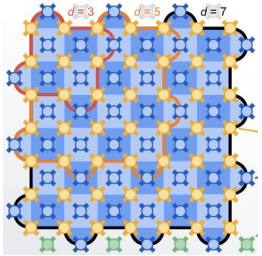
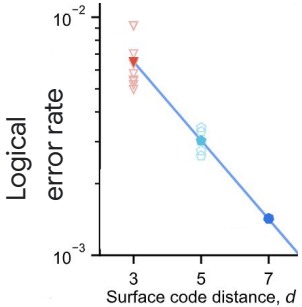
Exponential Quantum Communication Advantage in Distributed Inference and Learning

Dar Gilboa
Google Quantum AI

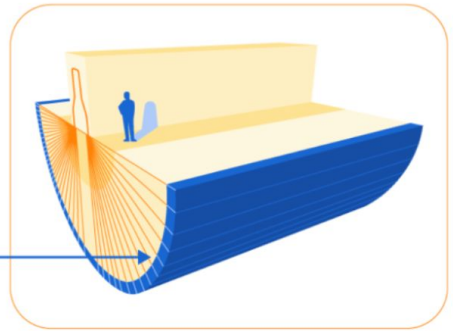
Joint work with Hagay Michaeli, Daniel Soudry and Jarrod McClean

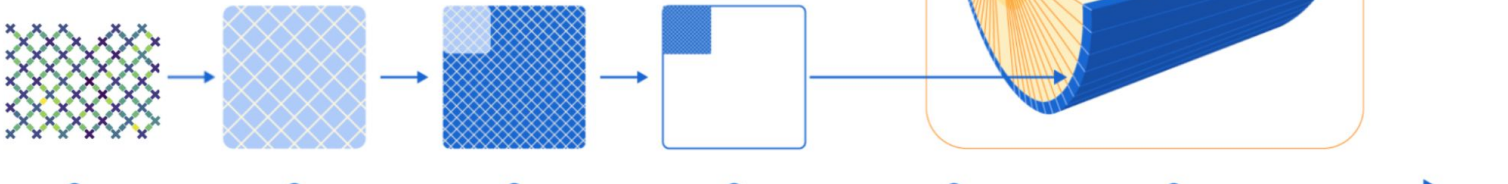
What could you do with a large fault-tolerant quantum computer?





[Google Quantum AI 2024]

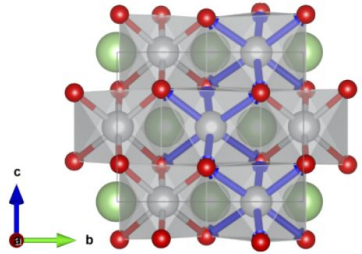




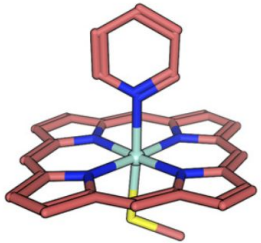
54	10^2	10^3	10^4	10^5	10^6	# physical qubits
Beyond classical ✔	**Logical qubit prototype** ✔	1 long-lived logical qubit	Tileable module (logical gate)	Engineering scale up	Error-corrected quantum computer	
M1 (2019)	**M2** (2023)	**M3** (2025+)	**M4**	**M5**	**M6**	

Applications of large fault-tolerant quantum computers

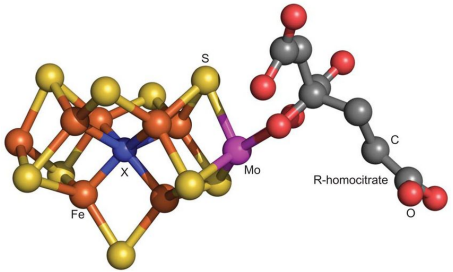
- Simulating highly-correlated systems



Battery design
$LiNiO_2$ (cathode material)
PRX Quant. 4, 040303 (2023)



Drug discovery
Cytochrome P450
(drug anti-target)
PNAS 119, 2203533119 (2022)



Homogeneous catalysis
FeMoCo (fertilizer production)
PRX Quant. 2, 030305 (2021)

- Exponential speedups for structured problems (e.g. factoring, discrete log).

Will quantum computers be useful more generally?

Quantum Machine Learning Algorithms: Read the Fine Print

Scott Aaronson

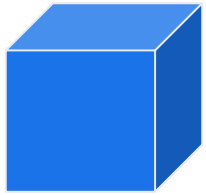
"Will the impact of quantum computers rival that of classical computers in the previous century" [Aaronson 2015]
/ GPUs in the last decade?

$\simeq$

Will quantum computers be useful for processing classical data?

The data loading problem

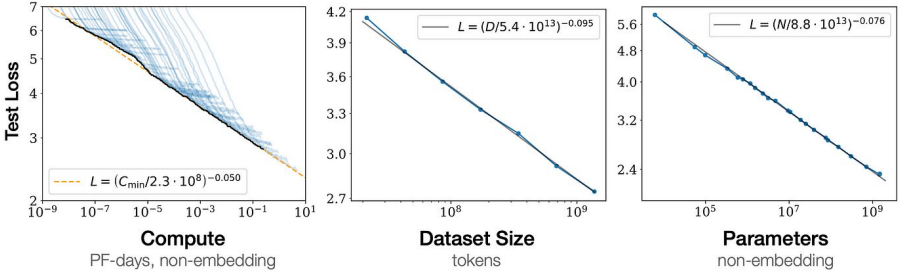
Classical ML (“big data”)



t = poly(input size)
e.g. neural network training and inference

Overparameterization:
Nonconvex problems become well-behaved once
(input size) $\lesssim$ # params ~ t.

Neural scaling laws [Kaplan et al. 2020]**:**



Loading N bits takes time ~N ⇒ log(N) runtime doesn’t give speedup

Potential for exponential speedup

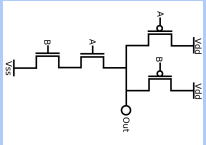
t = exp(input size)

e.g. factoring,
combinatorial optimization (?)

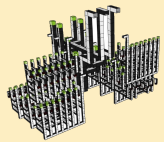
What about quadratic speedups?

