EmBestor

The Industrial Flash Storage Expert.

Multi-layered Data Storage Architecture for AI/ML Systems

By **Chanson Lin**

Email: Chanson.Lin@embestor.com

EmBestor Technology Inc.

http://www.embestor.com

FMS

the **Future** *of* **Memory** *and* **Storage**

OUTLINE

- Introductions: AI, AI Model, AI applications, industrial AIoT, efficiency, bottlenecks, …
- Computing with AI (AI Computing) vs. Memory and Storage (M&S).
- A novel interface bus for communicating between AI Computing with Memory and Storage.
- Speed and Responsitivity matching examples.
- An Example of Multi-layered Storage and System Application.
- Conclusions and Discussions.

Where / What is the AI?

Applications:

- Image Processing and Pattern Recognitions
- Medical Data Processing and Syndromes Identification.
- Robotics and Factory Automation:
- Autonomous Vehicles:
- …

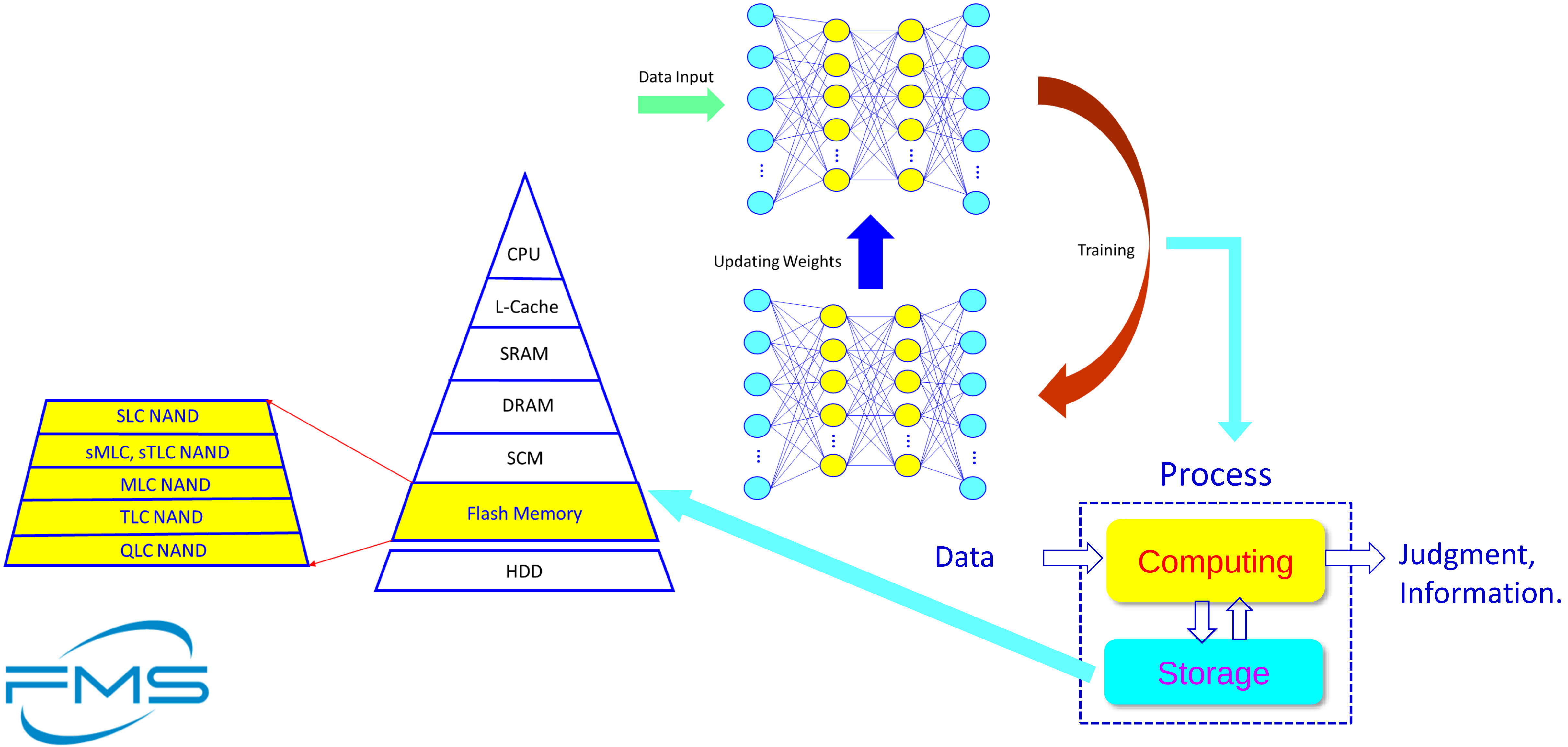
Locations:

- End-Points: simple and fast response.
- Edge: Local judgement.
- Cloud: Central Intelligent systems.
- …

Performance:

- Accuracy: 80%, 90%, 99%, …
- Speed (time): Latency, Throughput, …
- Transparency: The rule base principles.
- Scalability: Cost, Complexity, Efficiency, …
- …

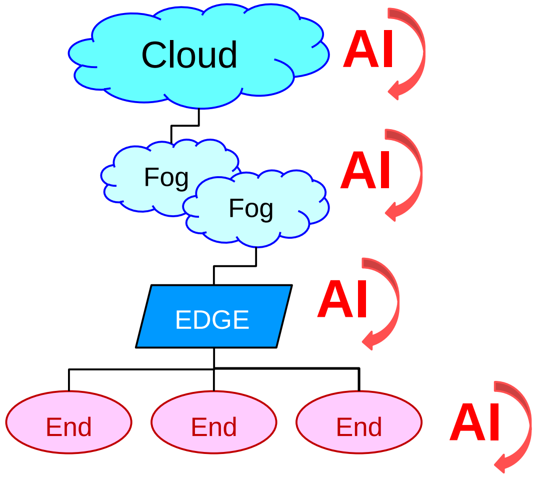
AI Model – the Storage & Memory for AI



AI Practical: Multi-structure

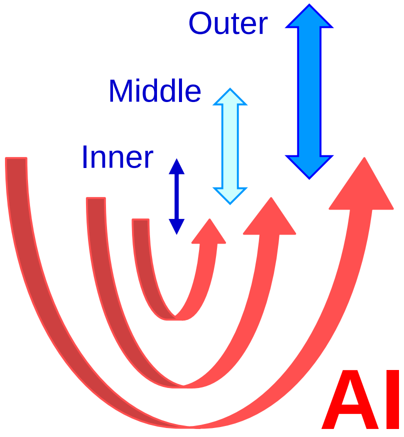
Multi-layer:

- AI on End-point:
- AI on Edge:
- AI on Fog:
- AI on Cloud:



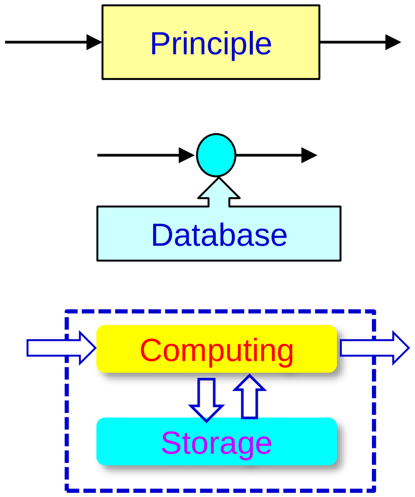
Multi-loop:

- Inner Loop: fastest response.
- Middle Loop: middle way
- Outer Loop: long / deep inferencing.



Multi-path:

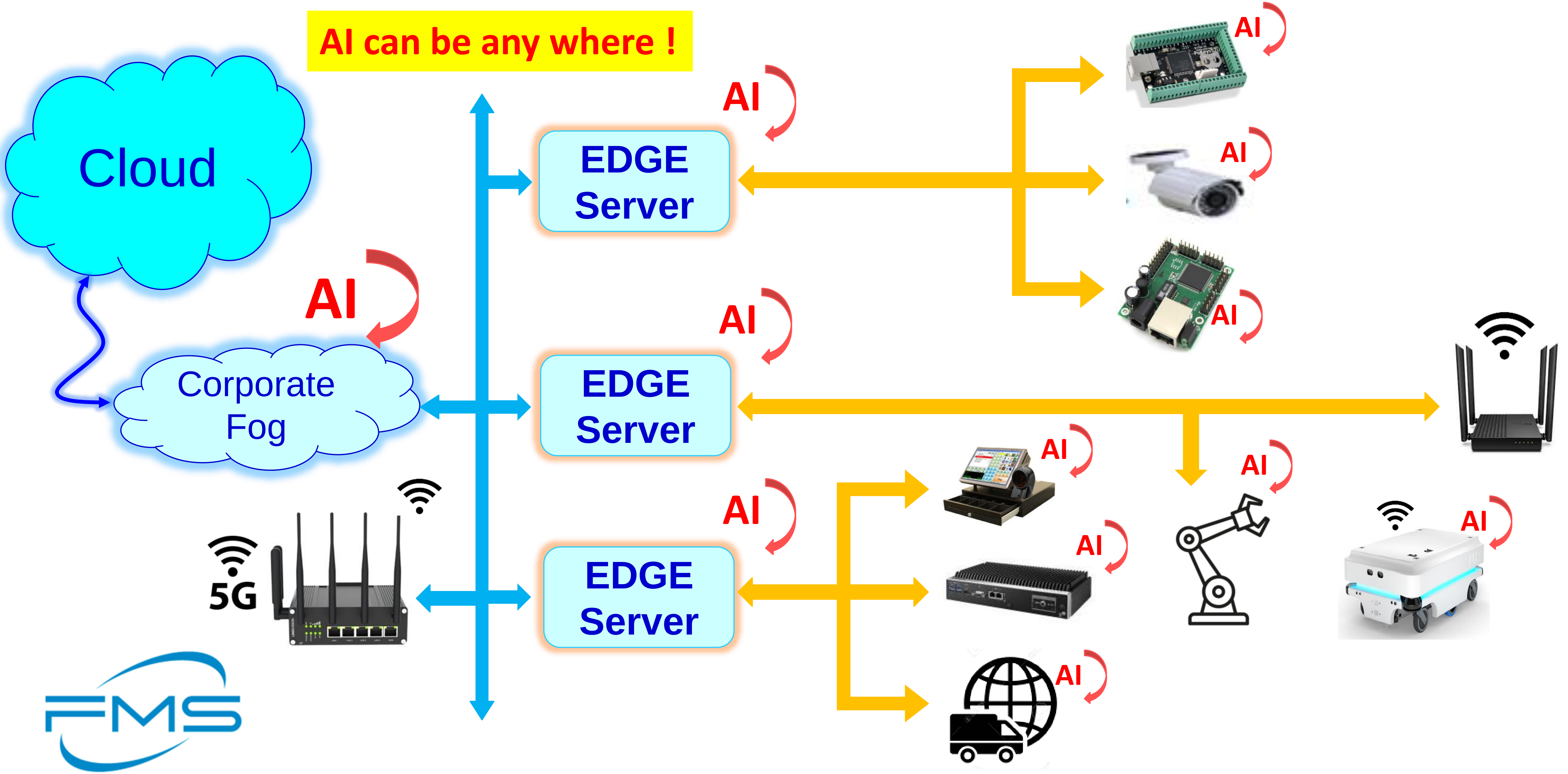
- Experience Principle:
- Expert database lookup:
- Simple Training:
- Machine Learning:
- Deep Learning:



EX: AI & IoT in a Smart Factory

EmBestor
The Industrial Flash Storage Expert.

AI can be any where !



