Sustainability – quo vadis?
The Journey to Sustainability

—
Alessandro Curioni, PhD

IBM Fellow
Vice President, IBM Research Europe and Africa
Director, IBM Research Europe – Zurich

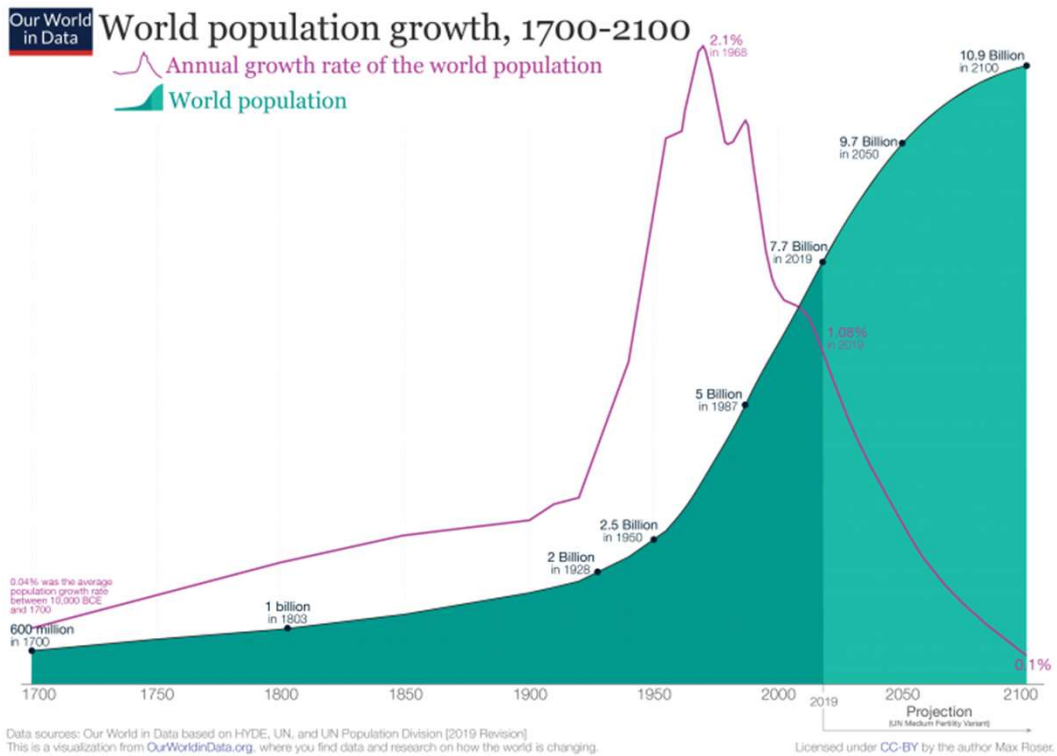
@ale_curioni

Our world in numbers: Population growth

10 Billion
in 2050s *(projected)*

8 Billion
in 2020s

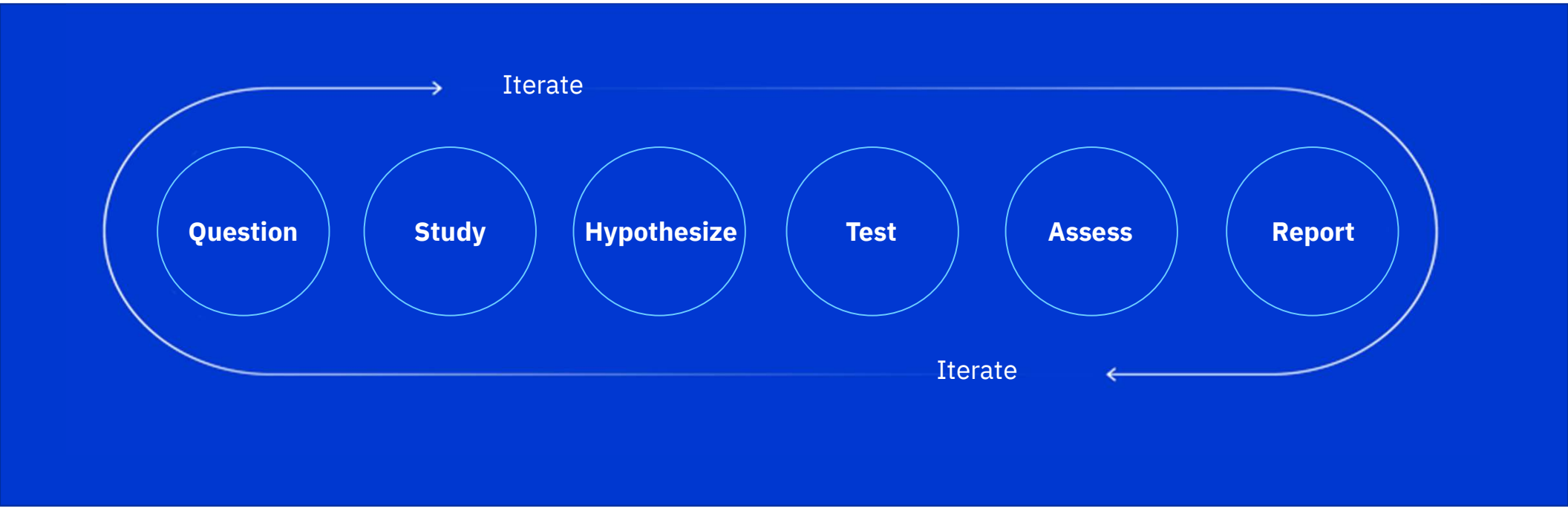
600 Million
in 1700s



The urgency of science has never been greater

How do we discover solutions to complex problems?

The digital journey to answering complex questions



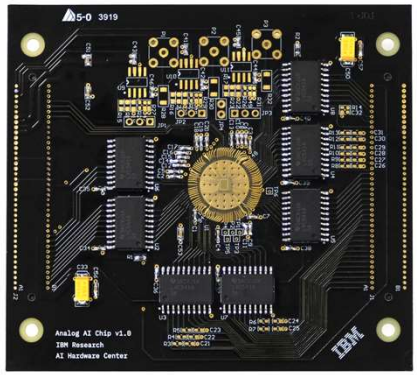
Sustainable material

It typically takes roughly
10 years and
$10 - $100 million to
discover a new material

We aim to cut down the time and cost by 90%.

