IBM Strategy

What’s next in computing

Vincent Perrin
IBM France Ecosystem CTO

Data, IT architectures & AI use multiply

Digital transformation leads to vast data and heterogenous IT, which is best navigated through hybrid cloud and AI

IT architectures become more heterogenous

– 90% of large companies use multi-cloud architectures[1]

– 72% of companies run on both private infra and public cloud[2]

– 75% of enterprise data will be created on the edge by '25[3]

+

Data volume, variety, and velocity soar in magnitude

– 181 zettabytes of data will be generated annually by 2025[4]

– 95% of businesses must manage unstructured data[5]

– Data volumes are exploding by 63% per month in organizations[6]

+

AI transforms business models and operations

– 60-70% of employee time today could be automated using gen AI[7]

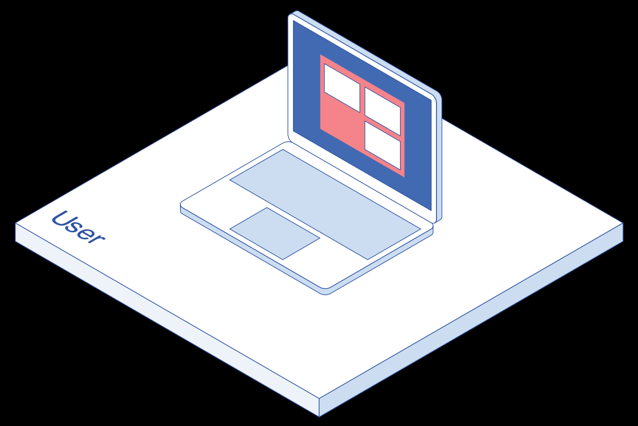
– 62% of execs say gen AI will disrupt how their org designs experiences[8]

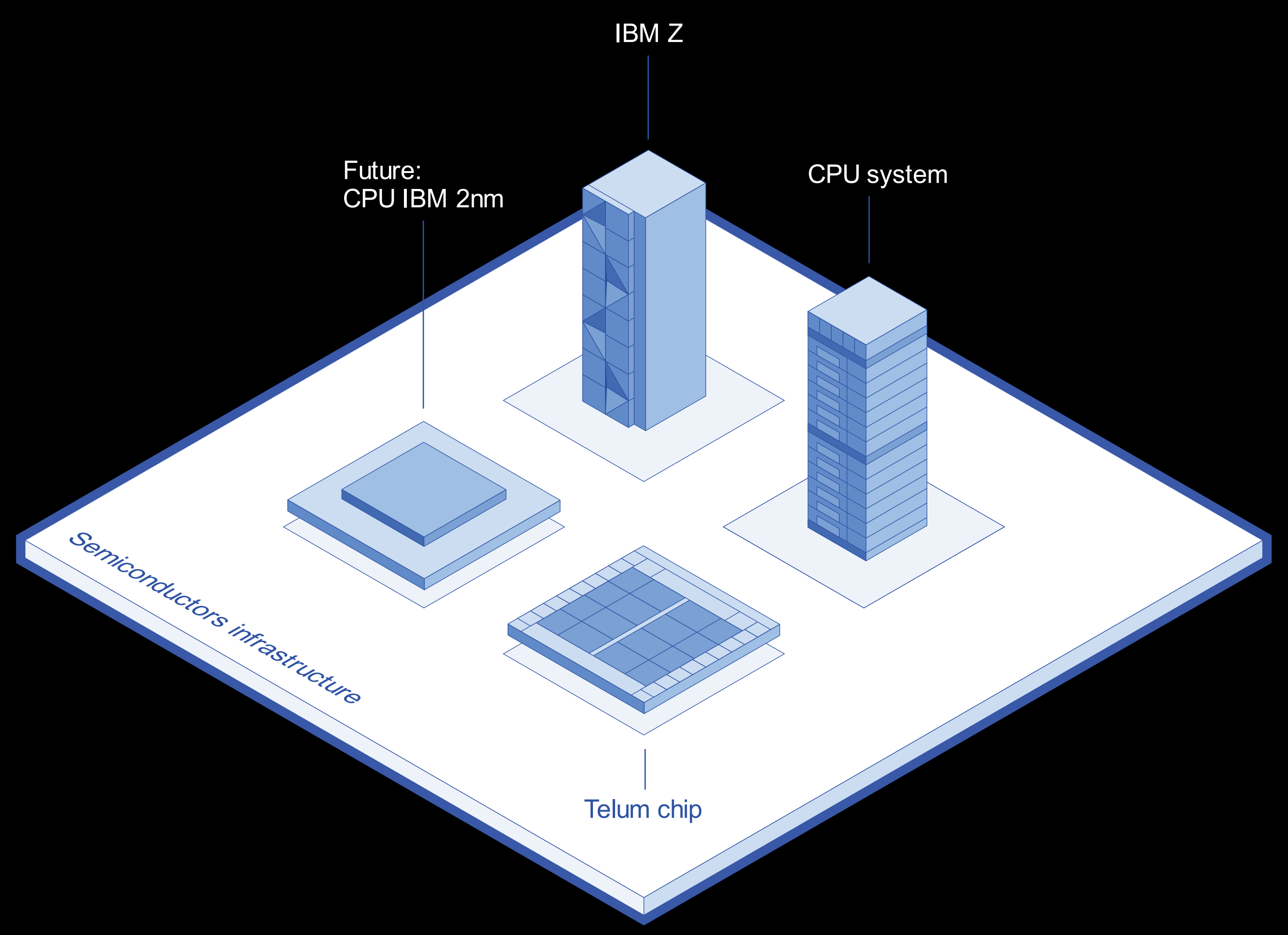
– 72% of enterprises are seeing value from their AI initiatives within 3 months[9]

▼

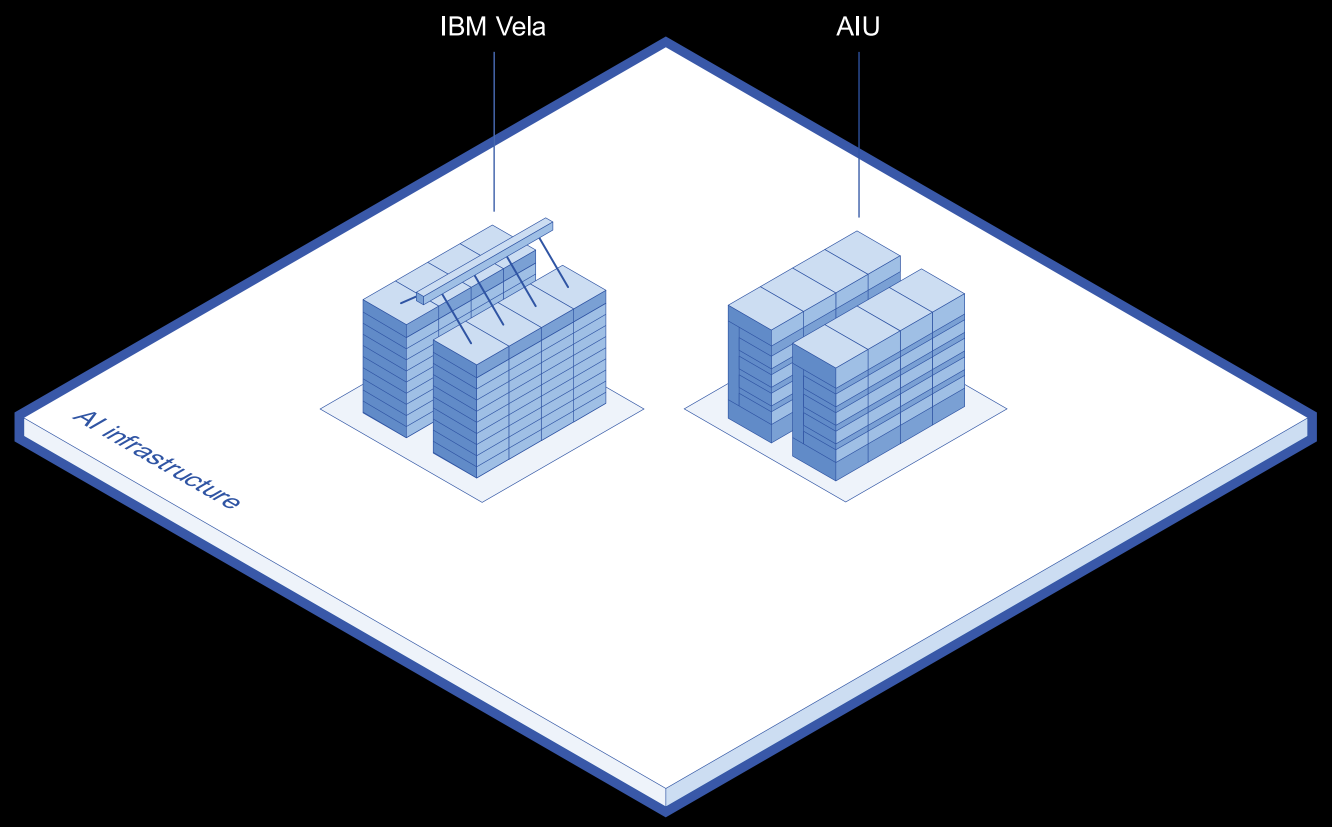
Enterprises need modern IT capabilities to manage this complexity and unlock innovation — and are turning to **hybrid cloud and AI as the universal standard**

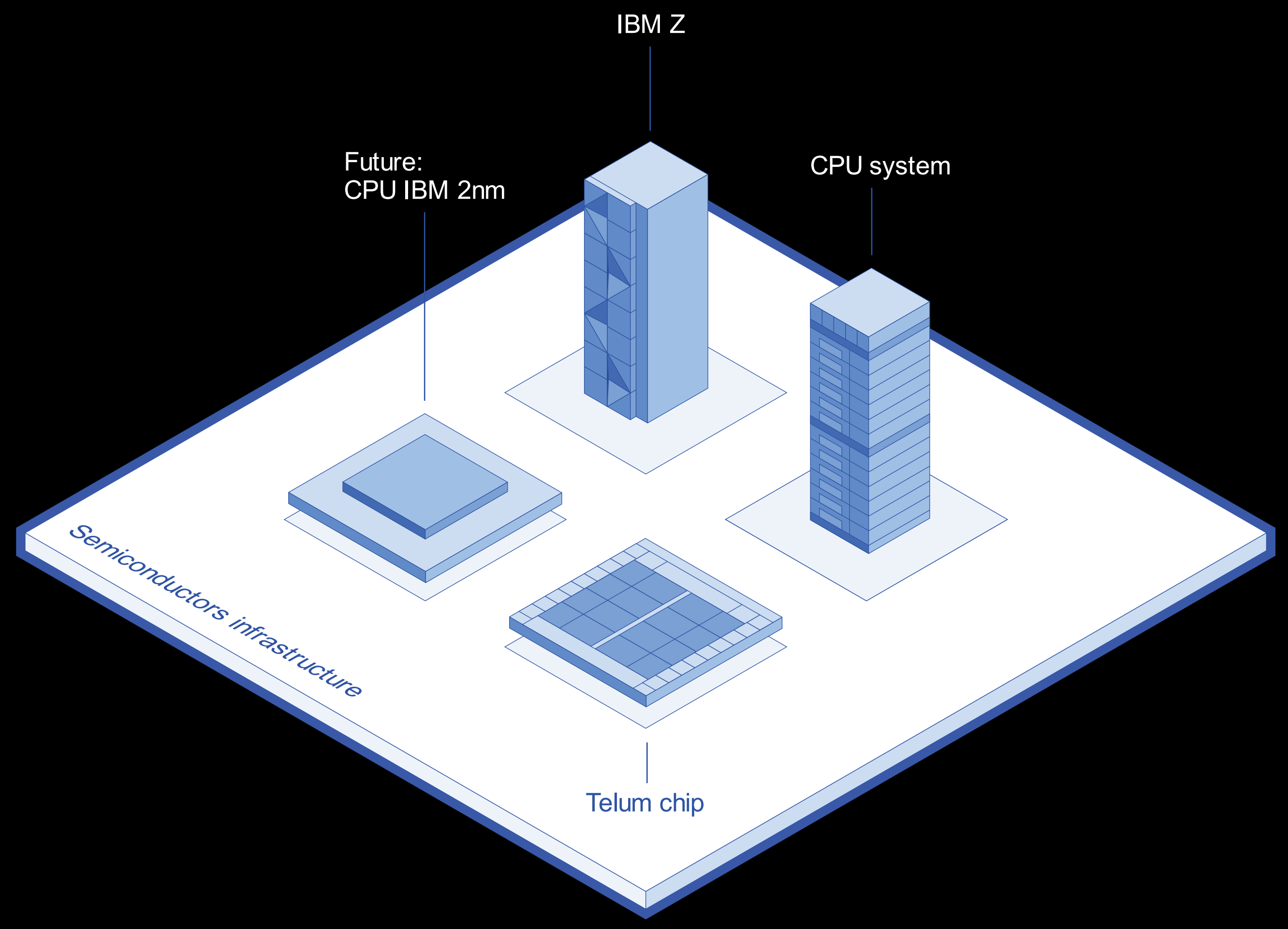
Source: 1) Forbes; 2) IBM IBV and Harris Poll; 3) Gartner; 4) Statista; 5) Forbes; 6) Matillion; 7) McKinsey; 8) IBM IBV; 9) TechTarget


