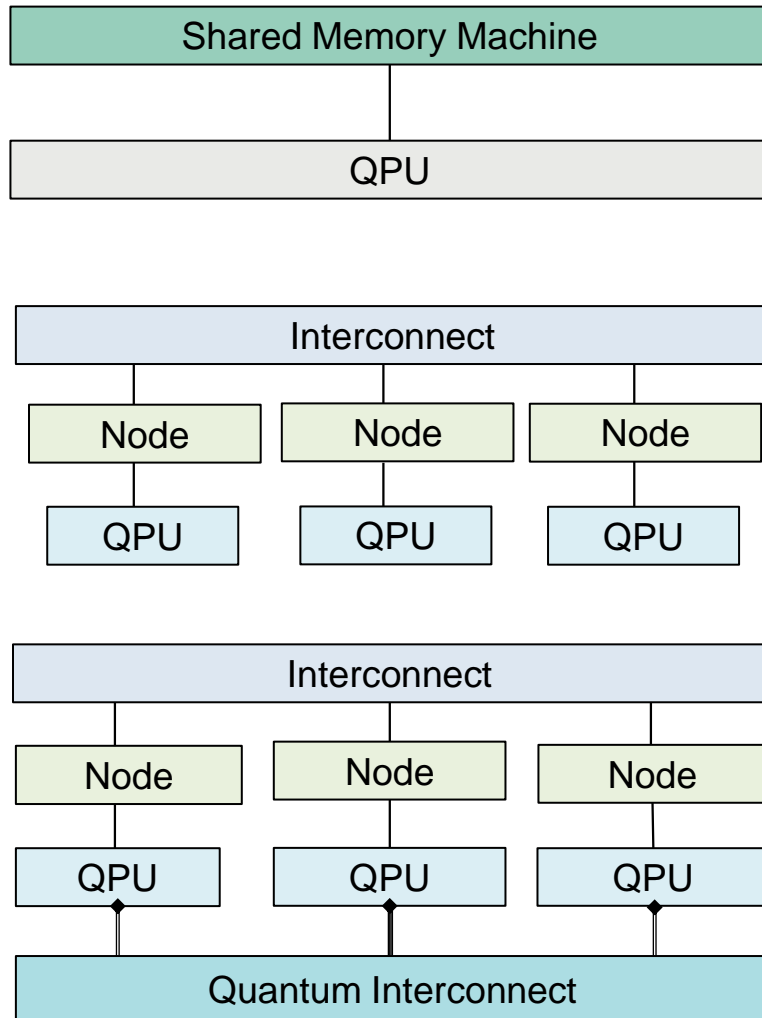


Quantum Node Run-time Environment

Program Component Interactions

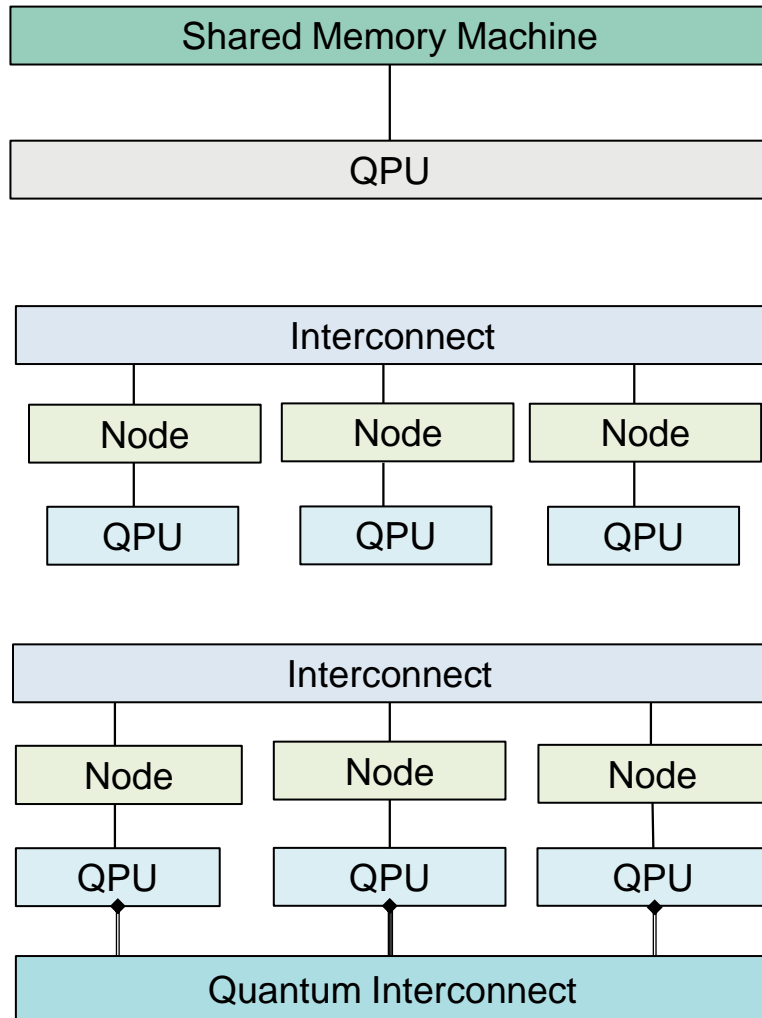


Quantum High-Performance Computing System Architecture



- Serial computing model
 - Off loaded task(s) return classical results
 - Single-QPU management policy, eg, queuing
 - Monolithic scaling of the architecture
- Conventional parallel computing
 - Embarrassingly parallel quantum computing
 - Example: Decompose parameter across nodes for parallelized classical algorithm
- Quantum parallel computing
 - Entangle quantum tasks between multiple nodes through quantum interconnect
 - Effectively a larger quantum computer but more granularity in programming and resource management

N.B. Domain decomposition and quantum programming



- Conventional parallel programming methods often partition a global domain into subdomains
 - **Domain decomposition** partitions the data and tasks represented by the program.
 - Parallel programming allocates subdomains across n processors, cf. **Amdahl's law**
- Domain decomposition for parallelized quantum computing is more subtle
 - Hilbert space decreases with decomposition