Supercomputers

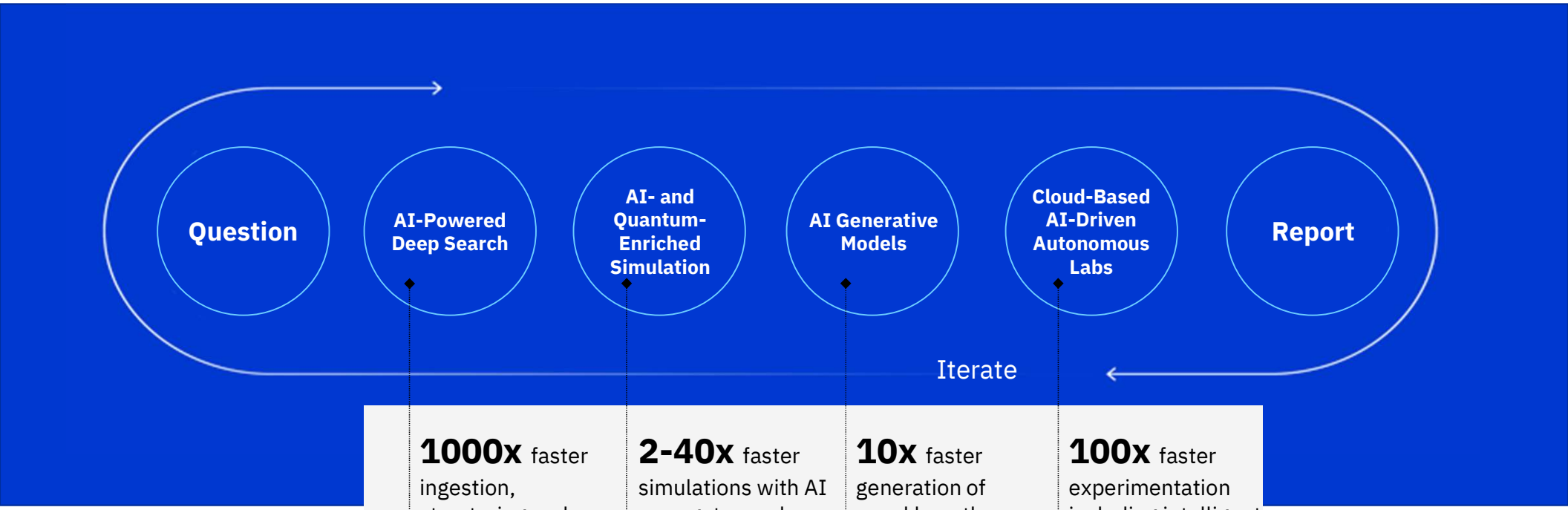
Quantum Computers



AI Systems

Hybrid Cloud

AI-powered Accelerated Discovery



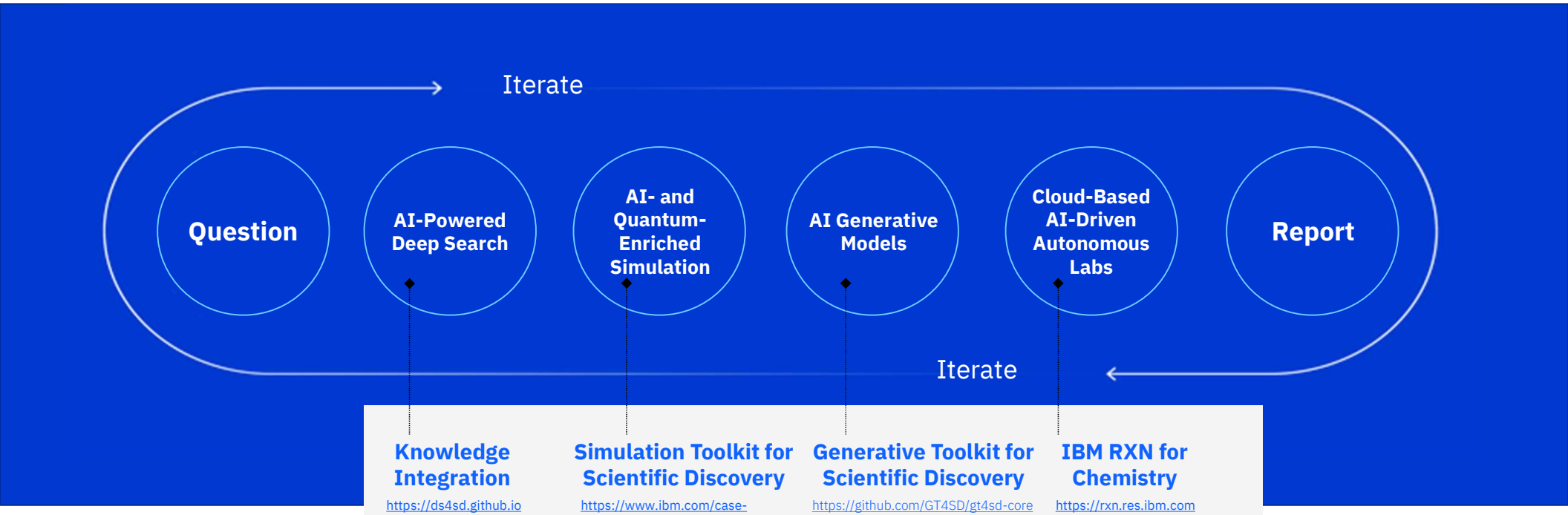
1000x faster ingestion, structuring and reasoning with growing scientific literature

2-40x faster simulations with AI surrogates and intelligent workflows to fill in knowledge gaps & augment data

10x faster generation of novel hypotheses including inverse design of materials and property prediction

100x faster experimentation including intelligent retrosynthesis, procedure prediction and automation

AI-powered Accelerated Discovery



Knowledge Integration
https://ds4sd.github.io

Simulation Toolkit for Scientific Discovery
https://www.ibm.com/case-studies/ibm-bayesian-optimization-accelerator-software

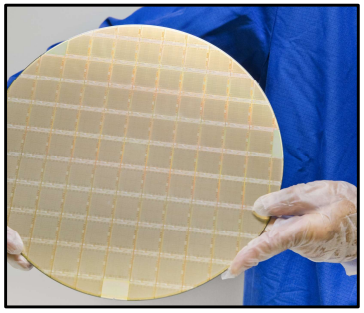
Generative Toolkit for Scientific Discovery
https://github.com/GT4SD/gt4sd-core

IBM RXN for Chemistry
https://rxn.res.ibm.com

Available online

Addressing needs in multiple domains

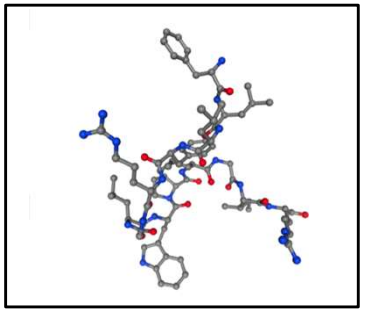
Sustainable semiconductors



Discovery and synthesis of a new photoacid generator molecule in less than 1 year

Pyzer-Knapp *et al. npj Comput. Mater.* **8** (2022) 84
https://research.ibm.com/science/photoresist

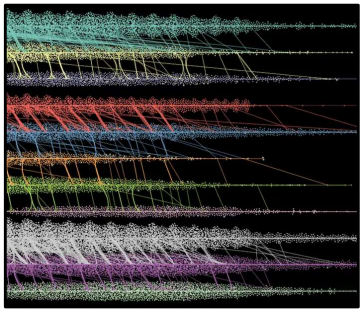
Therapeutics



Discovery and validation of two new antimicrobial compounds in 48 days instead of 2-4 years

Das *et al. Nat. Biomed. Eng.* **5** (2021) 613
https://research.ibm.com/publications/accelerated-antimicrobial-discovery-via-deep-generative-models-and-molecular-dynamics-simulations

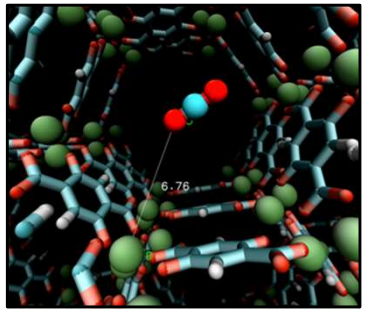
Biomarkers



Discovery of biomarkers and new disease trajectories for Type 1 diabetes

Kwon *et al. Nat. Commun.* **13** (2022) 1514
https://research.ibm.com/blog/ai-predicting-onset-of-type-1-diabetes

Climate & Sustainability



Discovery of 500 molecular candidates for membranes to better separate CO_2 from flue gas

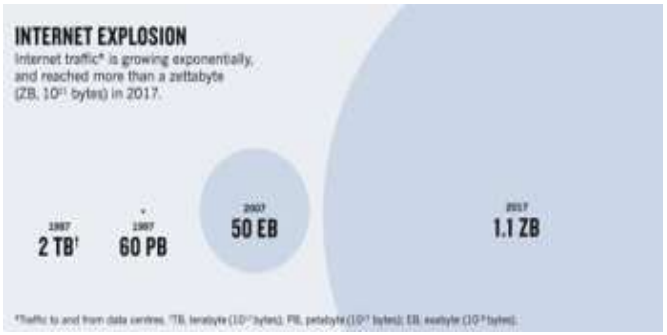
Hsu *et al. APS March Meeting* (2021)
https://research.ibm.com/blog/accelerating-materials-discovery

