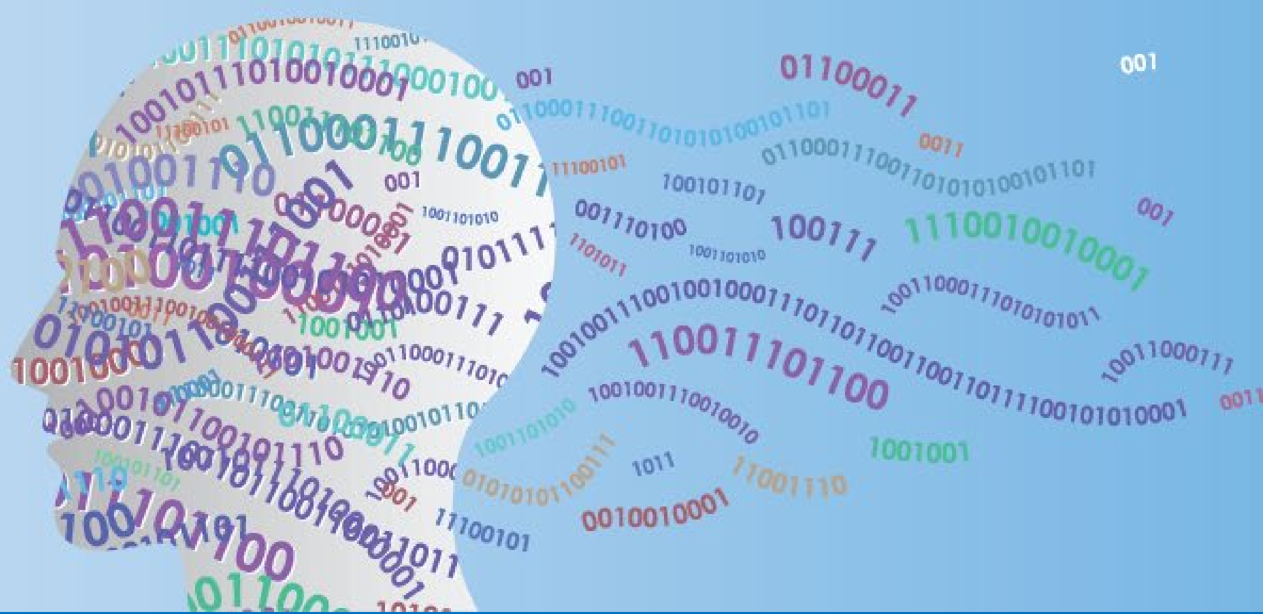
SNIA PERSISTENT MEMORY + SUMMIT 2021 COMPUTATIONAL STORAGE

FROM DATACENTER TO EDGE : VIRTUAL EVENT
APRIL 21-22, 2021



Distributed AI and Computational Storage

Michael Kagan
CTO, NVIDIA

The New Age of AI

Everything is Connected. Data Generated and Consumed Everywhere

Edge Data Centers

Millions of Edge DC
Each with Thousands of Nodes

2M Factories

13M Stores

570M Farms

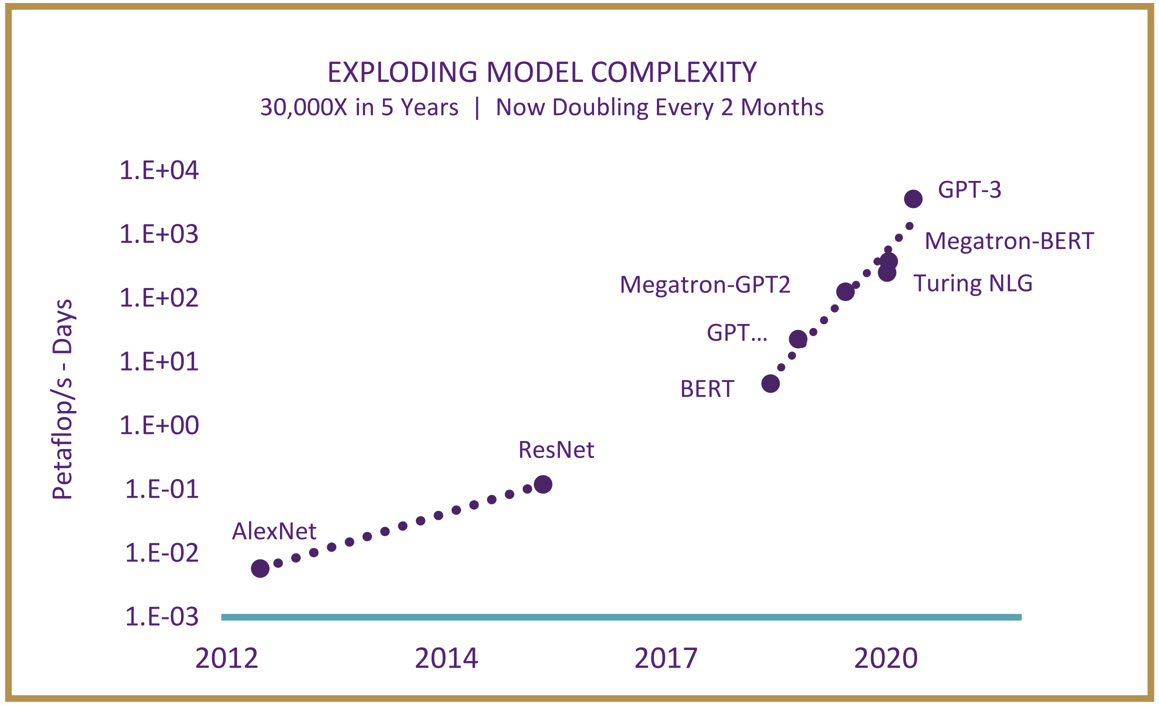
5M Call Center Agents

40M Miles of Road

10T Miles Driven per Year

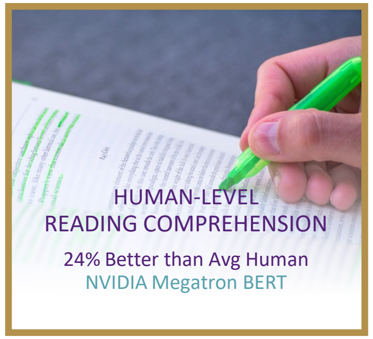
Exponential Growth in AI Model Complexity