- Quadratic speedups known for many problems relevant for classical data (unstructured search, MCMC, optimization, ...)

Overhead of error correction implies huge instance size is needed for practical advantage [Babbush et al. 2021].

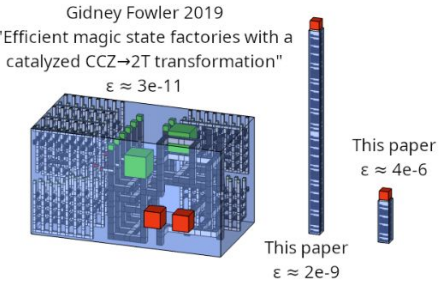


classical NAND gate (CMOS)
$<10^{-9}$ transistor x seconds



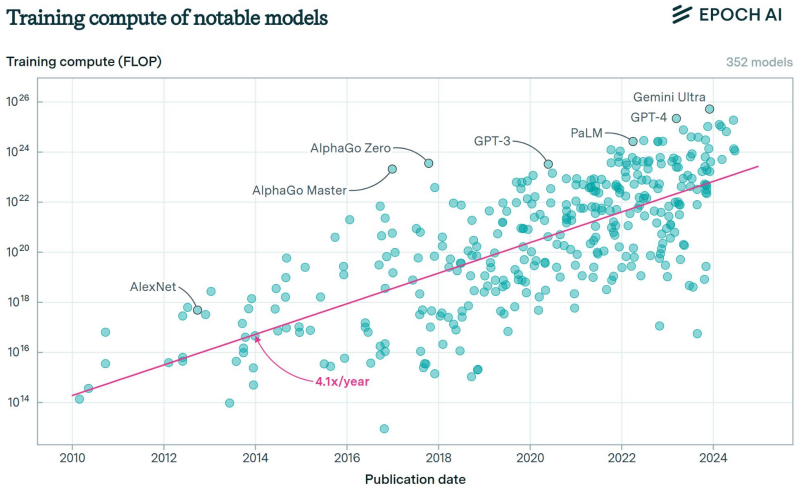
"quantum NAND" gate
(distillation of Toffoli state
in 2D surface code)
>10 qubit x seconds *

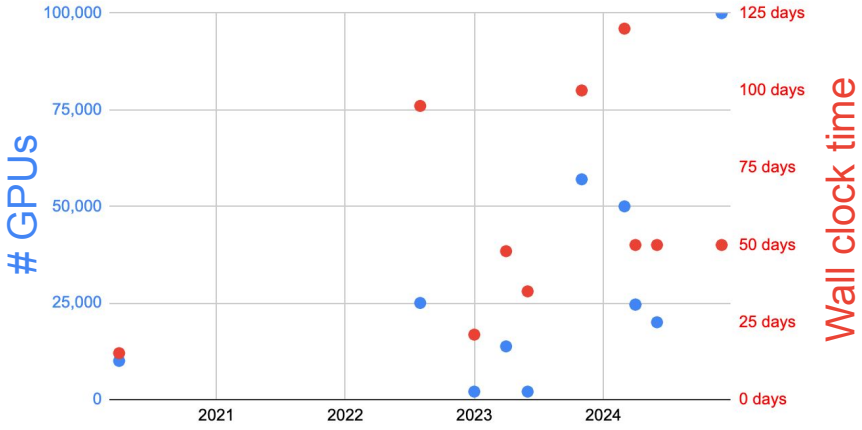
* Slightly outdated given T state cultivation [Gidney et al. 2024]:

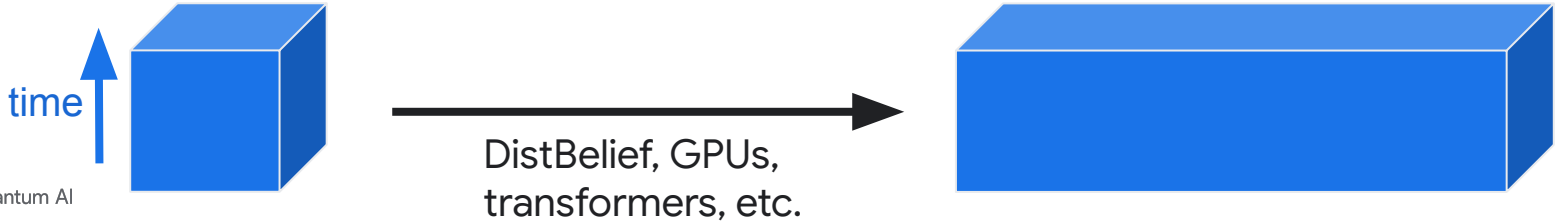


Beyond time/query complexity

- Progress in ML is often driven by better parallelization
- Computation at scale requires trade-offs between various resources





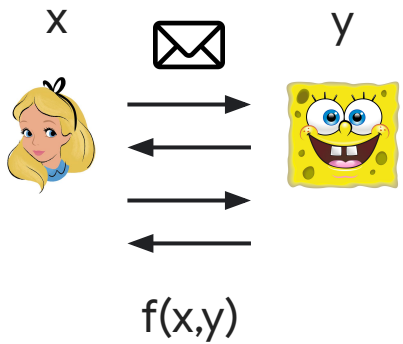


Quantum advantage beyond time/query complexity

Exponential quantum advantage is possible in terms of other resources.

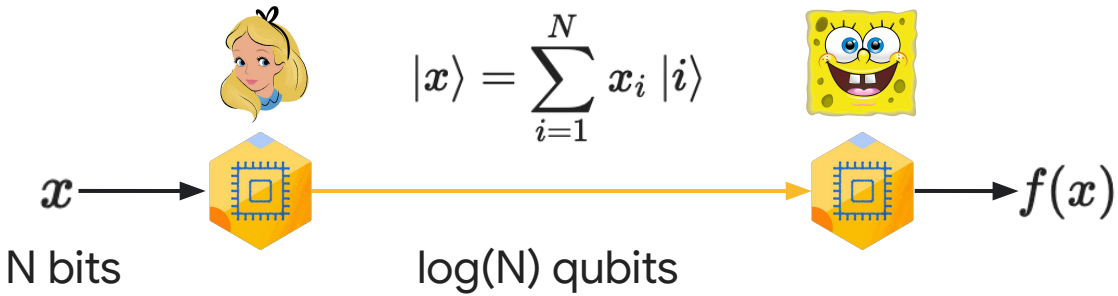
Could this be useful for classical ML problems?

