AI at the edge for Mobile & Wireless Communications

A Closed Loop Revolution of AIoT

Karl J. Weaver

APAC Biz Dev Director

Newport Technologies

www.Newport-technologies.com

Newport Technologies

美國新港科技

Edge AI Chips: The Quiet Revolution in Semiconductors

What's Driving Edge AI?

- **Latency** & **bandwidth**: Cloud-based AI can't respond fast enough for real-time decisions.
- **Power constraints**: Battery-powered devices demand efficient, low-heat compute.
- **Data privacy** & **compliance**: Processing on-device improves control and sovereignty.

As a result, there is a surge in demand for low-power, high-performance AI accelerators specifically for the edge.

Architectural Shifts in Edge Compute

Unlike datacenter chips, edge AI hardware focuses on:

- **On-device inference** with minimal latency
- **Efficiency** over peak FLOPs
- **In-memory compute**, **analog architectures**, and **#neuromorphic models**
- Support for **event-based vision** and **real-time response**

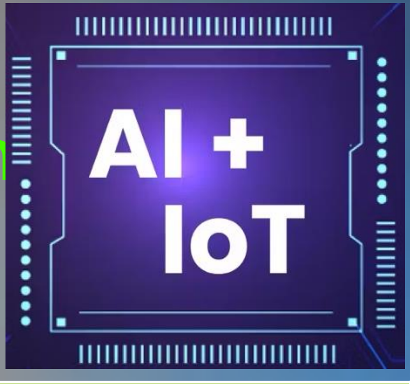
This isn't about training GPT-scale models. It's about perception, control, and autonomy at the edge.

What's Next for Edge AI?

Edge AI won't replace the datacenter; it will extend intelligence everywhere else. The next frontier is:

- Robust toolchains and deployment stacks
- Seamless edge-cloud integration
- Striking the right balance between vertical integration and open ecosystems

AI + IoT = AIoT - Building smarter devices with AI-enabled systems for real-time data analysis and decision making



From smart cameras and autonomous vehicles to industrial robots and wearables. AI is moving closer to where data is generated, fostering a new breed of highly efficient, task-specific chips are emerging

smart homes

Healthcare

eCommerce

manufacturing

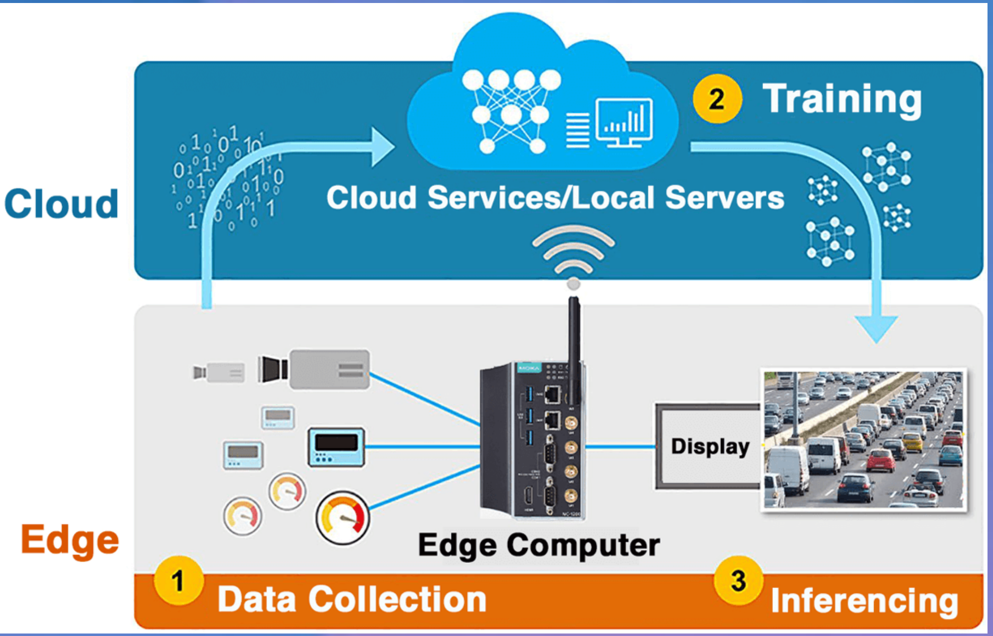
transportation

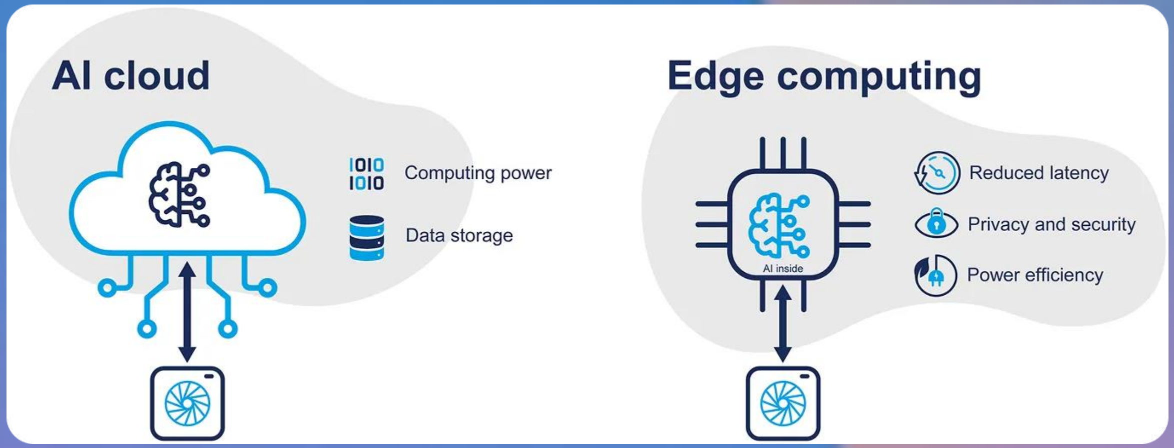
IOT system components	Key Description	Examples
IoT Devices	IoT devices are mostly used to collect large volumes of data Users can give instructions to IoT devices to perform an action	Sensors, Wearables, Smart Appliances
Connectivity	IoT devices are usually connected to the Internet using multiple communications protocols The gathered data is transmitted to other devices or cloud platforms	Bluetooth, WiFi
Data Processing	IoT Devices are also used for cleaning & Data aggregation tasks Edge devices are also used for data analytics and minimizing latency	Statistical Analysis, Deep Learning Model
Cloud Infrastructure	Data collection from IoT Devices are stored and organized in a database	Cloud Servers, Databases

