Computer Architecture

计 算 机 体 系 结 构

第24讲：Domain Specific Arch (3)

张献伟

xianweiz.github.io

DCS3013, 12/26/2022

Review Questions

- Why DSA?

  Ending of Moore’s law; limited perf impr of general-purpose.

- DSA design guidelines?

  Dedicated memories, larger ALUs, easy parallelism, smaller data size, domain-specific language.

- Target applications of TPU?

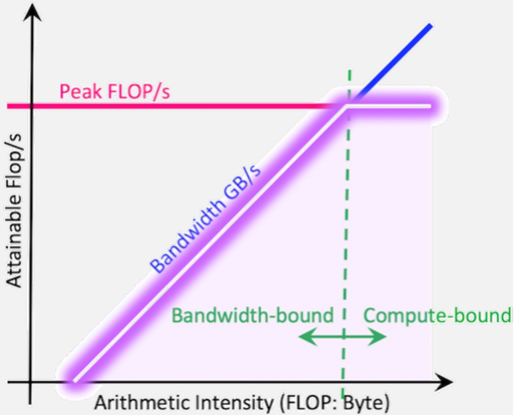
  DNN: MLP, CNN, LSTM.

- Why TPU uses ‘systolic execution’?

  Avoid repeated SRAM accesses, keeping matrix unit busy.

- Roofline model?

  Get perf bounds for compute and bandwidth limited applications.

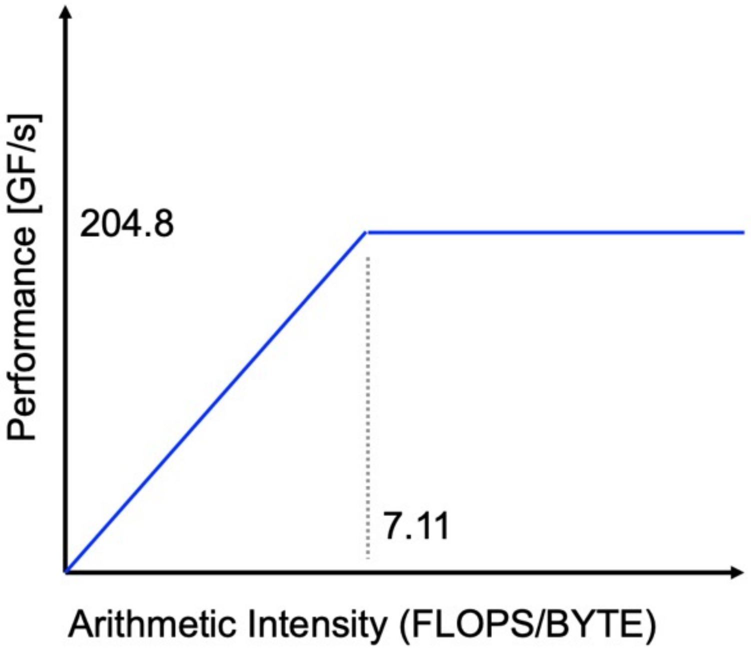


- X-axis of roofline graph?

  Computational/Arithmetic intensity.

Example

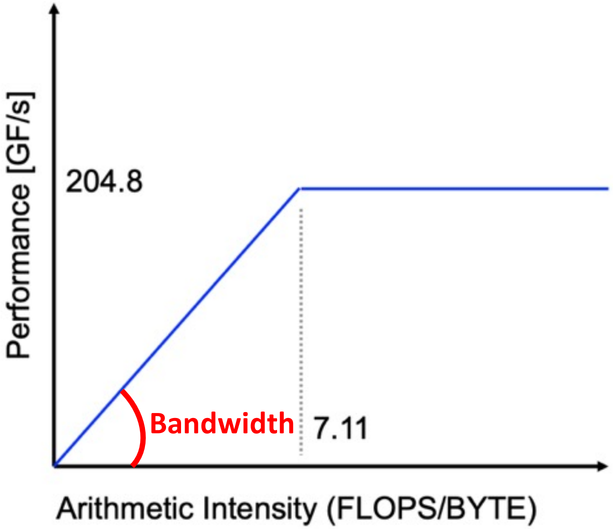
- Consider: for (i = 0; i < N; ++i) y[i] = a*x[i]+y[i]
  - For each "i" :
    - 1 addition, 1 multiplication
    - 2 loads of 8 bytes each
    - 1 store
- Execution on BlueGene/Q
  - Peak 204.8 GFLOP/node
- Performance estimates:
  - AI = 2/(3*8) = 1 / 12 1/12 < 7.11 → limited area on the Roofline plot
  - 7.11/(1/12)= 85.32
  - 204.8 / 85.32 = 2.4 GF/s



https://www.dam.brown.edu/people/lgrinb/APMA2821/Lectures_2015/APMA2821H-L_roof_line_model.pdf

Example (cont.)

- Peak double precision floating-point performance
  - 204.8 GFLOPS
- Peak memory bandwidth
  - 204.8/7.11 = 28.8 GB/s
  - The steady state bandwidth potential of the memory in a computer, <u>not</u> the pin bandwidth of the DRAM chips
  - Common way is to measure it with benchmarks like STREAM