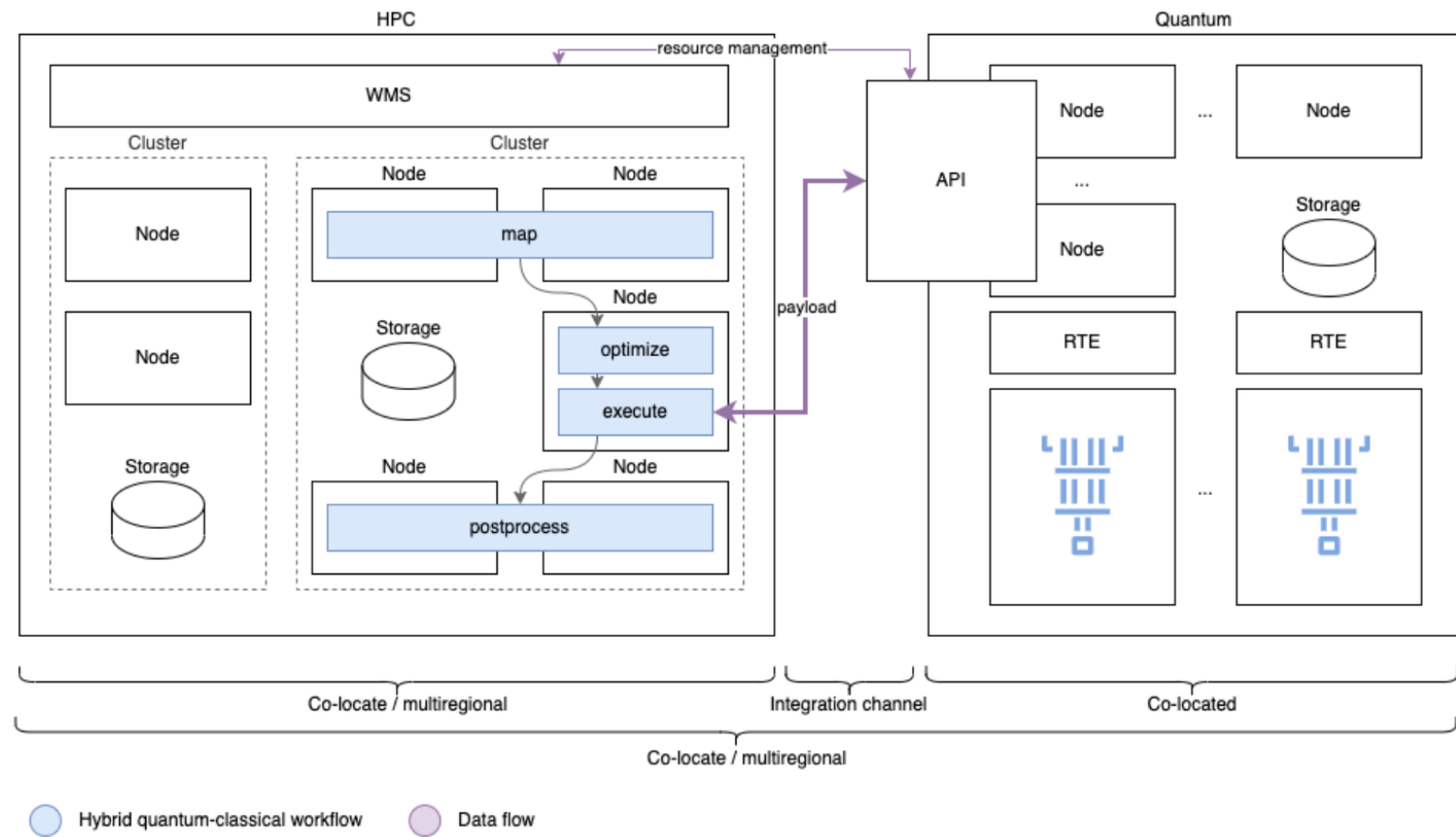
$$2^{nq} \text{ vs. } n2^q$$

- Communication between components is necessary to coordinate entanglement