User

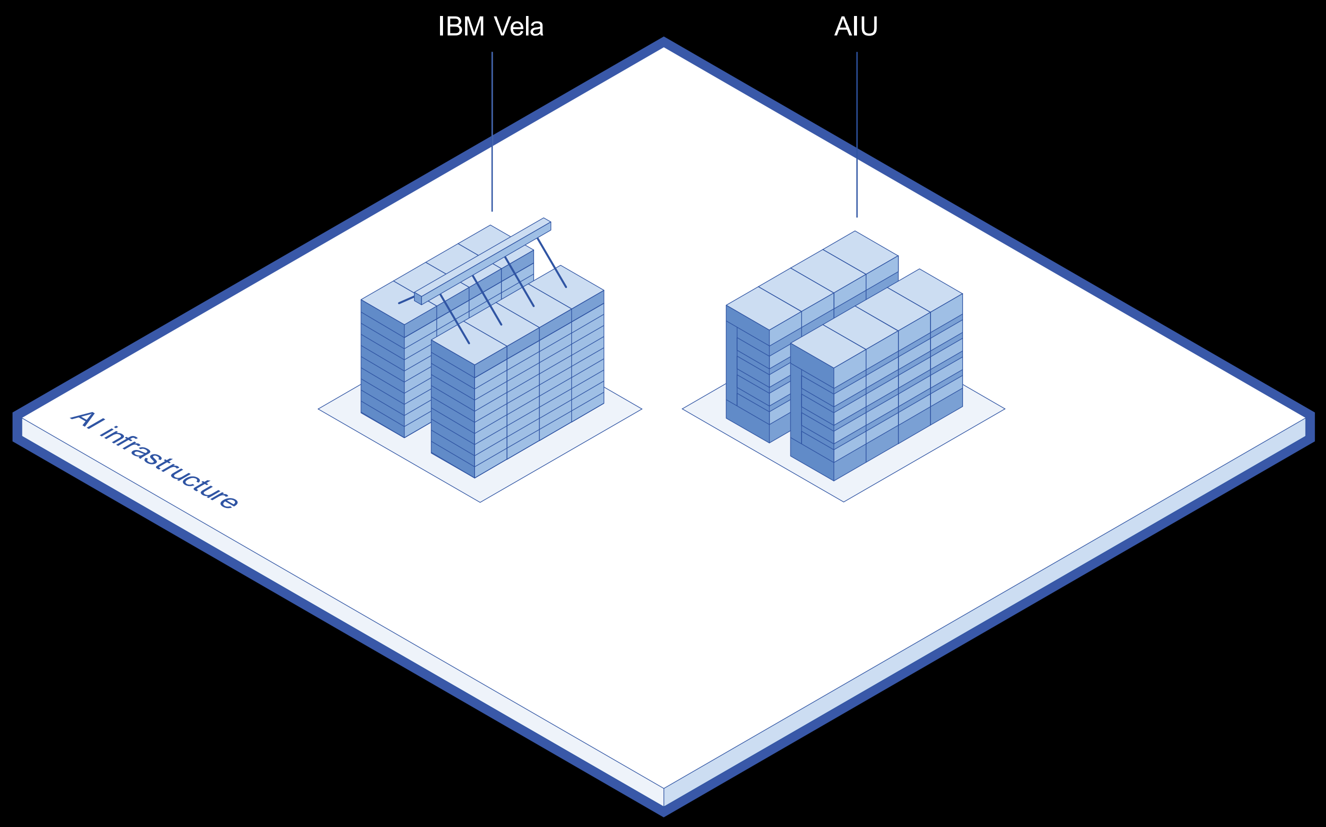

IBM Z
Future:
CPU IBM 2nm
CPU system
Semiconductors infrastructure
Telum chip

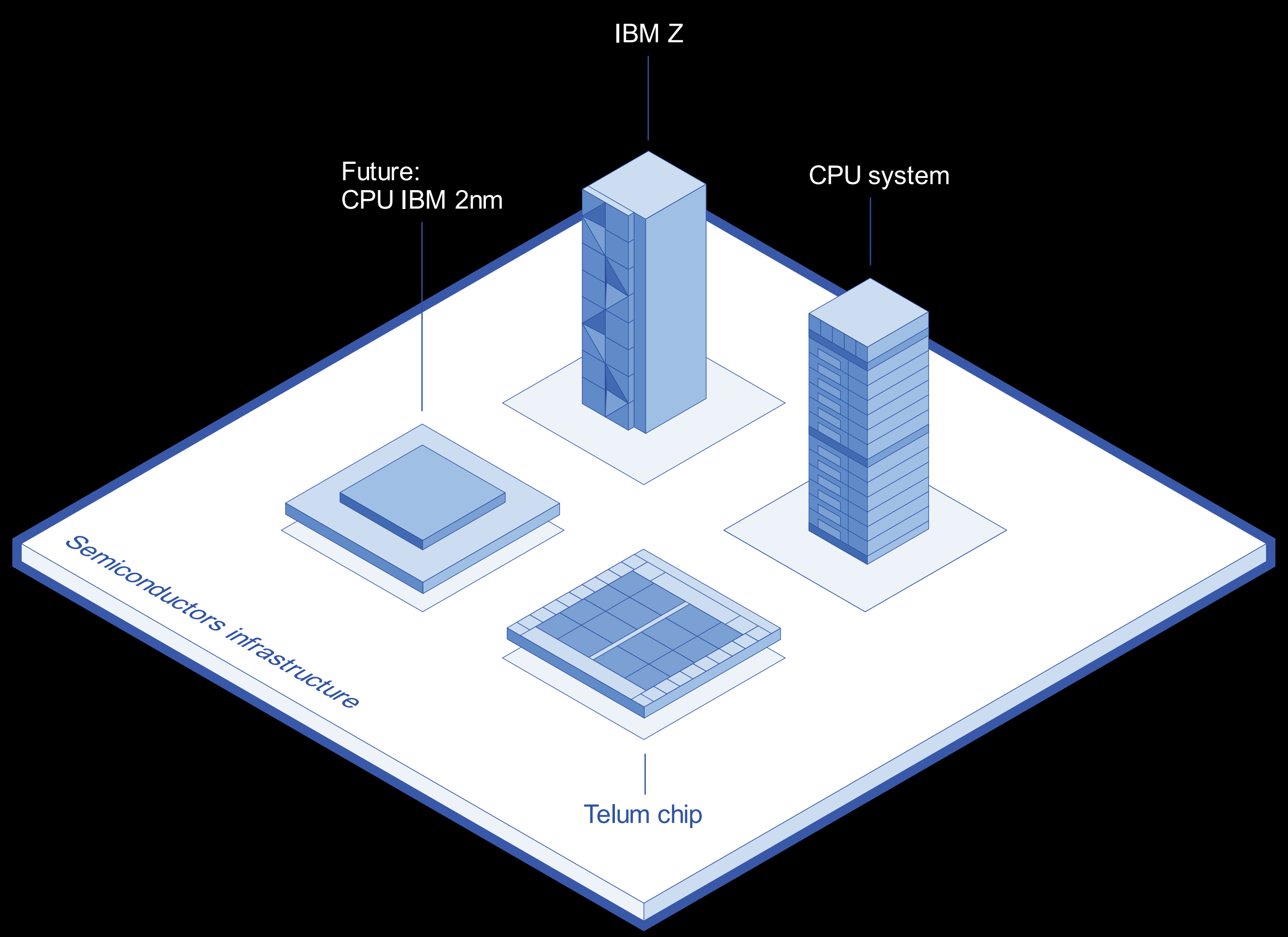
User


IBM Vela
AIU
AI infrastructure

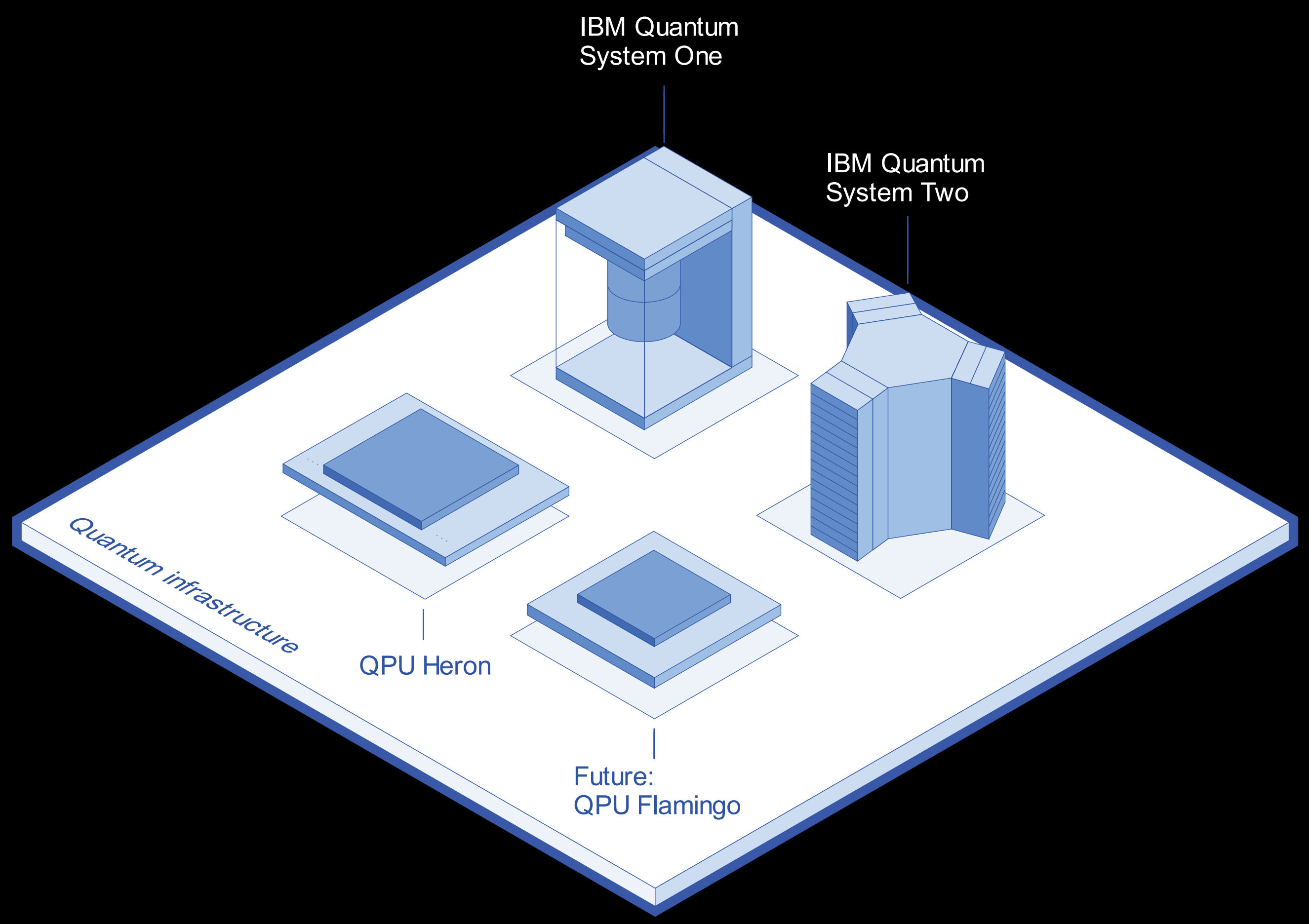

IBM Z
Future:
CPU IBM 2nm
CPU system
Semiconductors infrastructure
Telum chip

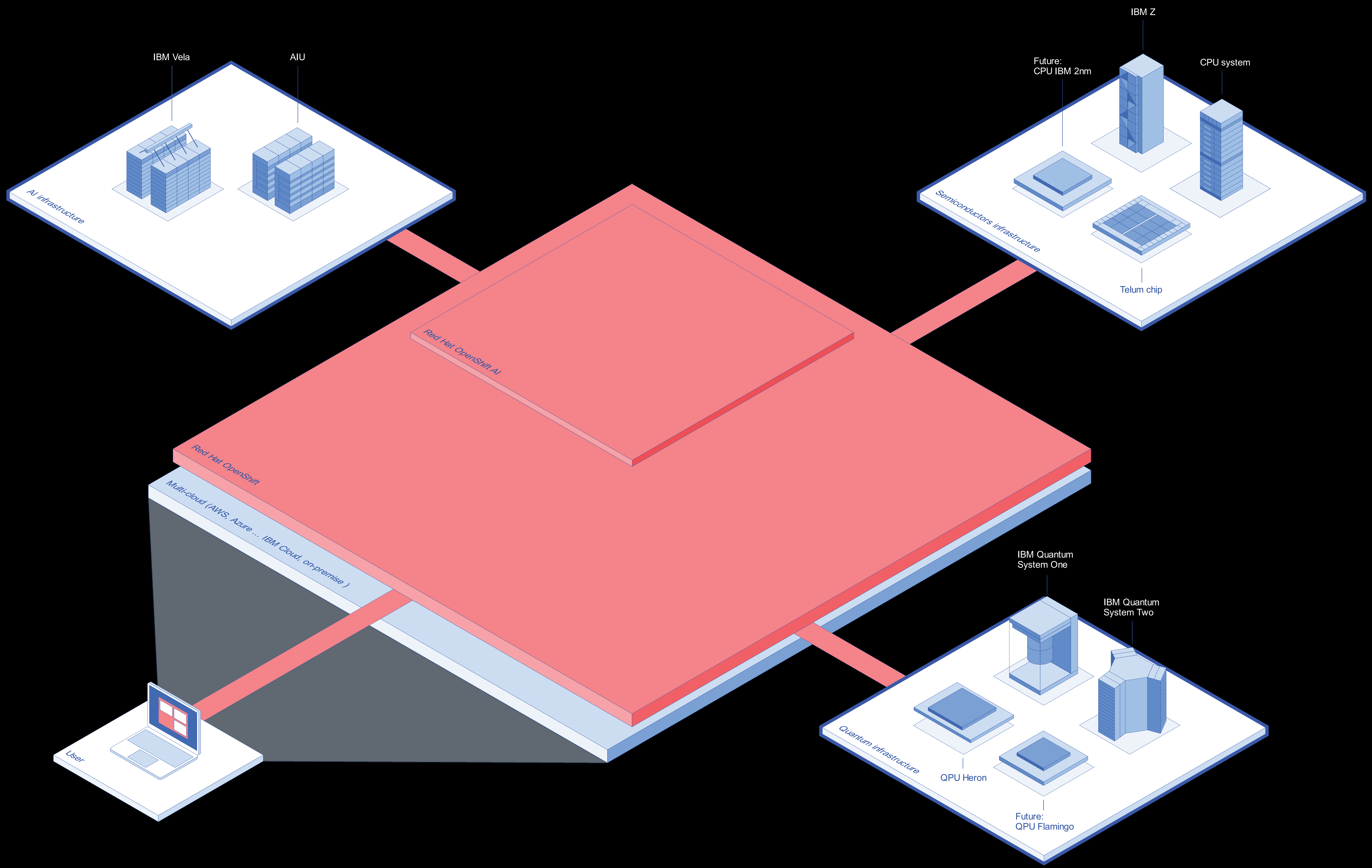
User


IBM Vela
AIU
AI infrastructure


IBM Z
Future:
CPU IBM 2nm
CPU system
Semiconductors infrastructure
Telum chip

User


IBM Quantum
System One
IBM Quantum
System Two
Quantum infrastructure
QPU Heron
Future:
QPU Flamingo


IBM Vela
AIU
AI infrastructure
IBM Z
Future:
CPU IBM 2nm
CPU system
Semiconductors infrastructure
Telum chip
Red Hat OpenShift AI
Red Hat OpenShift
Multi-cloud (AWS, Azure ... IBM Cloud, on-premise)
IBM Quantum
System One
IBM Quantum
System Two
QPU Heron
Future:
QPU Flamingo
Quantum infrastructure
User

Red Hat OpenShift

Leading hybrid cloud application platform built on open-source innovation that enables organizations to build, deploy, and run applications at massive scale, wherever they run

Open-source application platform leveraging leading projects including Kubernetes, Prometheus, Jenkins, and more

Mature, enterprise grade platform with run rate of $1.1B+; >4K global clients across industries

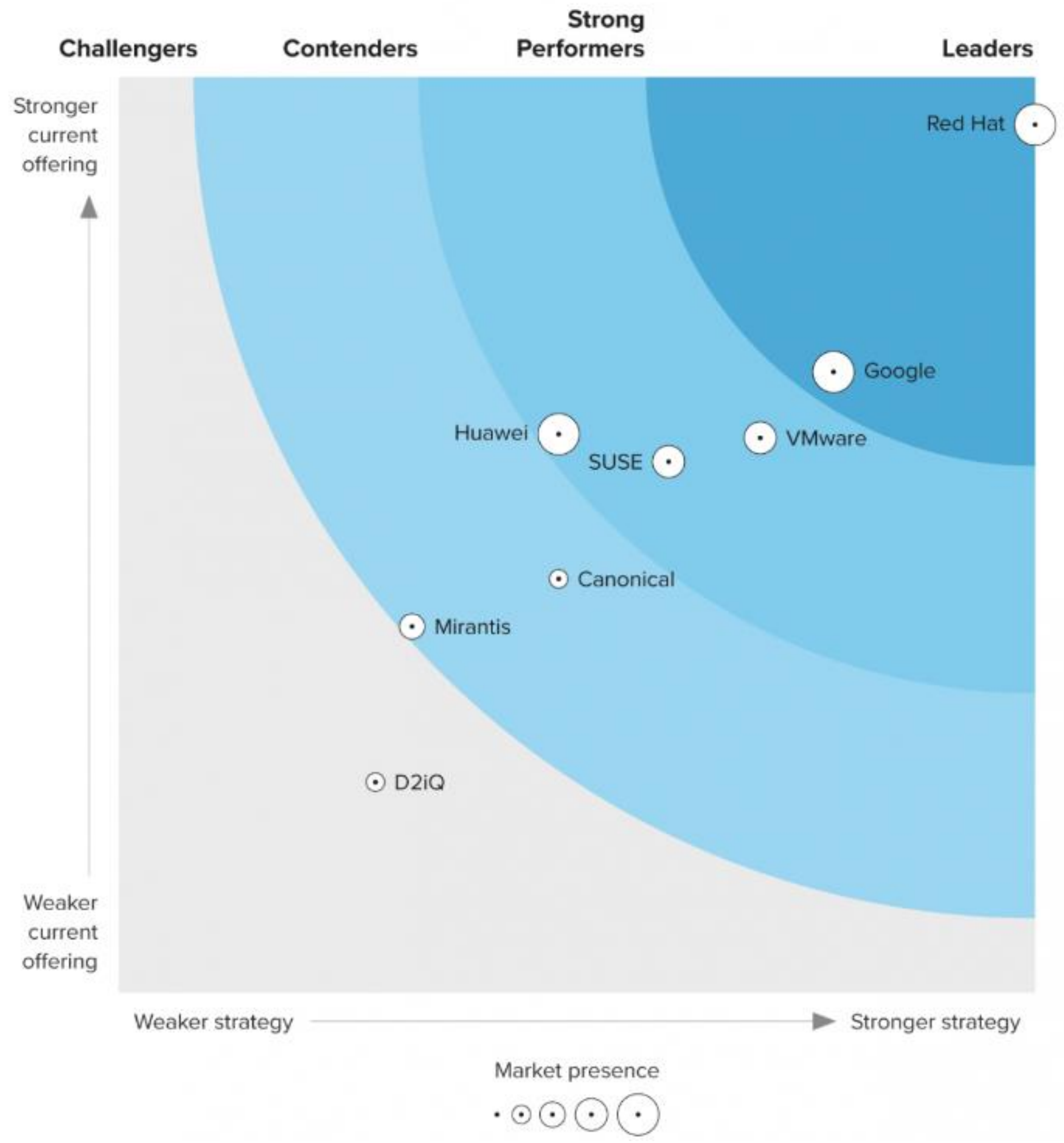
Provides a consistent platform for AI/ML workloads