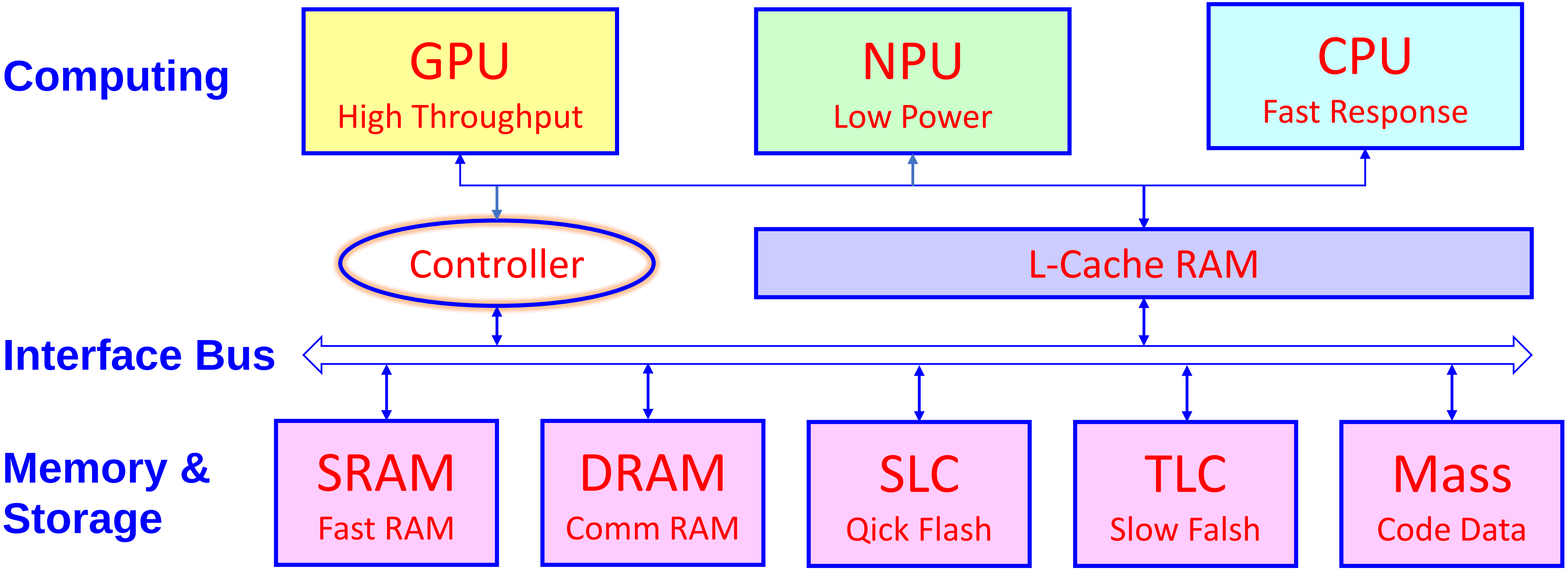
LLM and Multi-layered AI Model

- LLM (Large Language Model): LLMs can be used for text generation, a form of generative AI, by taking an input text and repeatedly predicting the next token or word.
- Some notable LLMs are OpenAI's GPT series of models (e.g., GPT-3.5, GPT-4 and GPT-4o; used in ChatGPT and Microsoft Copilot), Google's Gemini (the latter of which is currently used in the chatbot of the same name), Meta's LLaMA family of models, Anthropic's Claude models, OPT (Zhang et al., 2022b), PaLM (Chowdhery et al., 2022) and Mistral AI's models.
- Multi-layered AI Model:
  - To mimic human nural network and optimize the operation efficiency.
  - Grouping by the updating frequency of each node.
  - Grouping by the updating correlation of each node.

Bottlenecks and Inefficiencies

- **Good is enough**?: How fast and how accurate is good enough? there is no standard or analytical answes for users, …
- **Power consumptions**: Computation efficiency, Data transfer, Data amount, and iteration loops.
- **Material Cost**: Pursuing infinite computing power, memory speed and density, ….
- **Scope**: Universal or Specific, All-cover or Domain expert.