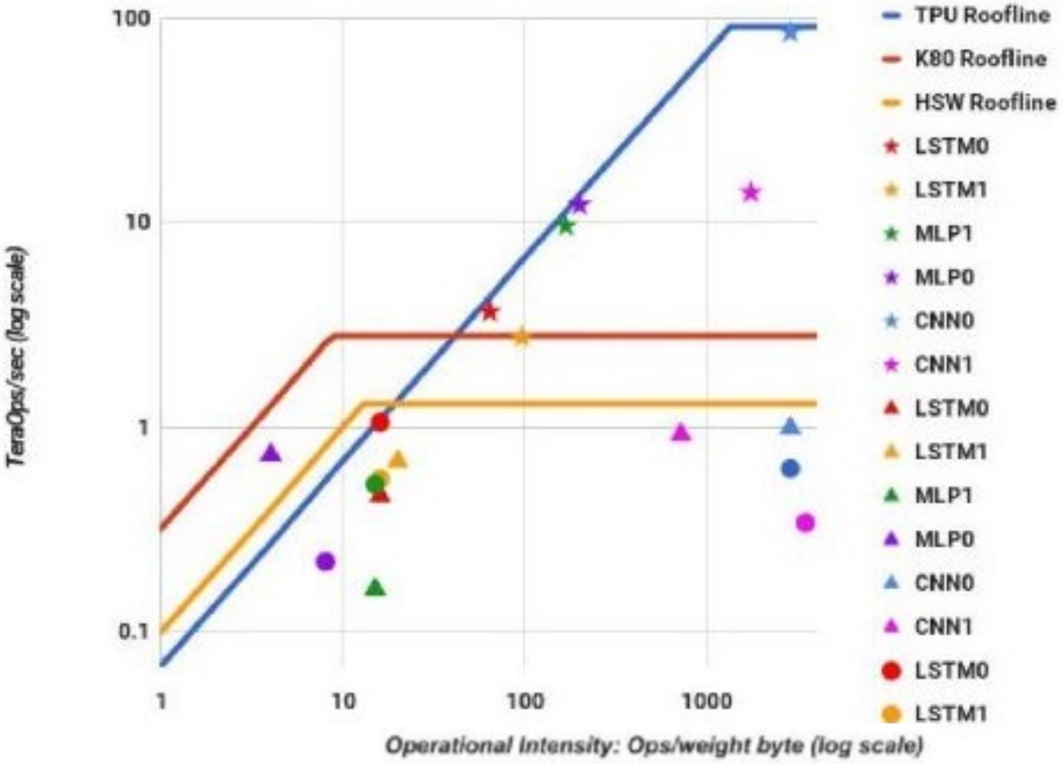


- Lawrence Livermore National Laboratory (LLNL)
- IBM BlueGene/Q
- Top500 2012/6 #1, 16.3PFLOPS (efficiency 81%), 1.57Mcore, 7.89MW, 2.07GFLOPS/W
- 4 BlueGene/Q were listed in top10
- 18core/chip, 1.6GHz, 4way SMT, 204.8GFLOPS/55W, L2:32MB eDRAM, mem: 16GB, 42.5GB/s
- 32chip/node, 32node/rack, 96rack

http://www.hpcs.cs.tsukuba.ac.jp/~taisuke/tmp/ACHPC-LN9-4cut.pdf

TPU Roofline Performance

- TPU: its ridge point is far to the right at 1350
  - CNN1 is much further below its Roofline than the other DNNs
    - ▫ Waiting for weights to be loaded into the matrix unit
  - Ridge point comparison:
    - ▫ CPU: 13, GPU: 9 → better balanced, but perf a lot lower



☆ Star = TPU
△ Triangle = GPU
○ Circle = CPU

Cost-Performance

- Cost metric: performance per watt
  - "Total": includes the power consumed by the host CPU server when calculating perf/watt for the GPU and TPU
  - "Incremental": subtracts the host CPU power from the total
- Total: GPU is 2.1x CPU, TPU is 34x CPU
- Incremental: TPU is 83x CPU, 29x GPU

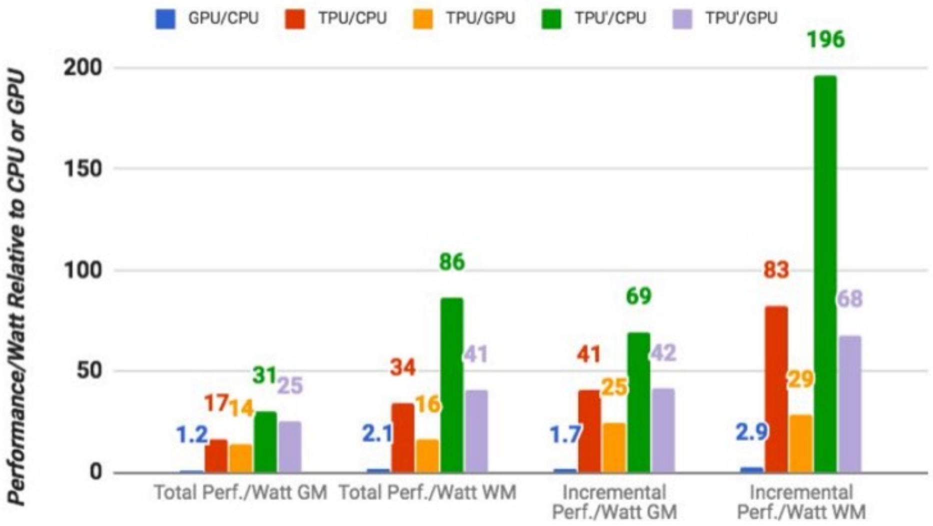
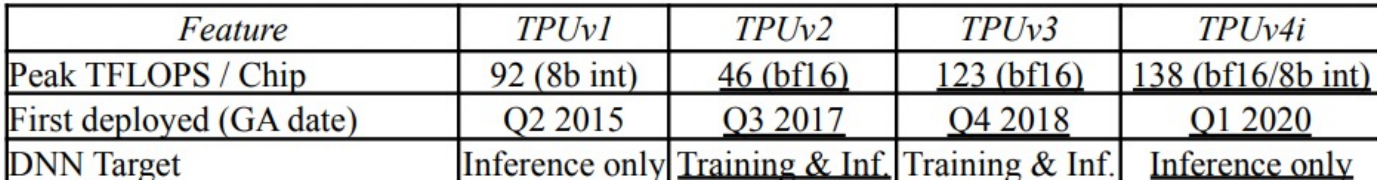


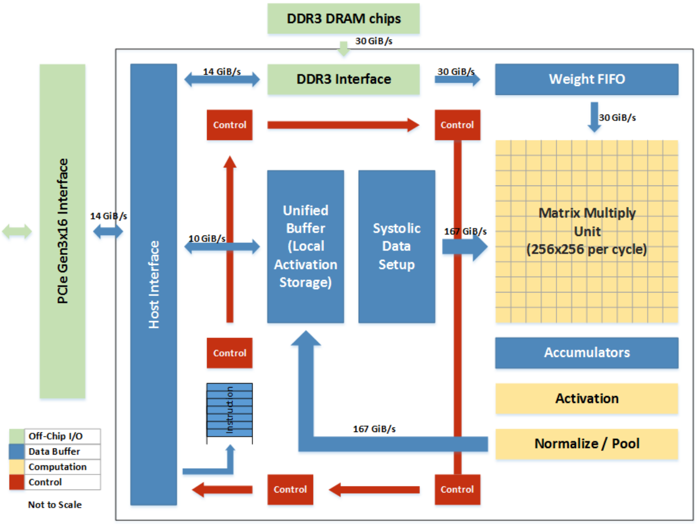
Figure 9. Relative performance/Watt (TDP) of GPU server (blue bar) and TPU server (red bar) to CPU server, and TPU server to GPU server (orange bar). TPU' is an improved TPU (Sec. 7). The green bar shows its ratio to the CPU server and the lavender bar shows its relation to the GPU server. Total includes host server power, but incremental doesn't. GM and WM are the geometric and weighted means.