Provides a proven, security-hardened (FIPS compliant) platform built on a core of Kubernetes and RHEL CoreOS

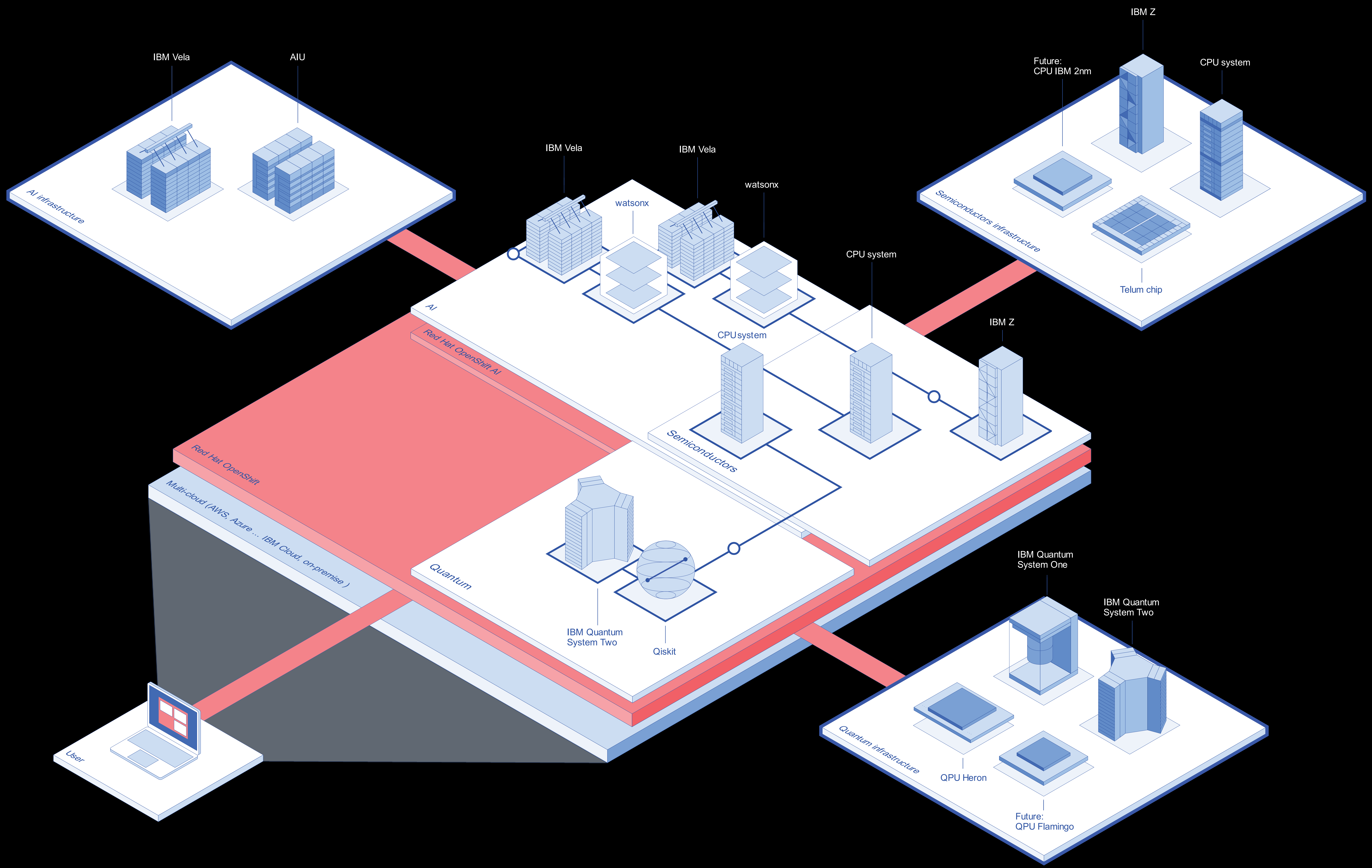
OpenShift Cloud Services help accelerate the move to public cloud with a fully managed service

The **Forrester Wave**™ recognizes OpenShift as the leading hybrid application platform

Multicloud container platforms[1]

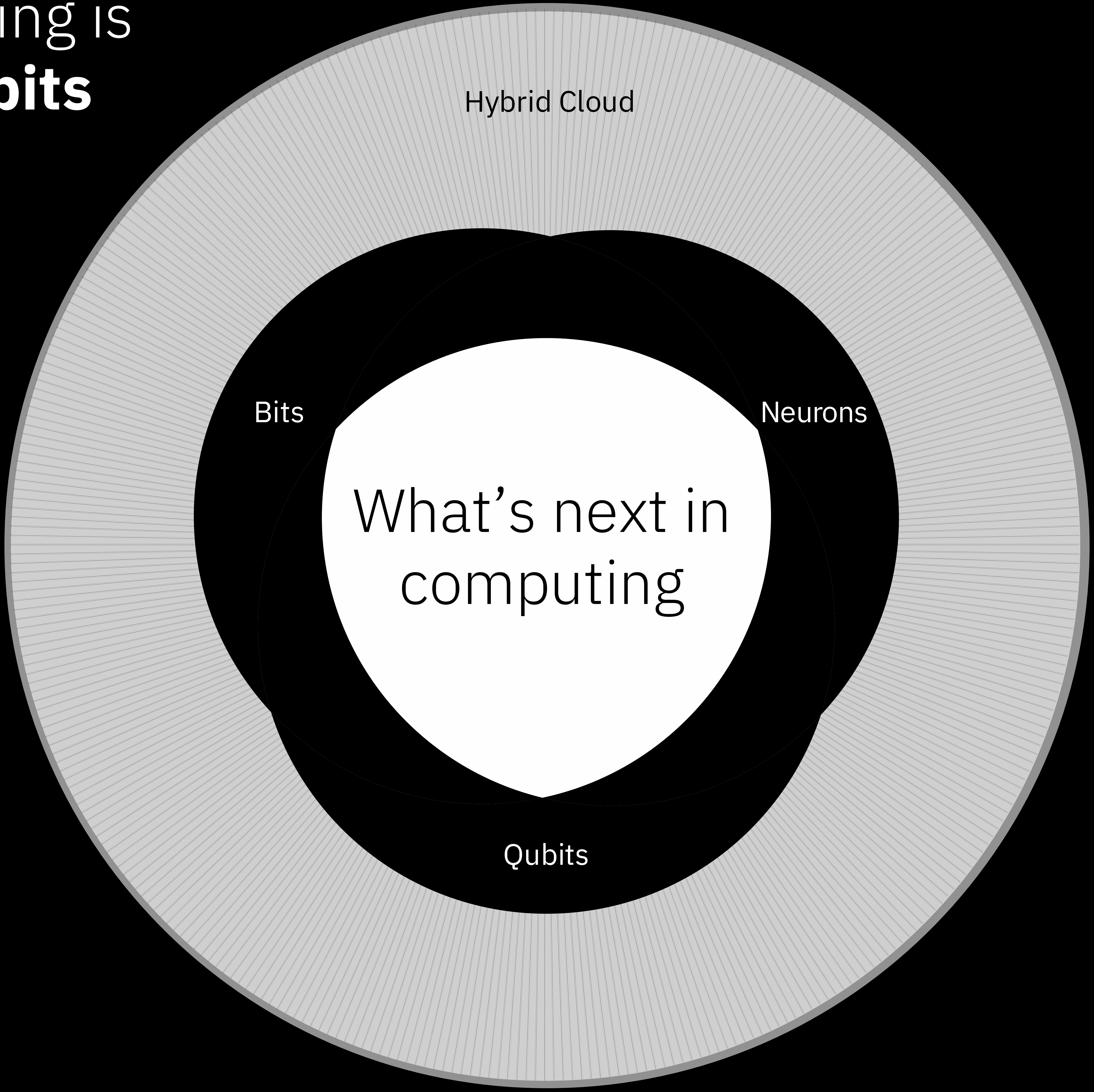


Source: 1) Red Hat-IBM recognized as a leader in Forrester Q4 2023 report; received highest score in market presence & strategy categories

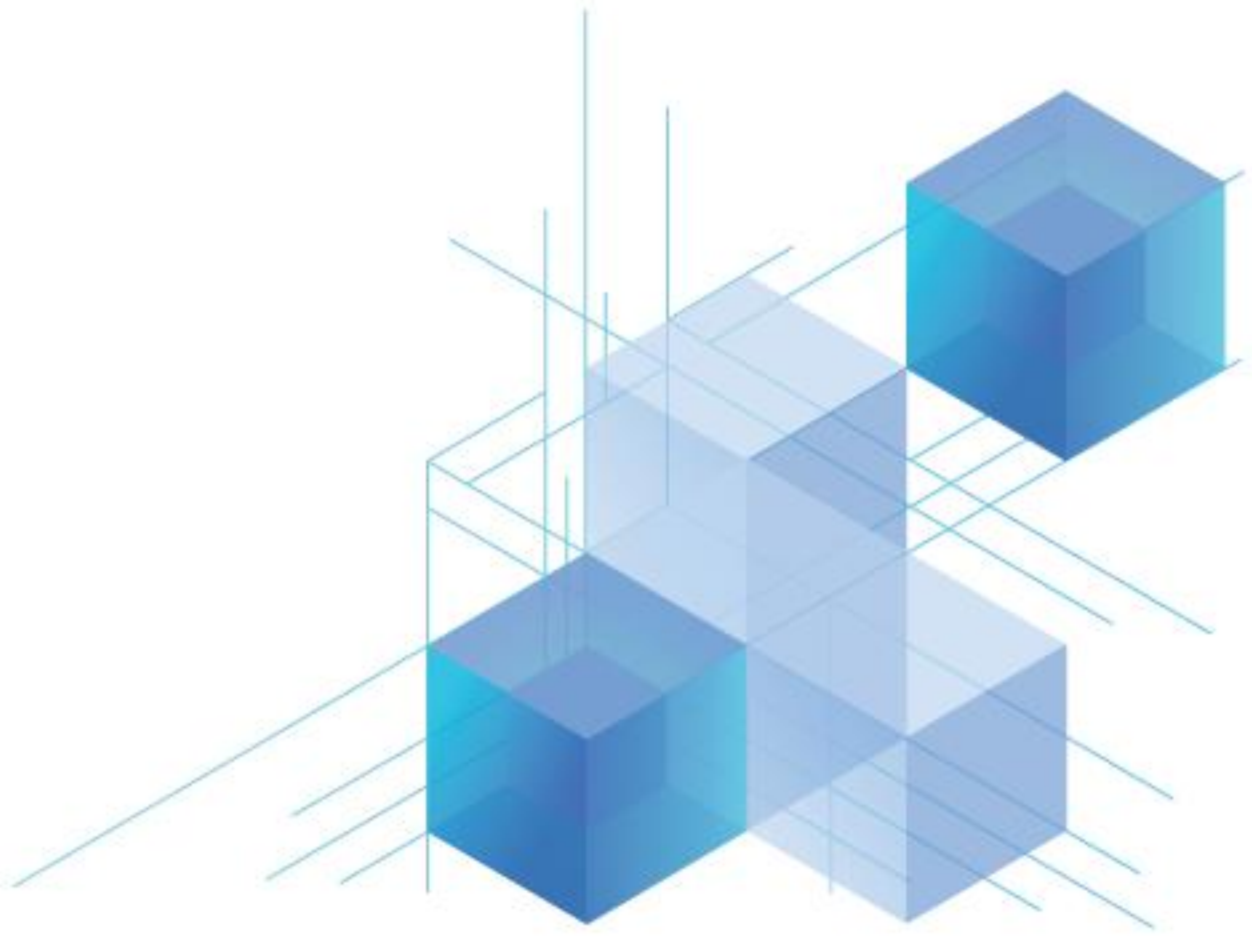

IBM Vela
AIU
AI infrastructure
IBM Vela
IBM Vela
watsonx
watsonx
AI
Red Hat OpenShift AI
CPU system
CPU system
IBM Z
Semiconductors
IBM Z
Future:
CPU IBM 2nm
CPU system
Semiconductors infrastructure
Telum chip
Red Hat OpenShift
Multi-cloud (AWS, Azure ... IBM Cloud, on-premise)
Quantum
IBM Quantum
System Two
Qiskit
User
IBM Quantum
System One
IBM Quantum
System Two
Quantum infrastructure
QPU Heron
Future:
QPU Flamingo