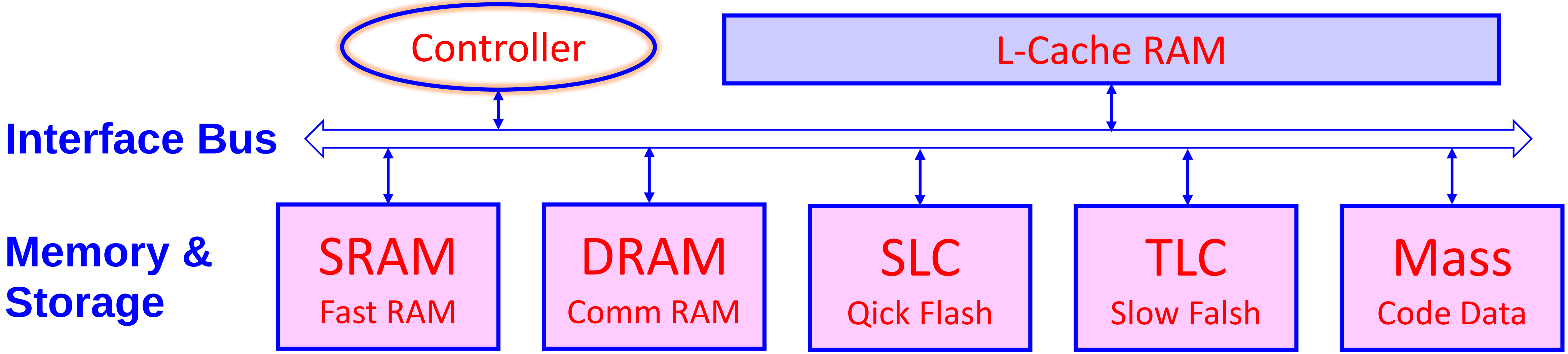
Computing vs. Memory & Storage



Interface Bus for Communicating

EmBestor

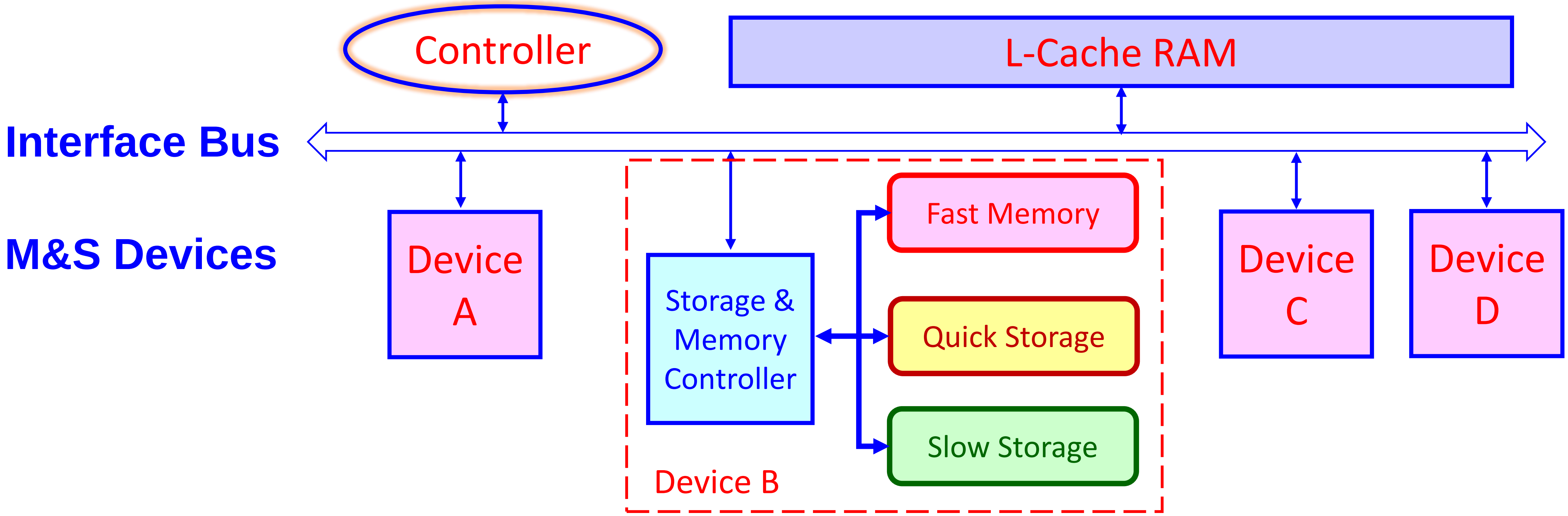
The Industrial Flash Storage Expert.



- A novel Interface bus with Respositivity garde as defining the Data I/O, Memory access commands and protocol.
- The bandwidth of the interface bus shall be high enough (e.g., PCIe Gen-5 or Gen-6) to provide highest response memory devices, like SRAM.
- An integrated Storage & Memory device can include the multiple storage media types to meet the requirement of the Multi-layer AI/ML system.

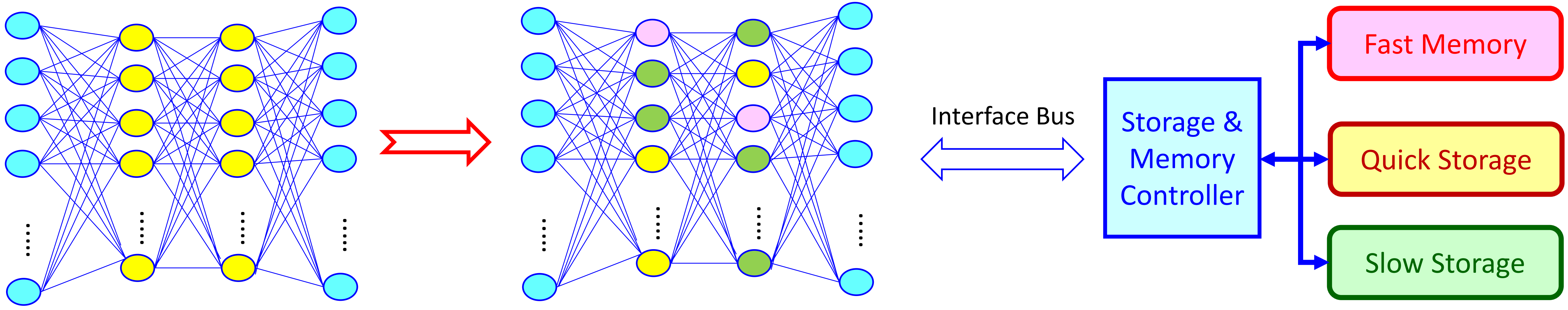
A Multi-layered Data Storage

EmBestor
The Industrial Flash Storage Expert.



- A Multi-layred Data Storage designed for the novel interface bus with respositivity garde.
- An example of 3 layered Data Storage device as: DRAM, SLC Flash, TLC Flash.