Communication complexity



- We care only about the total communication, not computation (computation will be comparable to classical)
- Unconditional, exponential quantum advantage is possible (dequantization is impossible, no hardness assumptions)
- Connection to memory/space complexity (Alice and Bob could be the same person at different times)

Amplitude Encoding

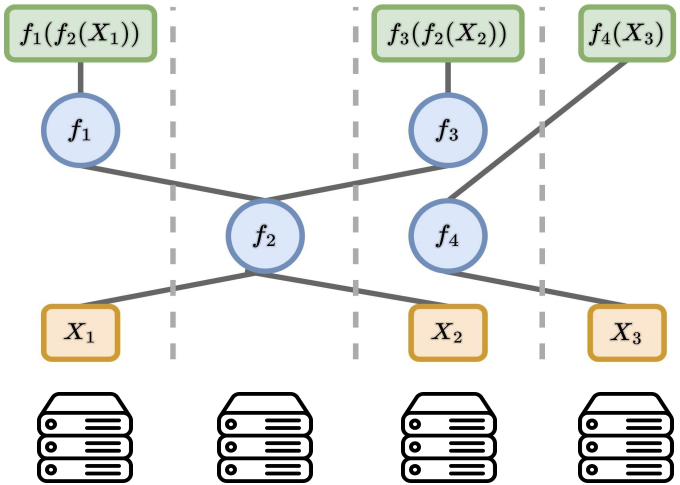


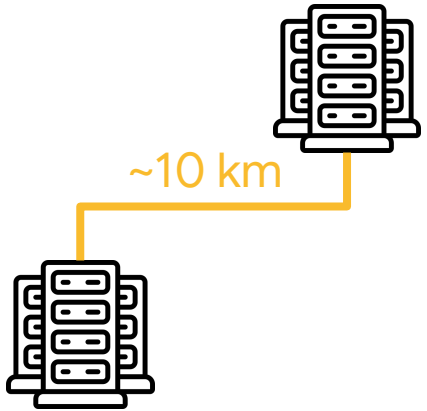
How does this compare to log(N) classical bits?

- By Holevo's bound, only log(N) bits of information can be transmitted.
- Can still be used for distributed computation with exponential advantage.

Distributed computation in machine learning

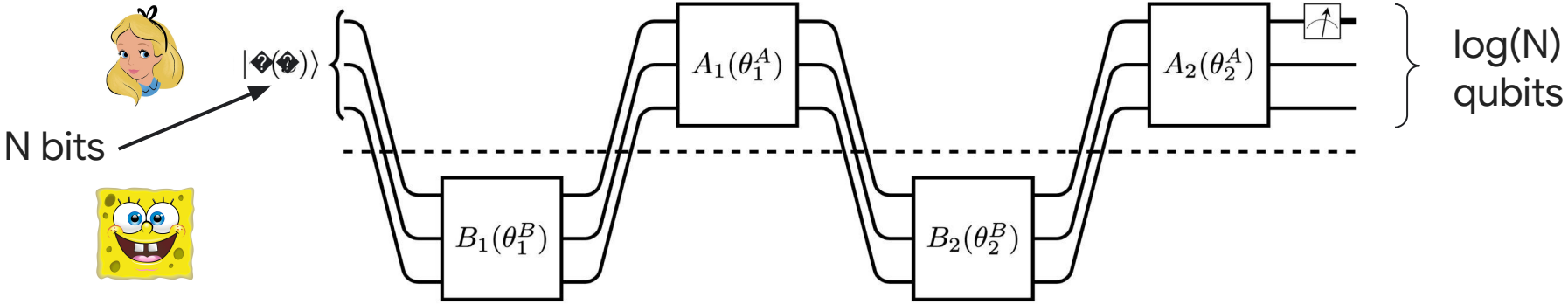
- Training and inference with large neural networks requires distributed computation.
- Nodes communicate features and gradients, typically scaling with input size.
  $\Longrightarrow$ communication bandwidth can become a bottleneck.





LLMs (e.g. Gemini) trained across multiple datacenters

A distributed quantum circuit



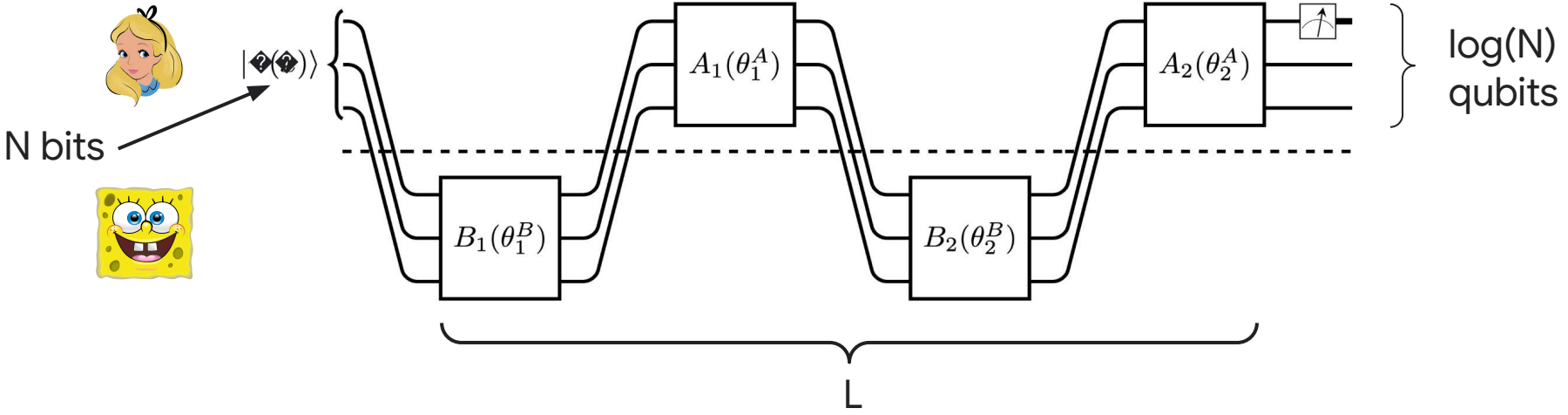
- Layers replaced by parameterized unitaries

$$|\psi(\theta)\rangle = U(\theta)\,|\phi\rangle = \prod_{\ell=1}^{M} U_\ell(\theta_\ell)\,|\phi\rangle$$