The future of Computing is
Bits + Neurons + Qubits

Hybrid cloud brings
them **together**



Artificial Intelligence



Principles for Enterprise-Grade Generative AI

Trusted

IBM’s AI is responsible and governed

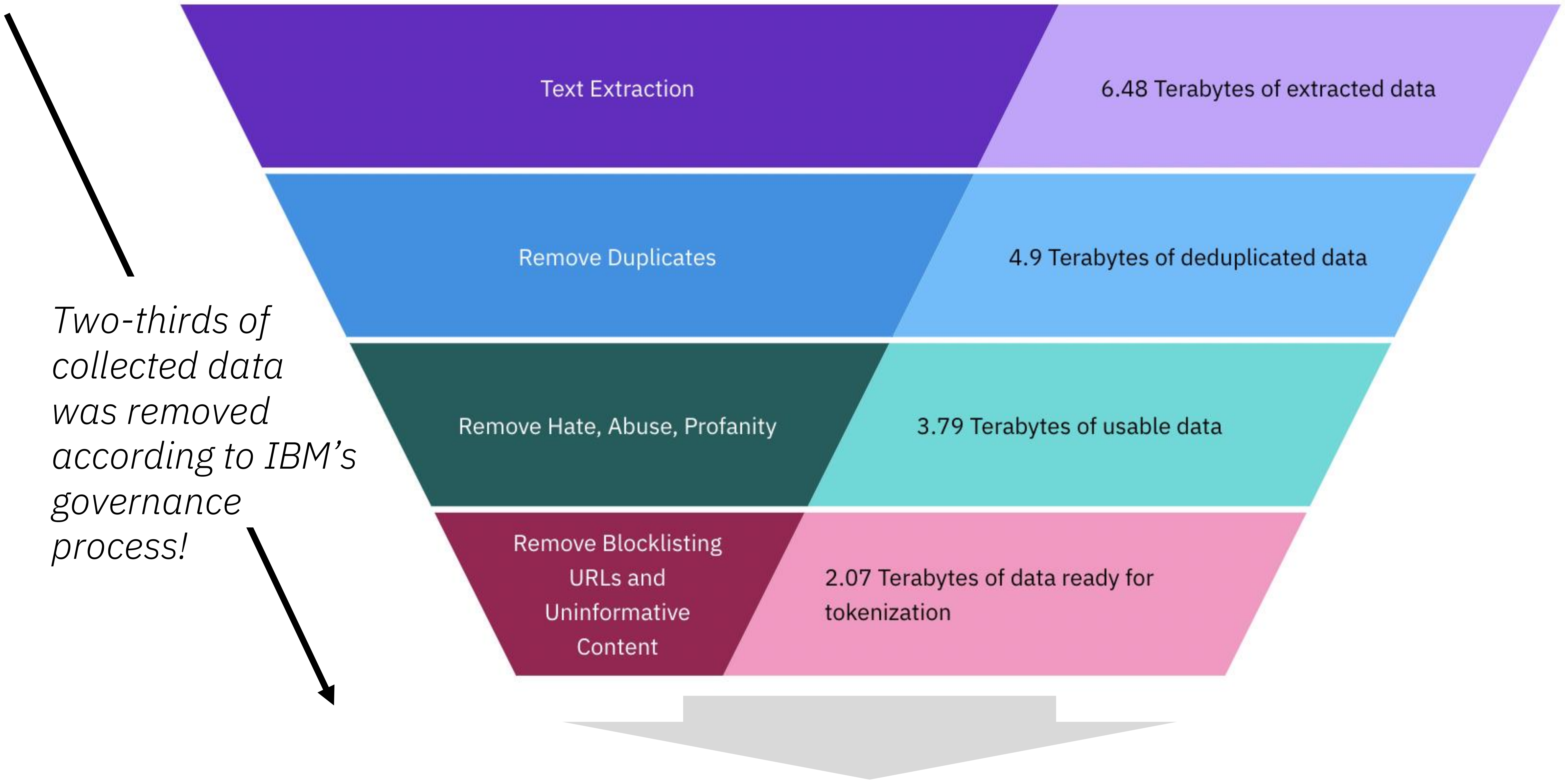
Targeted

IBM’s AI is designed for enterprise and targeted at business domains

Open

IBM’s AI is transparent, publishing key details such as training dataset names

Granite.13b:
Training data governance funnel



Text Extraction

6.48 Terabytes of extracted data

Remove Duplicates

4.9 Terabytes of deduplicated data

Remove Hate, Abuse, Profanity

3.79 Terabytes of usable data

Remove Blocklisting URLs and Uninformative Content

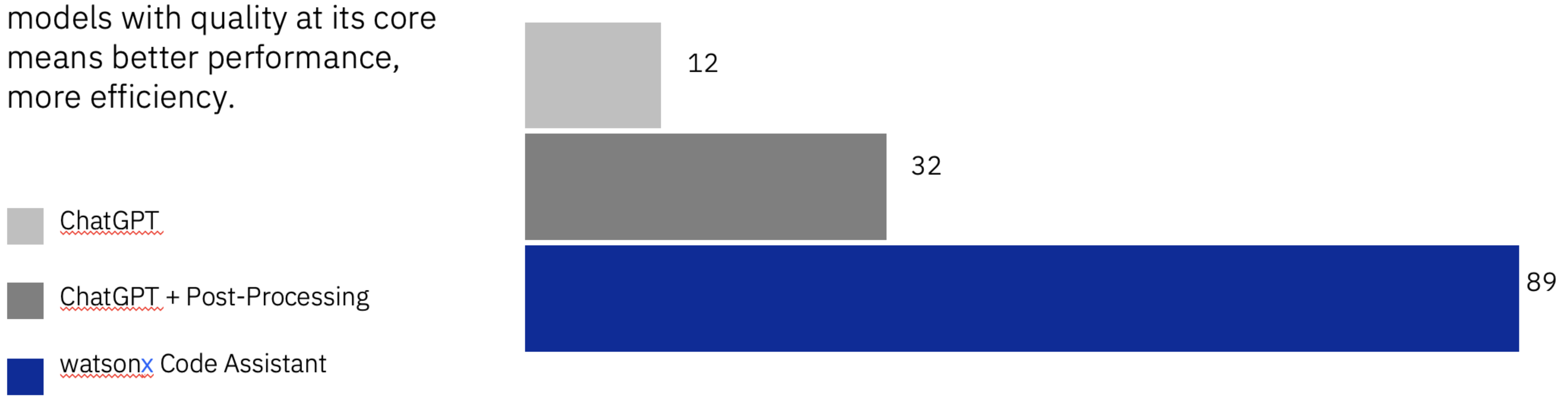
2.07 Terabytes of data ready for tokenization

Two-thirds of collected data was removed according to IBM's governance process!

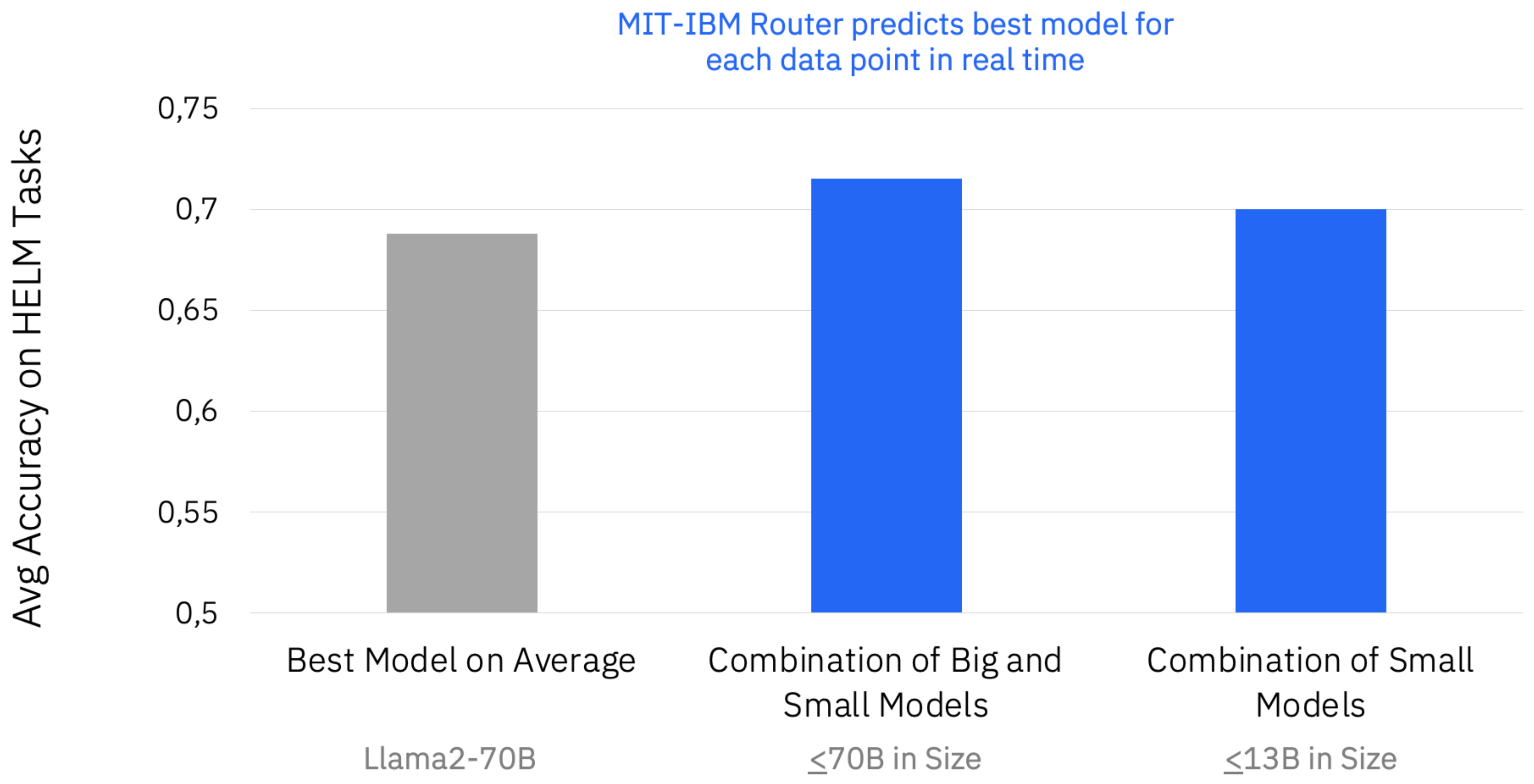
1T Tokens of data for training

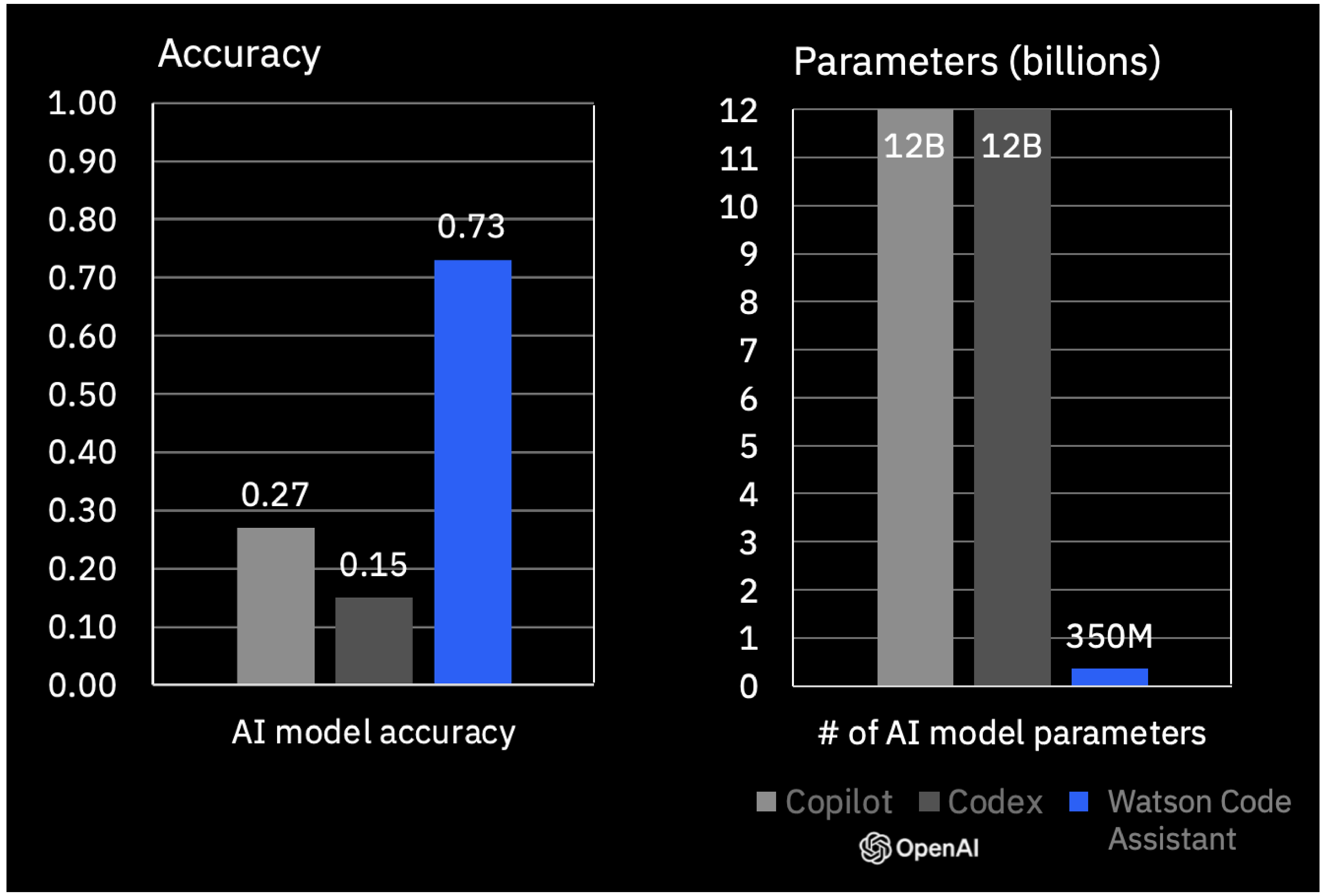
Small, specialized models can outperform large, generalist models

Purpose-built foundation models with quality at its core means better performance, more efficiency.