The demand keeps surging

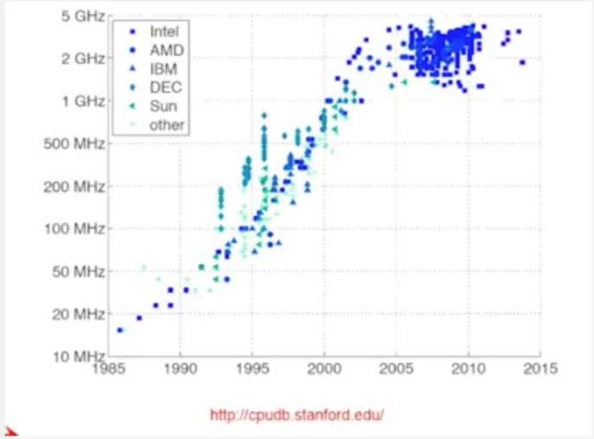
"Viewing 30 minutes of Netflix (1.6 kg of CO2) emits the same amount of CO2 as driving nearly four miles."[1]

We are at an inflection point :

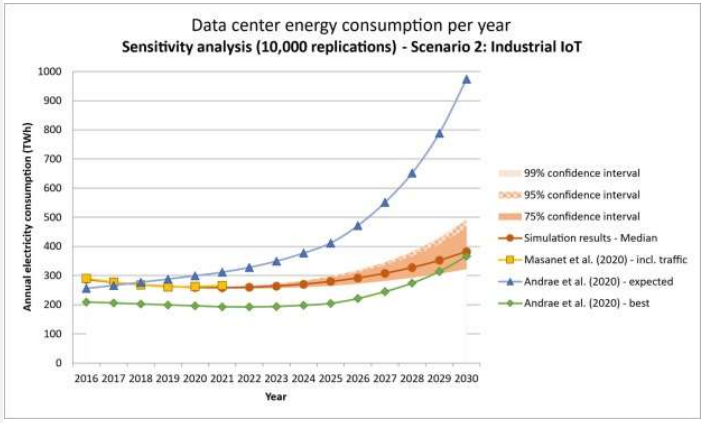
1) Demand is growing at exponential scale



2) The end of Dennard Scaling means we can't keep up



3) Electricity consumed by Data Centers will increase to 8% by 2030



How to stop data centers from gobbling up the world's electricity

https://www.nature.com/articles/d41586-018-06610-y

Koot, M. et al. Usage impact on data center electricity needs: A system dynamic forecasting model, Applied Energy Volume 291, 1 June 2021,

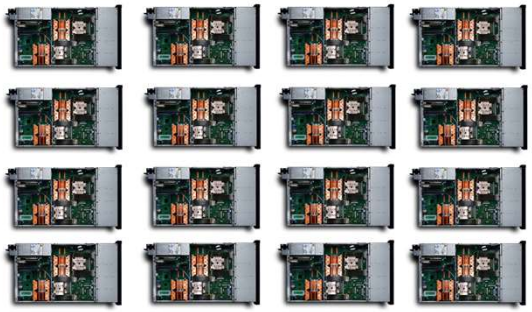
[1] https://www.akcp.com/blog/the-real-amount-of-energy-a-data-center-use/

Driven by AI Energy demand .....

Take for example: Training Image recognition model
Dataset: ImageNet-22K
Network: ResNet-101

4 GPUs
16 days
~385 kWh

256 GPUs
7 hours
~450kWh



For reference: 1 model training run is
~2 weeks of home energy consumption

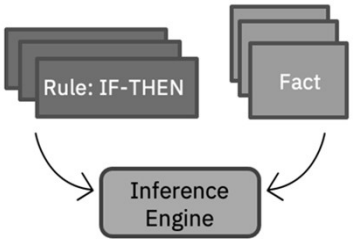
https://arxiv.org/abs/1708.02188

- Deep learning is computationally intensive
- Time consuming even with high-performance computing resources
- Power consumption prohibitive for applicability in domains such as internet of things

An emerging paradigm is changing the AI landscape

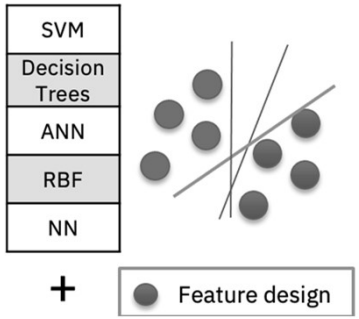
Expert Systems

Hand-crafted symbolic representations



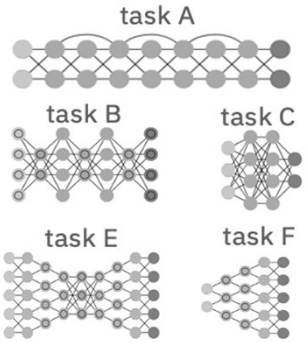
Machine Learning

Task-specific hand-crafted feature representations



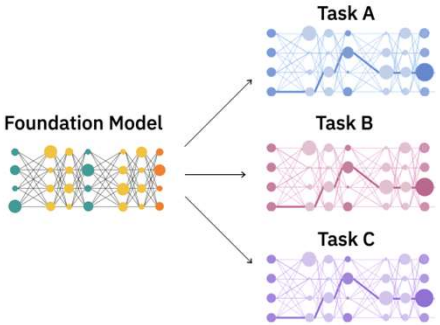
Deep Learning

Task-specific learnt feature representations



Foundation Models

Generalizable & adaptable learnt representations



1980s

limited data

1980s to ~2010

Big data

2010+

large amounts of labeled data
+
compute

2017+

self-supervision at scale
+
massive unlabeled data
+
compute

Making AI more effective

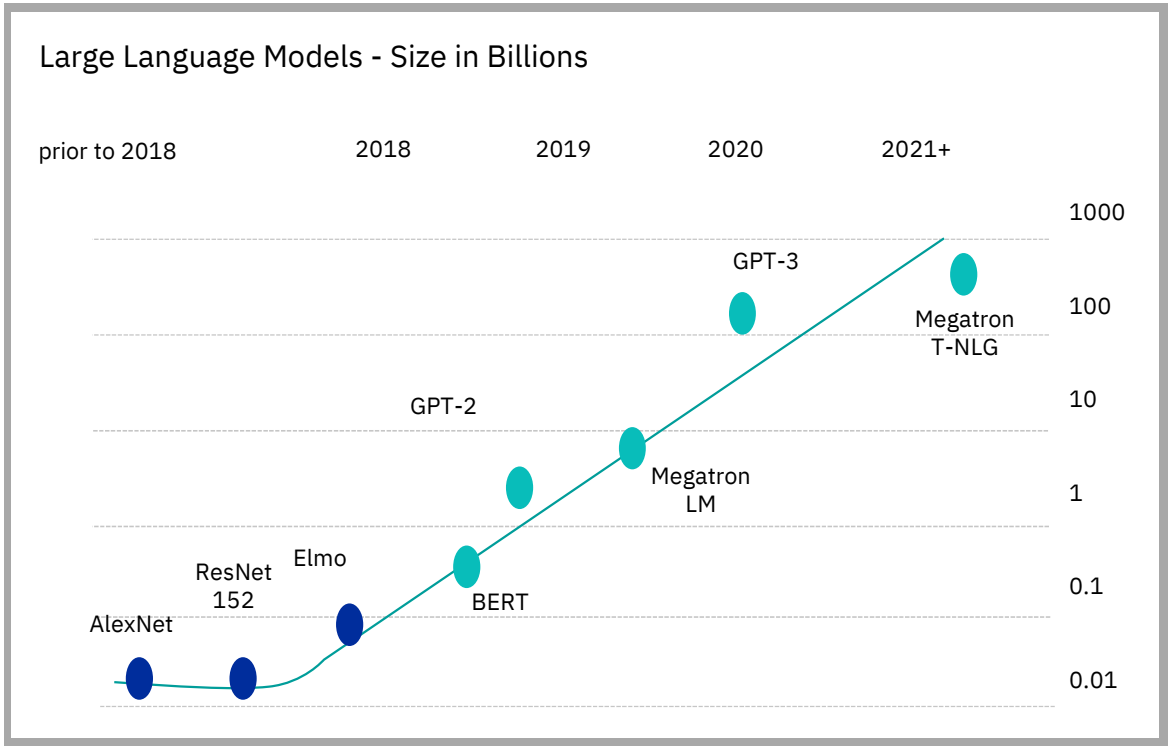
Self supervised learning	Before foundation models	After foundation models
6X faster 5X cheaper	7 years = 12 languages	1 year = 13 languages