- Output is an expectation value

$$F(\theta) = \langle\psi(\theta)|\,O\,|\psi(\theta)\rangle$$

ML scaling

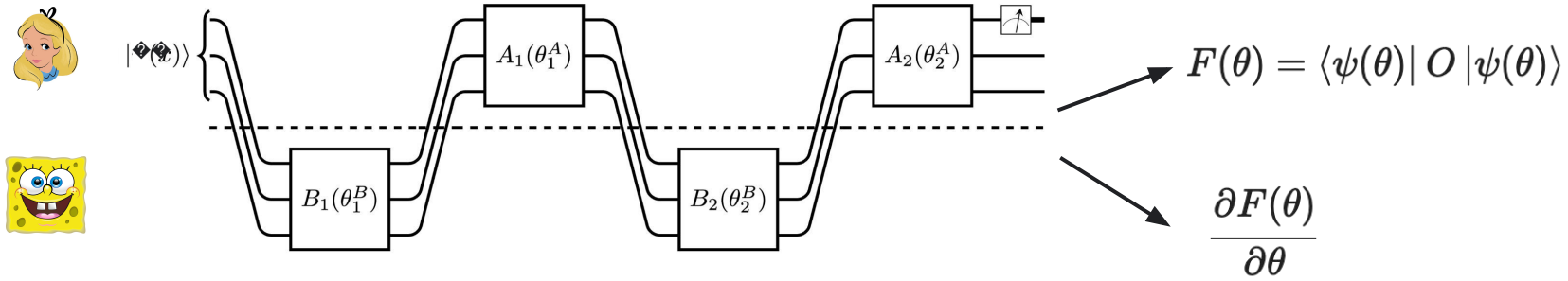


- P = # Params = poly(N)
- L = O(log(N))

$$|\psi(\theta)\rangle = U(\theta)\,|\phi\rangle = \prod_{\ell=1}^{M} U_\ell(\theta_\ell)\,|\phi\rangle$$

An exponential communication advantage

Theorem:
Inference and gradient estimation require Ω(poly(N,P)) bits of communication but only O(polylog(N,P)) qubits.



An exponential communication advantage: Proof sketch (classical lower bounds)

Vector In Subspace
Inputs: Alice: x, $||x||_2=1$, Bob: N/2 dim orth. subspaces S_1, S_2.
Promised that either $x \in S_1$ or $x \in S_2$, determine which is the case.

- $\Omega(\sqrt{N})$ bits of communication required [Raz 1999].
- Quantum algorithm: Alice sends $|x\rangle$, Bob measures.
- Reduces to inference or gradient estimation.

An exponential communication advantage: Proof sketch (quantum upper bounds)

Estimating gradients w.r.t. P parameters reduces to estimating P expectation values.

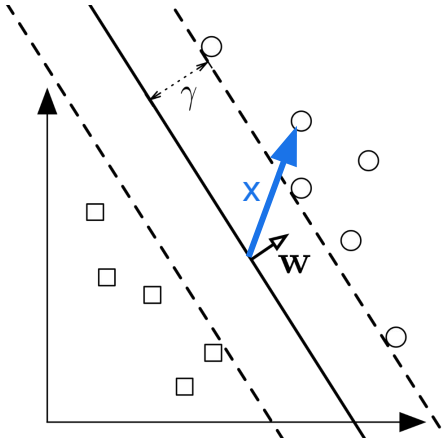
$$F(\theta) = \langle\psi(\theta)| O |\psi(\theta)\rangle \qquad F'_\ell = \langle\beta_\ell| \frac{\partial U_\ell}{\partial \theta_\ell} |\alpha_\ell\rangle + h.c.$$

Shadow Tomography [Aaronson 2017]
Given P observables O_1, ..., O_P and copies $|\phi\rangle$, estimate $\langle\phi| O_i |\phi\rangle$.

O(polylog(P)) copies suffice [Aaronson 2017].
⇒ Gradient computation requires O(polylog(P,N)) qubits [Abbas et al. 2023].
We extend this to the distributed setting.

Exponential advantage is impossible for some inference problems

Linear Classification with Margin
Inputs: Alice: w , Bob: x. $\|x\|_2 = \|w\|_2 = 1$.
Promised that $|x^T w| > \gamma$, determine $\text{sign}(x^T w)$.



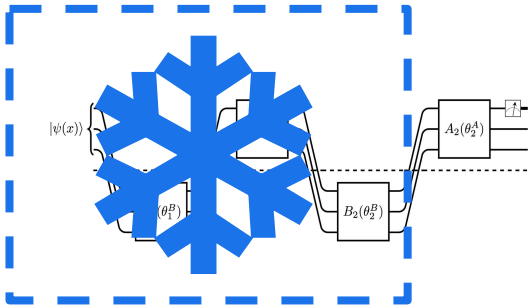
Lemma:
Linear classification requires $\Omega(\sqrt{(N/\max\{1,\lceil \gamma N \rceil\})})$ qubits of communication but only $O(\min(N, 1/\gamma^2))$ bits.

⇒ Classically easy for large γ, quantumly hard for small γ.