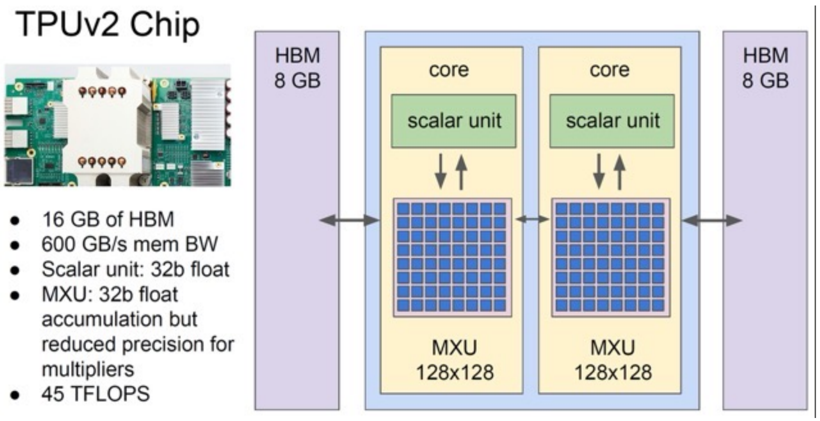
https://www.extremetech.com/computing/247199-googles-dedicated-tensorflow-processor-tpu-makes-hash-intel-nvidia-inference-workloads

TPU Generations



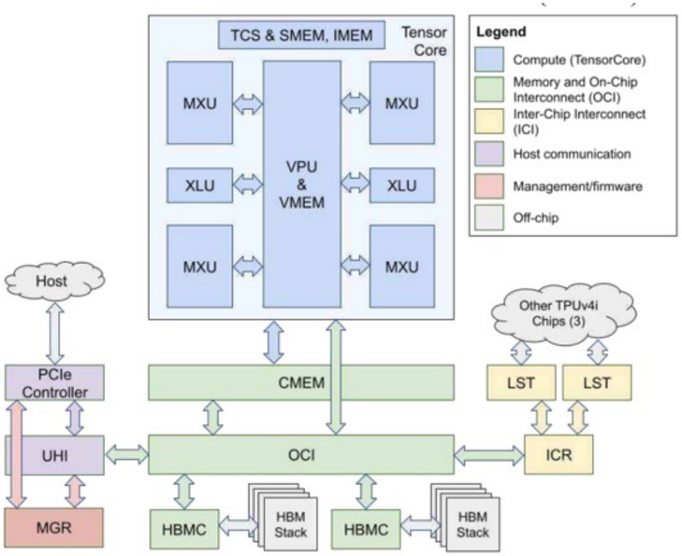
Feature	*TPUv1*	*TPUv2*	*TPUv3*	*TPUv4i*
Peak TFLOPS / Chip	92 (8b int)	46 (bf16)	123 (bf16)	138 (bf16/8b int)
First deployed (GA date)	Q2 2015	Q3 2017	Q4 2018	Q1 2020
DNN Target	Inference only	Training & Inf.	Training & Inf.	Inference only





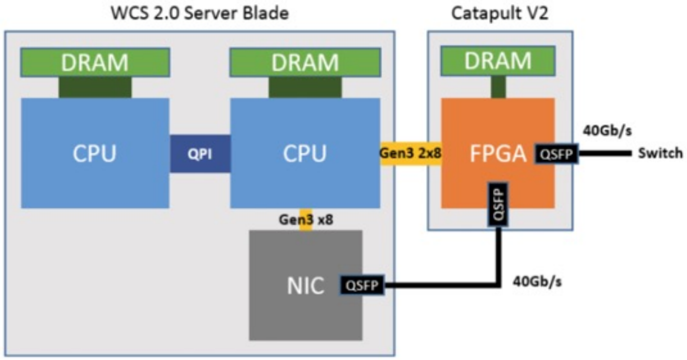
TPUv2 Chip

- 16 GB of HBM
- 600 GB/s mem BW
- Scalar unit: 32b float
- MXU: 32b float accumulation but reduced precision for multipliers
- 45 TFLOPS



Microsoft Catapult

- Project Catapult
  - To transform cloud computing by augmenting CPUs with an interconnected and configurable compute layer composed of programmable silicon
- FPGAs offer a unique combination of speed and flexibility
  - FPGAs could deliver efficiency and performance without the cost, complexity, and risk of developing custom ASICs
  - The FPGA can act as a local compute accelerator, an inline processor, or a remote accelerator for distributed computing