But exponential problems require exponential computation

Largest models:
Training compute ~2x every 3 months

Example: GPT-3 language model
Size: ~175 billion parameters
One training run: ~552.1 tons of CO_2

D.. Patterson et al. (Google/UC Berkeley), arXiv 2104.10350 (2021)



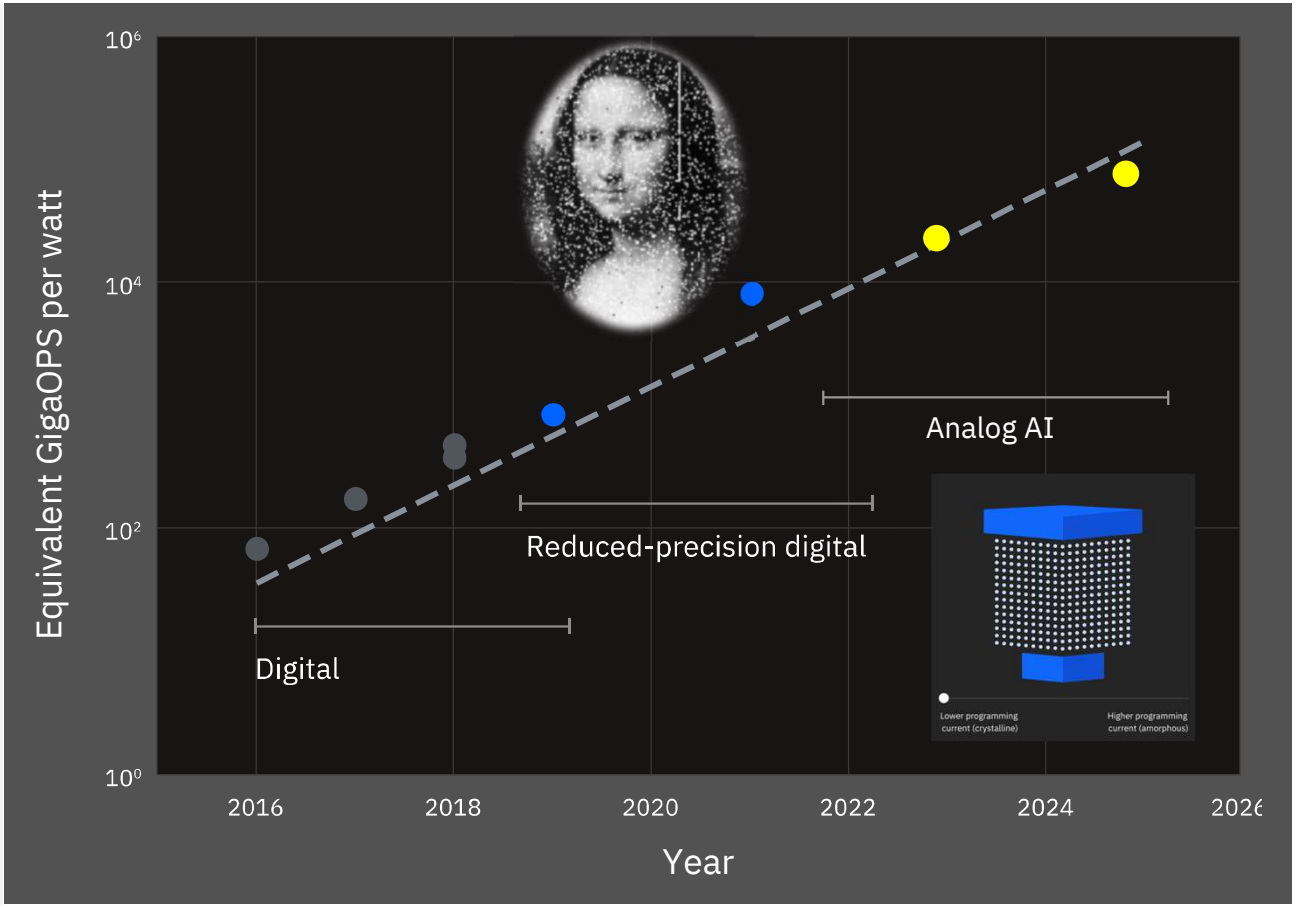
Training FLOPs :

Transformer: 750x / 2 yrs
CV/NLP/Speech: 15x / 2yrs
Moore's Law: 2x / 2yr

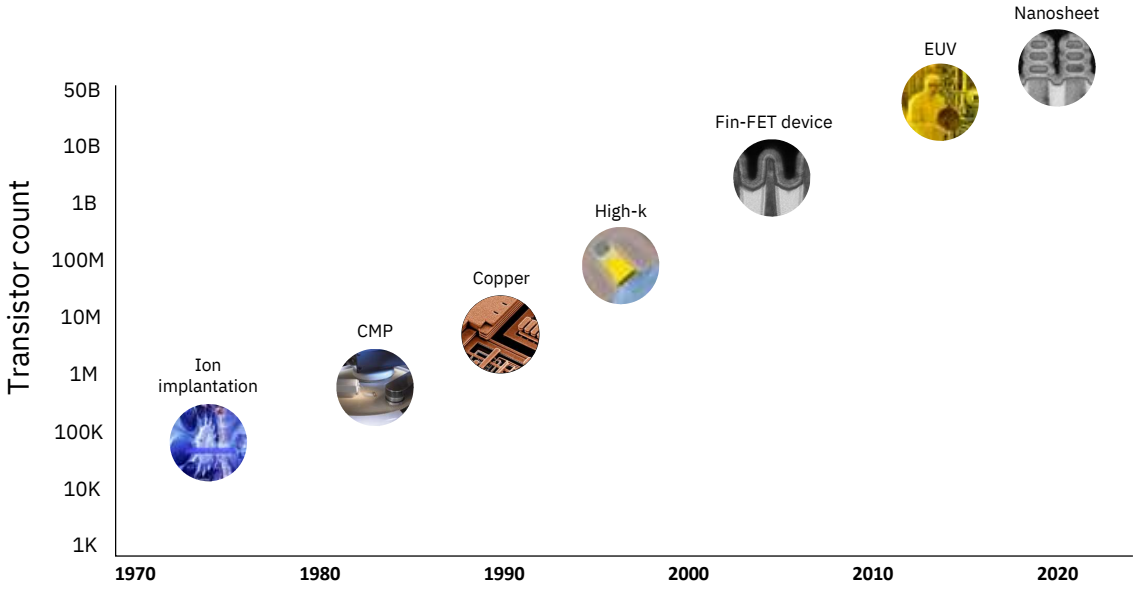
https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8/

Tackling large models through AI Hardware innovation

- Extending performance by 2.5X / year through 2025
- Approximate computing principles applied to Digital AI Cores with reduced precision, as well as
- Analog AI Cores, which could potentially offer another
- 100x in energy-efficiency



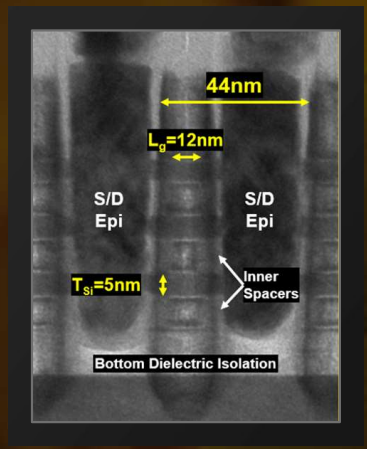
Improvements in semiconductor technologies

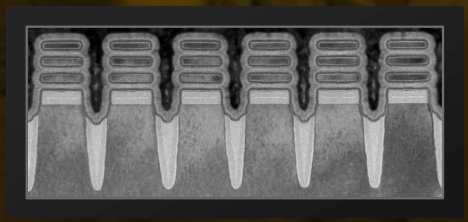


Technology	Density	Transistor/Standard Cell Performance	Cost / Chip (Fixed Size)	Cost / Transistor	Power	Power / Chip	Chip Performance (by adding cores)
7nm	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5nm	1.40	1.15	1.40	1.00	0.79	1.11	1.61
3nm	1.89	1.27	1.96	1.04	0.58	1.09	2.39
2nm	2.46	1.45	2.74	1.12	0.37	0.92	3.57

Source: Huiming Bu, Dechao Guo, IBM

IBM Research produces the world's first 2 nm technology node





45% better performance or

75% less power consumption compared to 7 nm technology.