⇒ Advantage is not an immediate consequence of amplitude encoding

Beyond a single step of gradient descent

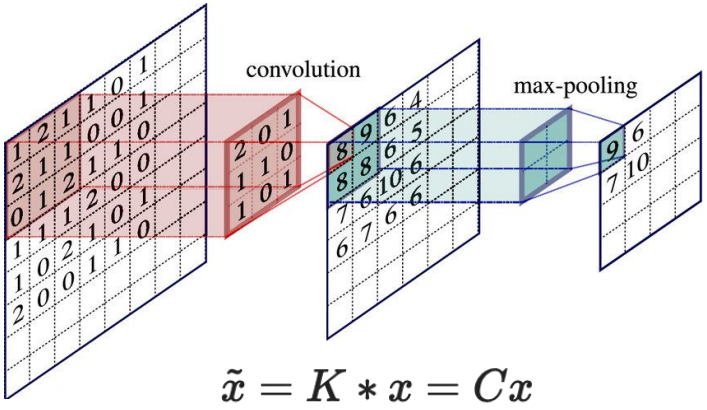
- When training only the last layer for T steps, only polylog(T) qubits are needed.



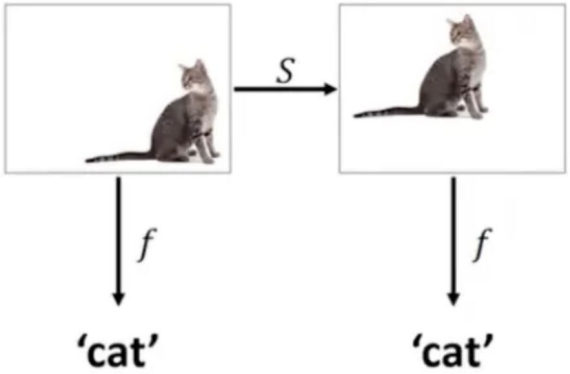
- Advantage holds for other (more efficient) algorithms, e.g. stochastic coordinate descent [Harrow & Napp 2021].
- Does advantage hold for useful functions?

$$|\psi(\theta)\rangle = U(\theta)\,|\phi\rangle = \prod_{\ell=1}^{M} U_\ell(\theta_\ell)\,|\phi\rangle$$

Convolutional neural networks



$$\tilde{x} = K * x = Cx$$



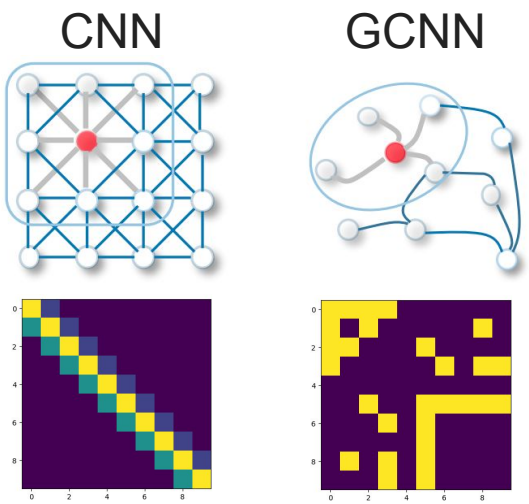
Translation symmetry of natural images baked into the architecture

⇒ Leads to learning of smoother functions that generalize better

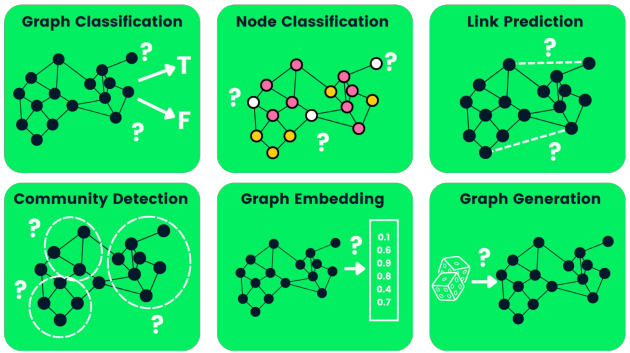
Generalizing CNNs to graphs

- Convolution has a natural analog in non-Euclidean geometry:
  Replace circulant C with adjacency matrix A

$$\tilde{x} = Ax$$



- Used to solve local and global problems on graphs [Kipf & Welling 2016].



Communication advantage in graph CNNs

- We study shallow, polynomial graph CNNs:

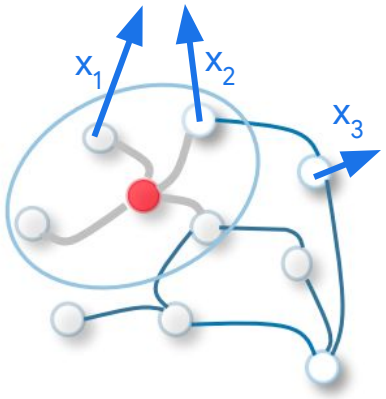
$$\varphi(X) = \mathrm{tr}\left[\mathcal{P}\left(\sigma(AXW_1)\right)W_2\right]$$

X: matrix of node features

A: message passing operator

σ: quadratic nonlinearity

P: pooling operator



(One-way) exponential communication advantage in graph CNN inference

$$\varphi(X) = \mathrm{tr}\left[\mathcal{P}\left(\sigma(AXW_1)\right)W_2\right]$$

Theorem: On a graph with N nodes, inference requires
i) O(log(N)) qubits of communication.
ii) $\Omega(\sqrt{N})$ bits of communication.

: X (node features)

: A, W_1 ,W_2 (Connectivity, weights)

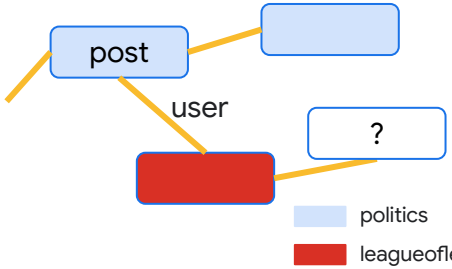
ii): Reduction from f-Boolean Hidden Partition [Doriguello & Montanaro 2020].

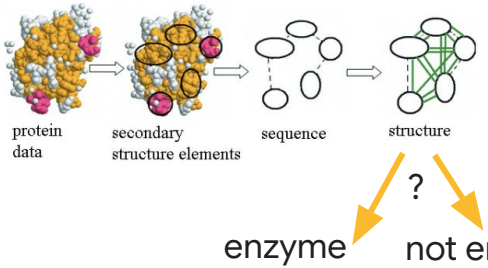
Matching performance of classical NNs with exponential communication advantage

Model performs well on standard benchmarks.