The Types of AIoT - Open and Closed Loop

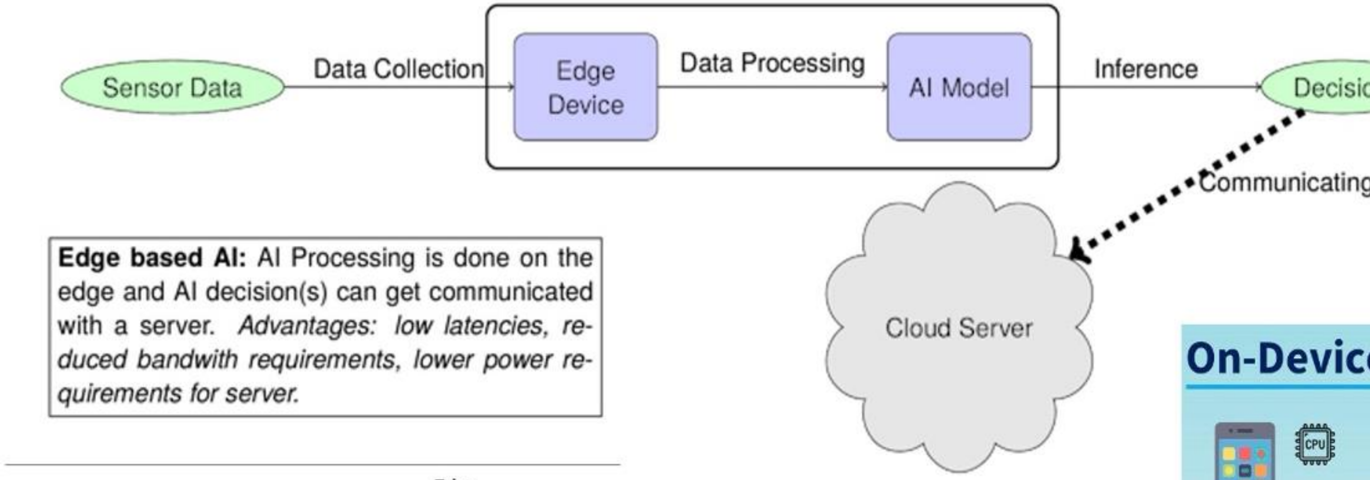
Open-Loop:
Data is transmitted from sensors on devices to the cloud, by a human or a cloud-based system that makes a decision.

Closed-loop (AIoT Revolution): On-device AI enables an "on-site perception -> real-time decision -> intelligent service" cycle. This decentralization of decision-making to the edge reduces latency, improves responsiveness, and allows the system to operate with a high degree of independence.

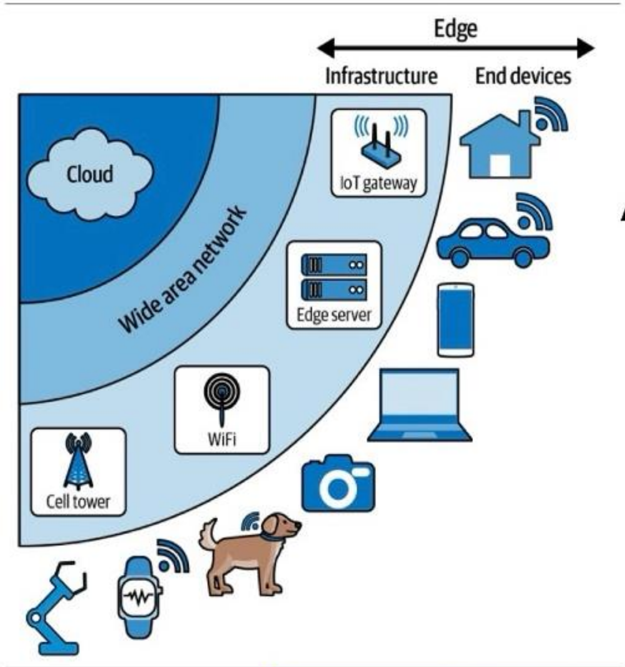




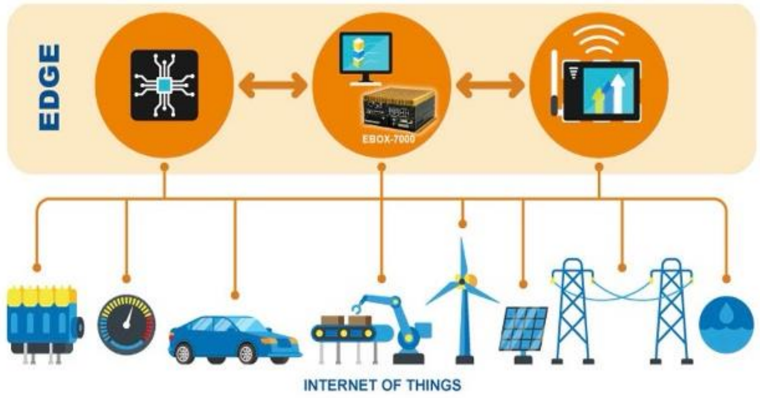
Inference is the future of AI at the edge