SNIA PERSISTENT MEMORY + SUMMIT 2021 COMPUTATIONAL STORAGE



EXPLODING MODEL COMPLEXITY

30,000X in 5 Years | Now Doubling Every 2 Months



HUMAN-LEVEL READING COMPREHENSION

24% Better than Avg Human

NVIDIA Megatron BERT

HUMAN-LIKE CHATBOTS

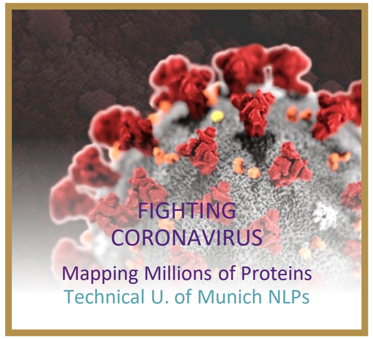
Half of Users Preferred AI

Facebook BlenderBot

DRONE DELIVERIES

6X Safer Landings

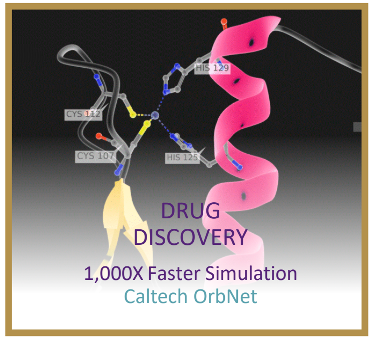
Caltech Neural Lander



FIGHTING CORONAVIRUS

Mapping Millions of Proteins

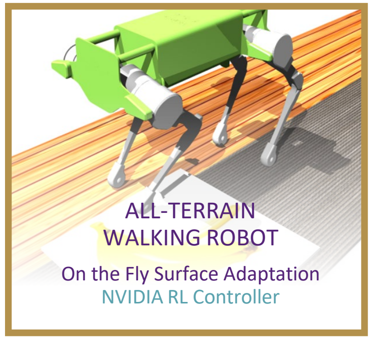
Technical U. of Munich NLPs



DRUG DISCOVERY

1,000X Faster Simulation