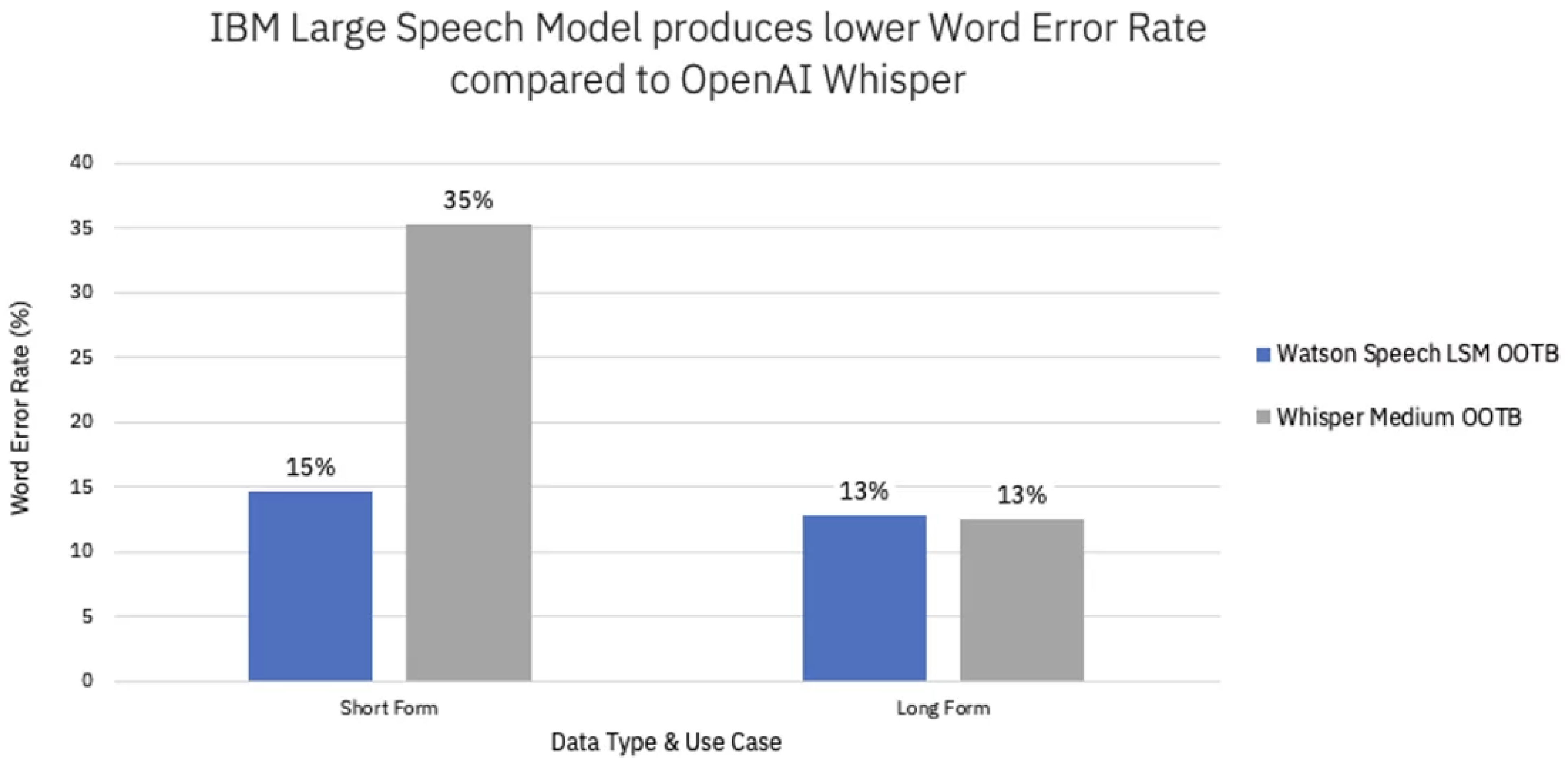


Test accuracy (%): zero-shot performance with CodeNet* benchmark





Watsonx Code Assistant

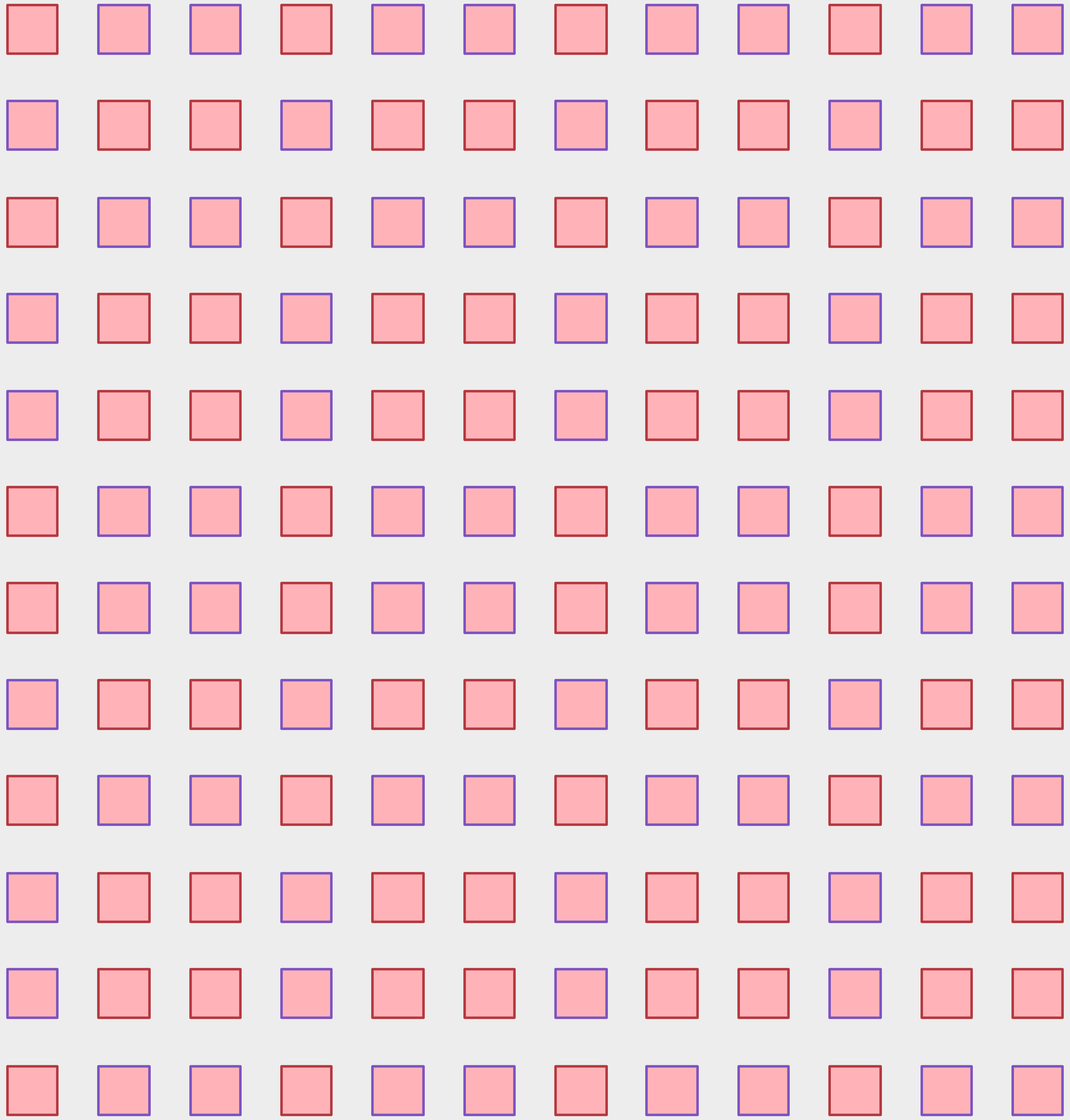


watsonx

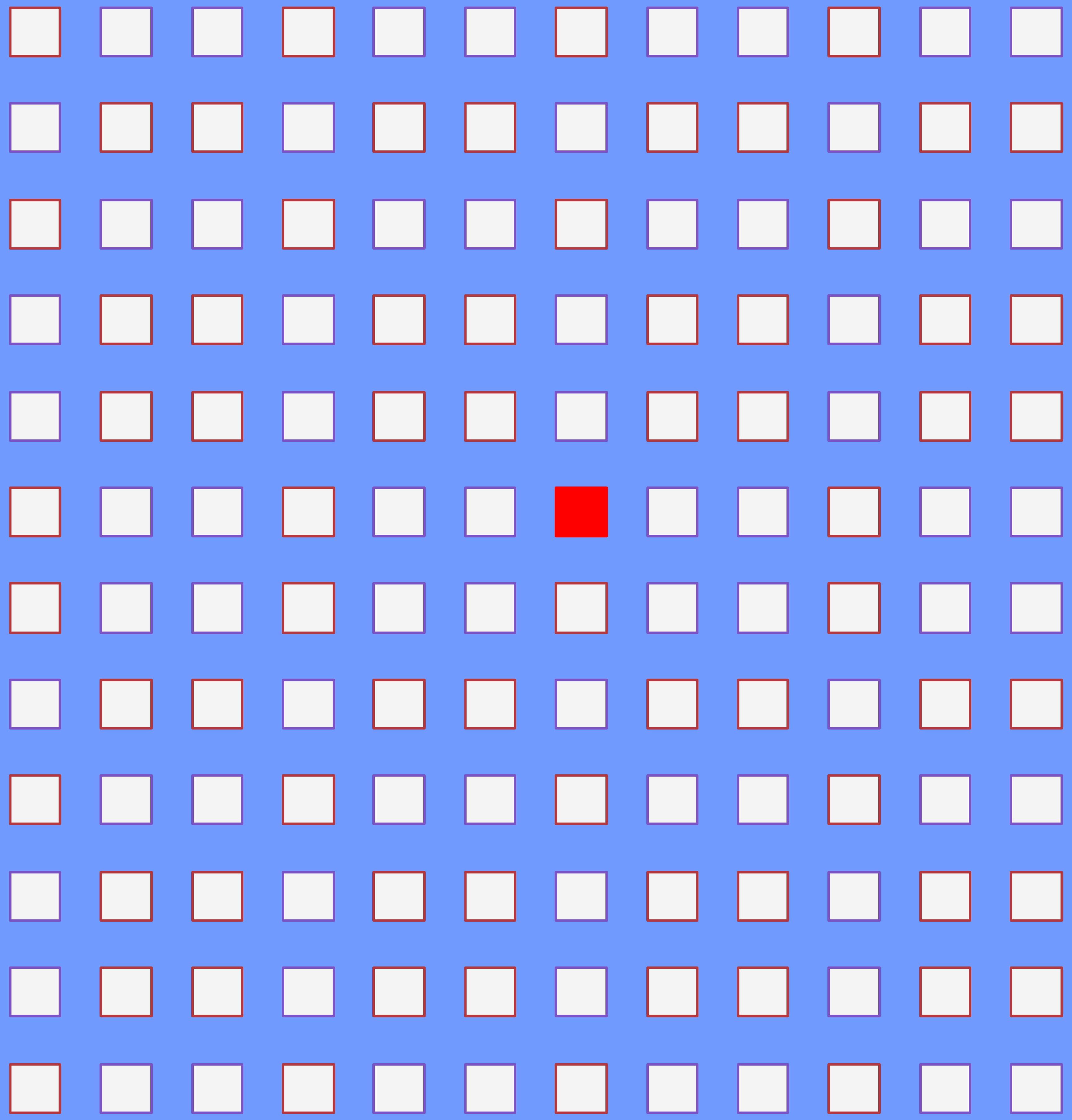
Scale and accelerate the impact of AI with trusted data, an open architecture, and seamless integration

AI assistants	Empower individuals to do work without expert knowledge across a variety of business processes and applications	**watsonx** Orchestrate **watsonx** Assistant **watsonx** Code Assistant **watsonx** Orders	
SDKs and APIs	Use programmatic interfaces to embed watsonx platform capabilities in assistants and applications	**Ecosystem integrations**	
AI and data platform	Leverage generative AI and machine learning — tuned with your data — with responsibility, transparency and explainability	**watsonx** watsonx.ai watsonx.governance watsonx.data	**Foundation models** Open Source | *Hugging Face* Llama 2 | *Meta AI* Geospatial | *IBM + NASA* Granite | *IBM*
Data services	Access data fabric services to define, organize, manage, and deliver trusted data to train and tune models	**Data fabric services**	
Hybrid cloud AI tools	Build on a consistent, scalable foundation based on open-source technology	**Red Hat** OpenShift AI (*e.g.*, Ray, Pytorch)	

Nearly all available public data is now represented in foundation models



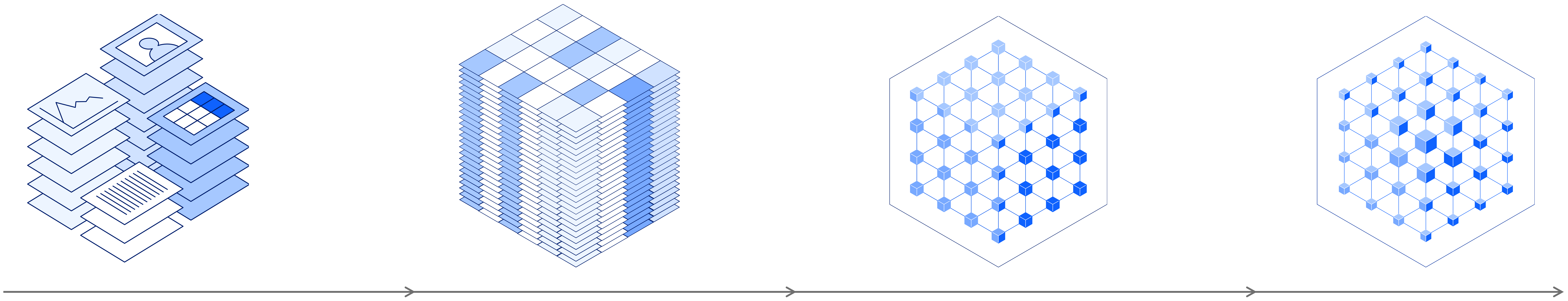
Less than 1% of all enterprise data is represented in foundation models



We’ve invented a new methodology: InstructLab

- Makes LLMs truly open-source with collaborative mode development
- Allow LLM to learn as humans do, using knowldege and skils
- Enable incremental skill teaching

https://instructlab.ai/



Generate examples

High quality, hand-curated knowledge sources, plus a curated taxonomy of tasks with human-generated examples for each.

Teacher model(s)

A teacher model generates a “curriculum” of millions of questions and answers for the taxonomies.

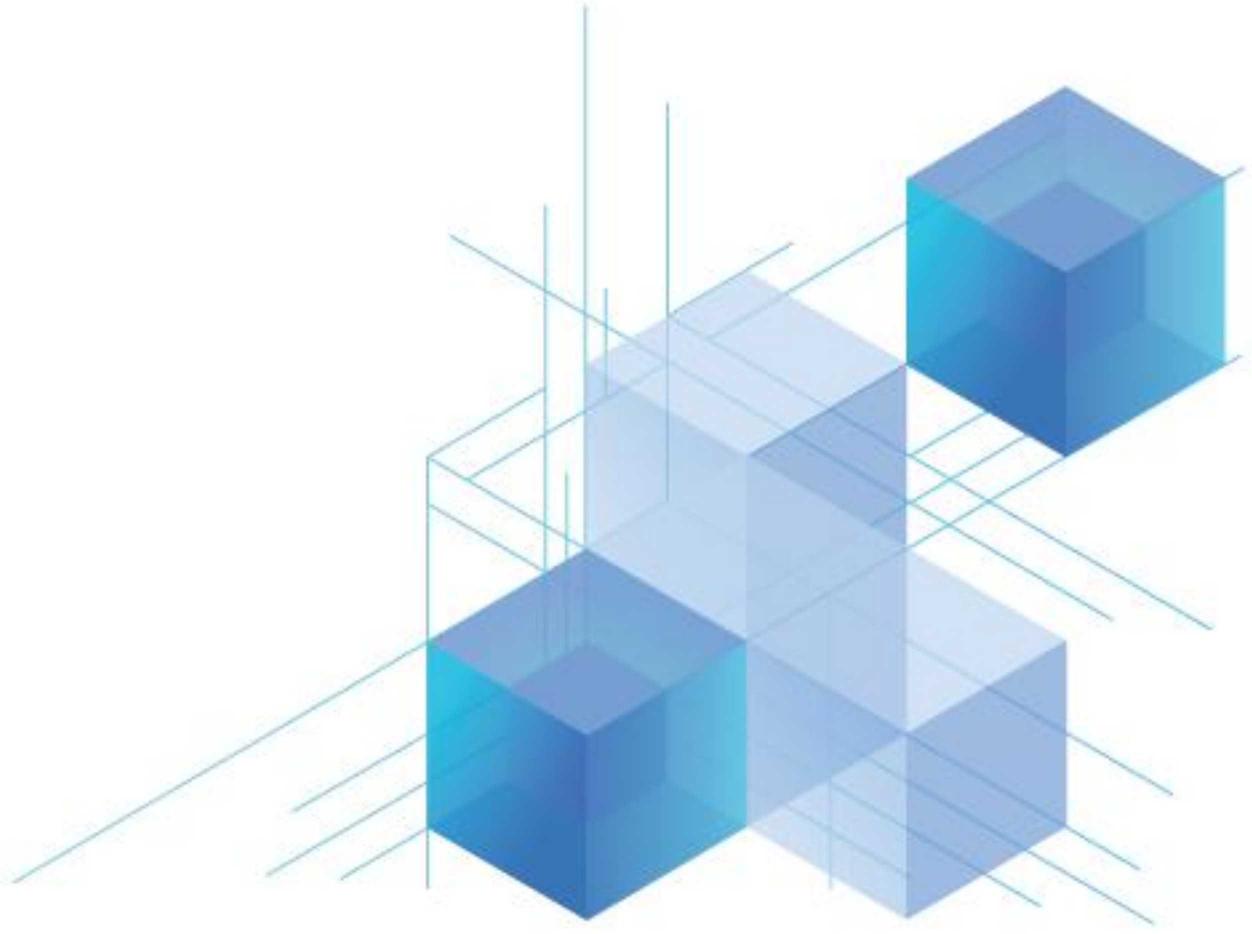
Critic model(s)

Critic models filter the questions for correctness and quality. Synthetic data is scanned for prohibited material.

Student model(s)

The student model is trained with the curriculum using a novel training approach.