1. Real-Time AI Processing (Low Latency)
2. Reduced Cloud Costs & Bandwidth Usage
3. Improved Privacy & Data Security
4. Energy Efficiency & Sustainability
5. Growth of AI-Powered Edge Devices (IoT, 5G, & DePIN)



AI @ the Edge: Smartphones & IoT Devices



✅ **Real-time AI** → No latency, instant decisions.

✅ **Lower costs** → No expensive cloud inference fees.

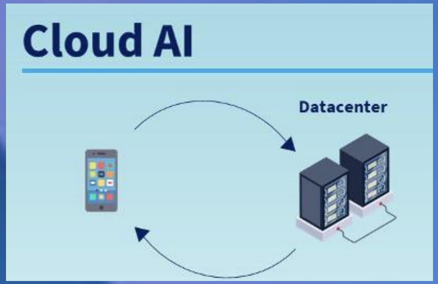
✅ **Better privacy** → Data stays on-device.

✅ **Energy-efficient AI** → Longer battery life for IoT & mobile AI devices.

✅ **Scalability** → AI-powered edge devices will drive **the next wave of computing**.

Source: Feb. 2025 Karl Weaver – AI Chip Supply Chain in Asia-Pacific & GenAI Smartphone Revolution

Cloud AI vs 'AI-at-the-edge'

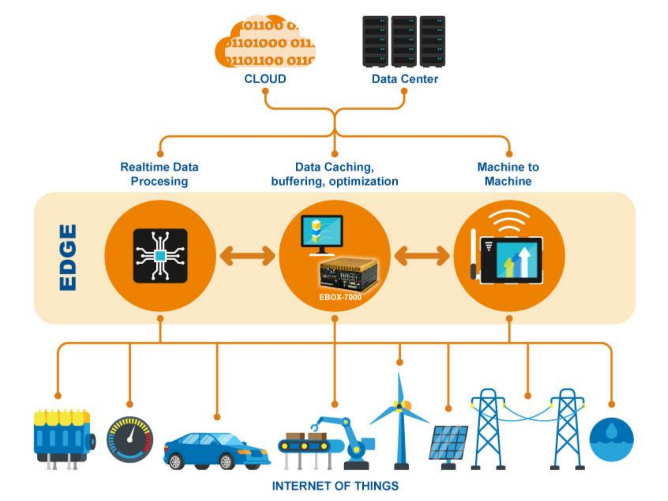


Types of Ai Chips
Graphics Processing Units (GPUs)
Field Programmable Gate Arrays (FPGAs)
Application-Specific Integrated Circuits (ASICs)
Tensor Processing Units (TPUs)

Key AI Chip components: Memory (dram) NPU, GPU and CPU

- Compute occurs in a server; results are sent back.
- Slower Response time
- More privacy concerns, data needs to travel vast distances
- More battery consumption, but more powerful,
- No need for connectivity

- Compute takes place in the device's hardware
- Faster response time
- Better for privacy, Data stays in the device
- Less battery used; less powerful processing,
- Needs connectivity



AI Chip ecosystem, who does what?

Ecosystem Role	Key Players	Core Contribution
Designer (Fabless)	Nvidia, Apple, AMD, Qualcomm	Chip architecture (product IP)
Manufacturer (Foundry)	TSMC, Samsung Foundry	Chipset fabrication (process IP)
Architect (IP Provider)	ARM	Licensed CPU/GPU cores, high-speed interfaces, standard cells
Toolmaker (EDA)	Synopsys, Cadence, Siemens EDA	Design, simulation, & verification software

The Steps to Transform IoT into AIoT

1. Data collection through IoT sensors and devices

2. Data transmission to cloud or edge computing systems

Data from IoT devices is sent to edge or cloud computing systems to enable edge intelligence. The former enables faster real-time data analysis,