Caltech OrbNet



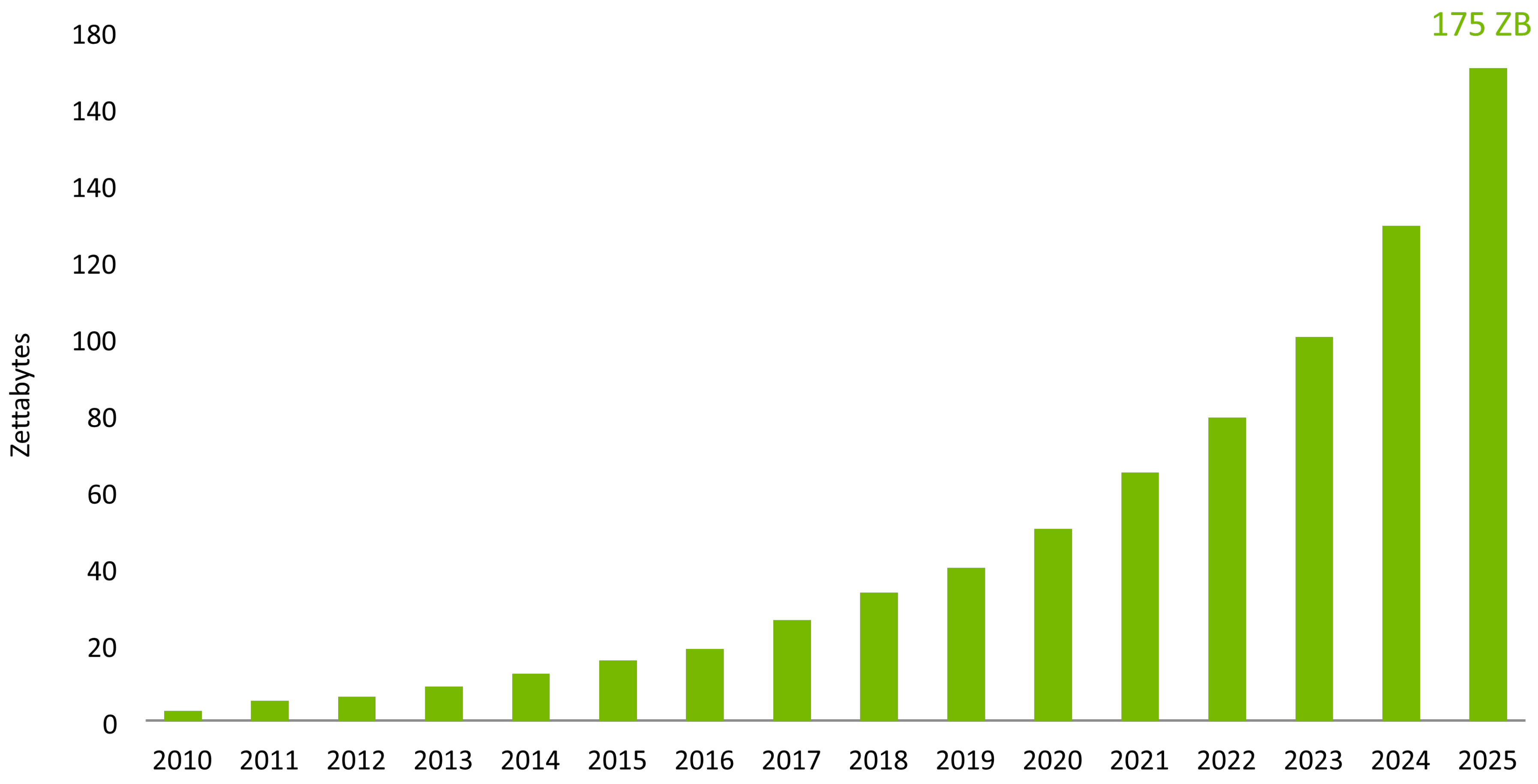
ALL-TERRAIN WALKING ROBOT

On the Fly Surface Adaptation

NVIDIA RL Controller

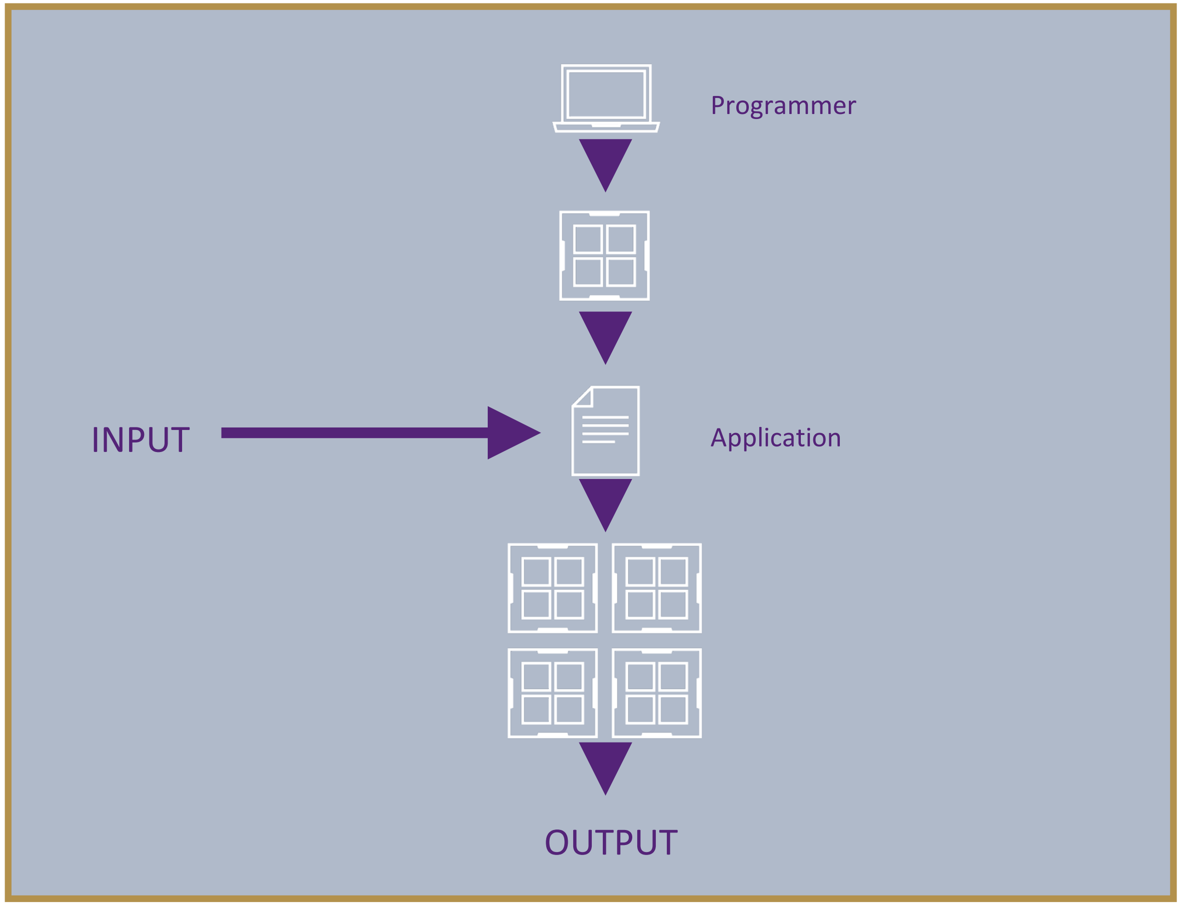
Data Grows Exponentially

More Data → New Models → New Applications → Even More Data

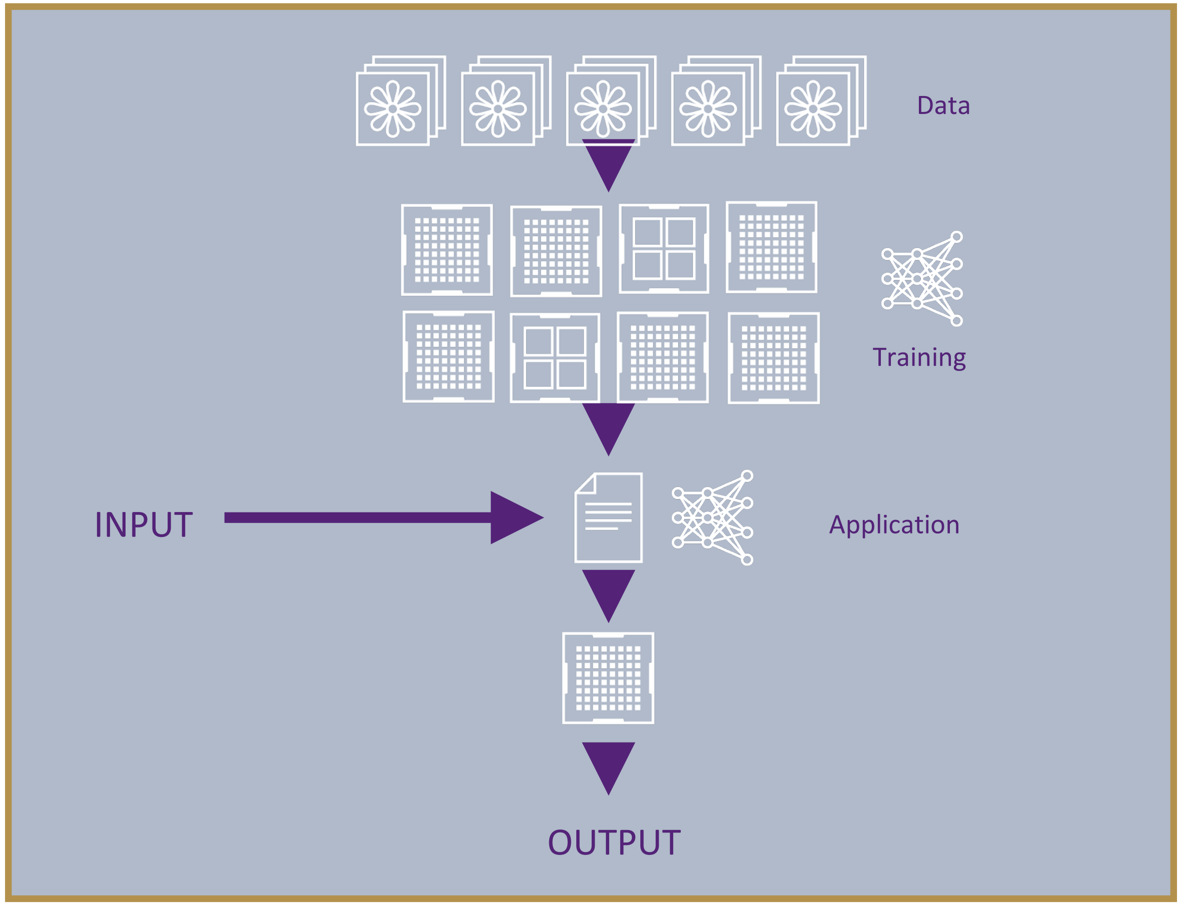


Software Writes Software

Traditional Model: Programmer uses small computer to create an application that runs on a large centralized computer

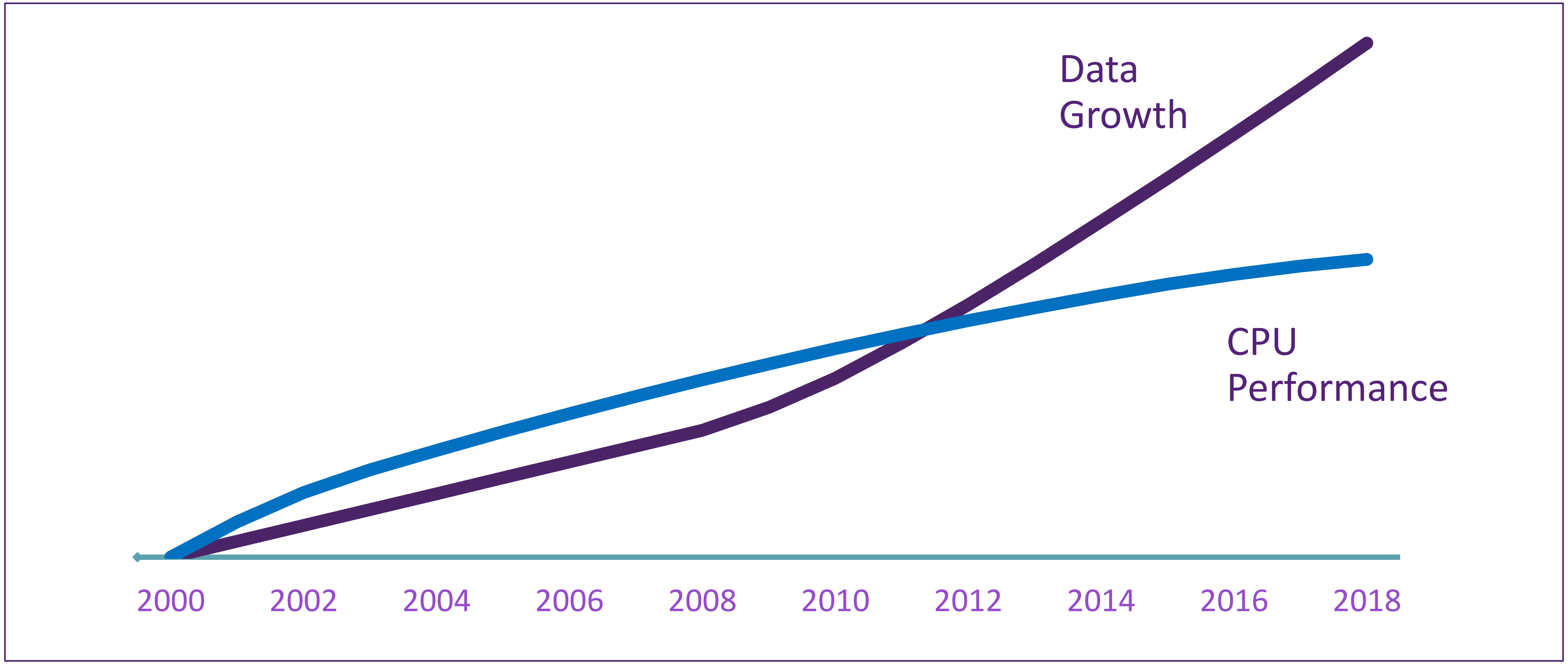


New AI Model: A large computer uses data to train applications that run on distributed small computers