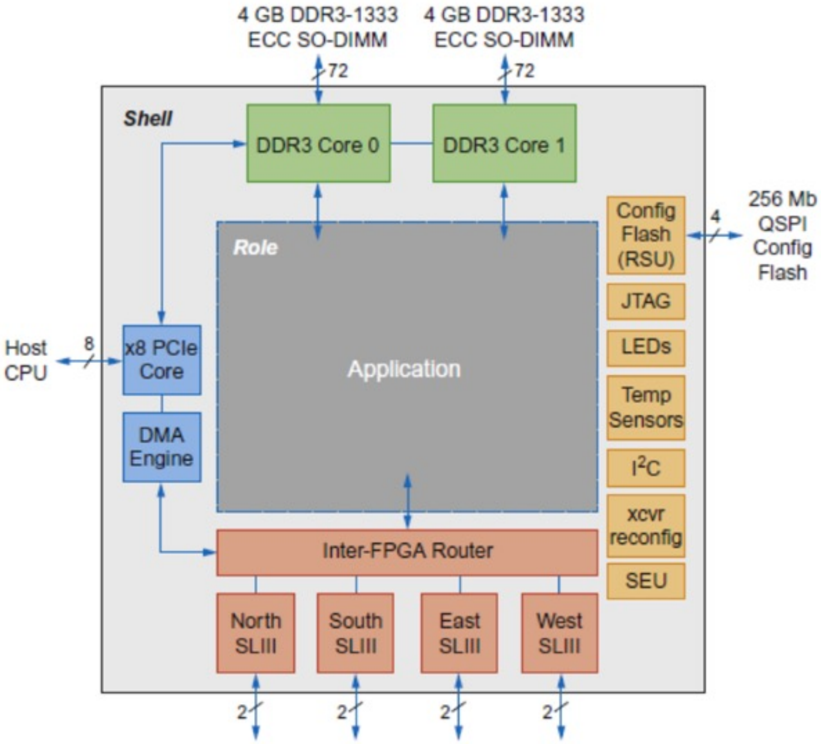


Catapult v2 Mezzanine card

https://www.microsoft.com/en-us/research/project/project-catapult/

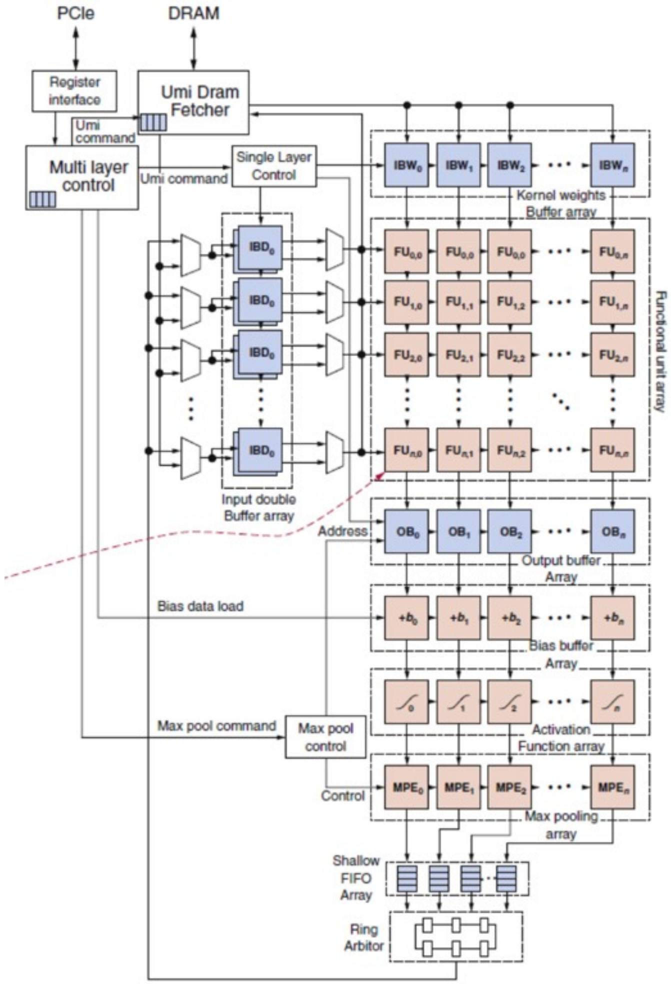
Microsoft Catapult (cont.)

- Needed to be general purpose and power efficient
  - Uses FPGA PCIe board with dedicated 20 Gbps network in 6 x 8 torus
  - Each of the 48 servers in half the rack has a Catapult board
  - Limited to 25 watts
  - 32 MB Flash memory
  - Two banks of DDR3-1600 (11 GB/s) and 8 GB DRAM
  - FPGA (unconfigured) has 3962 18-bit ALUs and 5 MB of on-chip memory
  - Programmed in Verilog RTL
  - Shell is 23% of the FPGA

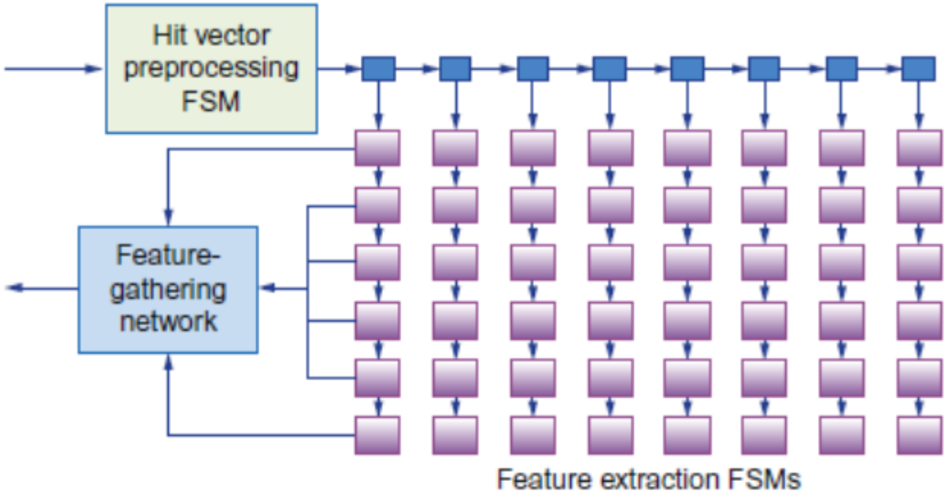


Catapult Applications

- The processing element (PE) of the CNN Accelerator for Catapult



- The architecture of FPGA implementation of the Feature Extraction stage in search acceleration