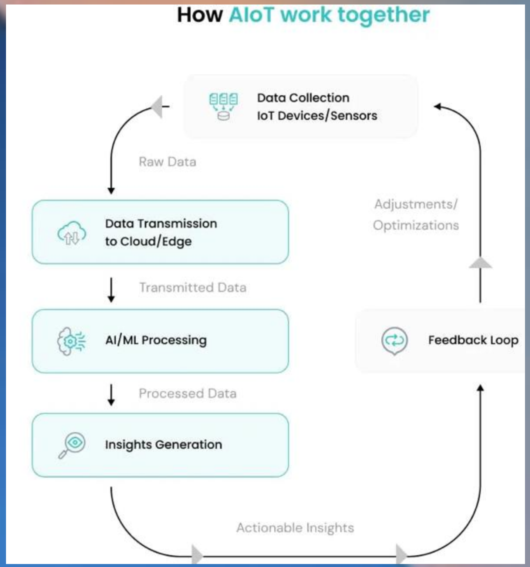
3. AI/ML algorithms processing the data

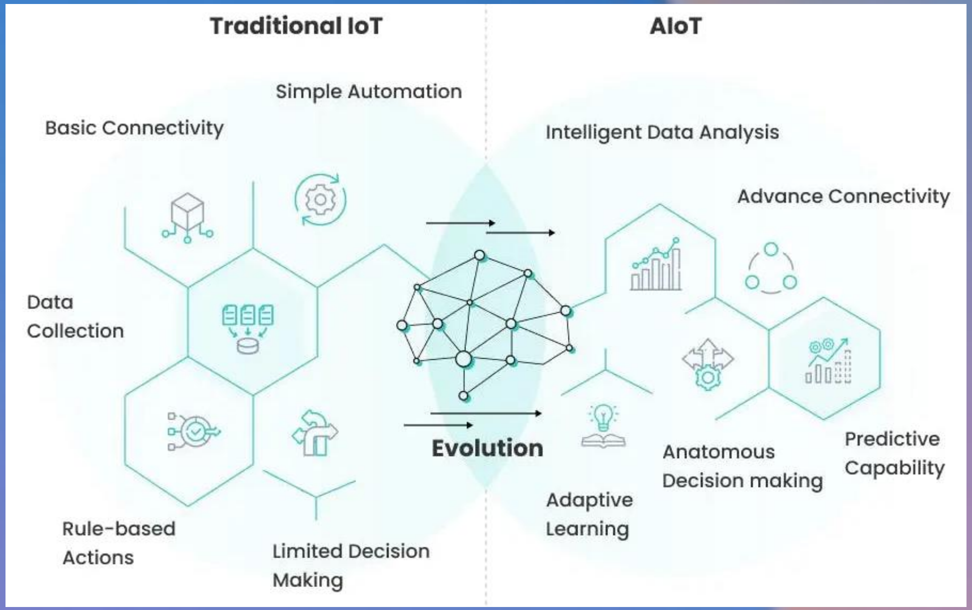
4. AI using the data to deliver insights and recommendations

Your AI systems deliver actionable recommendations and insights

5. Feedback loop for continuous improvement

designed to learn continuously from the outcomes of its recommendations and insights, becomes more robust,





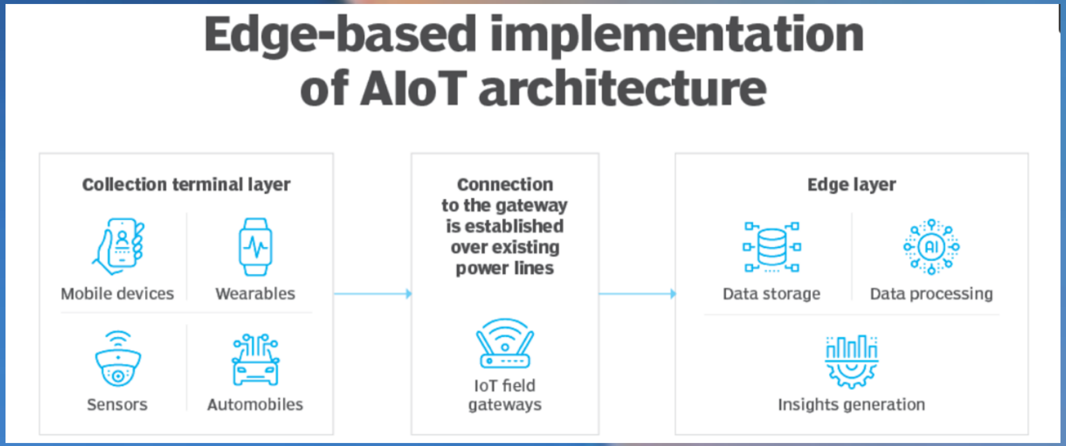
Edge-based AIoT Architecture

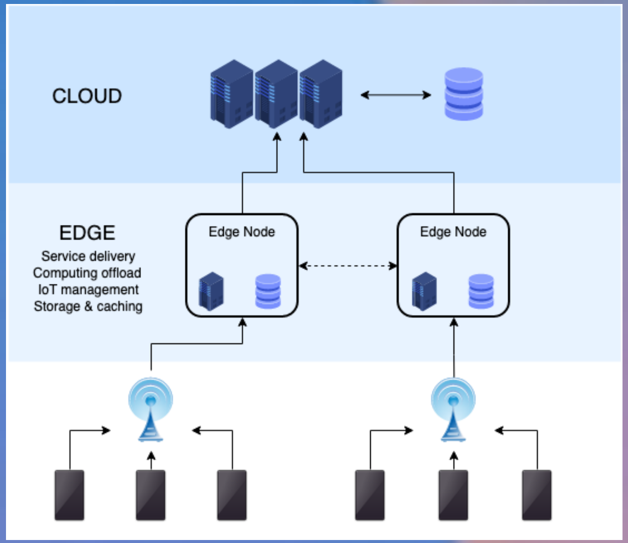
Data from IoT devices is processed as close to the devices as possible to minimize bandwidth needed to move data while avoiding delays in data analysis = AIoT

Edge-based AIoT consists of the following three layers:

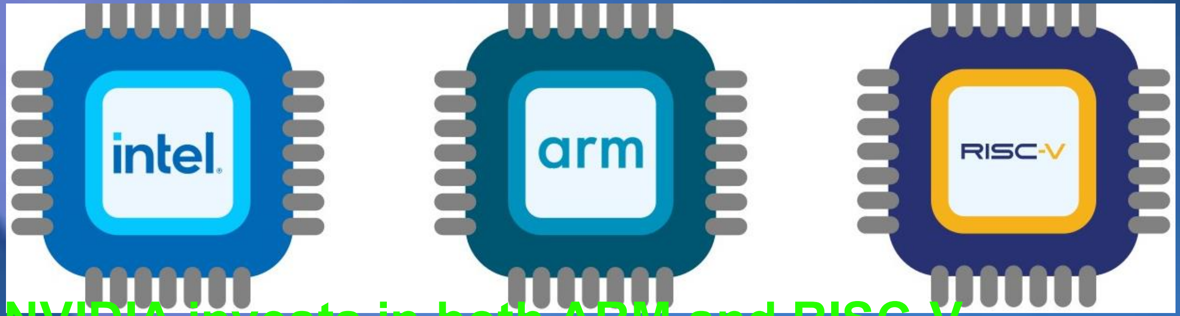
1. **Collection terminal layer.** This covers a range of hardware devices, such as embedded devices, cars, manufacturing equipment, tags, beacons, sensors, mobility devices, and health and fitness equipment, that are connected to the gateway over existing power lines.
2. **Connectivity layer.** This consists of the field gateways that the collection terminal layer is connected to over existing power lines.
3. **Edge layer.** This layer includes facilities for data storage, data processing and insight generation.

Figure 2 depicts an edge-based AIoT implementation.





Competition in the processor architecture market



- ✓ NVIDIA invests in both ARM and RISC-V
- ✓ **Apple's M-series processors (Laptops)** are based on **ARM**
- ✓ **RISC-V** has **customizability** and **open-source innovation;** RISC-V is gaining popularity as a viable alternative to ARM and x86, using **an open-standard ISA**
- ✓ ARM's **ARMv9 architecture** uses **Confidential Computing**, **a hardware-based security** for cloud security.
- ✓ **Chinese semiconductor industry** has adopted RISC-V to reduce dependency on Western IP.
- ✓ **NVIDIA** will use some **RISC-V cores** for its future **GPU architectures,** testing a flexible ISA, but its main GPU compute cores will remain proprietary. ISA is important

➢ **x86**: A complex instruction set computer (CISC) architecture developed by Intel, primarily used in most personal computers and servers.

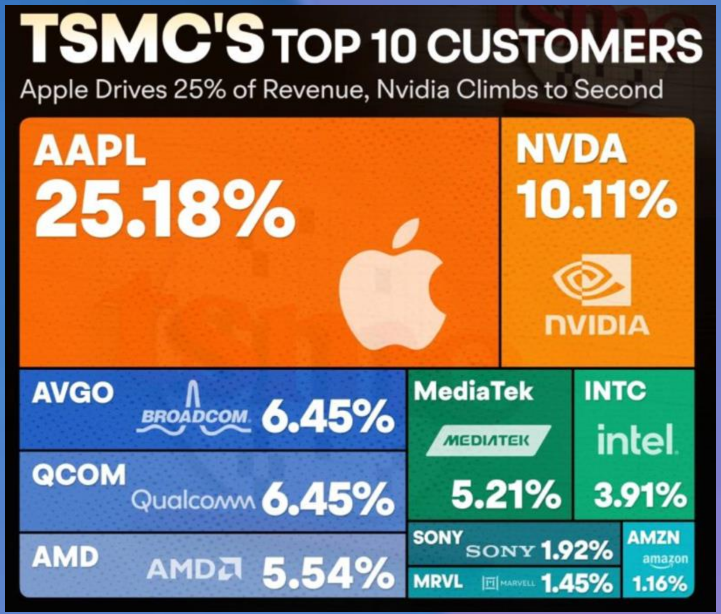
➢ **Arm**: A reduced instruction set computer (RISC) architecture known for power efficiency, dominating the mobile device market and increasingly used in servers and laptops.

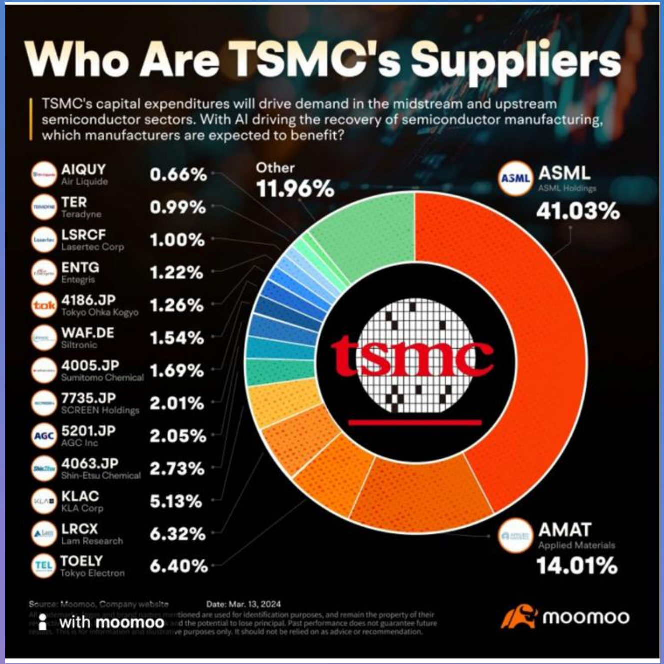
➢ **RISC-V**: An open-source RISC instruction set architecture that allows for full customization and is gaining traction as a royalty-free alternative to proprietary designs.