Model	Node Classification			Decision Problem		
	ogbn-products	Reddit	Cora	ogbn-products	Reddit	Cora
SIGN (PReLU)	79.48 ± 0.07	96.55 ± 0.02	78.84 ± 0.37	84.39 ± 1.73	90.33 ± 0.33	88.10 ± 5.61
Our Model	78.51 ± 0.05	96.31 ± 0.03	78.69 ± 0.26	83.70 ± 1.48	89.37 ± 0.60	87.14 ± 3.92

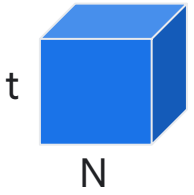
	Model	Dataset						
		MUTAG	PTC	NCI1	PROTEINS	COLLAB	IMDB-M	REDDIT-M
SOTA models	GIN	89.40±5.60	64.60±7.0	82.17±1.7	76.2 ±2.8	80.2 ±1.90	52.3 ±2.8	57.5±1.5
	DropGIN	90.4 ±7.0	66.3 ±8.6	-	76.3 ±6.1	-	51.4 ±2.8	-
	DGCNN	85.8 ±1.7	58.6 ±2.5	-	75.5 ±0.9	-	47.8 ±0.9	-
	U2GNN	89.97±3.65	69.63±3.60	-	78.53±4.07	77.84±1.48	53.60±3.53	-
	HGP-SL	-	-	78.45±0.77	84.91±1.62	-	-	-
	WKPI	88.30±2.6	68.10±2.4	87.5 ±0.5	78.5±0.4	-	49.5 ± 0.4	59.5 ± 0.6
	Our Model	92.02±6.45	68.0 ±8.17	77.25±1.42	76.55±5.10	81.82±1.42	53.13±3.01	54.09±1.76



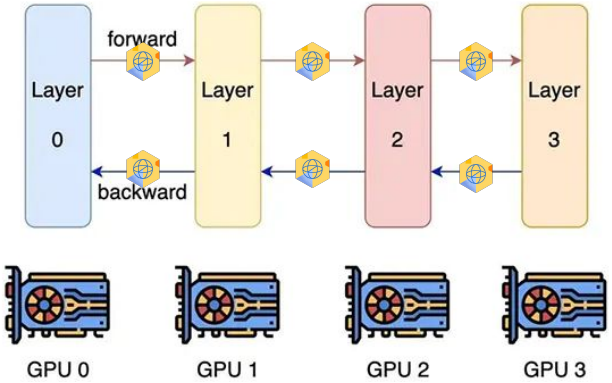


(10^6 nodes, 10^2 features per node)

The bottom line

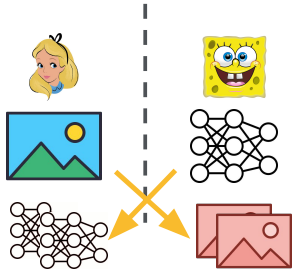


- Exponential speedups seem tricky for ML problems, where t = poly(N).
- Exponential, unconditional *communication* advantage for generic ML tasks is possible.



- Applies to circuits that look like realistic classical NNs and perform well on benchmarks.

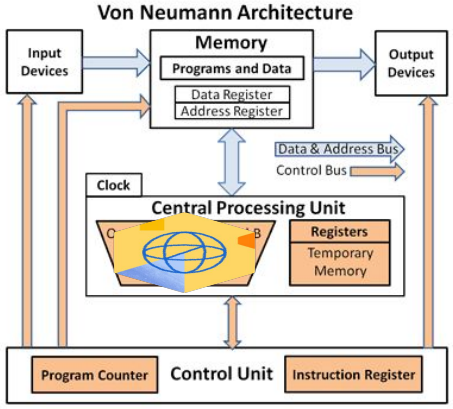
The frontier beyond speedups



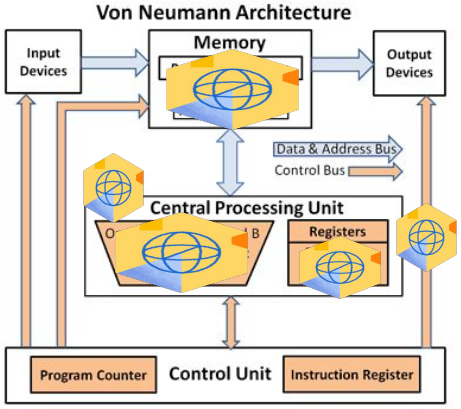
- Destructive measurement can prevent data re-use for computation

  Consumable Data via Quantum Communication arXiv: 2409.08495
  With Siddhartha Jain and Jarrod McClean

- What about other types of non-computational quantum advantage (e.g. space, latency, privacy, amortization, synchronization…)?



V.S.



(workshop in ~1 year)

Hagay Michaeli (Technion)

Daniel Soudry (Technion)

Jarrod McClean (Google)

Siddhartha Jain (UT Austin)

Thanks

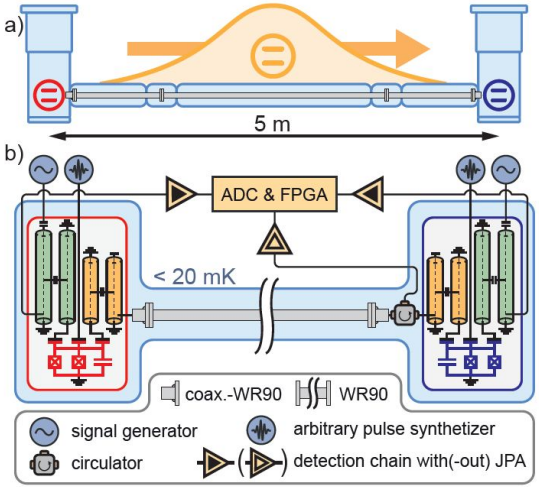
What would a quantum internet look like?

- Communicating arbitrary states reduces to
  1) sharing n Bell pairs
  2) O(n) bits of classical communication
  3) Local operations
  [Bennett 1993, Gordon et al. 2006].
- Fidelity of Bell pairs can be amplified exponentially.