$$m_q n \text{ vs. } m_q n^2$$

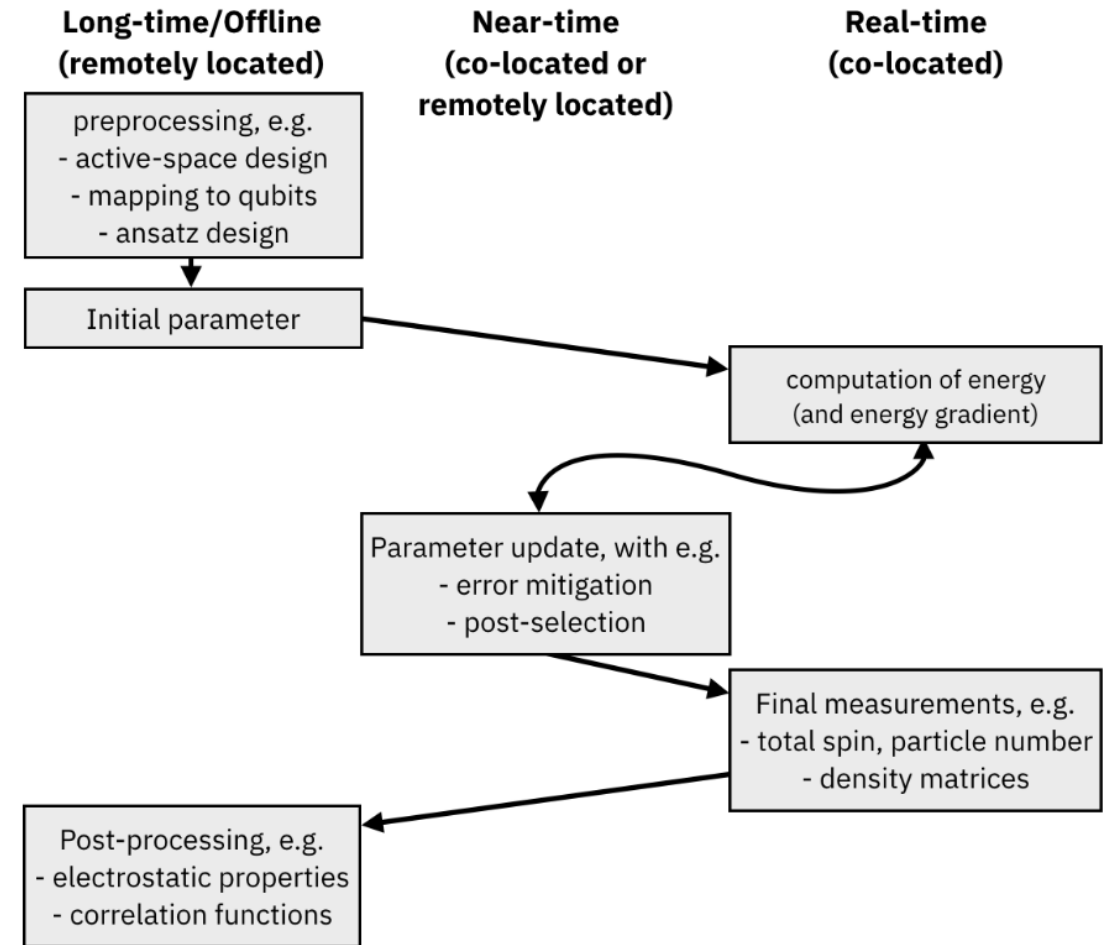
Quantum High-Performance Computing System Integration