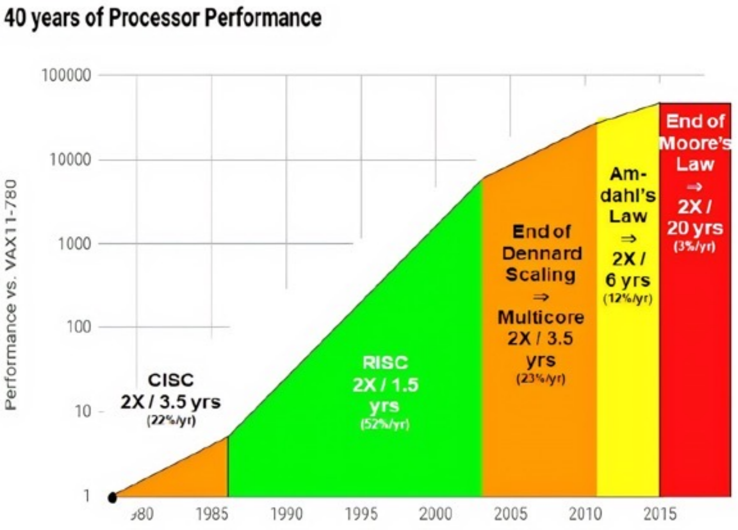
How Catapult Follows the Guidelines

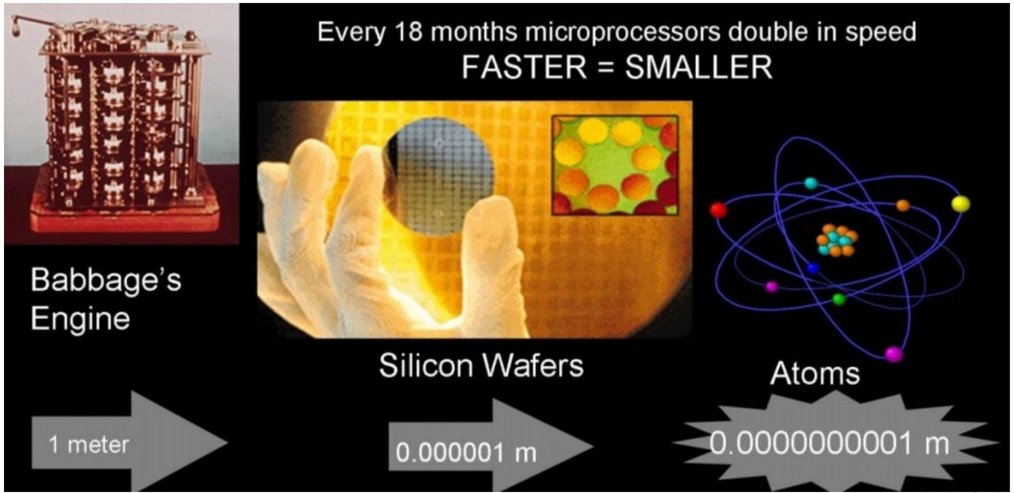
- *Use dedicated memories*
  - 5 MB dedicated memory
- *Invest resources in arithmetic units and dedicated memories*
  - 3926 ALUs
- *Use the easiest form of parallelism that matches the domain*
  - 2D SIMD for CNN, MISD parallelism for search scoring
- *Reduce the data size and type needed for the domain*
  - Uses mixture of 8-bit integers and 64-bit floating-point
- *Use a domain-specific programming language*
  - Uses Verilog RTL; Microsoft did not follow this guideline

Quantum Computing

The End of Moore’s Law

- Today’s computers work based on a “classical mechanics” framework where 1 is 1 and 0 is 0[经典力学]
  - Chip components were continuously shrinking and added
  - But, physical components cannot be reduced in size infinitely
- At atomic level, particles behave according to the laws of quantum mechanics rather than classical[量子力学]
  - Even defining 1s and 0s becomes a major problem at this level

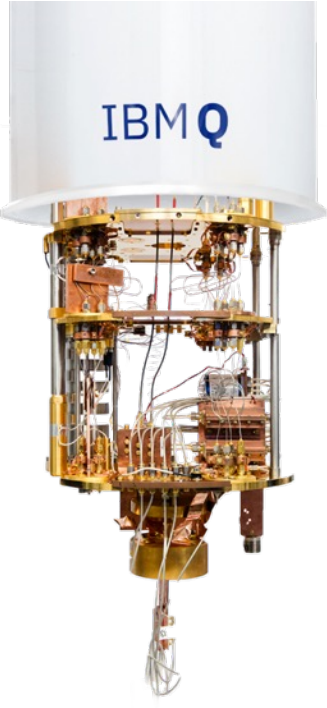


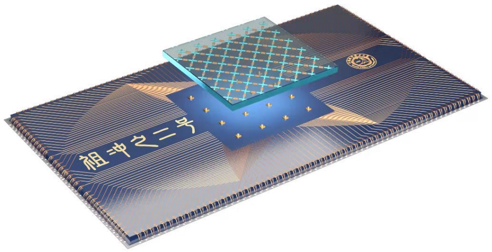


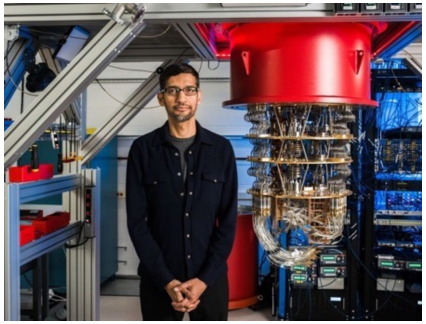
https://www.huawei.com/en/huaweitech/publication/86/quantum-computing-ai

Quantum Computer[量子计算机]

- 2017, D-Wave 2000Q, 2000 qubit
  - Designed to solve opt problems (used by NASA)
- 2019, IBM Q System One, 20 qubit
  - General purpose
- 2019, Google quantum computer, 53 qubit
  - Claimed “quantum supremacy”[量子霸权]
- 2018, Intel/Rigetti/IonQ quantum computer
  - Azure Quantum, Amazon Braket
- 2021, 66-qubit two-dimensional superconducting quantum processor







Quantum supremacy or quantum advantage is the goal of demonstrating that a programmable quantum device can solve a problem that no classical computer can solve in any feasible amount of time.

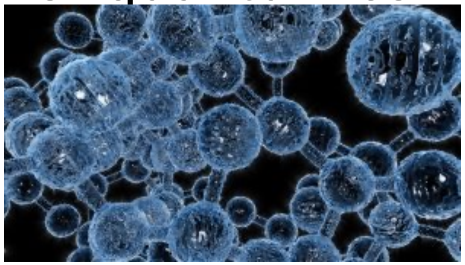
Quantum Supremacy

- Google declared Quantum Supremacy in 2019
  - The Sycamore superconductive quantum computer is over a billion times faster than Summit (comparing 200 seconds against 10,000 years in the task of measuring/simulating one million samples)
- IBM challenged the supremacy in 2019
  - An ideal simulation of the same task can be performed on a classical system in 2.5 days and with far greater fidelity. This is in fact a conservative, worst-case estimate, …
- Zhejiang Lab closes the supremacy gap in 2021
  - A random quantum circuit simulator on the Sunway exascale system. Reduced the simulation sampling time to 304 seconds from that previous estimate of 10,000 years

https://www.ibm.com/blogs/research/2019/10/on-quantum-supremacy/
https://www.hpcwire.com/2021/11/18/2021-gordon-bell-prize-goes-to-exascale-powered-quantum-supremacy-challenge/