Forbes

Big Blue Goes Tiny With World's First 2nm Chip Tech

WIRED

To Make These Chips More Powerful, IBM Is Growing Them Taller

The company reveals a process that it says can cram two-thirds more transistors on a semiconductor, heralding faster and more efficient electronic devices.

The New York Times

IBM on Thursday announced another leap in miniaturization, a sign of continued U.S. prowess in the technology race.

AI Accelerator integration for IBM Z Systems

- Focus: Enterprise on-CPU (central processing unit) inference requirements
- AI accelerator integrated into Telum processor
- 8X – 12X overall inference performance
- On-chip AI accelerator enables real-time data insights for applications such as fraud detection

IBM z16

IBM's New Telum Chip Reboots the Mainframe ›Big Blue's z16 computer—and the cache-savvy design at its core—gives new relevance to the platform

BY DEXTER JOHNSON | 1 **April 29, 2022** **IEEE Spectrum**

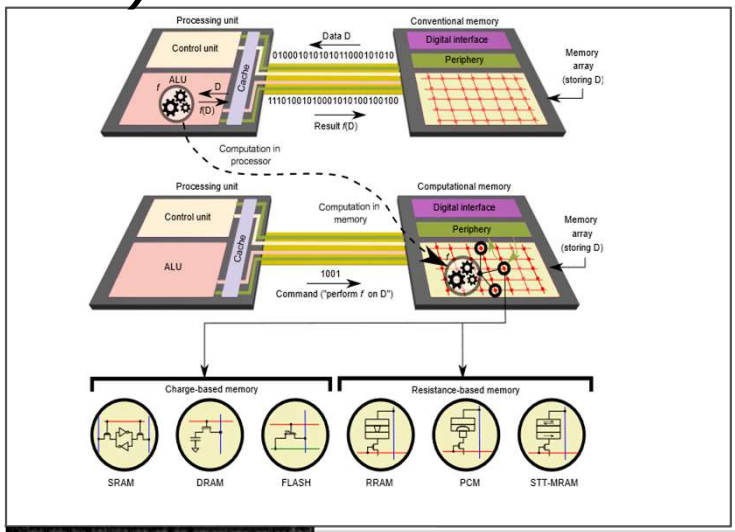
IBM Artificial Intelligence Unit (AIU)

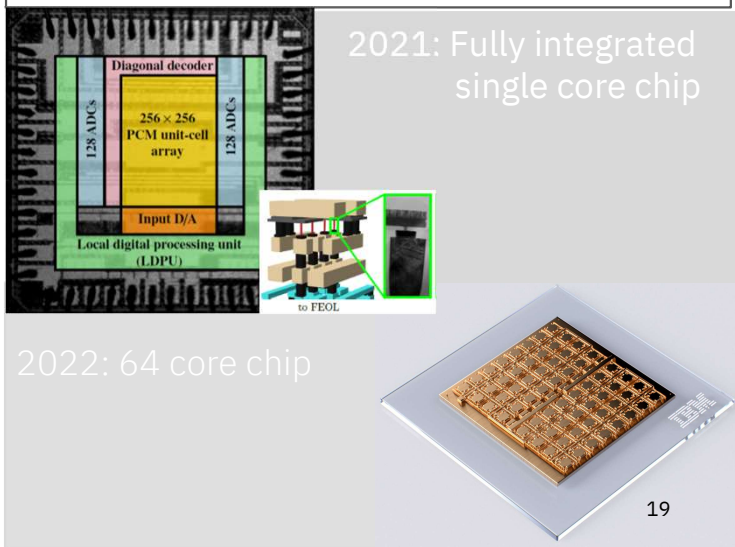
SoC implements IBM's leadership innovations in low-precision AI arithmetic and algorithms

- Chip architecture optimized for enterprise AI workloads
- Enabled for Foundation Models
- Enabled in the Red Hat software stack
- Integration into the IBM Watson software stack underway
- Supports multi-precision inference (& training)
  - FP16, FP8, INT8, INT4, INT2
- Implemented in leading edge 5nm technology

Analog In-Memory Computing (AIMC)

- Break the current “memory wall”: from >>300fJ/Operation to 1fJ/Op
- 2021 FOAK: Fully integrated AIMC core based on PCM, with
  - 10.5 TOPS/W and 1.6 TOPS/sq.mm
- 2022 FOAK: 64 cores achieving:
  - 63.1 TOPS at an energy efficiency of 9.76 TOPS/W for 8-bit input/output matrix-vector multiplications
  - MVM throughput per chip area > 15x higher than comparable multi-core resistive memory AIMC chips
  - highest accuracy on the CIFAR-10 benchmark





Khaddam-Al et al., Proc. VLSI (2021) (highlight paper)

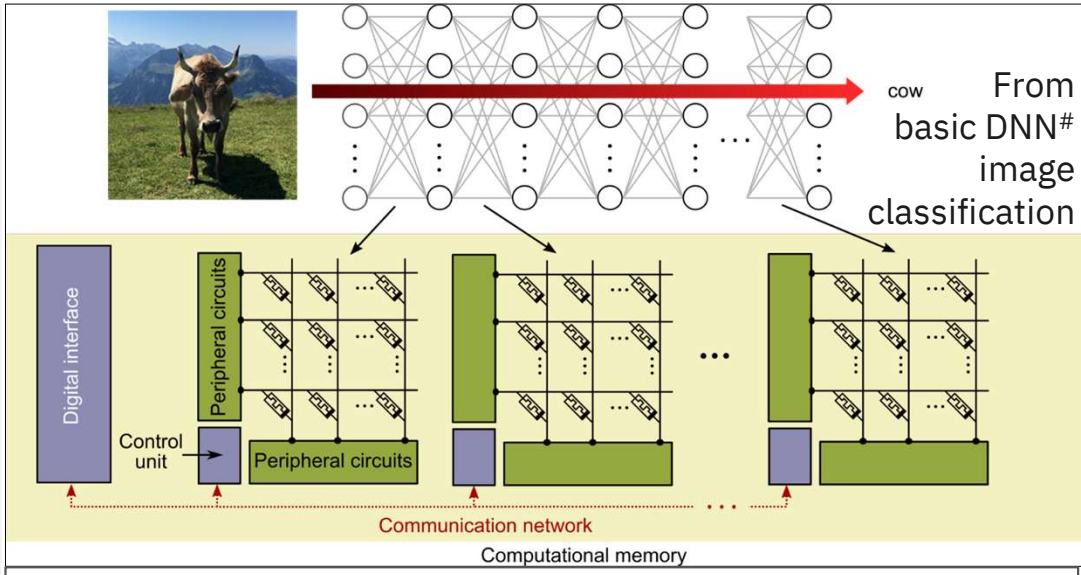
Khaddam-Al et al., JSSC (2022)