Updating Freq. Filtering/Ranking for CNN



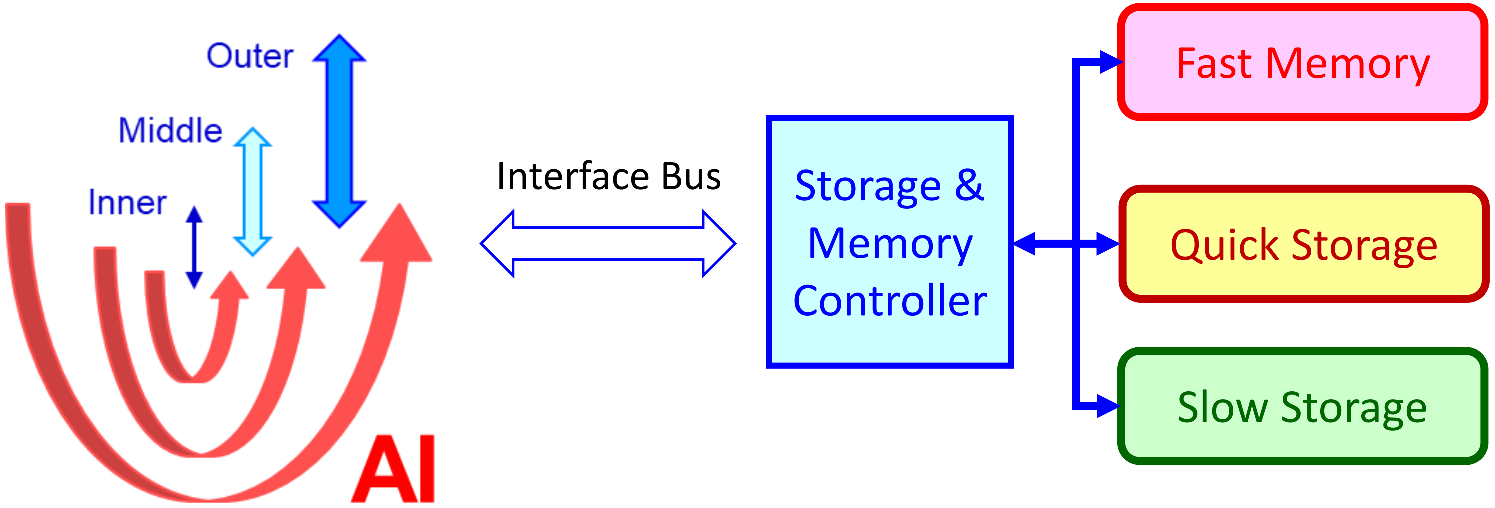
- CNN is Natural Sparsity (around 1~2%).
- Grouping by the updating frequency: Hot nodes or Cold nodes.
- Frequently updating nodes matching to Fast Memory.

Layered Matching with CNN Architecture

EmBestor
The Industrial Flash Storage Expert.

Multi-loop:

- Inner Loop: fastest response.
- Middle Loop: middle way
- Outer Loop: long / deep inferencing.



- CNN: different architecture for different application scenrio.
- Inner nets matching to fast-response memory. (Fast Memory)
- Middle nets matching to quick-response storage. (Quick Storage)
- Outer nets matching to long-latency storage. (Slow Storage)