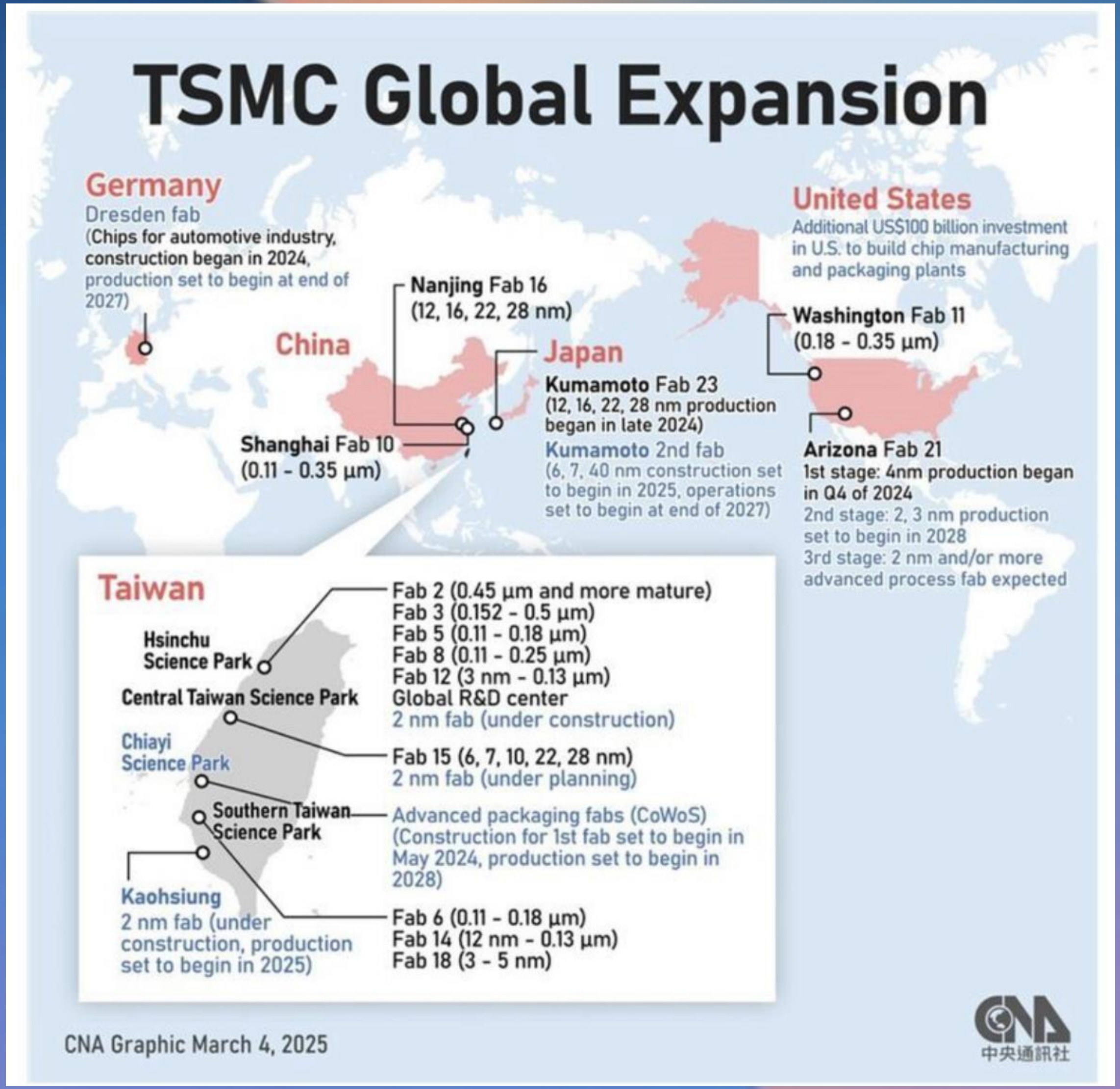
Types of AI Hardware

1) GPUs – for deep learning & neural network computation for LLMs
2) TPUs - designed to accelerate machine learning tasks, developed by Google,
3) AI Server – Equipped with high-performance computing chips to support large-scale AI model training and Inference
4) ASIC Chips – tailored specifically for AI applications, such as algorithm acceleration
5) FPGAs – reconfigurable hardware chips used to accelerate AI Inference and training
6) – NPUs – Designed specifically to accelerate neural network