AIMC to power emerging applications

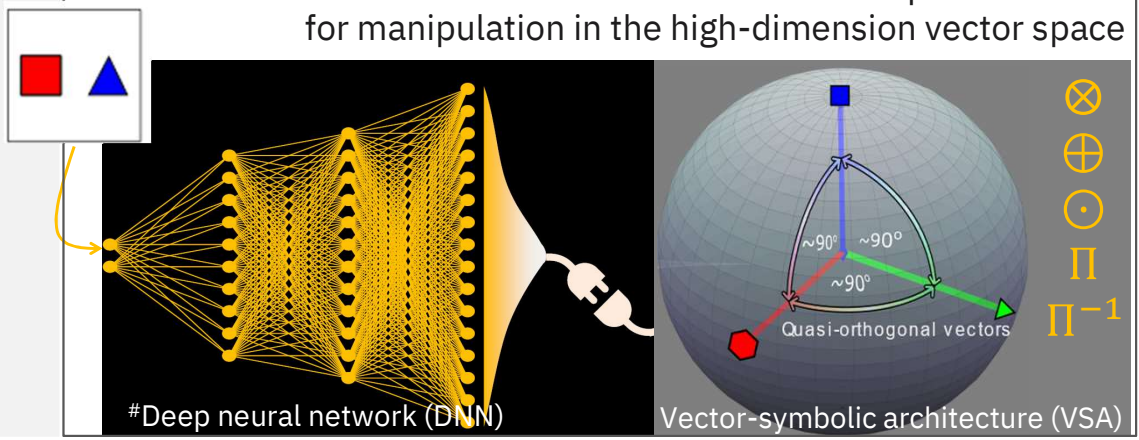
- AIMC can be used to efficiently realize neural networks, and beyond

- To address the main shortcomings of today's AI such as:
  - few-shot learning,
  - continual learning,
  - visual abstract reasoning
  - …

Karunaratne et al., Nature Communications, 2021
Hersche et al., CVPR, 2022
Karunaratne et al., ESSDERC, 2022



…to vector representations for manipulation in the high-dimension vector space



Tape Sustainability-Saving the planet 27 PB at a time

10 Years of Archival data
11-nines data durability

Bryce Canyon
OCP Commodity HDD Racks

- Energy 74%
- Infrastructure 24%
- End of Life 2%

10-Year
1954.3
mtons CO2e

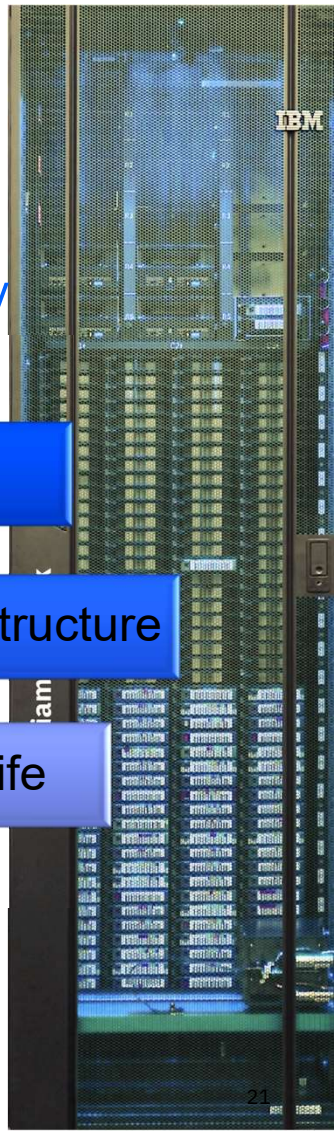
- 5-year life
- 18TB Archive HDDs
- 3 JBOD per Controller
- 1.26 Erasure Coding
- Full replacement cycle yea
  - Assumes 36TB drives

IBM Diamondback
Newest single rack tape library

- 82% Energy
- 16% Infrastructure
- 2% End of Life

10-Year
78.1
mtons CO2e

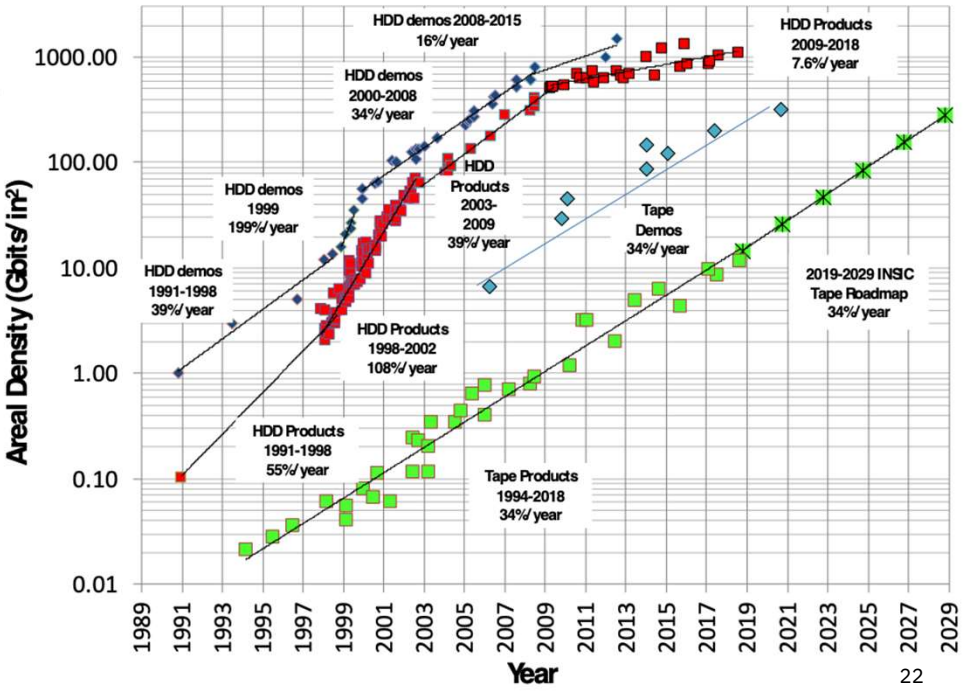
- 1 Frame
- 14 – LTO-9 Tape drives
- 1500 – LTO-9 Cartridges
  - No refresh required



Tape Capacity Scaling: Sustainable roadmap for the next decade and beyond

Product Year	IBM 726 1952	LTO9 2021	**TS1160 2018**	Demo 2017 Sputtered Tape	Demo 2020 SrFe Tape
Capacity	2.3 MB	18 TB	**20 TB**	330 TBytes	580 TBytes

Areal Density >8.5M X



Requires sustained breakthroughs in:

- Physics of magnetism
- Materials Science
- Signal Processing
- Error correction coding
- Mechatronics and Control

Moving the needle forward : Quantum Computing

WHAT IS SPECIAL ABOUT QUANTUM COMPUTING?

Quantum computing is a fundamentally different model of computation that exploits the laws of quantum mechanics to process information and provide advantages that classical computing cannot.

Applications

Simulating Nature

Physics
Chemistry
Materials Science

Data with Structure

Machine Learning
Ranking in groups
Factoring

Non-exponential Speedup

Sampling problems
Optimization
Risk analysis and option pricing

For example, simulating carbon capture with 52 – 65 molecular orbitals would take ...