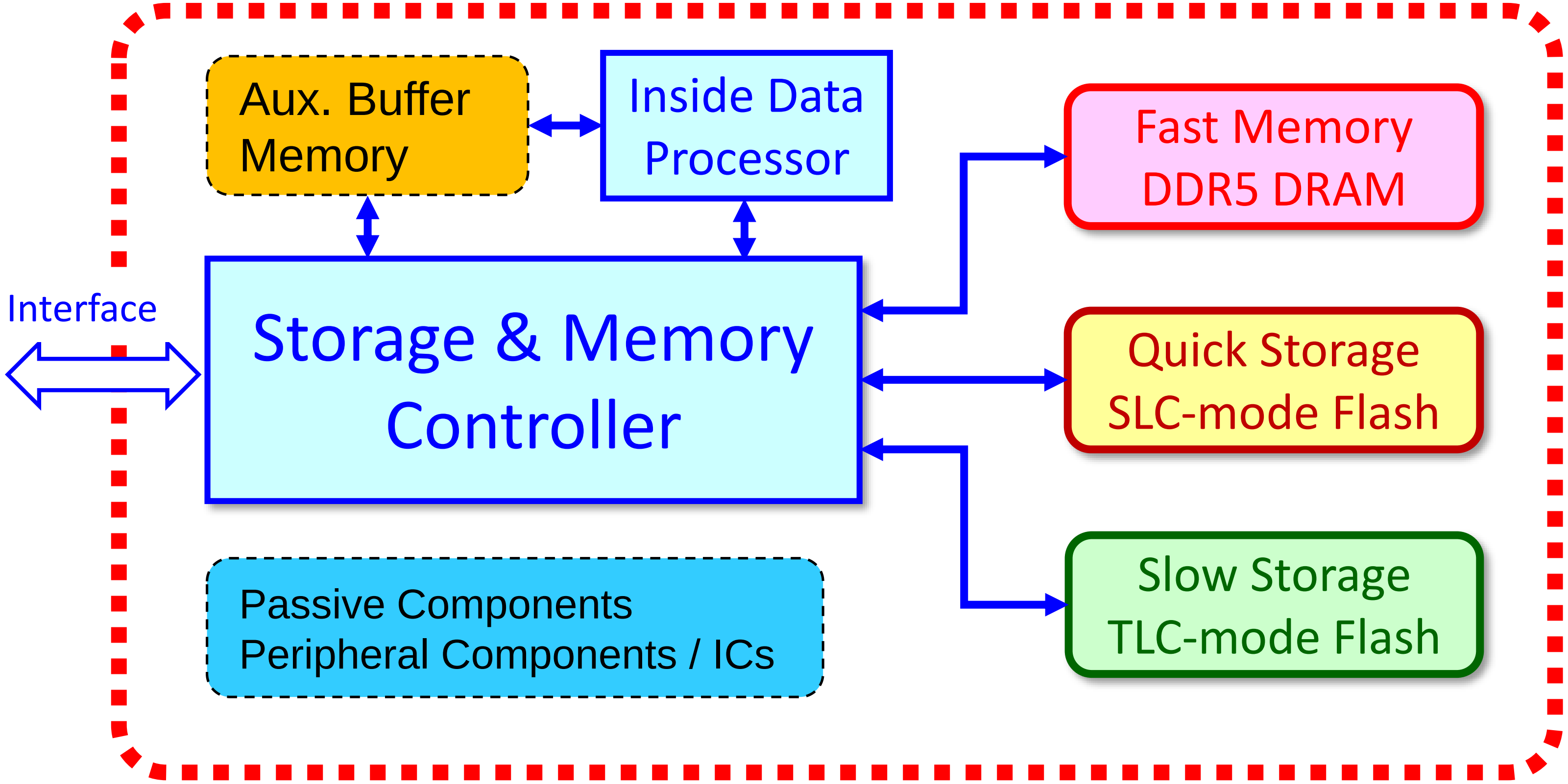
FMS

An Example of Multi-layered Storage

EmBestor
The Industrial Flash Storage Expert.

Host (AI Computing):

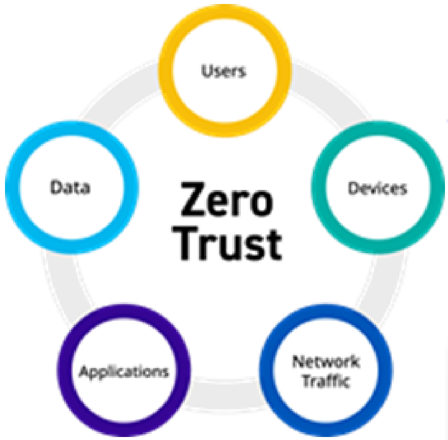
- NVMe/PCIe Gen 5 standard commands.
- Vendor commands with responsitivity ranks.
- Vendor commands supporting Multi-layered features and functions.
- Multiple Ports optional.
- SR-IOV (Single-Root I/O Virtualization) optional.

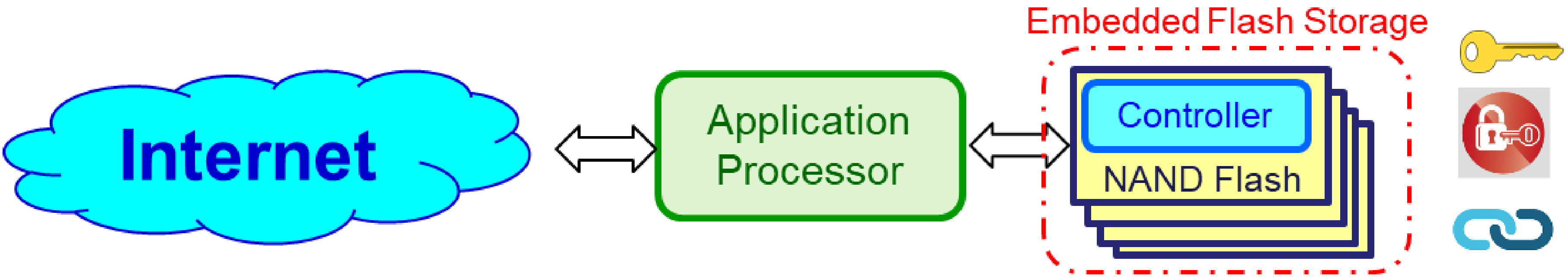


Interface
Aux. Buffer Memory
Inside Data Processor
Storage & Memory Controller
Fast Memory DDR5 DRAM
Quick Storage SLC-mode Flash
Slow Storage TLC-mode Flash
Passive Components Peripheral Components / ICs

Security is always Essential in AI, IoT.

EmBestor
The Industrial Flash Storage Expert.