Semiconductors Infrastructure



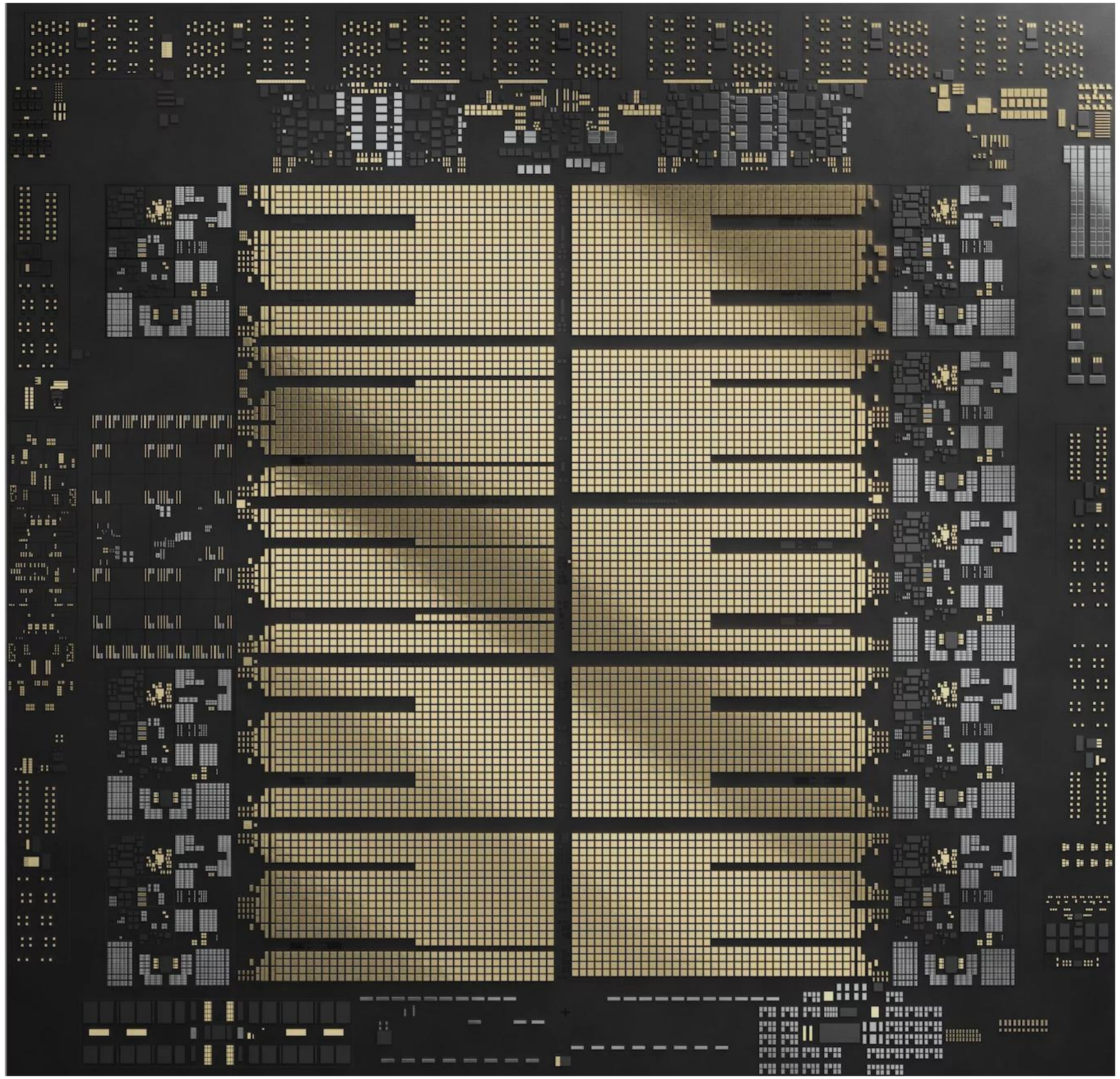
IBM Telum Family

Telum 1 (2020)

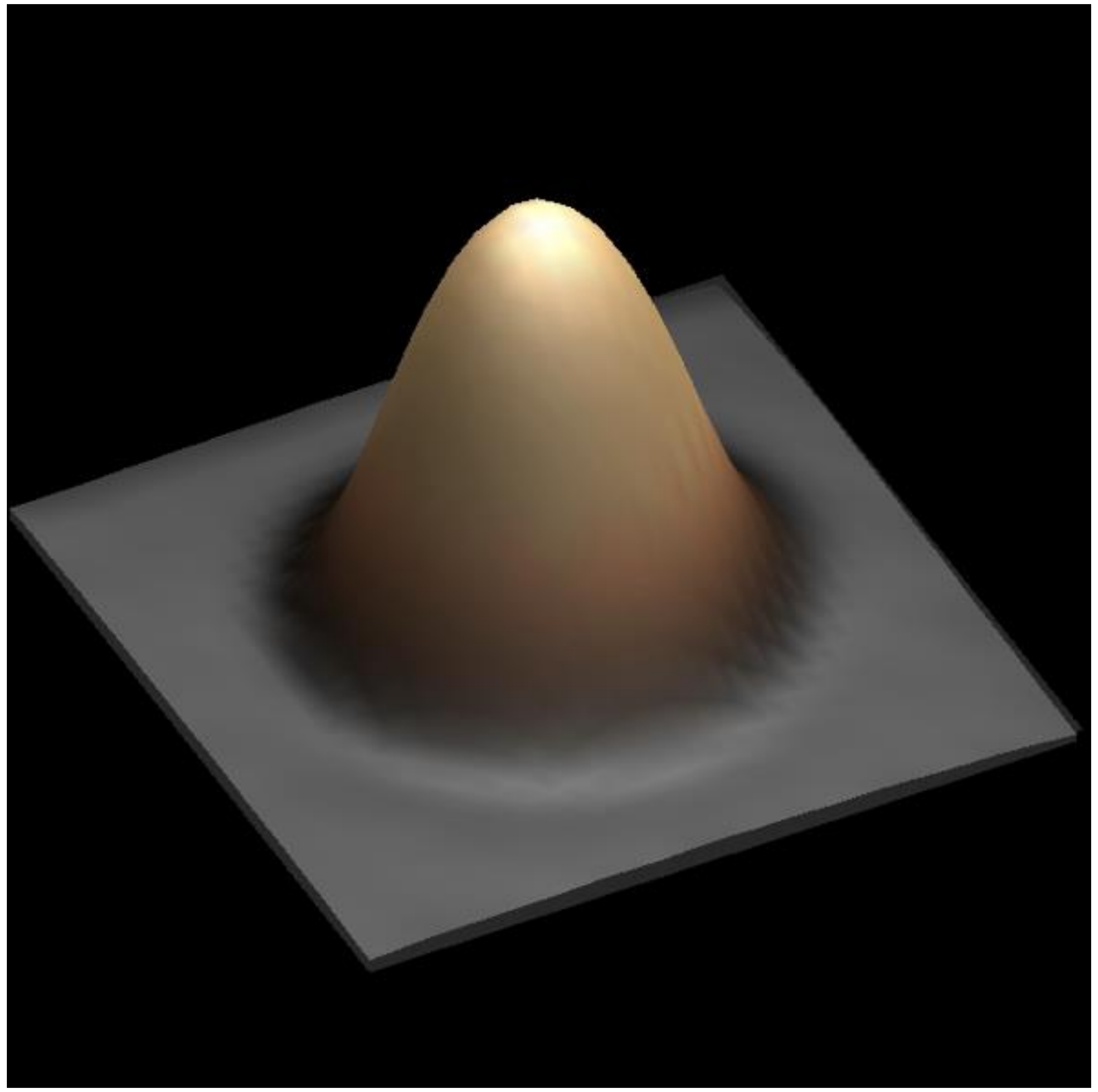
- Response time less than a millisecond for a volume of 100,000 transactions per second
- 7 nm
- Explosion of digital payments, Detecting fraud

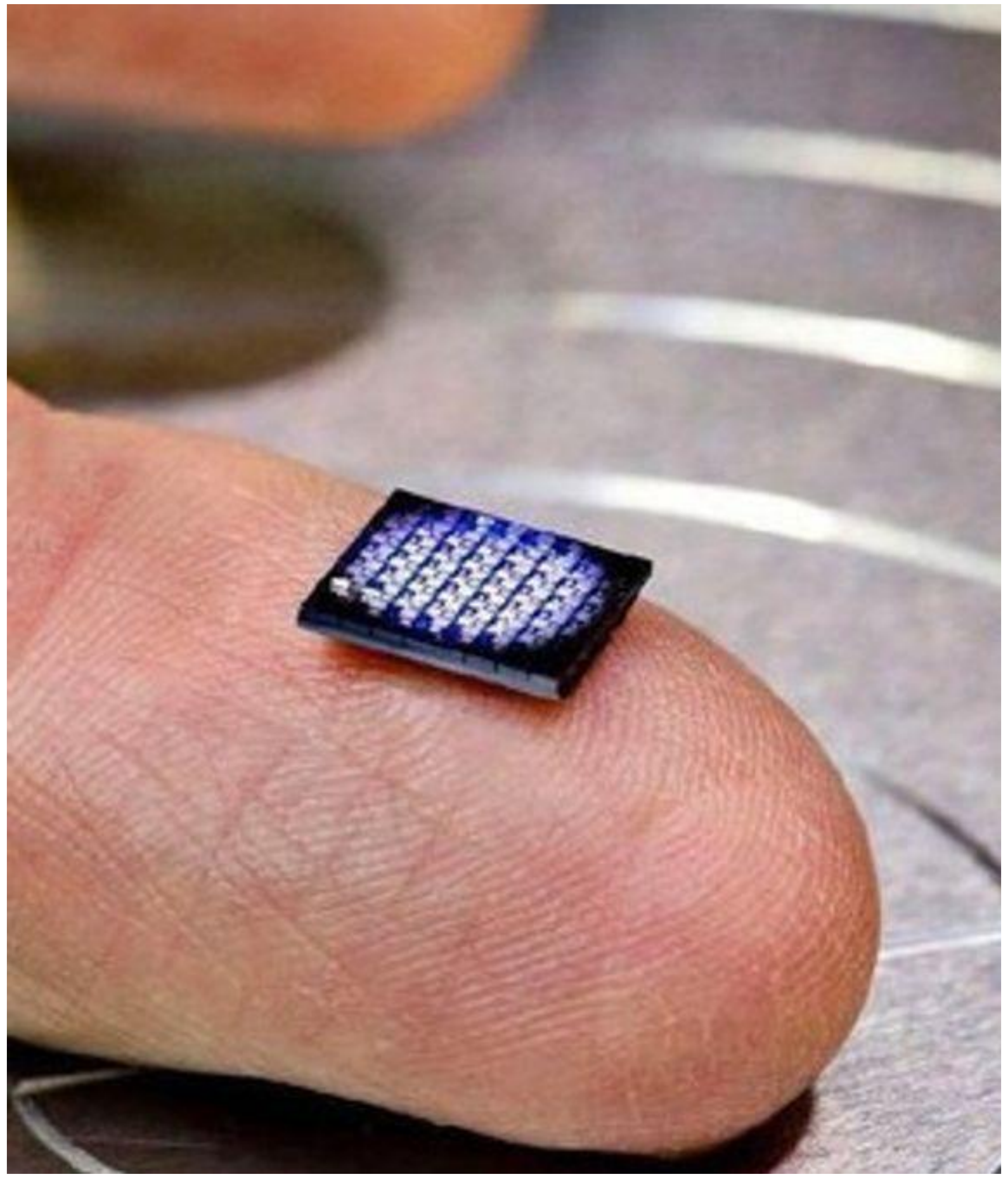
Telum 2 (2024)

- 32 AI cores
- While Telum I offered 32MB of L2 cache per core, Telum II increased this by 40%, with virtual L3 and L4 caches growing to 360MB and 2.88GB, respectively