Quantum Computer

- Quantum computers behave very differently from ordinary computers
  - Quantum computing is inherently **probabilistic**, which means it solves challenges based on the most probable outcome while using several dimensions simultaneously
- Usages
  - A promising computing paradigm with great potential in cryptography[Shor 1999], database[Grover 1996], linear systems[Harrow 2009], chemistry simulation[Peruzzo 2014], etc
  - Several quantum program languages[Abraham 2019; Google 2018; Green 2013; Paykin 2017; Rigetti Forest team 2019; Svore 2018] have been published to write quantum programs for quantum computers

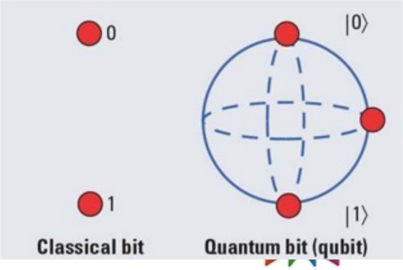


https://dl.acm.org/doi/pdf/10.1145/3428218

https://www.redhat.com/architect/quantum-computing

Qubit[量子比特]

- Quantum computing stores and transfers digital info using
  - A microscopic object (e.g., electron, photon, ion) as the medium[微观物体]
- One bit info (i.e., 0/1) can be encoded using two orthogonal states of a microscopic object
  - 微电子：自旋向上或向下，光子：极化方向水平或垂直
- The quantum two-state system is called a ***quantum bit*** (or ***qubit***)
- A quantum computer solves a problem by setting qubits in initial states and then manipulating the states so that an expected result appears on the qubits
  - Quantum mechanics is used to describe the states



Qubit (cont.)

- The state of a qubit can be written as a vector
  - 在|0⟩在振幅为α (一次测量中，结果为|0⟩概率α²)，在|0⟩上振幅为β

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \qquad |\alpha|^2 + |\beta|^2 = 1$$

- The orthonormal basis 0 and 1 can be written as

$$|0\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad |1\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

- Qubit basis states can also be combined to form product basis states

$$|00\rangle = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, |01\rangle = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, |10\rangle = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \text{ and } |11\rangle = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

- A **quantum state** is defined by a gathering of all physical properties of a quantum system, which includes four main properties: 1) position, 2) momentum, 3) spin, 4) polarization

https://en.wikipedia.org/wiki/Qubit

Bra-Ket Notation

- To describe a quantum state in terms of vectors
- Ket[右矢]: $|v\rangle$
  - The column vectors on the right in Dirac form (|>)
- Bra[左矢]: $\langle v|$
  - The row vectors on the left in Dirac form (<|)
  - The conjugate transpose of ket

$$\langle 0| = [1\ 0]$$

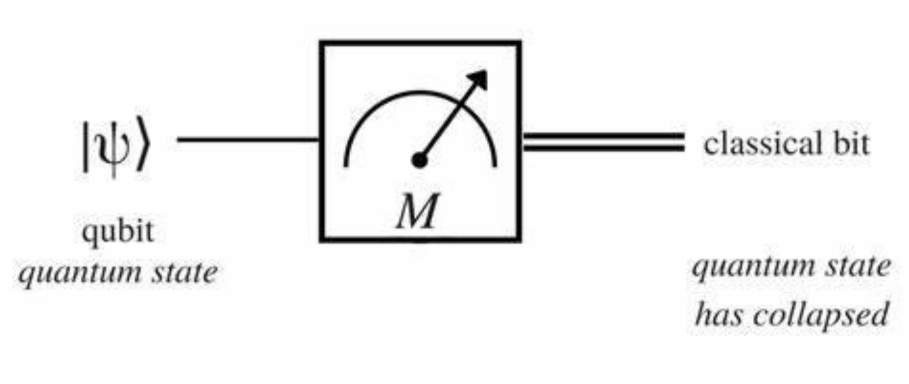
$$\langle 1| = [0\ 1]$$

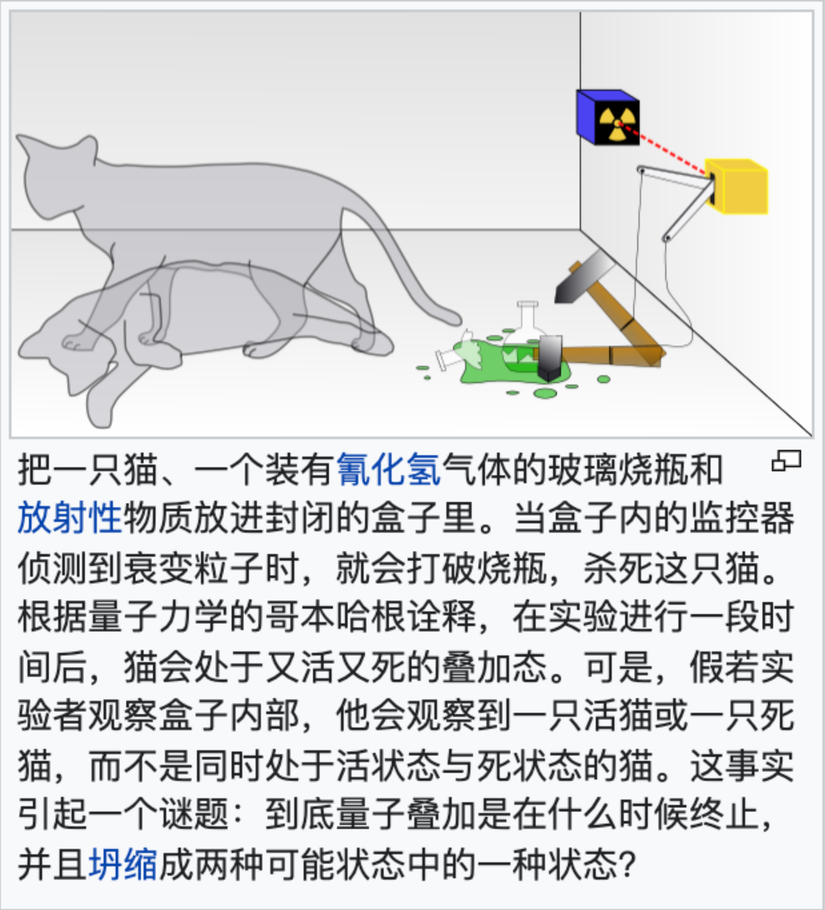
$$|0\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$|1\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Schrödinger’s Cat[薛定谔的猫]