Classical computer

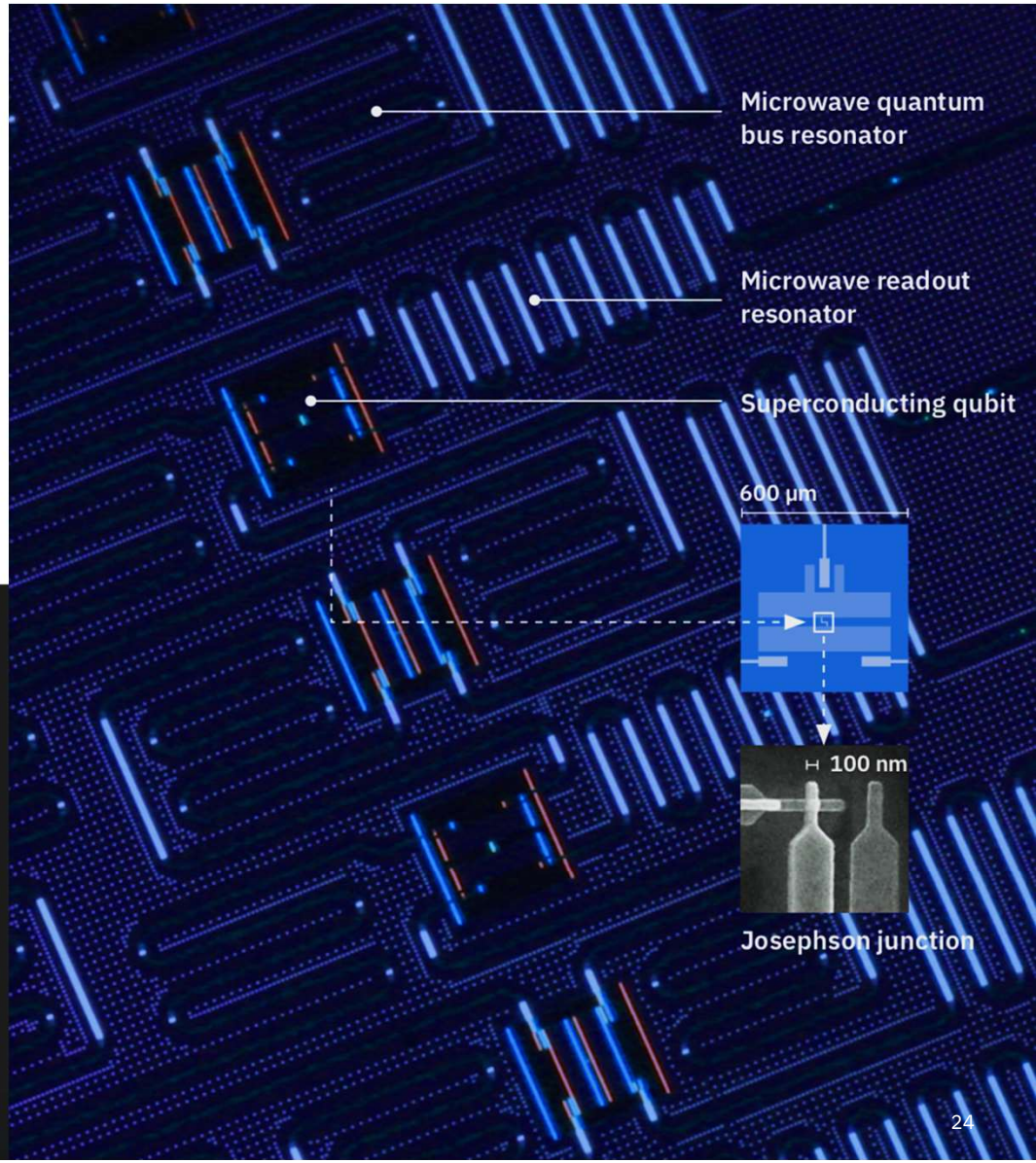
Out of reach

Quantum computer

~3.5 days

Inside an IBM Quantum superconducting quantum processor

10^{-16} W per two-qubit gate at 5 GHz for 100 ns/gate



Driving the roadmap for quantum computing

2021:

127-qubit Eagle processor

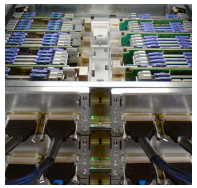
Qiskit Runtime

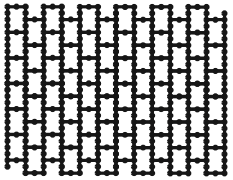
Qiskit

2022:

433-qubit Osprey processor

Dynamic Circuits

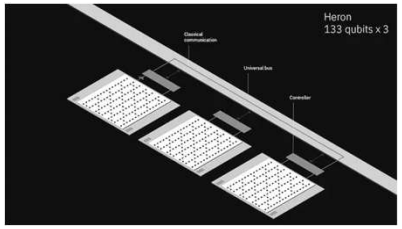




2023:

1,121-qubit Condor processor

Circuit knitting with 133-qubit Heron processor



IBM Quantum Network

A collaborative community of discovery

20+
Operational systems

410k+
Users

200+
IBM Quantum Network members

3.5B
Daily executions

280+
Universities

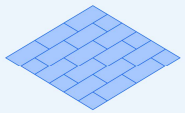
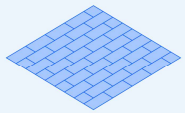
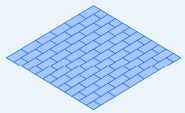
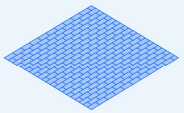
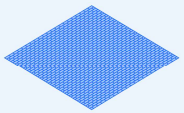
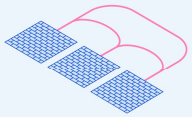
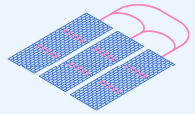
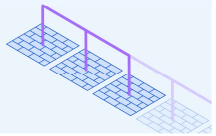
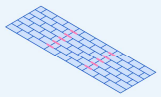
1400+
Papers written