TSMC - Produces 92% of all AI chips in Taiwan & Phoenix





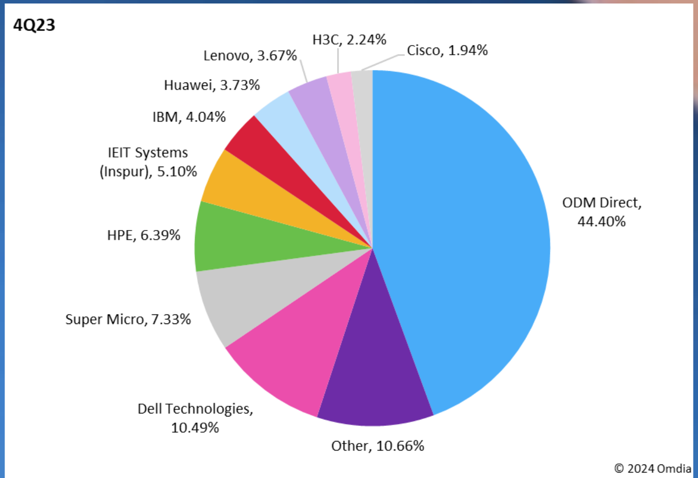


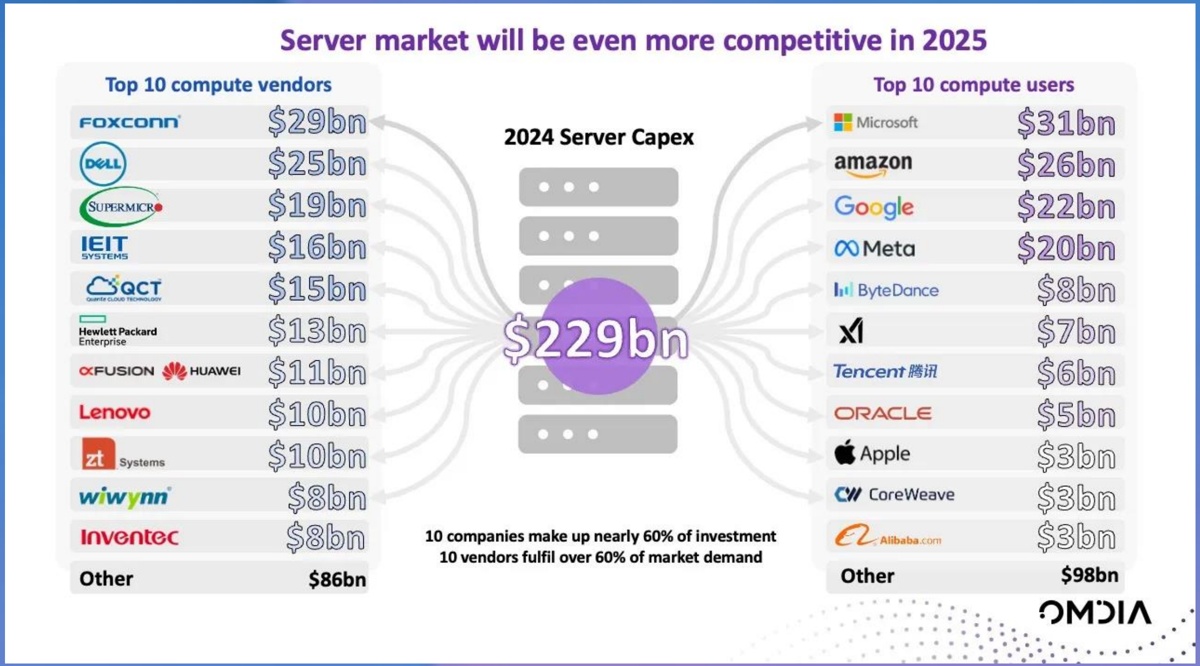
Taiwan ODMs design 90% of Servers for Data Centers

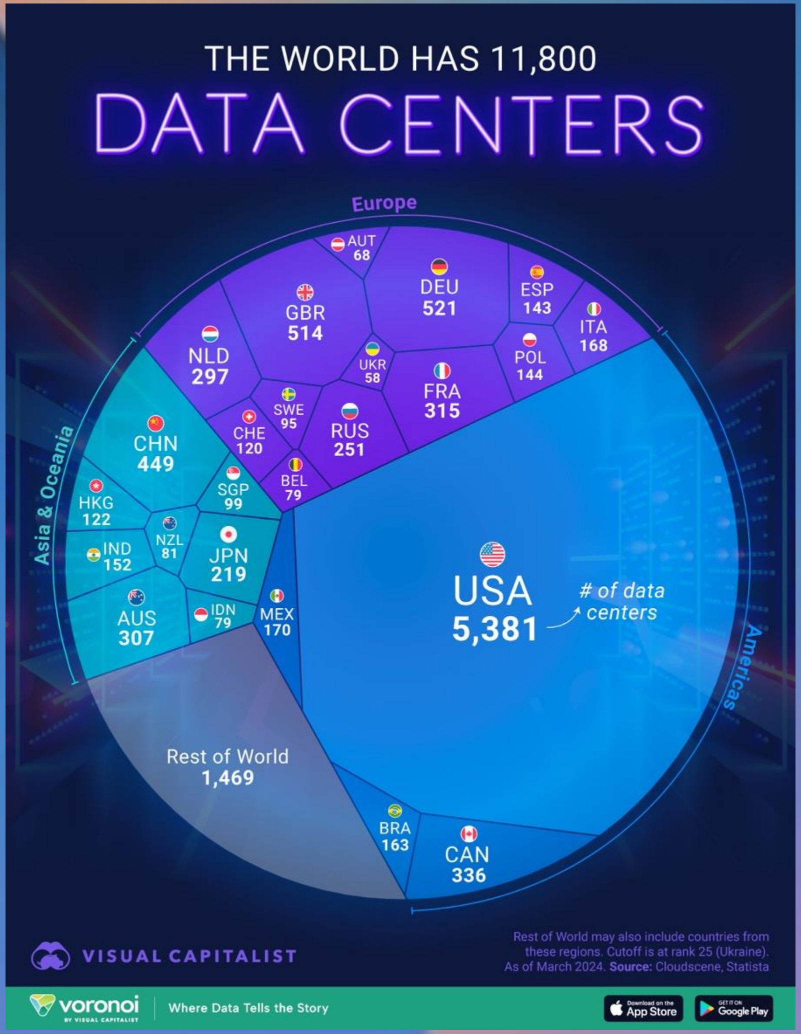
Taiwan-Based Server ODM's L10 and L6 Production Bases and Future Layout