- The cat after a while is dead and alive at the same time, but when one looks into the box, it collapses into one state, where the cat is either dead or alive (but not both)
- Quantum measurement

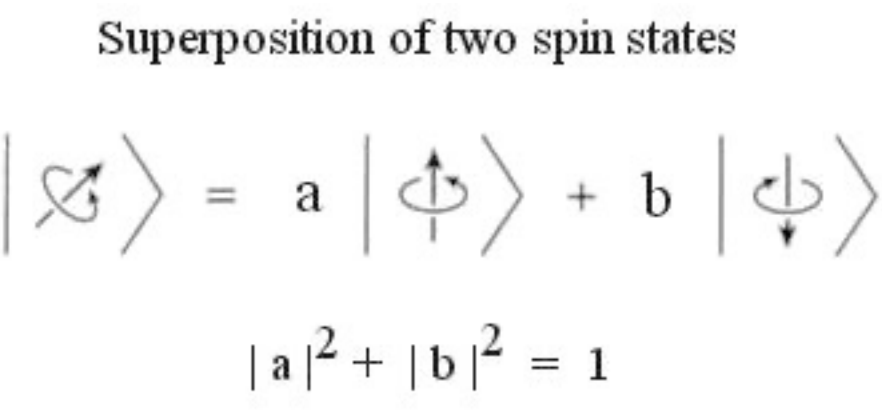




把一只猫、一个装有氰化氢气体的玻璃烧瓶和放射性物质放进封闭的盒子里。当盒子内的监控器侦测到衰变粒子时，就会打破烧瓶，杀死这只猫。根据量子力学的哥本哈根诠释，在实验进行一段时间后，猫会处于又活又死的叠加态。可是，假若实验者观察盒子内部，他会观察到一只活猫或一只死猫，而不是同时处于活状态与死状态的猫。这事实引起一个谜题：到底量子叠加是在什么时候终止，并且坍缩成两种可能状态中的一种状态?

Superposition State[叠加态]

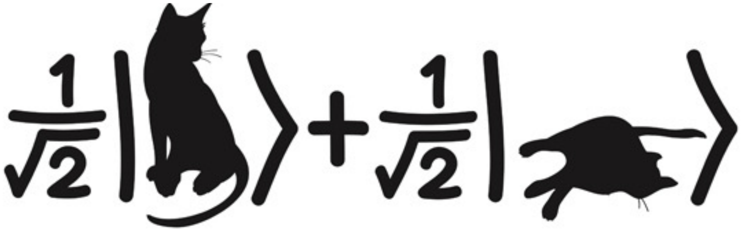
- The states “0” and “1” exists at the same time



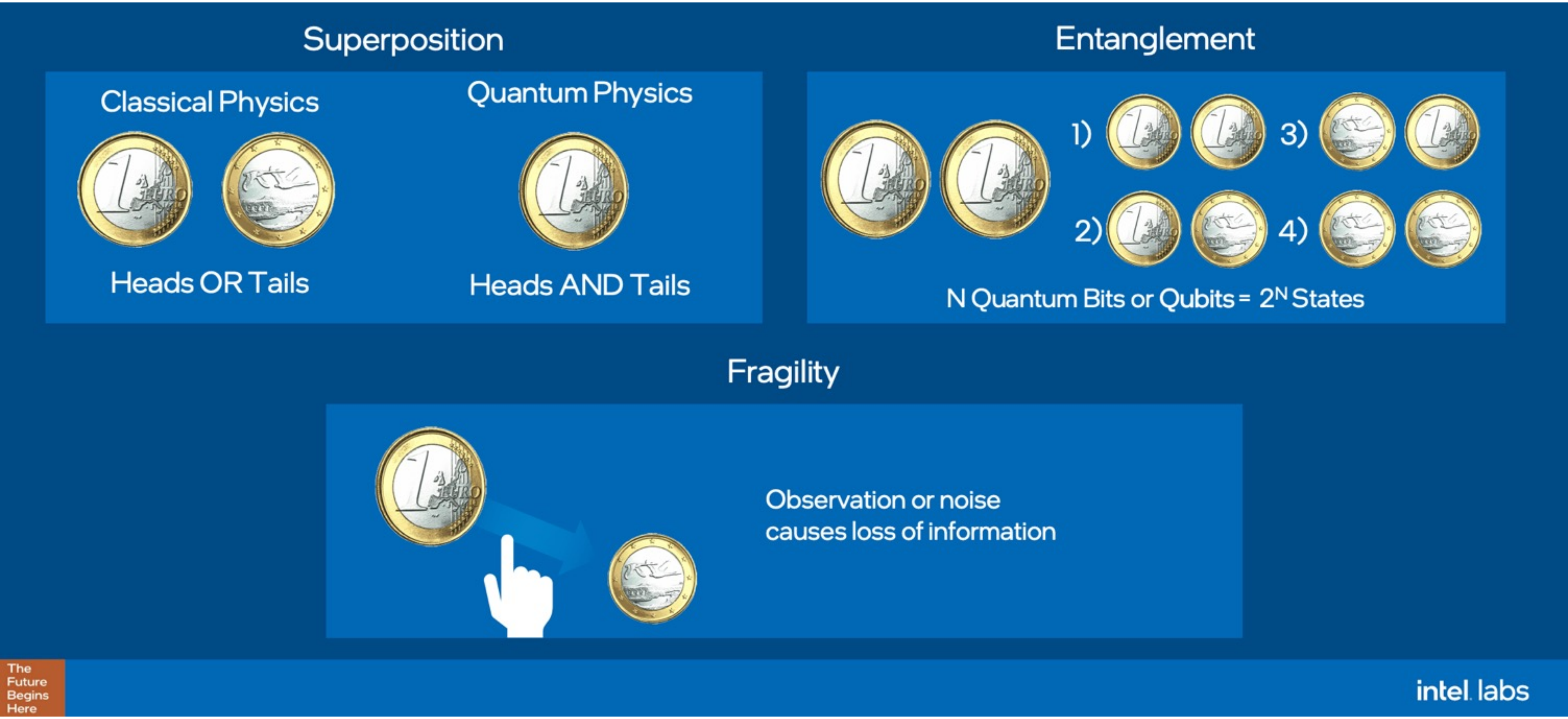
Superposition of two spin states

$$|a|^2 + |b|^2 = 1$$

- A quantum system consisting of *n* qubits can exist in a linear superposition of 2^n basis states
  - In contrast to a classical system of *n* bits which can exist as exactly a single of these states