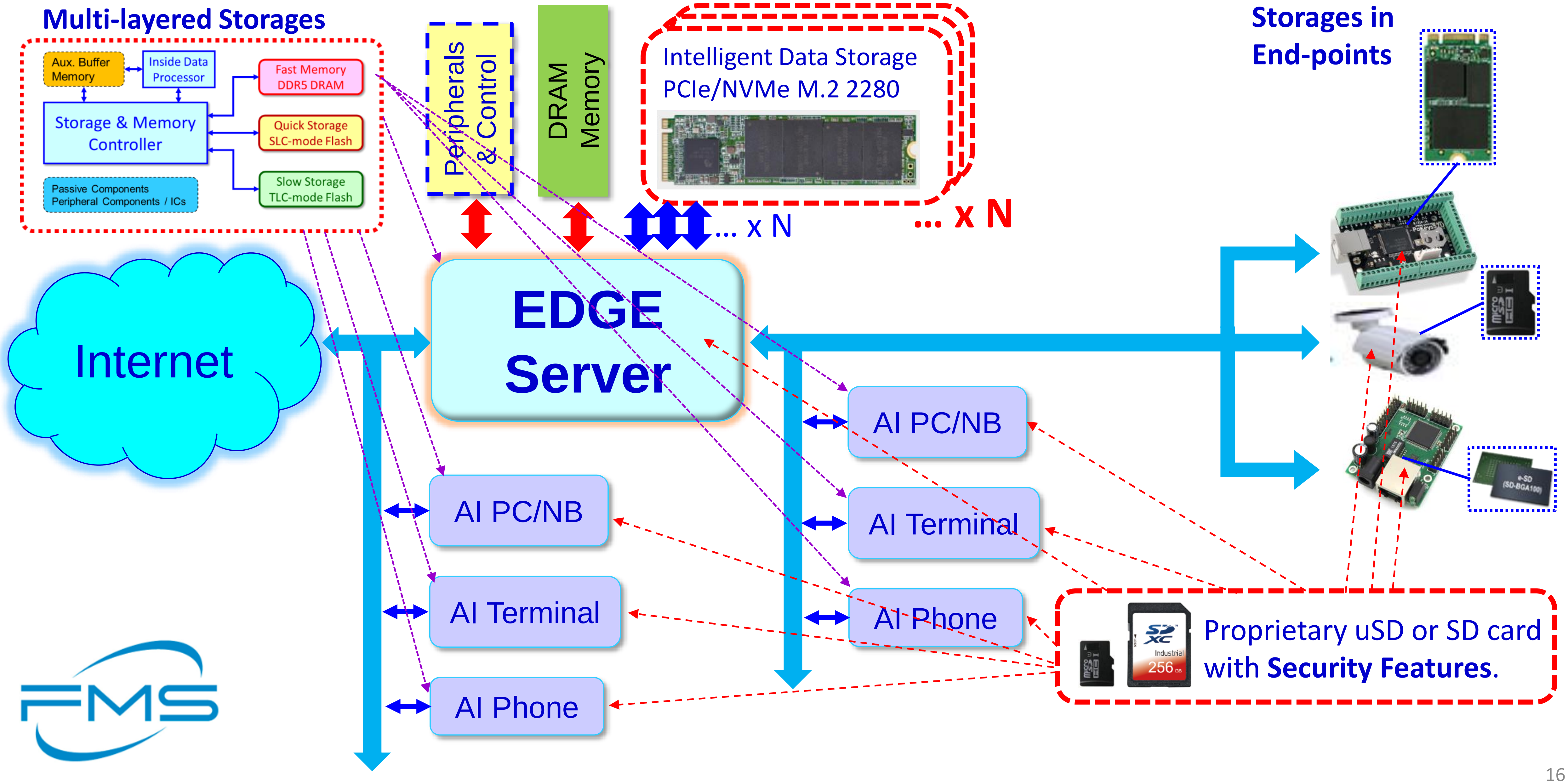
Digital Signature	• Point identification and authentication by Private Key. • Key management with Security Module.
Data Crypto	• Data Encryption and Decryption. • Data hidden and Data encrypted.
WORM	• WORM: Write Once Read Many. • Data printed and secure the Data chain.

FMS

Example: System View for Applications

EmBestor

The Industrial Flash Storage Expert.



Multi-layered Storages

Aux. Buffer Memory

Inside Data Processor

Storage & Memory Controller

Fast Memory DDR5 DRAM

Quick Storage SLC-mode Flash

Slow Storage TLC-mode Flash

Passive Components Peripheral Components / ICs

Peripherals & Control

DRAM Memory

Intelligent Data Storage PCIe/NVMe M.2 2280

... x N

... x N

Storages in End-points

Internet

EDGE Server

AI PC/NB

AI Terminal

AI Phone

AI PC/NB

AI Terminal

AI Phone

Proprietary uSD or SD card with **Security Features.**

FMS

Conclusions

- AI Model architecture for better operational efficiency would be worth to having further research.
- A novel interface bus is pesented for matching the speed and responsitivity between AI Computing and Memory & Storage devices.
- A good enough AI application ECO systems is still on the way.
- An example of Multi-layered data storage device and system application is illustrated for the application scenario.

Thank You !!

EmBestor Technology Inc.

http://www.embestor.com

Enjoy Best Service !!