- System integration requires addressing concerns for execution and resource management as well as performance criteria.
 - HPC systems are multi-user environments that require on-demand coordination of system resources through centralized management.



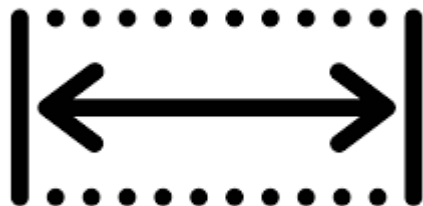
Quantum High-Performance Computing System Scheduling

- Scheduling jobs for a system with different components requires considering the locality and timescale of the task.
 - **Long-time** tasks include pre- and post-processing of application data as well as static program compilation
 - **Near-time** tasks include just-in-time compilation of quantum programs, e.g., based on parameter selection
 - **Real-time** tasks include processing measurement data for coherent circuit operations, e.g., error correction, state preparation.



System performance is defined by multiple metrics

Area



Error



Latency



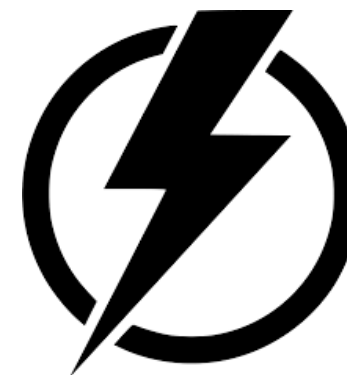
Accuracy



Time-to-solution



Power



System performance is defined by multiple metrics

<p>Area</p> <p>How much space does a program on the system require?</p> <ul style="list-style-type: none">• Size of the register	<p>Error</p> <p>How much error does the system generate?</p> <ul style="list-style-type: none">• Logical error rate	<p>Latency</p> <p>How quickly is information generated by the system?</p> <ul style="list-style-type: none">• System clock speed
<p>Accuracy</p> <p>How accurate and precise is the result computed?</p> <ul style="list-style-type: none">• Statistical variance	<p>Time-to-solution</p> <p>How long does system take to complete a calculation?</p> <ul style="list-style-type: none">• Sample size and rate	<p>Power</p> <p>How much power does the system consume?</p> <ul style="list-style-type: none">• Energy per operation



Quantum Computing User Program

Enable Research

Provide a broad spectrum of user access to the best available quantum computing systems



Evaluate Technology

Monitor the breadth and performance of early quantum computing applications



Engage Community

Support growth of the quantum ecosystem by engaging with users, developers, vendors, and providers

What you should learn from this presentation

- Motivation for integrating quantum computing with high-performance computing, aka, QHPC
- Terminology and techniques for evaluating QHPC
- Priority research areas to advance QHPC design and development
- Advance energy-efficient computation for value creation with quantum computing
- Leverage unique computational models for problem-specific advantages
- Language for describing algorithms, applications, and architectures
- Components and interfaces for QHPC systems using latest technologies
- Metrics for testing and evaluation of system performance

Questions?