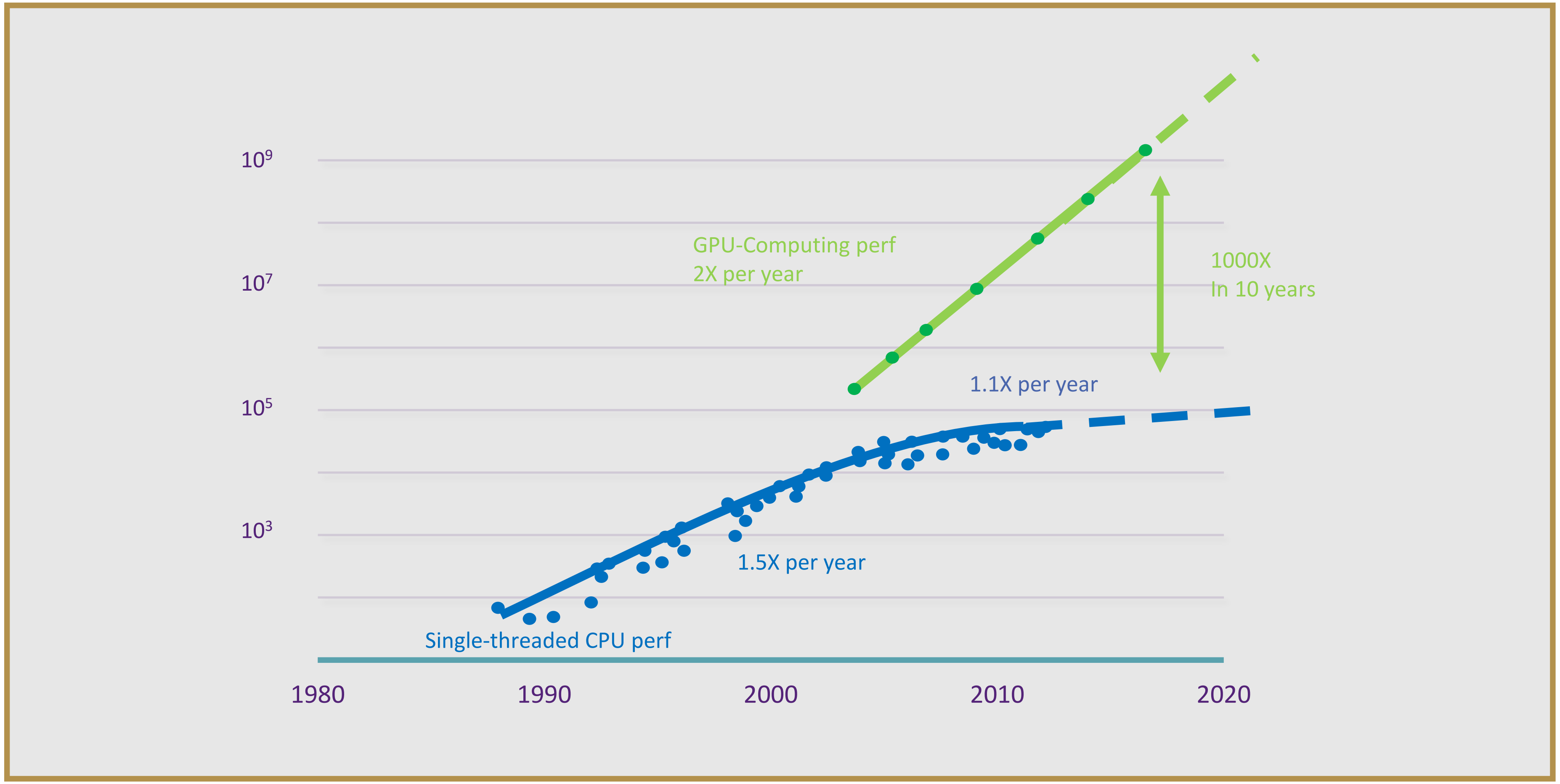
Now Moore’s Law Is Slowing Down

Fails to Keep up with Data Growth, Model Complexity, Memory Speeds, etc.