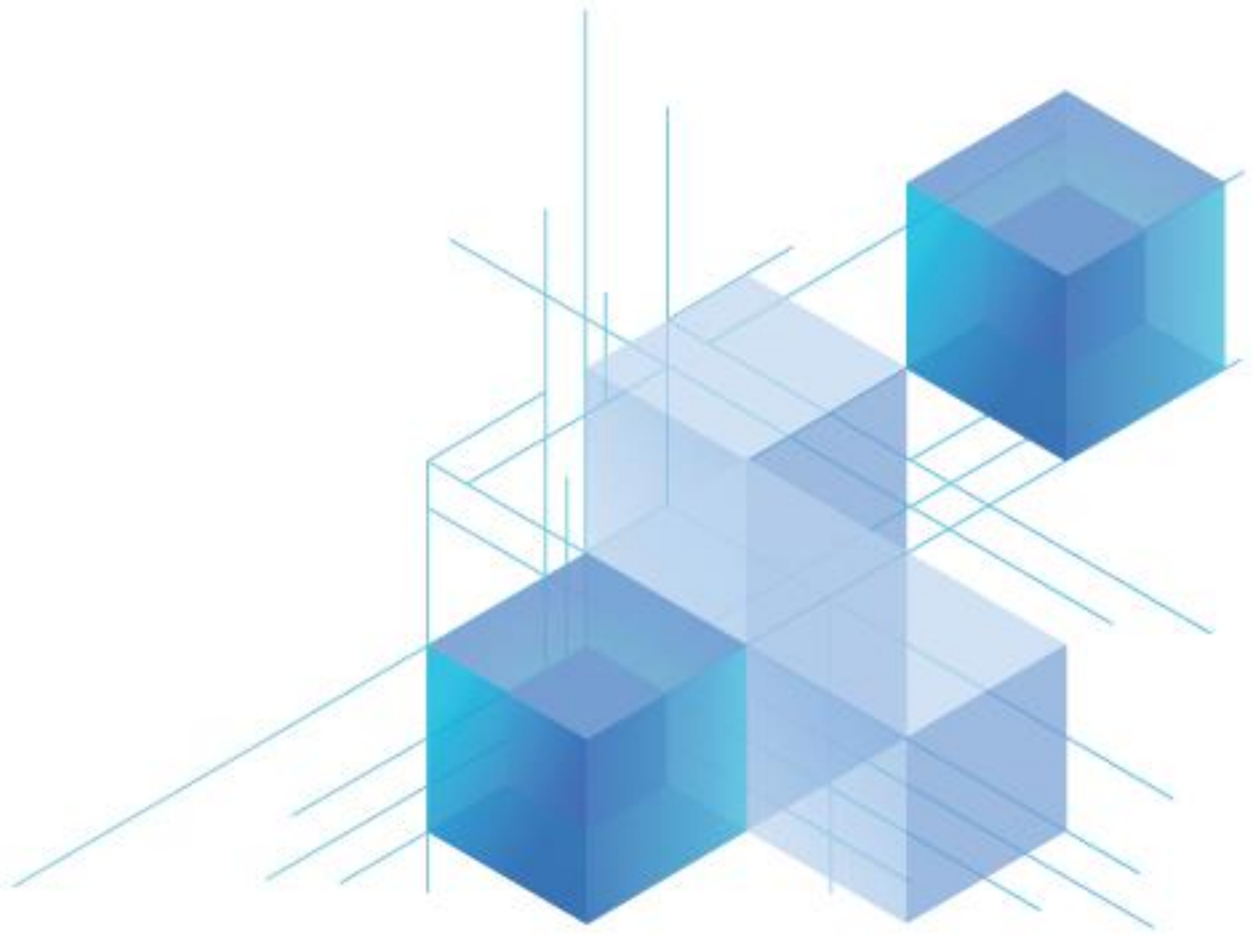


Exploring the future of hybrid cloud infrastructure



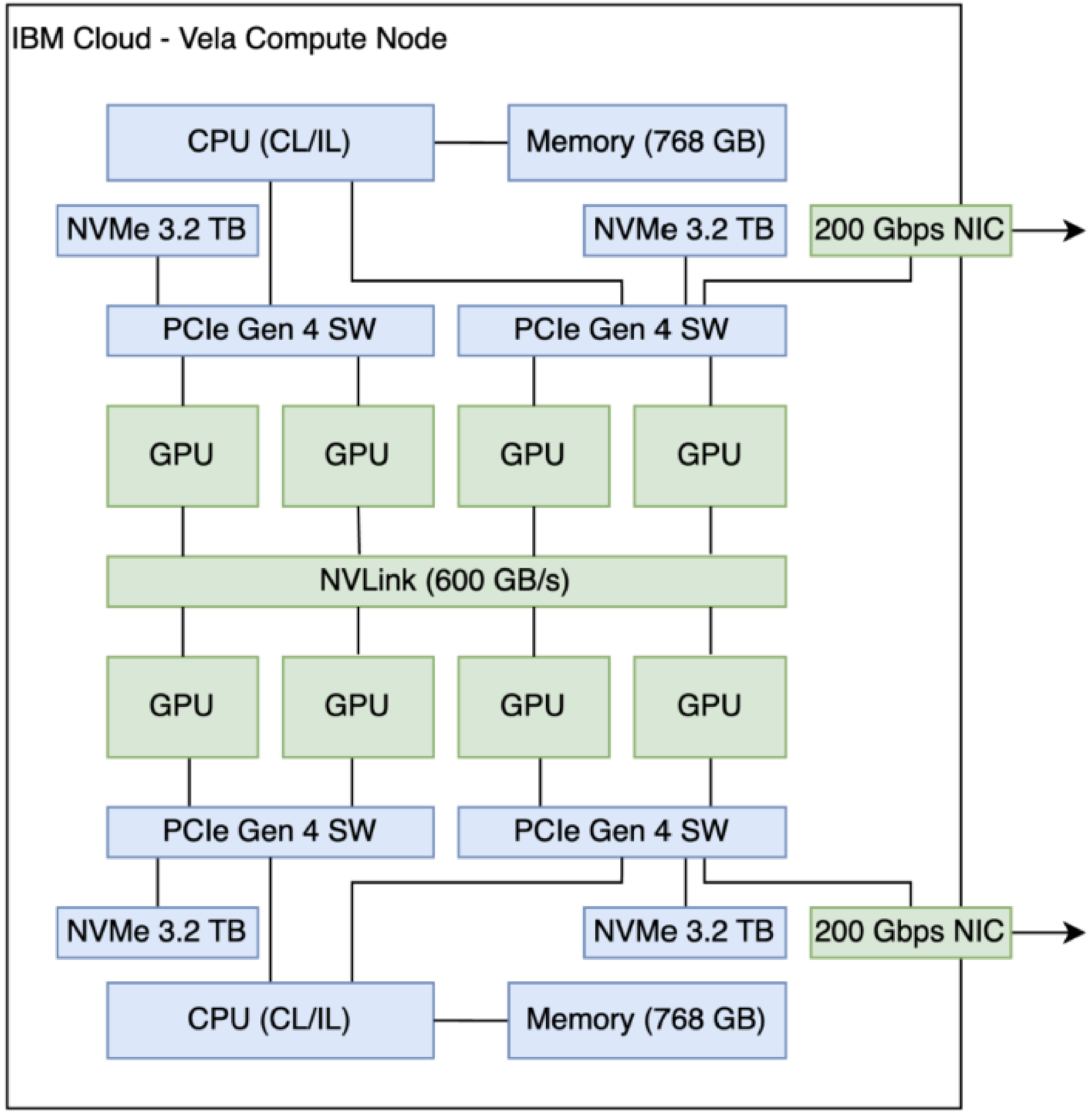


AI Infrastructure



IBM Vela: Our high-performing, cloud-native AI training stack running on the Red Hat OpenShift Container Platform

- ~900 petaflop fully software defined AI system with near (within ~5%) bare metal performance
- IBM's contributions to PyTorch enable 4.5x efficient training on large models (10B parameters) on commodity Ethernet

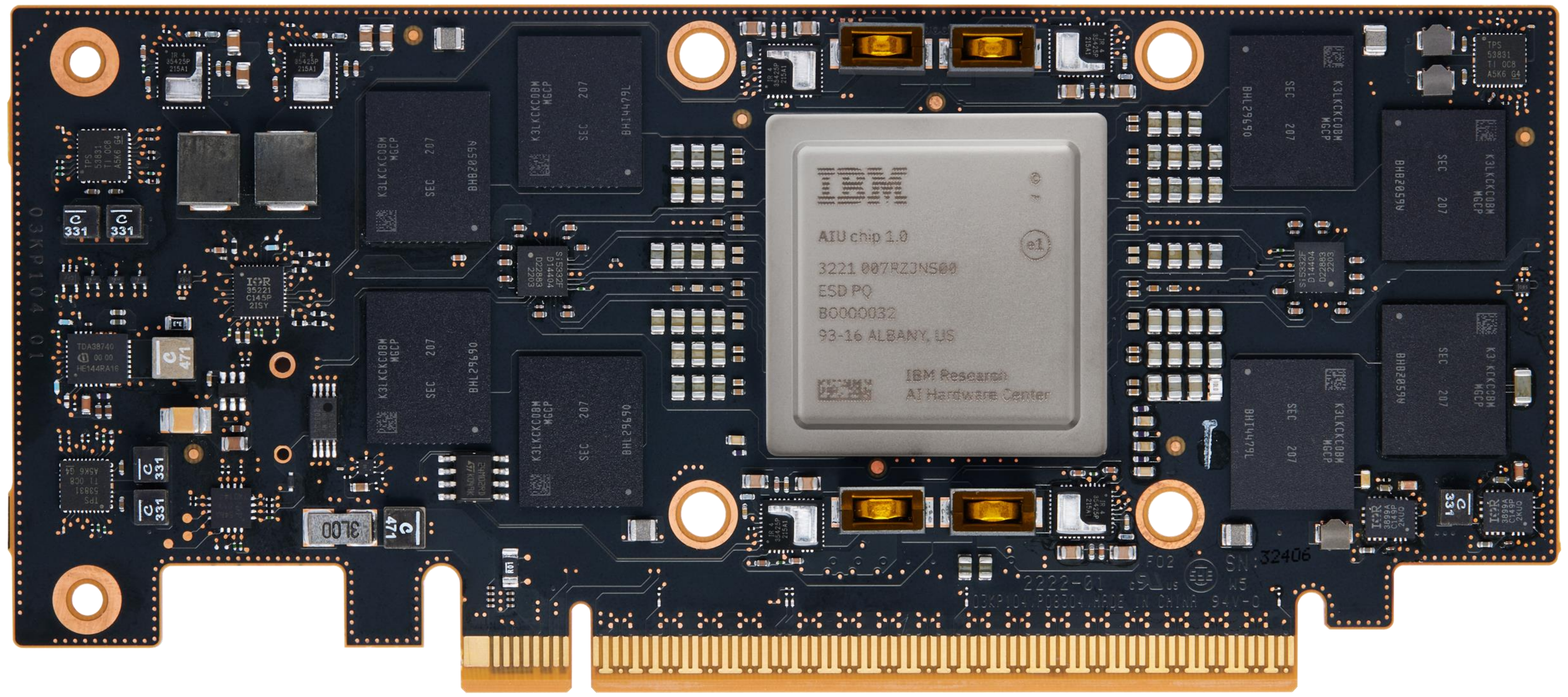


IBM Artificial Intelligence Unit

An entire chip dedicated to AI

- 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)

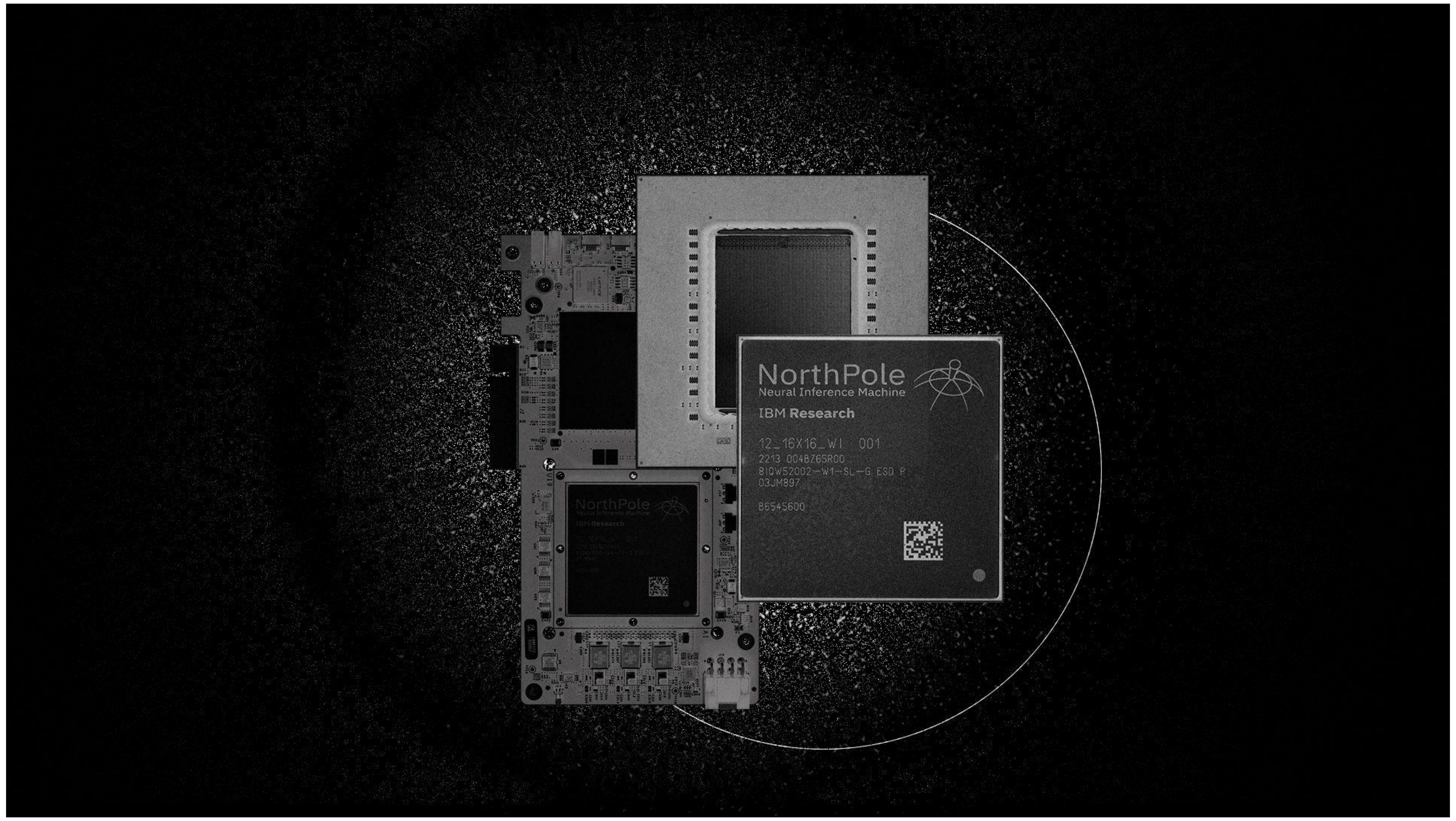
  FP16, FP8, INT8, INT4, INT2
- Implemented in leading edge 5nm technology
  - 32 processing cores
  - 23 billion transistors