Server ODM	Assembly House (L10)	SMT Line (L6)	SMT Line Being Planned
Quanta	US (California, Tennessee), Germany	Shanghai, Taiwan, Mexico	Thailand
Inventec	Czech Republic, Mexico, Shanghai	Shanghai, Taiwan, Thailand	Thailand
Mitac	Foshan, Czech Republic, Mexico	Foshan, Taiwan	Southeast Asia
Foxconn	Czech Republic, Mexico, Wisconsin	Tianjin, Taiwan, Vietnam	Vietnam (Hanoi), US (Wisconsin)
Wistron/ Wiwynn	Czech Republic, Mexico, Malaysia	Zhongshan, Taiwan, Malaysia	Malaysia

Source: TrendForce, Aug., 2023



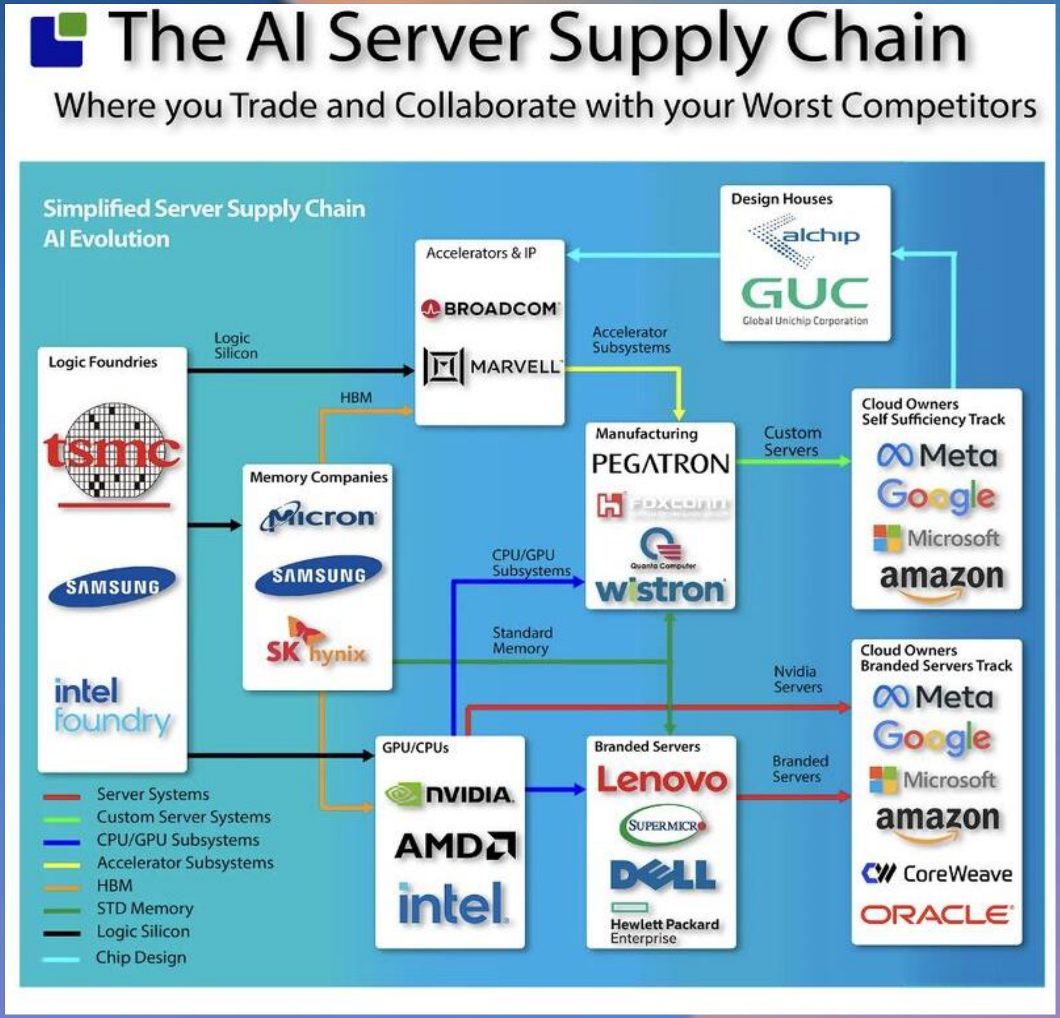




The AI Server Supply Chain

(Chips to Servers)

- Nvidia sells its own servers manufactured by the EMS/ODMs. It also sells its GPU subsystems to Server manufacturers, who rebrand them as their own servers.
- Nvidia sells GPU subsystems to end customers who use EMS/ODMs to create server systems that fit their needs.
- Largest server customers (Broadcom, Marvel, Qualcomm) design their own accelerator chips for different workloads, which the Nvidia servers excel at using Silicon from the logic foundries.
- Like the GPU/CPU subsystems, these accelerator subsystems include memory & networking components, which are installed into custom server systems controlled by the principal owner of the architecture.
- The best-known principals are: Amazon (Inferential, Trainium) 2) Google (TPU), Microsoft (Maia) 4) Meta (MTIA). Hyperscalers mostly!



Newport Technologies
美國新港科技

Executive Summary

Karl J. Weaver 魏卡爾 is a bilingual Mandarin Chinese/English-speaking business development and technology executive with more than three decades of Ex-Pat work experience to drive growth, partnerships, and market expansion across the global mobile, wireless, semiconductor, AI, IoT, Web3, cybersecurity, and satellite communications industries. He has built strategic relationships with Tier-1 OEMs, mobile network operators (MNOs), semiconductor vendors, cloud providers, and ecosystem partners throughout Asia-Pacific, North America, and Europe.