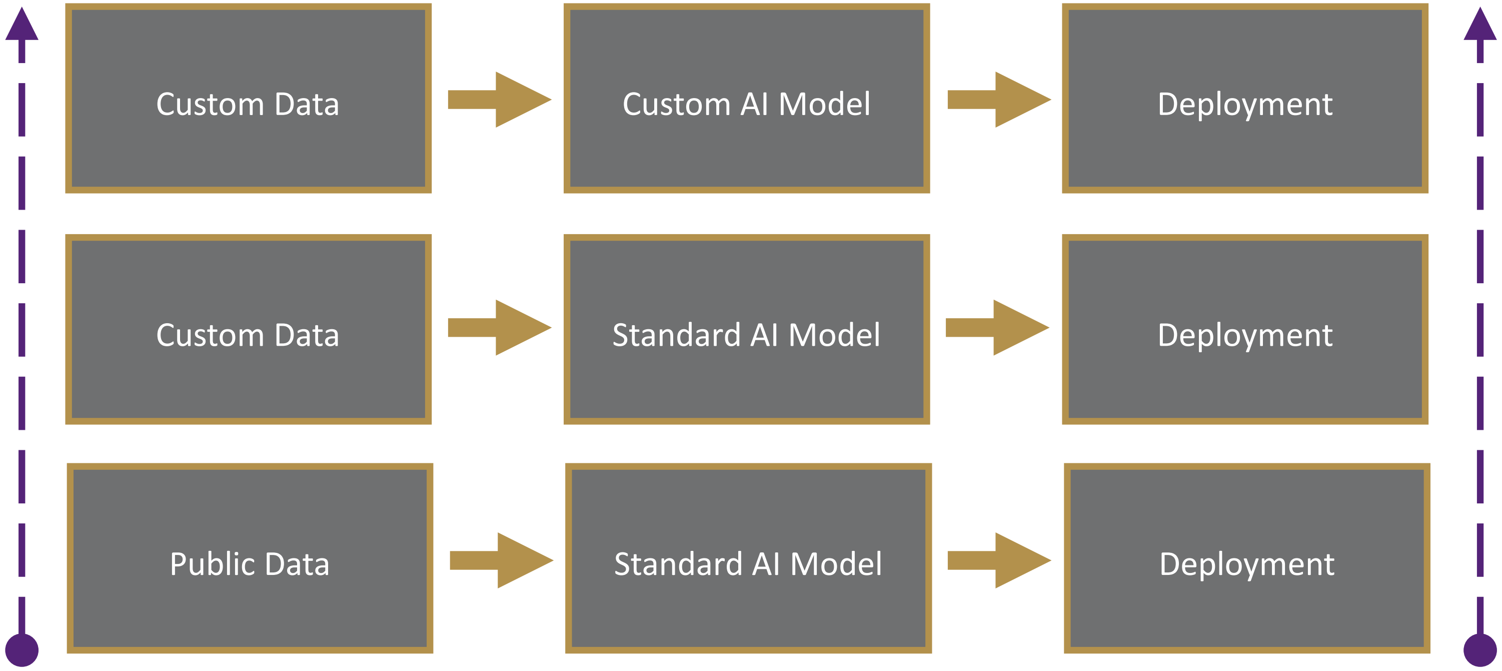


New Compute Growth Engines

GPU Becomes the General Processing Unit

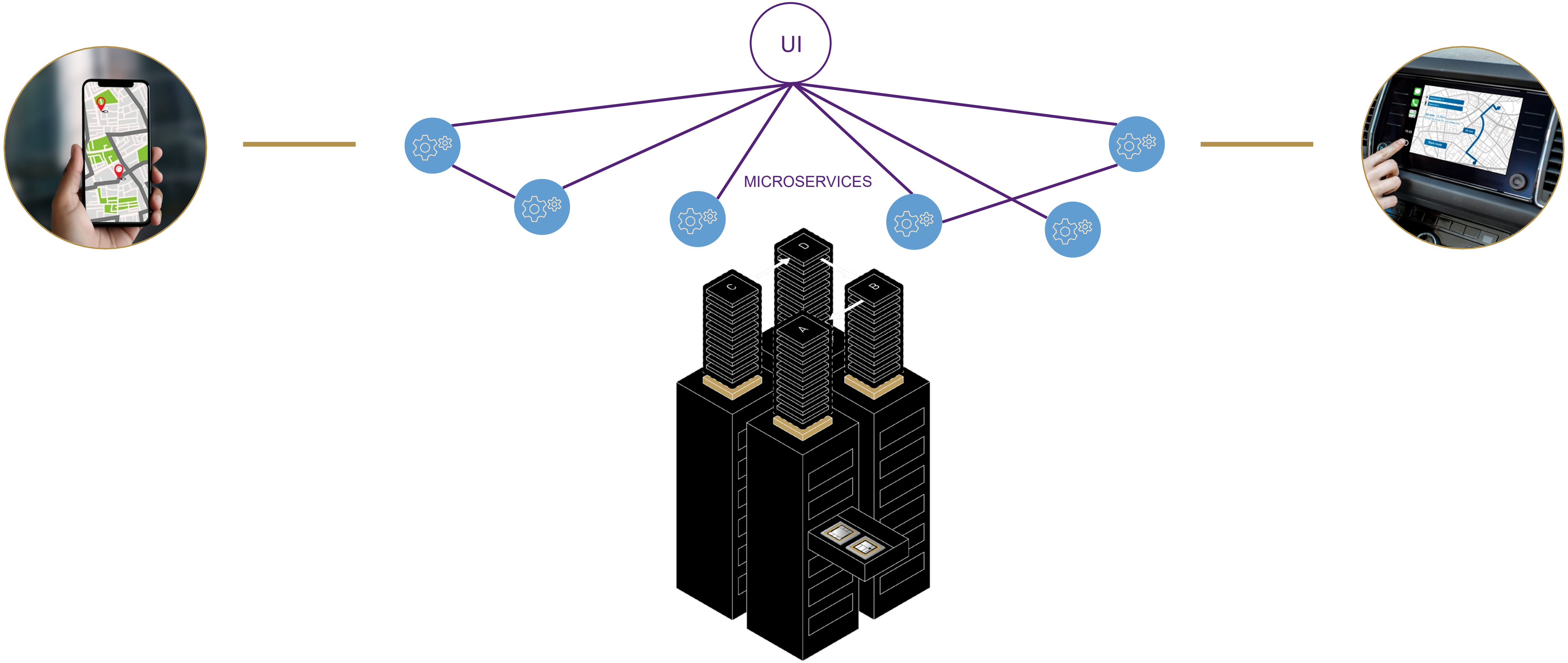


Democratization of AI



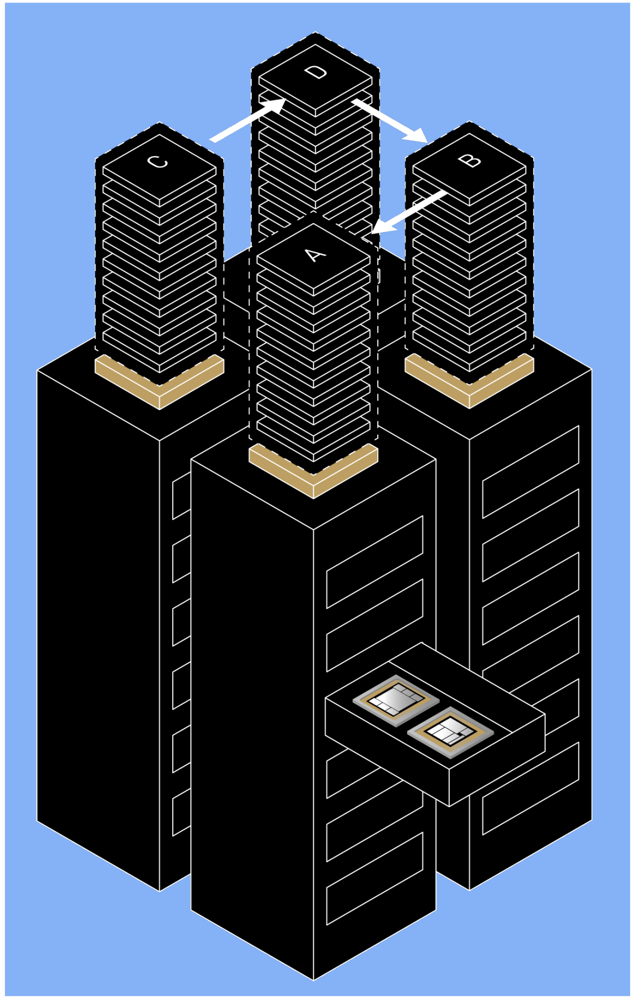
Data Center Is the New Unit of Computing

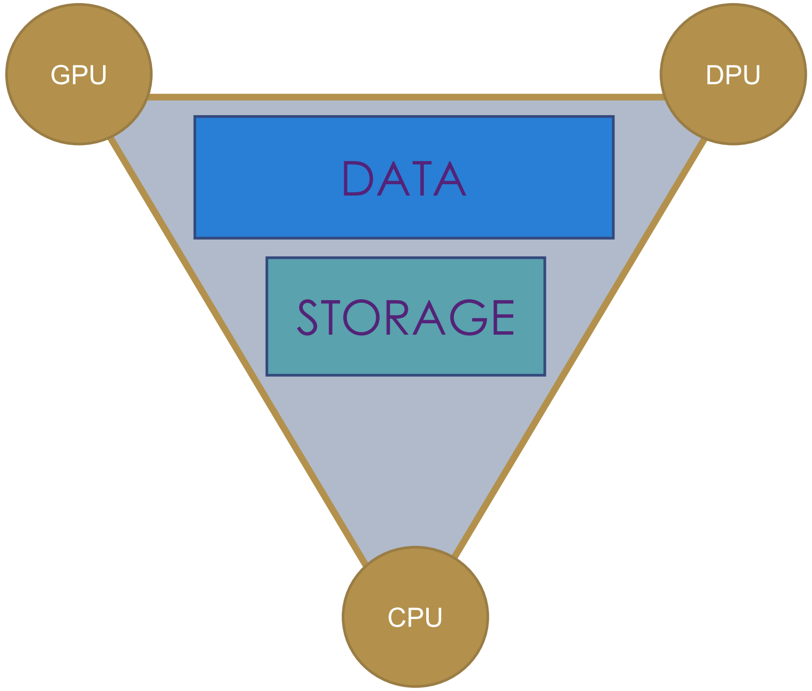
Distributed Compute with Scale-Out Microservices



Powering the New Unit of Computing

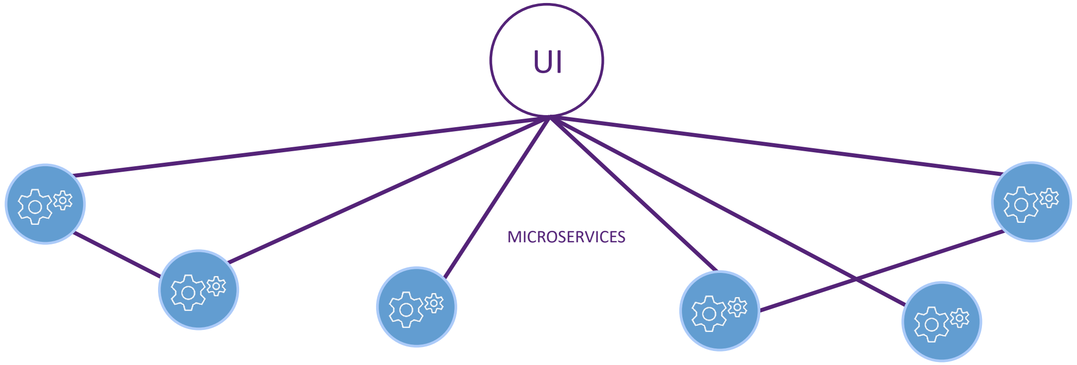
GPU, CPU and DPU are the Engines; Data is the Fuel