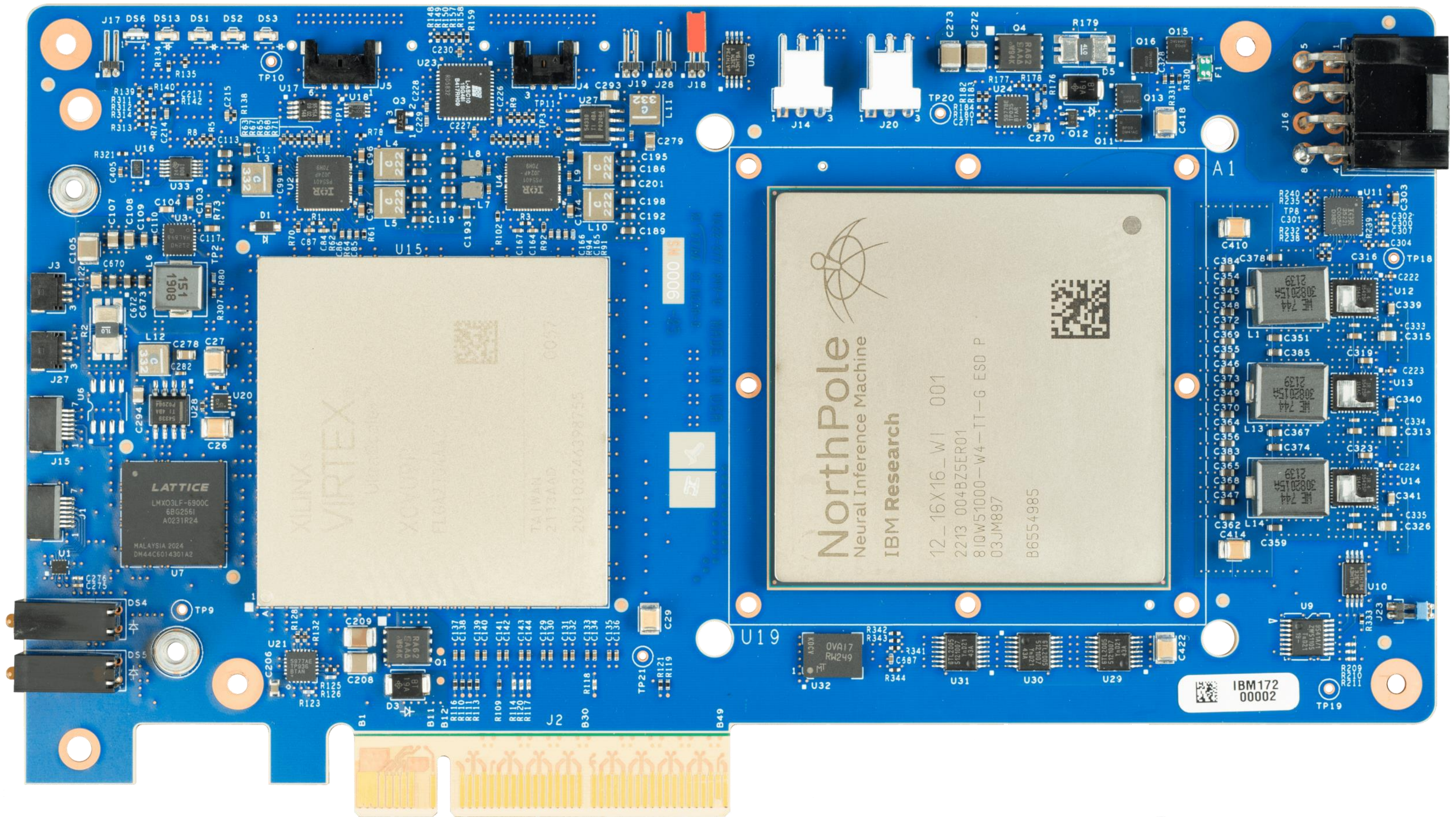


IBM NorthPole

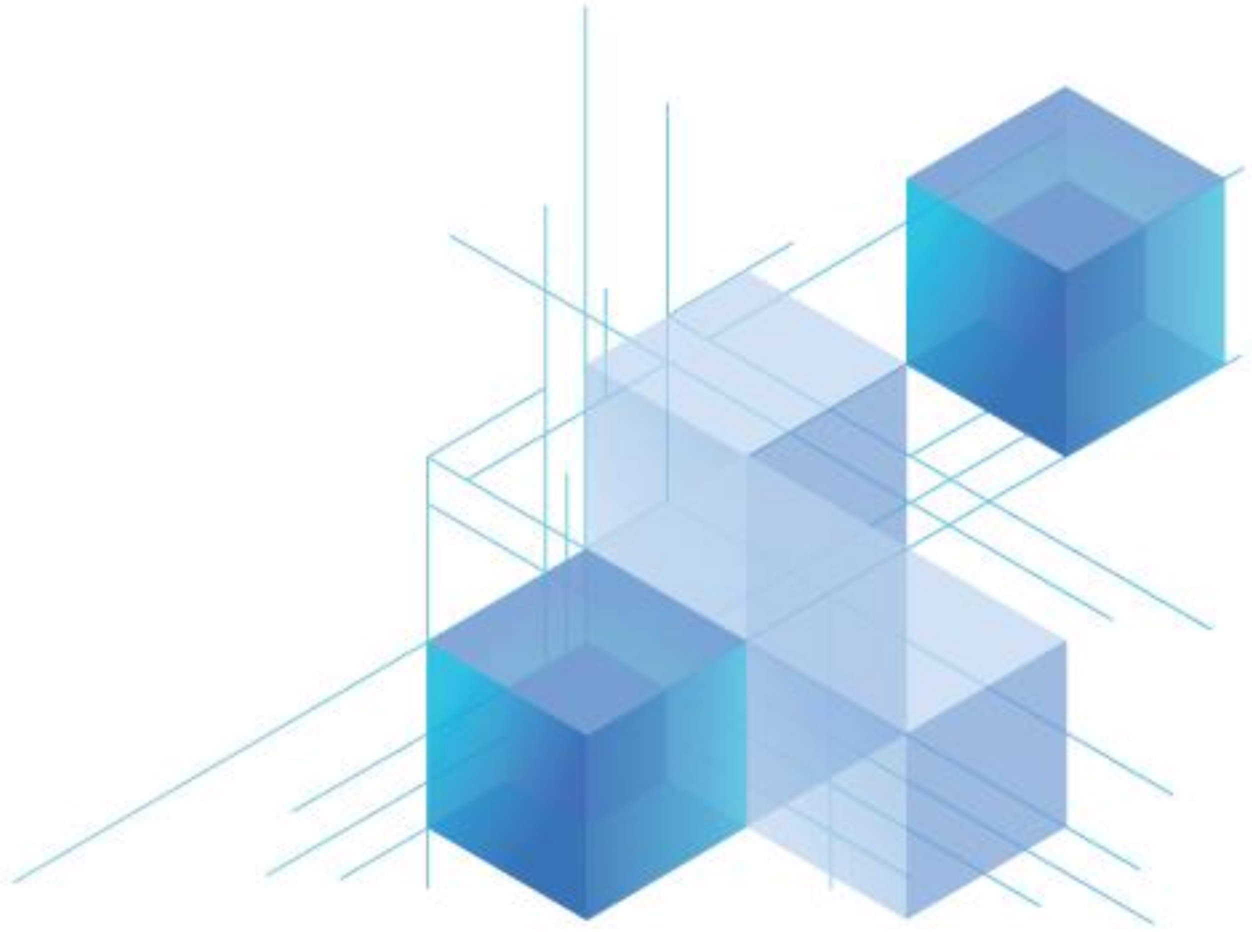
- Brain inspired chip (Published in Science)
- Compared common 12nm GPUs and 14-nm CPUs, NorthPole is 25 times more energy efficient

- On ResNet-50, NorthPole outperforms all major prevalent architectures — even those that use more advanced technology processes, such as a GPU implemented using a 4 nm process.





Quantum Infrastructure



IBM has the strongest quantum ecosystem advancing the field of quantum computing

since 2016

3T+

Circuits run on our systems

250+

IBM Quantum Network members

75+

Systems deployed worldwide since 2016

8+

Quantum systems at IBM Quantum data centers

2

Global quantum data centers

5

Quantum systems at client-locations
(by end of 2024)

IBM
IBM Quantum
System One

BAY 01
IBM Quantum
System Two

Heron

133 / 156

qubit count

Tunable coupler architecture

I/O complexity on par with Osprey

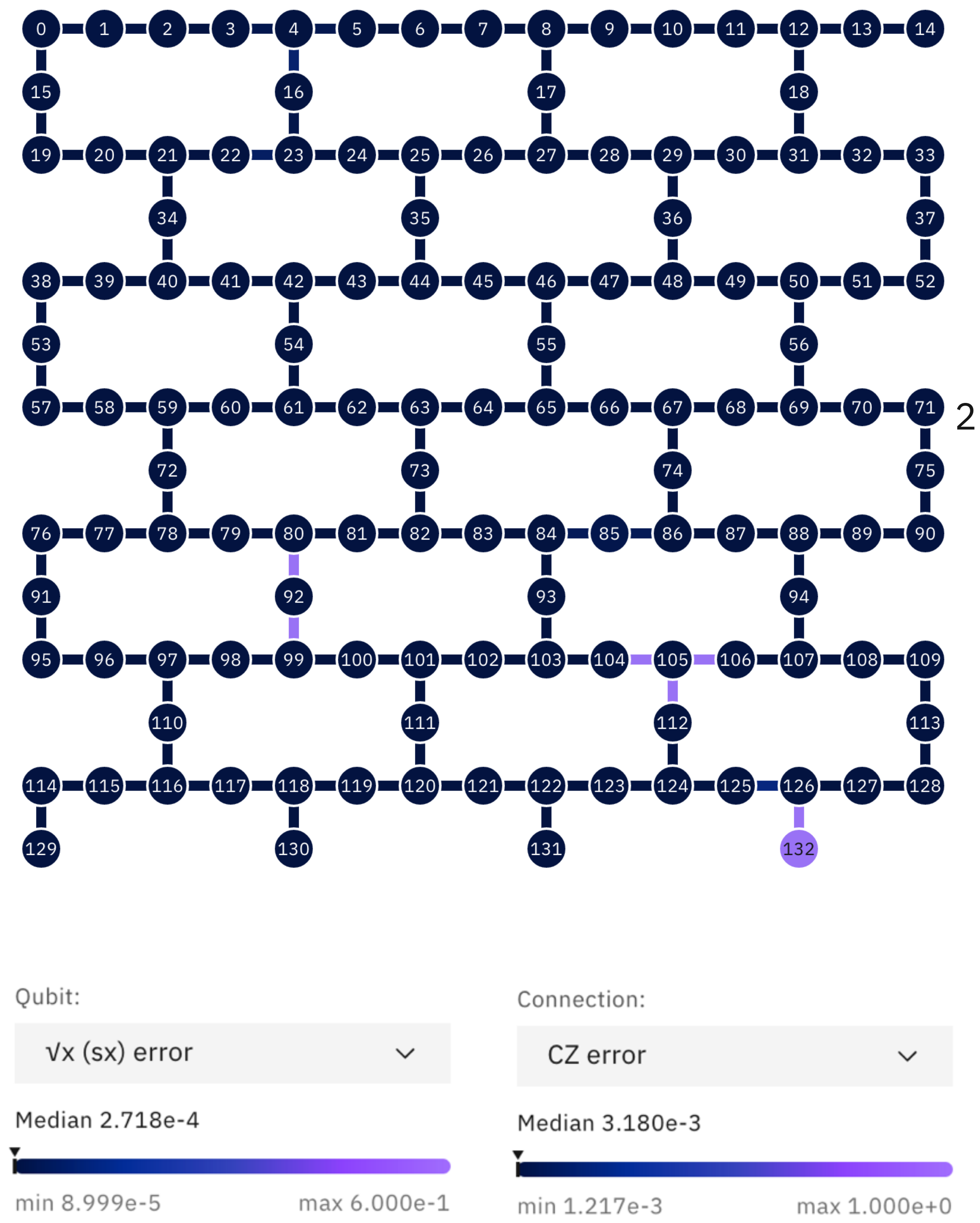
Heron

R1: 133-qubit systems
R2: 156-qubit systems

Tunable coupler architecture

R2: Includes ability to tune away two-level system defects during calibration

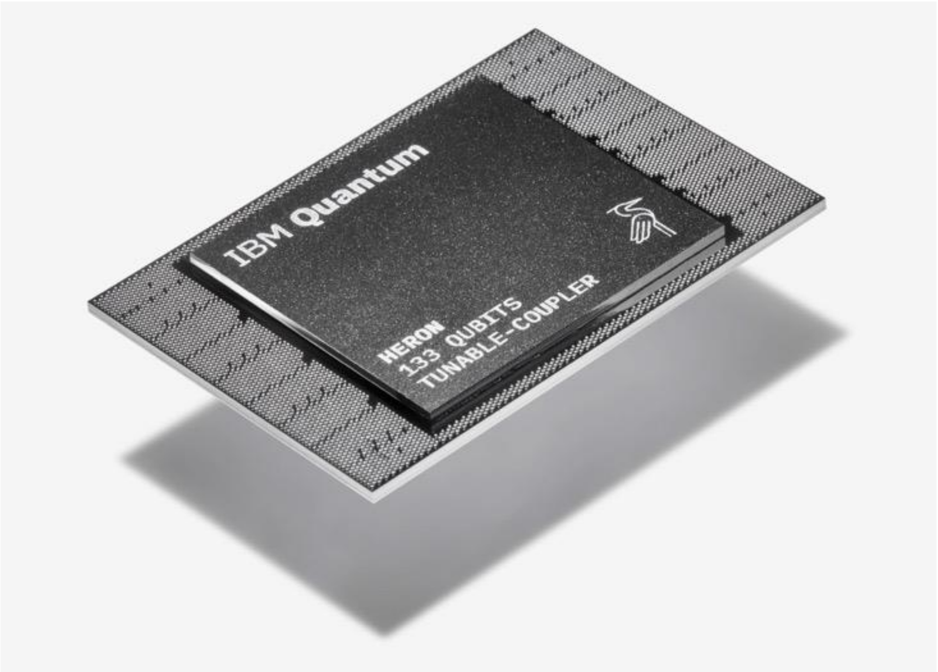
IBM Monte Carlo simulation



Qubit:
√x (sx) error
Median 2.718e-4
min 8.999e-5 max 6.000e-1

Connection:
CZ error
Median 3.180e-3
min 1.217e-3 max 1.000e+0

	IBM Sherbrooke Eagle	IBM Monte Carlo (Heron)
Gate error (best system)	0.6%–0.7%	0.3% – Best ~ 0.1%
Crosstalk	High (qubit-qubit collisions)	Almost zero!
Gate time	500–600ns	90–100ns



Condor

Pushing the limits of scale & yield

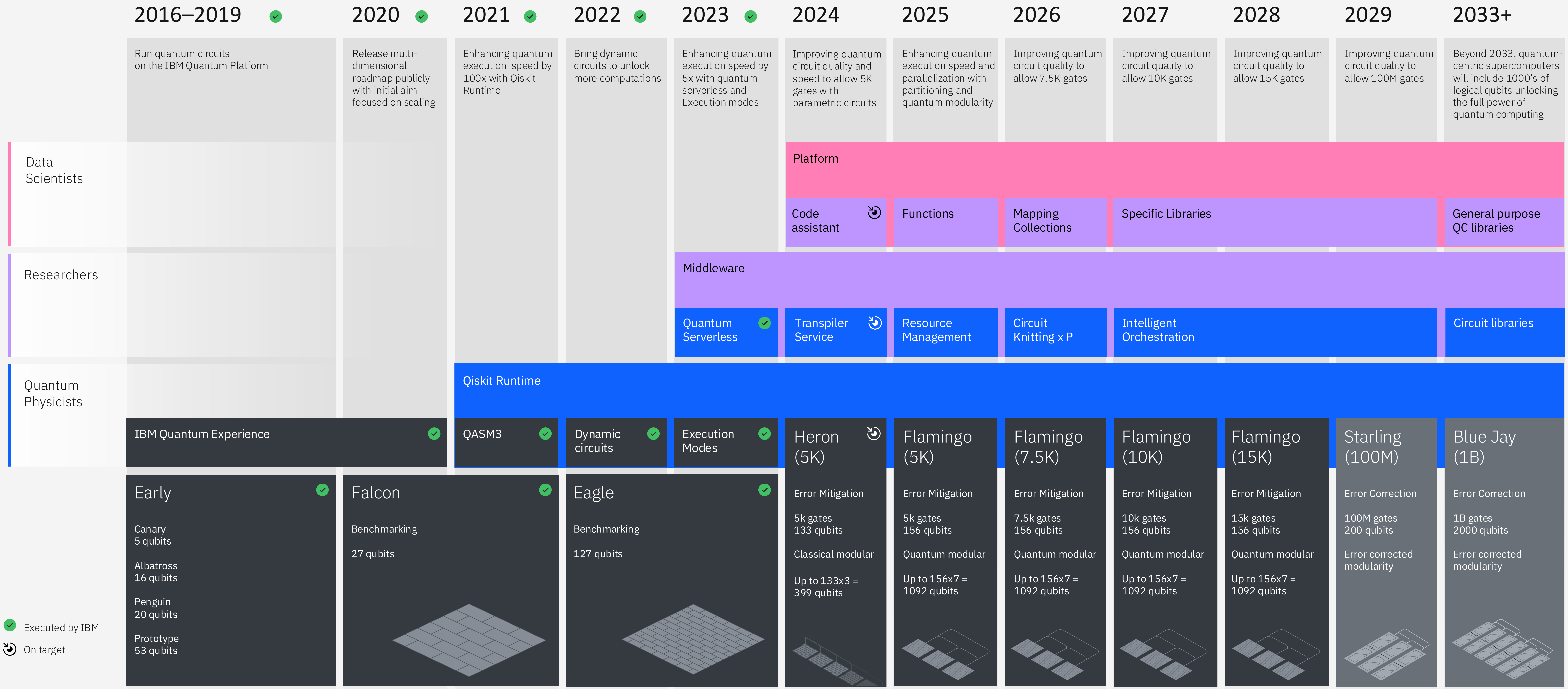
1,121

Superconducting metal qubits

Chip wiring and layout enhancements

Predictive simulation enhancements

Development Roadmap



	2016–2019 ✓	2020 ✓	2021 ✓	2022 ✓	2023 ✓	2024	2025	2026	2027	2028	2029	2033+
	Run quantum circuits on the IBM Quantum Platform	Release multi-dimensional roadmap publicly with initial aim focused on scaling	Enhancing quantum execution speed by 100x with Qiskit Runtime	Bring dynamic circuits to unlock more computations	Enhancing quantum execution speed by 5x with quantum serverless and Execution modes	Improving quantum circuit quality and speed to allow 5K gates with parametric circuits	Enhancing quantum execution speed and parallelization with partitioning and quantum modularity	Improving quantum circuit quality to allow 7.5K gates	Improving quantum circuit quality to allow 10K gates	Improving quantum circuit quality to allow 15K gates	Improving quantum circuit quality to allow 100M gates	Beyond 2033, quantum-centric supercomputers will include 1000's of logical qubits unlocking the full power of quantum computing
Data Scientists						Platform						
						Code assistant ⟳	Functions	Mapping Collections	Specific Libraries			General purpose QC libraries
Researchers					Middleware							
					Quantum Serverless ✓	Transpiler Service ⟳	Resource Management	Circuit Knitting x P	Intelligent Orchestration			Circuit libraries
Quantum Physicists			Qiskit Runtime									
	IBM Quantum Experience ✓		QASM3 ✓	Dynamic circuits ✓	Execution Modes ✓	Heron (5K) ⟳	Flamingo (5K)	Flamingo (7.5K)	Flamingo (10K)	Flamingo (15K)	Starling (100M)	Blue Jay (1B)
						Error Mitigation 5k gates 133 qubits Classical modular Up to 133x3 = 399 qubits	Error Mitigation 5k gates 156 qubits Quantum modular Up to 156x7 = 1092 qubits	Error Mitigation 7.5k gates 156 qubits Quantum modular Up to 156x7 = 1092 qubits	Error Mitigation 10k gates 156 qubits Quantum modular Up to 156x7 = 1092 qubits	Error Mitigation 15k gates 156 qubits Quantum modular Up to 156x7 = 1092 qubits	Error Correction 100M gates 200 qubits Error corrected modularity	Error Correction 1B gates 2000 qubits Error corrected modularity
	Early ✓ Canary 5 qubits Albatross 16 qubits Penguin 20 qubits Prototype 53 qubits	Falcon ✓ Benchmarking 27 qubits		Eagle ✓ Benchmarking 127 qubits								

✓ Executed by IBM

⟳ On target

Merci