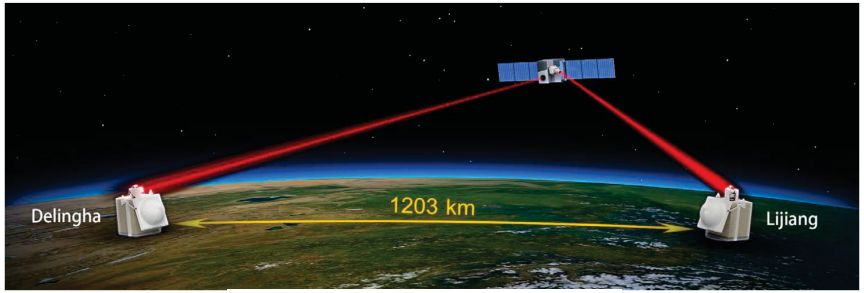
$$\frac{1}{\sqrt{2}}(|0\rangle\,|0\rangle + |1\rangle\,|1\rangle)$$

What would a quantum internet look like?

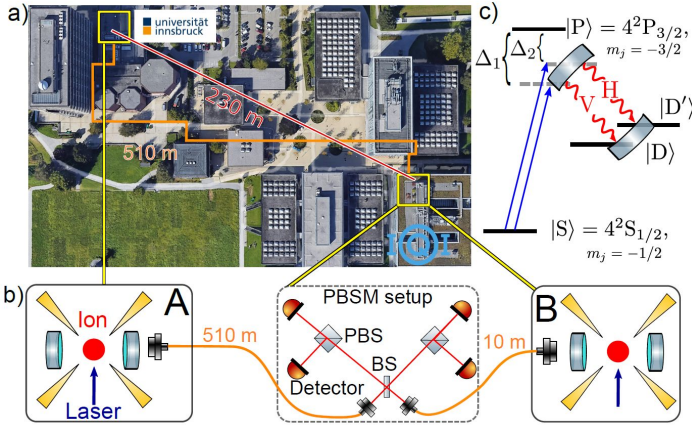
Sharing Bell pairs demonstrated across multiple media.



P. Magnard et al., Phys. Rev. Lett. 125, 260502 (2020)



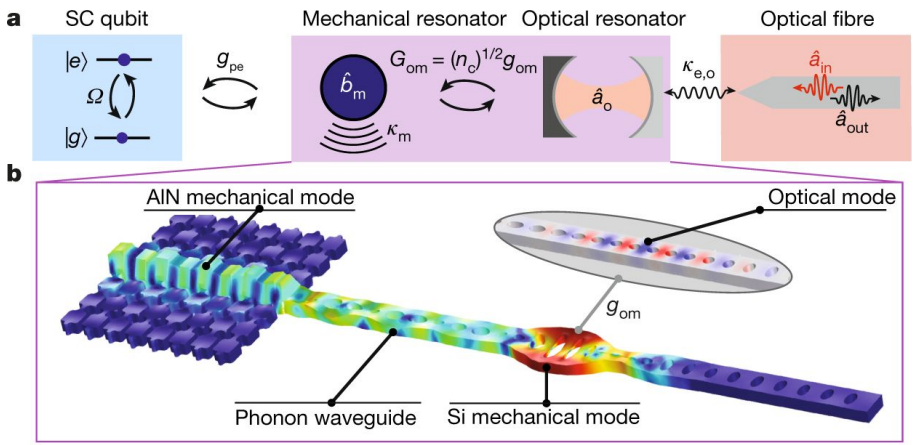
Nature 558, 264-267 (2018)



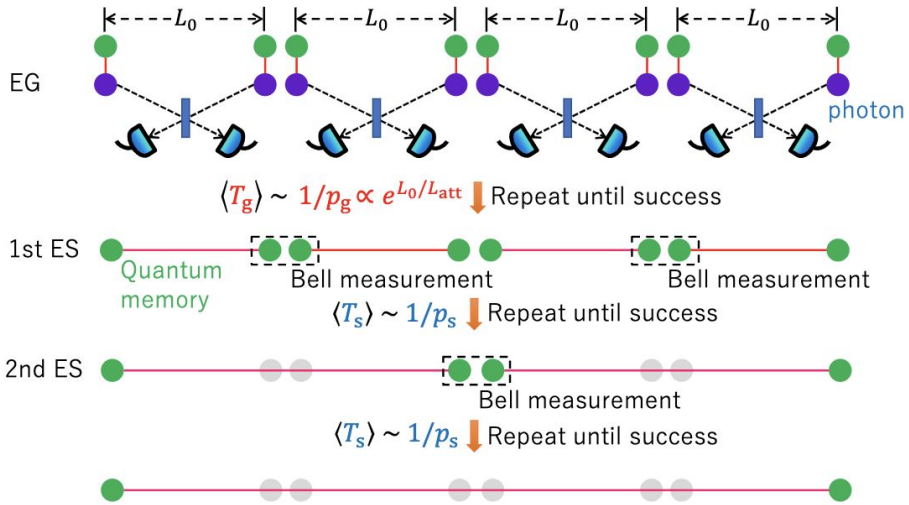
Phys. Rev. Lett. 130, 050803

What would a quantum internet look like?

- Challenges such as building repeaters, transduction between optical and “computational” (e.g. superconducting) qubits remain.