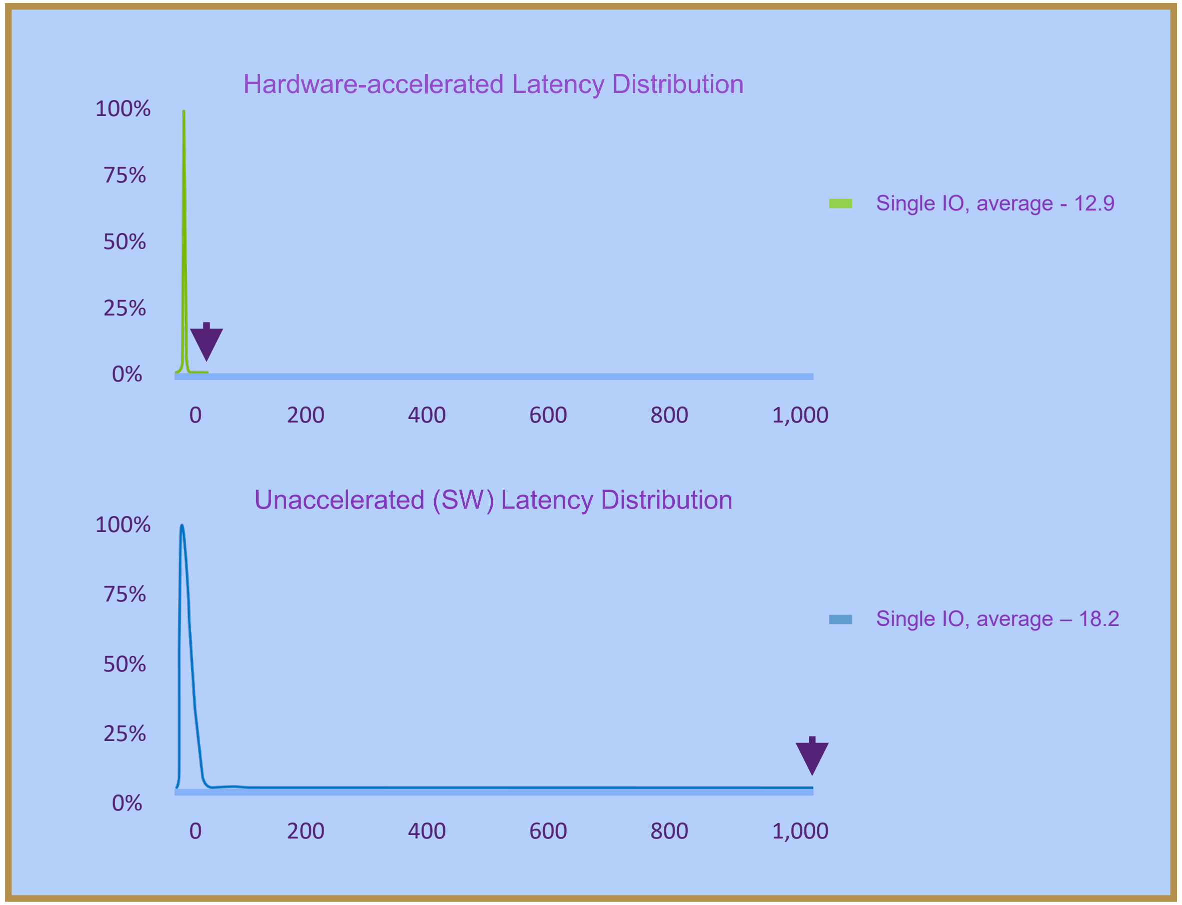


Latency Challenge for Distributed Computing

Average Latencies are Misleading

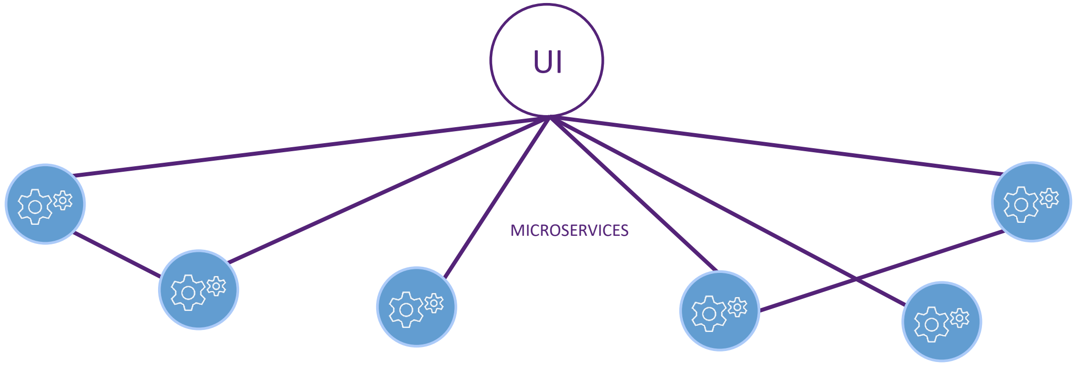


Storage Tail Latency Can be Much Higher than Average Latency When Microservices and Infrastructure Tasks Run Entirely in Software

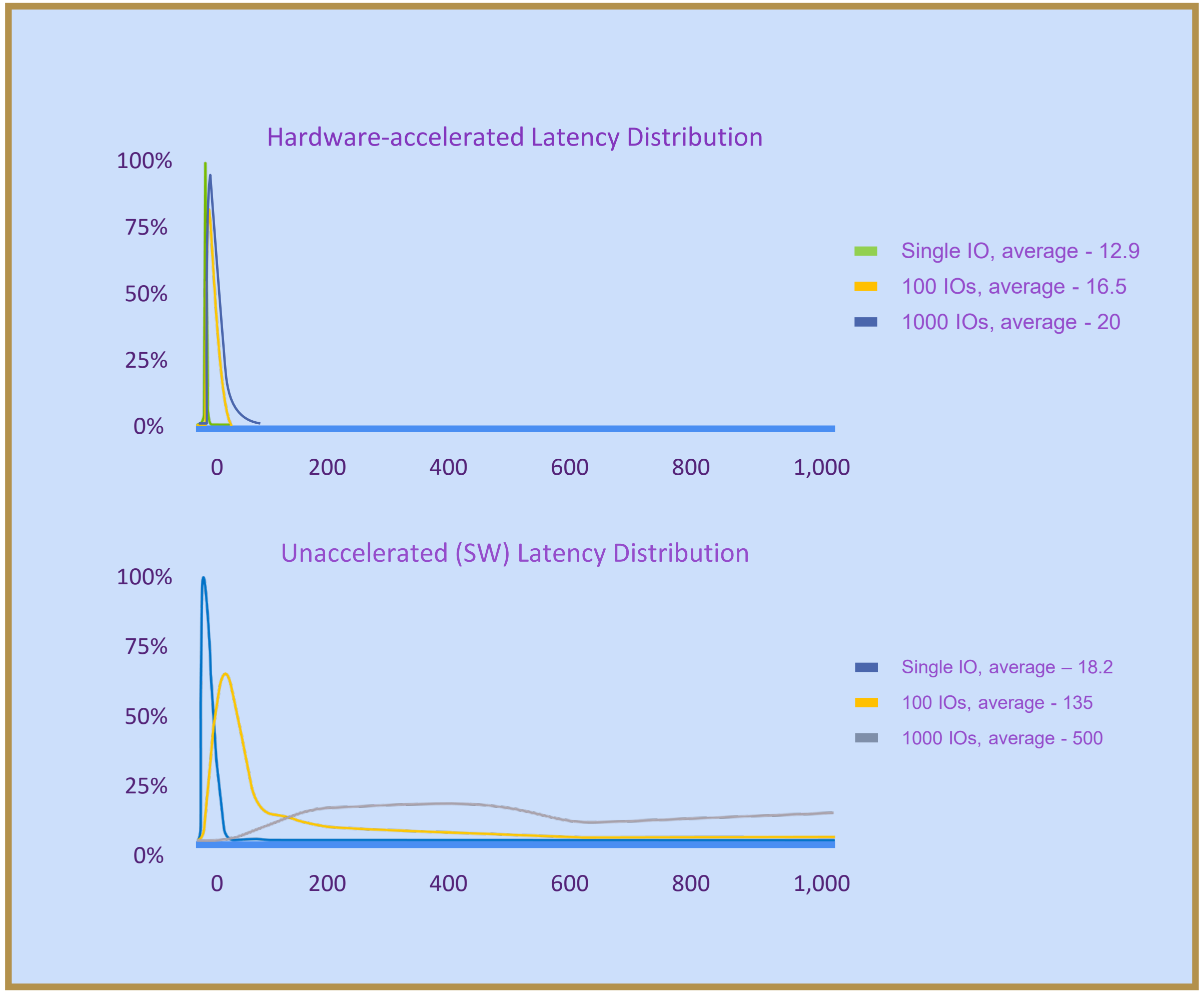


The Tail at Scale

Distributed Application Response Time Limited by Tail Latency of the Slowest Microservice