Quantum Computing: Key Concepts



Superposition

Classical Physics

Heads OR Tails

Quantum Physics

Heads AND Tails

Entanglement

1) 2) 3) 4)

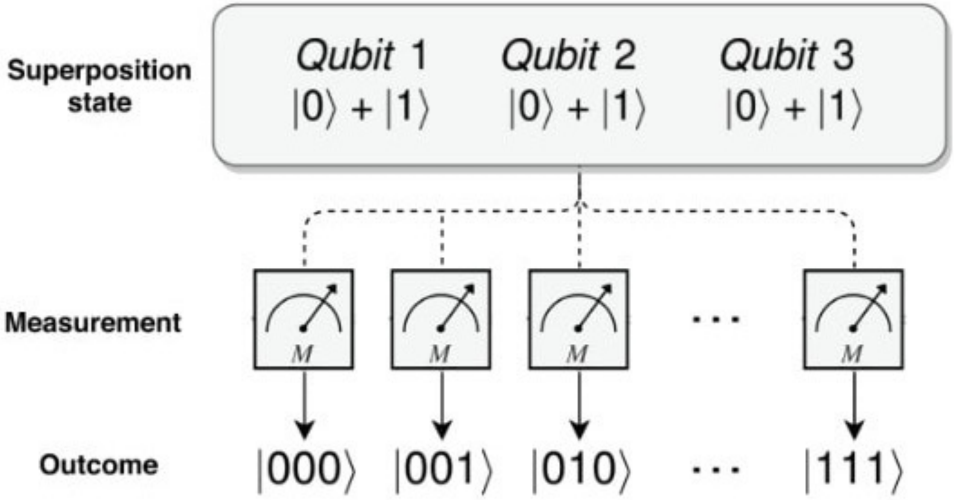
N Quantum Bits or Qubits = 2^N States

Fragility

Observation or noise
causes loss of information

Measurement[测量]

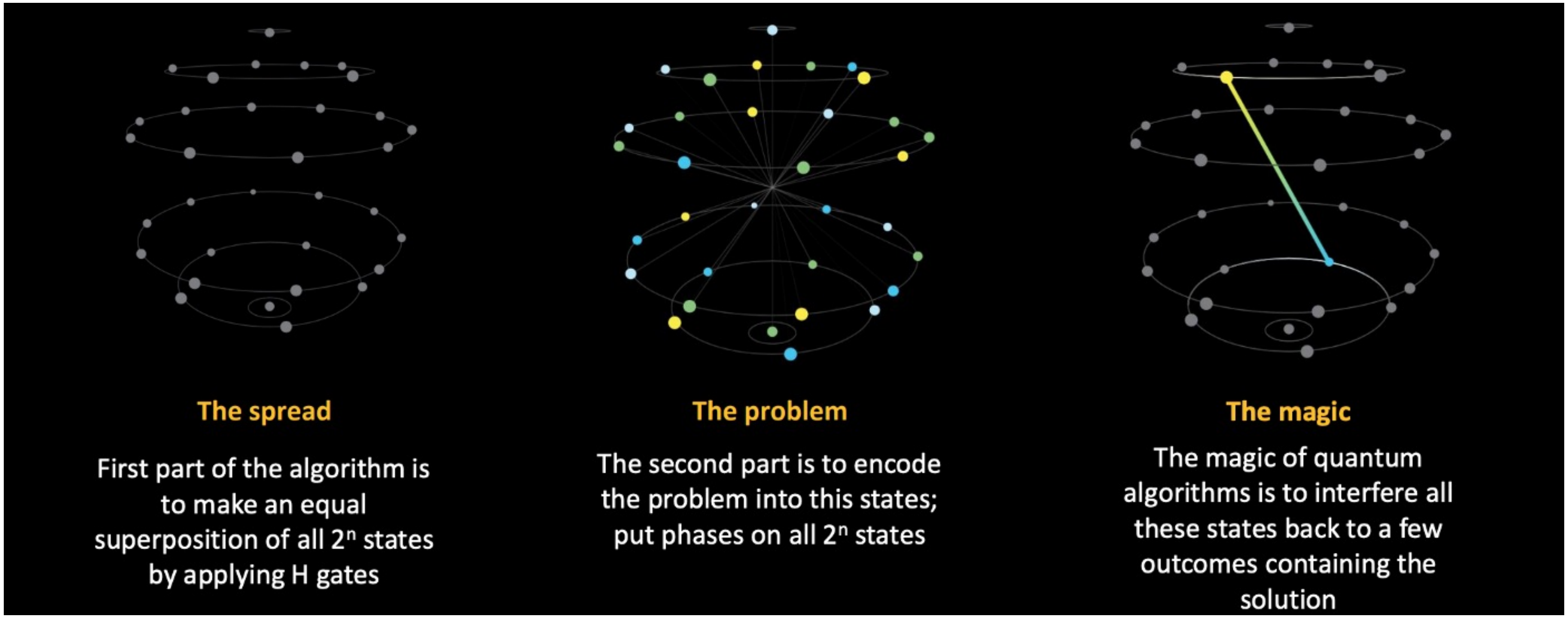
- While a qubit system can exist in these superpositions during computation, at the end of the computation, the qubits are measured producing a classical binary outcome
  - The probability of each outcome depends on the amplitude[振幅] of each basis state[基态] (the values of α, β, γ, …)
- **Measurement** = projection of state to a basis vector
  - Changes the state: superposition is destroyed



ASPLOS'2021, Orchestrated Trios
https://arcb.csc.ncsu.edu/~mueller/qc/qc18-2/qc18/readings/quantum_circuits_part1.pdf

Quantum Algorithm[量子算法]

- General flow
  - First, we prepare the superposition
  - Then, we encode the problem info into the superposition and manipulate it in a high dimensional space
  - Finally, we apply interference to consolidate the superposition into fewer outcomes



The spread

First part of the algorithm is to make an equal superposition of all 2^n states by applying H gates

The problem

The second part is to encode the problem into this states; put phases on all 2^n states

The magic

The magic of quantum algorithms is to interfere all these states back to a few outcomes containing the solution

medium