SC to optical transduction [Mirhosseini et al. 2023]



Quantum Repeaters [Azuma et al. 2023]

Quantum Sneakernet

Implementation	Qubit pitch (m)	Gate time (s)	Physical error rate	(d, N)	Memorystick capacity	Bandwidth (Hz)
NV^- (optical)	$6.6 \times 10^{-4\,33}$	$3.5 \times 10^{-6\,30}$	1×10^{-3}	(33, 4225)	12.7 KEb	7.3×10^{1}
Trapped ions	$1.5 \times 10^{-3\,36}$	$1.0 \times 10^{-4\,35}$	1×10^{-5}	(11, 441)	32 KEb	1.9×10^{2}
Transmons	$3.0 \times 10^{-4\,39}$	$4.0 \times 10^{-8\,31}$	1×10^{-5}	(13, 625)	2.4 MEb	1.4×10^{4}
Quantum dots	$1.0 \times 10^{-6\,28}$	$3.2 \times 10^{-8\,28}$	1×10^{-3}	(36, 5041)	2.8 TEb	1.6×10^{10}
NV^-	$3.0 \times 10^{-7\,29}$	$1.0 \times 10^{-3\,29}$	1×10^{-3}	(29, 3249)	200 TEb	1.6×10^{12}
silicon	$2.0 \times 10^{-7\,38}$	$5.0 \times 10^{-8\,37}$	1×10^{-3}	(36, 5041)	350 TEb	2.0×10^{12}

Table 1. Effective bandwidth of a transoceanic link. Effective bandwidth achieved using a single VLCS-class container ship transporting error-corrected quantum memories between Japan and the United States, estimated for a range of qubit implementations for a fixed infidelity of $1 - F = 10^{-10}$.

[Devitt et al. 2016]

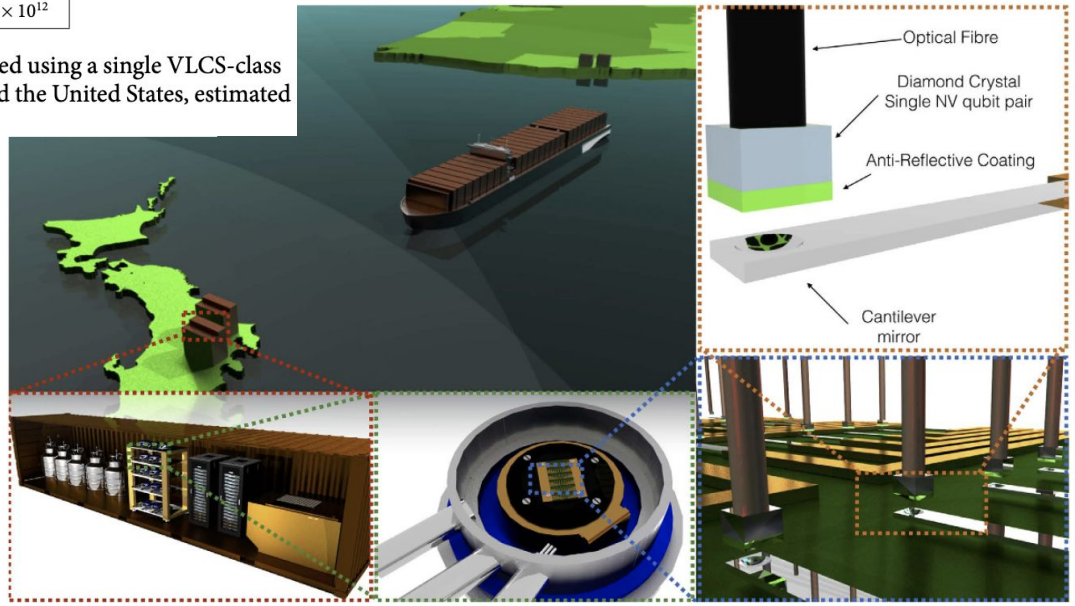
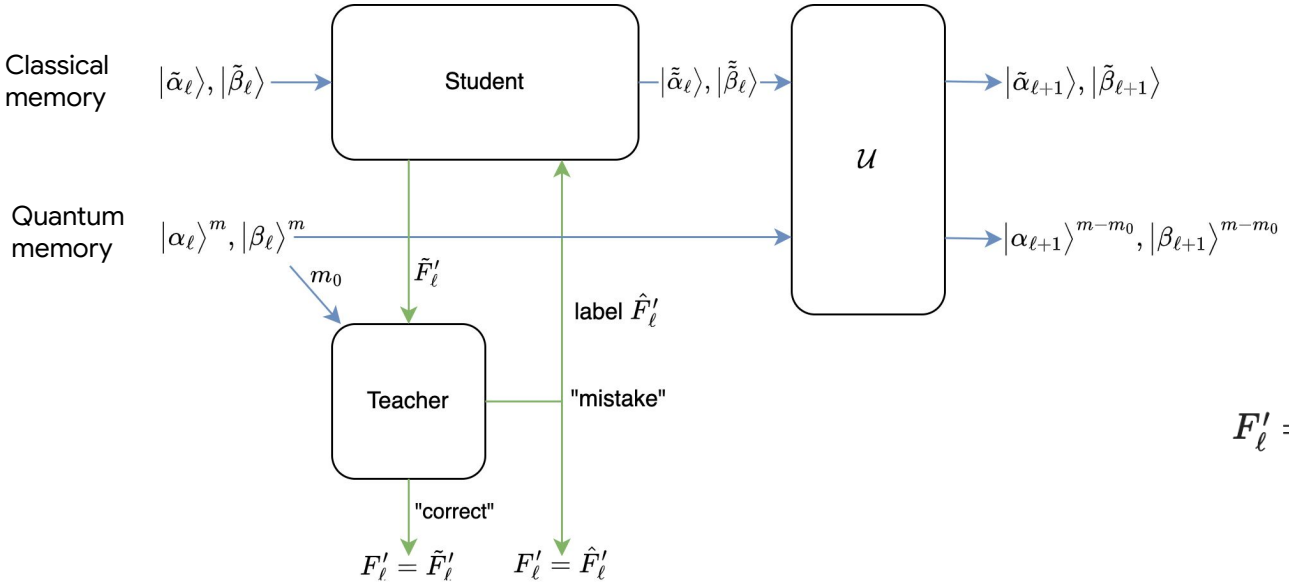


Figure 1. Physical transport protocol for a transpacific sneakernet using a single ship.

An exponential communication advantage - Proof sketch (quantum upper bounds)



$$F'_\ell = \langle\beta_\ell| \frac{\partial U_\ell}{\partial \theta_\ell} |\alpha_\ell\rangle + h.c.$$

(# mistakes)*(cost(Teacher)) = O(polylog(P,N)) [Aaronson 2017].
⇒ Gradient computation requires O(polylog(P,N)) qubits [Abbas et al. 2023].
We extend this to the distributed setting.