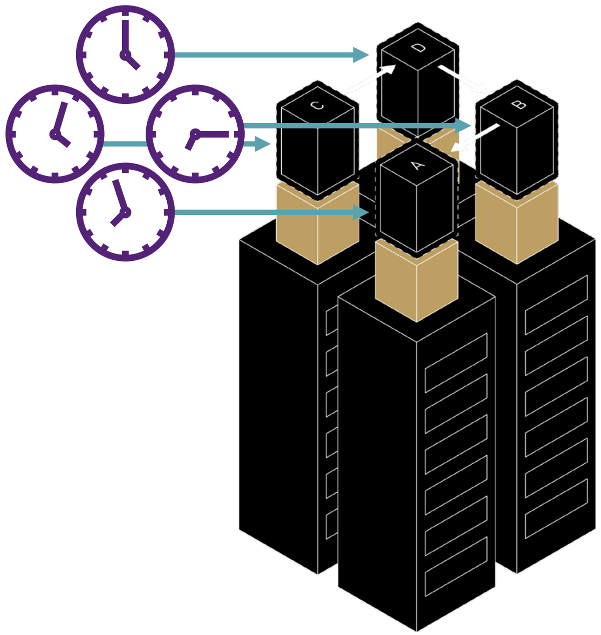


Hardware Accelerations Cuts Storage Tail Latency so Distributed Applications Can Meet SLAs



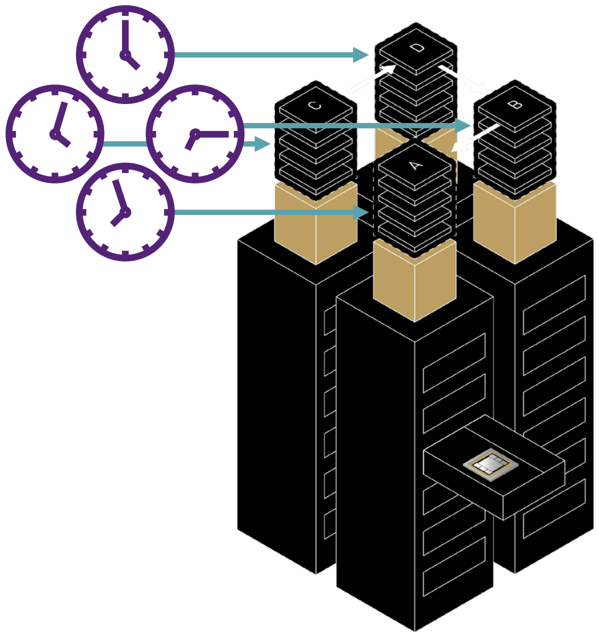
Accelerating Compute, Infrastructure, Storage

CPU



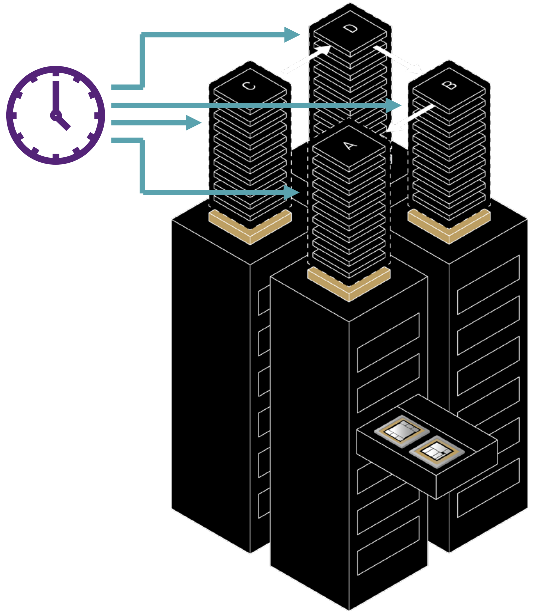
Disaggregated,
Microservices, Scaled Out

CPU + GPU



GPU-Accelerated
Computing

CPU + GPU + DPU + Faster Storage



GPU- and DPU-Accelerated
Data Center Also Needs Accelerated
Storage Access

Must Bring Compute and Data Closer Together

AI Challenges

- More data, more complex models
- AI growth, End of Moore’s Law
- Hungry CPUs and GPUs
- Tail latency, Distributed compute

Data Solutions

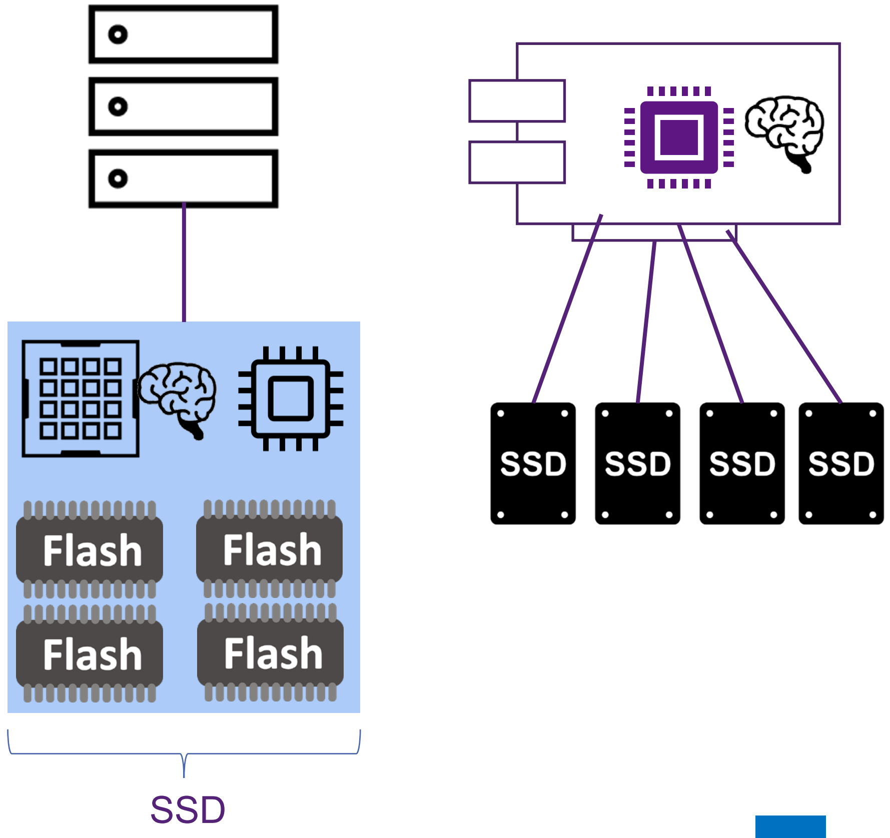
- Computational storage devices
- Smarter storage controllers
- Faster interconnects
- In-network computing

Computational Storage Devices

- Storage device with embedded compute
- Move results instead of moving data
- Details in earlier sessions