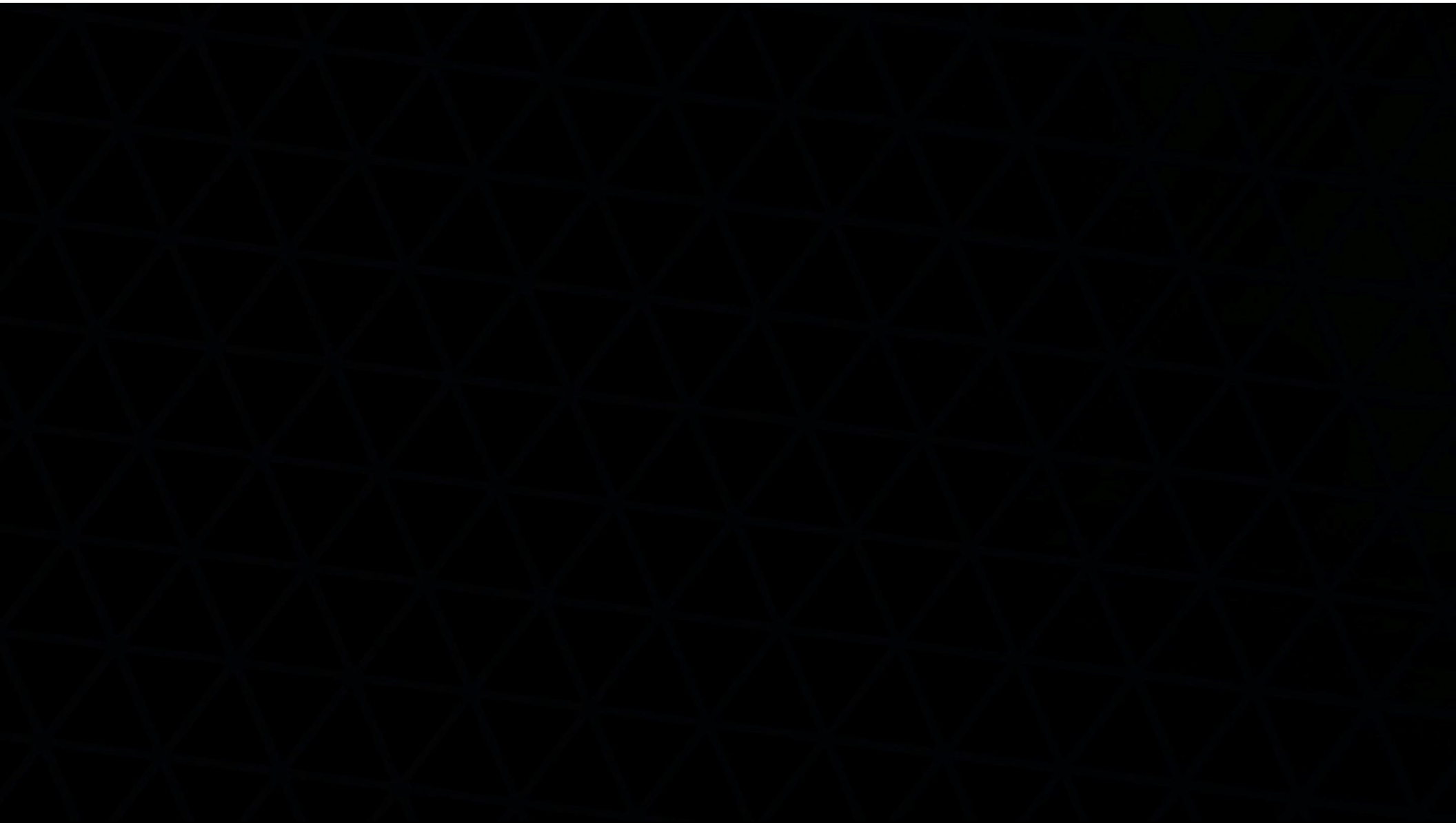
Development Roadmap

Executed by IBM On target

IBM **Quantum**

	2019	2020	2021	2022	2023	2024	2025	Beyond 2026
	Run quantum circuits on the IBM cloud	Demonstrate and prototype quantum algorithms and applications	Run quantum programs 100x faster with Qiskit Runtime	Bring dynamic circuits to Qiskit Runtime to unlock more computations	Enhancing applications with elastic computing and parallelization of Qiskit Runtime	Improve accuracy of Qiskit Runtime with scalable error mitigation	Scale quantum applications with circuit knitting toolbox controlling Qiskit Runtime	Increase accuracy and speed of quantum workflows with integration of error correction into Qiskit Runtime
Model Developers					**Prototype quantum software applications** → **Quantum software applications**		Machine learning | Natural science | Optimization	
Algorithm Developers		**Quantum algorithm and application modules**			**Quantum Serverless**			
		Machine learning | Natural science | Optimization				Intelligent orchestration	Circuit Knitting Toolbox	Circuit libraries
Kernel Developers	**Circuits**		**Qiskit Runtime**					
				Dynamic circuits	Threaded primitives	Error suppression and mitigation		Error correction
System Modularity	**Falcon** 27 qubits	**Hummingbird** 65 qubits	**Eagle** 127 qubits	**Osprey** 433 qubits	**Condor** 1,121 qubits	**Flamingo** 1,386+ qubits	**Kookaburra** 4,158+ qubits	Scaling to 10K-100K qubits with classical and quantum communication
					Heron 133 qubits x p	**Crossbill** 408 qubits		



Sustainability is a Business Imperative

- Investor pressure
- Consumer pressure
- Policy landscape

Blackrock Doubles Down On Climate Pressure In The Midst Of Global Crisis

Climate Changed
Large Exxon Shareholder Starts Divesting Over Climate Change
Bloomberg

Exxon Directors Face Shareholder Revolt Over Climate Change
Bloomberg

Tesla's Sustainability Cred Is Being Challenged With Shareholder Proposals at Annual Meeting
BARRON'S

Shareholder climate rebellions surge despite coronavirus crisis
Investors pile pressure on companies including JPMorgan and Rio Tinto over global warming
FINANCIAL TIMES

40%
Purpose-driven consumers who seek products and services aligned with their values.

57%
Consumers willing to change purchasing habits
to help reduce negative environmental impact.

75%
Consumers across generations state sustainability as a very important attribute (Gen Z, Millennials, Gen X, and Boomers)

A European Green Deal
Striving to be the first climate-neutral continent

Ratified by EU parliament, Jan. 2020
Investment: €260B (2030), €1T (2050)

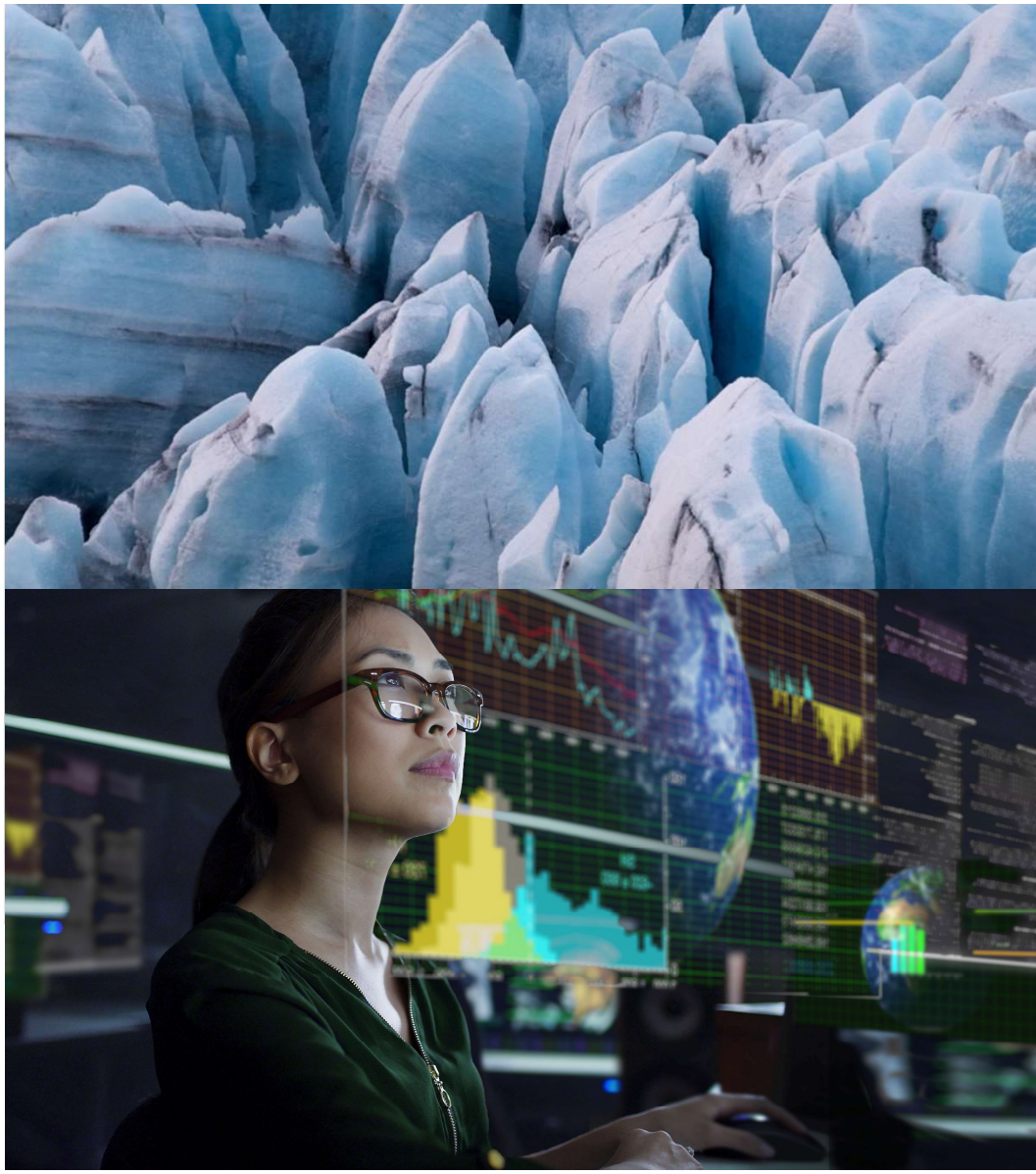
China's new climate pledge could cut emissions everywhere else too
Xi Jinping has announced the country's goal of going carbon neutral by 2060, but China's manufacturing heft will mean other nations will reap benefits too
WIRED

NEWSROOM
A "New Day for Climate Action in the United States" as U.S. Congress Passes Historic Clean Energy and Climate Investments
Bill will provide the most ambitious funding ever for tackling climate change.
August 11, 2022 | Arlington, VA

Biden signed the Inflation Reduction Act

The world is changing rapidly …

- Social & economic dependence on sustainable solutions is increasing
- Power requirements growing unsustainably
- AI & AI HW and Quantum reshape computing & business




IBM