Energy-efficient Arm CPUs make it easier to distribute compute

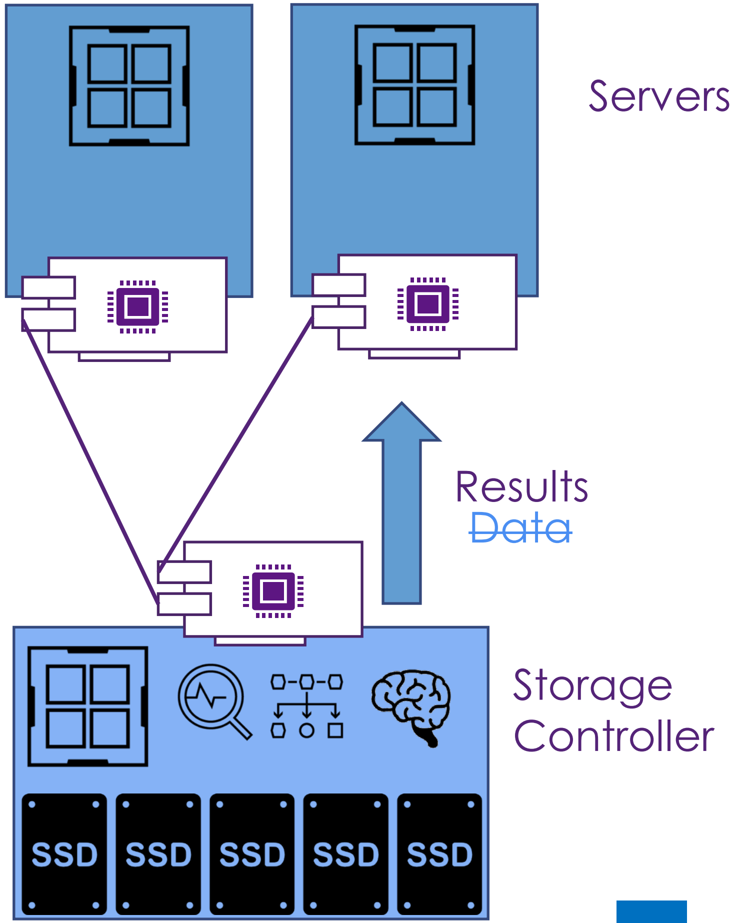
DPU can control SSDs directly



Adding Intelligence to Storage Controllers

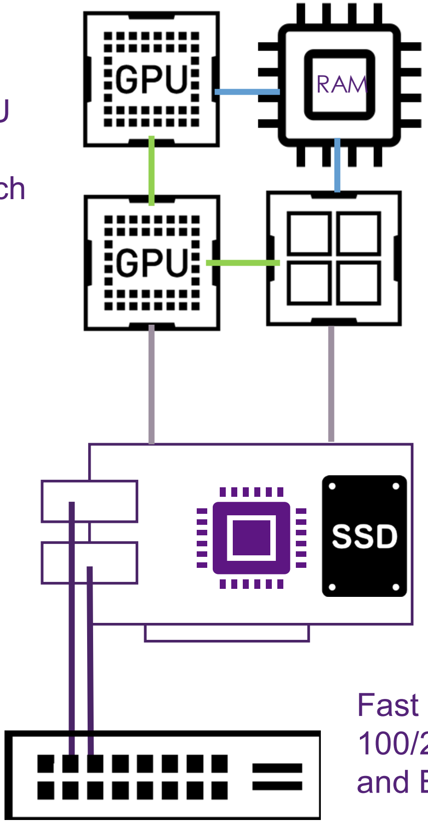
- Storage controllers perform application compute
  - Not just storage task compute
  - Return results/analysis, not just data
- It moves compute closer to storage
  - Distributing computing to the storage
  - Can debate if it's truly computational storage
  - Solves part of the distributed AI problem

Data Processing Units (DPUs) add compute to storage controllers



Faster Interconnects Remove Data Bottlenecks

- CPU-to-GPU, GPU-to-GPU
- Memory
- Peripherals
- Networking
- Converged hardware
  - CPU + GPU in one package
  - DPU + Flash in one card



Faster GPU/CPU interconnects: NVLInk, NVSwitch

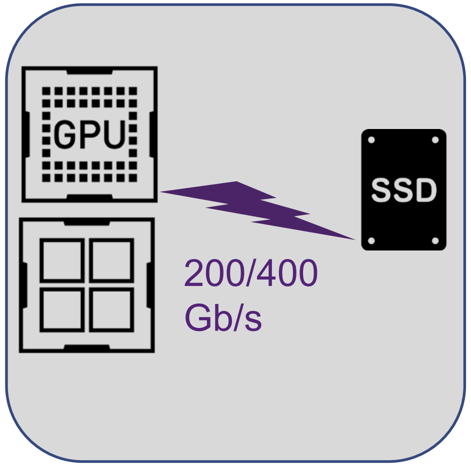
Fast memory interconnects: HBM2, DDR5, LPDDR5

Fast peripheral interconnects: PCIe Gen4, PCIe Gen5, CXL

Fast network interconnects: 100/200/400 Gb/s InfiniBand and Ethernet

Faster connections from compute to storage (or persistent memory) create distributed proxy for computational storage

In-Network Computing

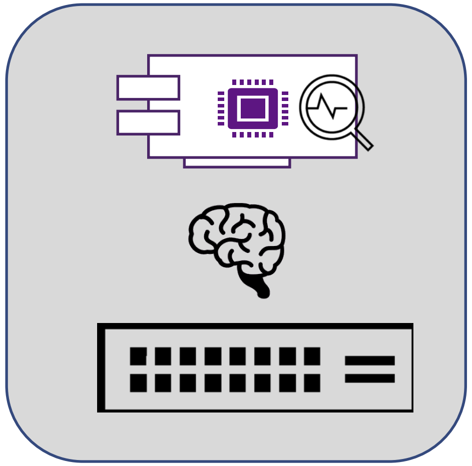


Network moves data faster

Faster data access for CPUs, GPUs and storage

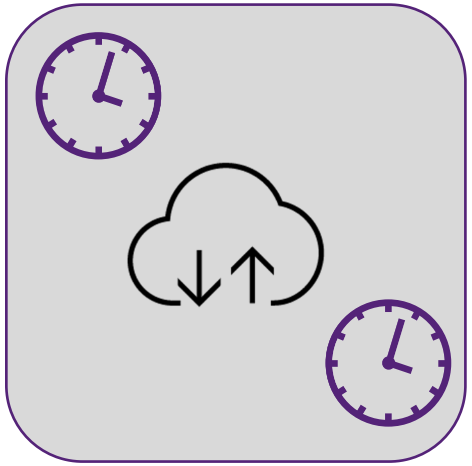
DPU accelerates infrastructure

Improved efficiency and security with reduced tail latency



Network Performs Compute

Faster AI / ML / HPC problem solving; reduced tail latency



Time-Synchronized Datacenter

Faster distributed compute; improved telemetry & security

Key Takeaways

- AI is everywhere and requires more data
- Data growing faster than CPU power
- New data center distributes compute
- Reducing tail latency becomes critical
- Move compute and storage closer together
  - Add compute to storage devices or controllers
  - Accelerate storage access
  - Compute in the storage network

Distributed application performance

Tail latency

Thank You

Please visit www.snia.org/pm-summit for presentations

Title and Abstract

- Date: Thursday April 22, 11:15-11:45am
- Title: Why Distributed AI Needs Computational Storage
- Abstract: Artificial Intelligence is increasingly being used in every type of business and industry vertical including finance, telco, healthcare, manufacturing, automotive, and retail. The nature of AI is becoming distributed across multiple nodes in the data center but also across the cloud and edge. Traditional local and networked storage solutions often struggle to meet the needs of AI running on many different types of devices in various locations. Computational storage can solve the challenge of data locality for distributed AI. These solutions include smart storage devices, adding data processing to storage arrays, and deploying new types of compute processors in the storage, next to the storage, or even in the storage network.