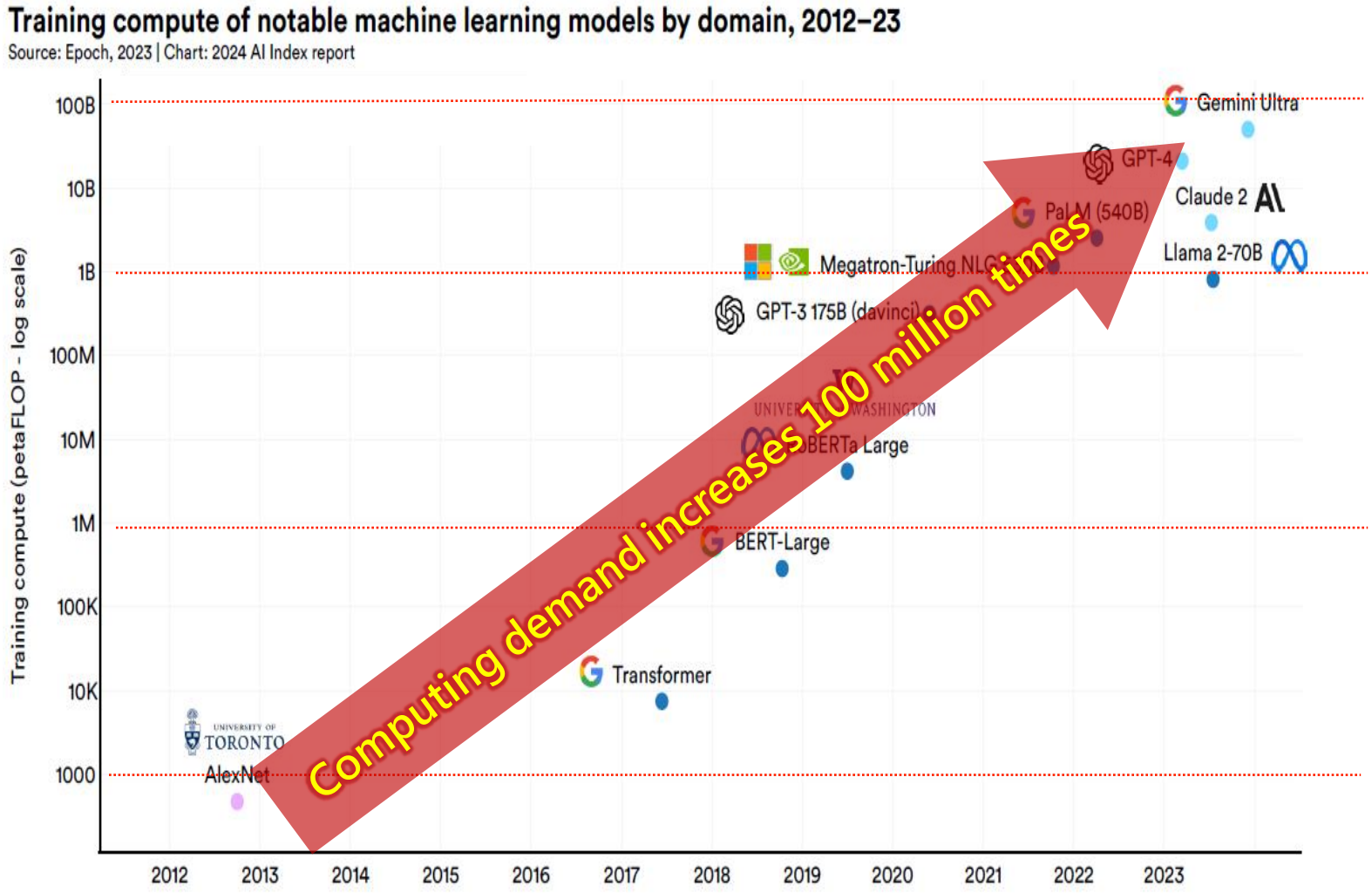
AI Computing Design Trends for LLMs in the Generative AI Era

Bor-Sung Liang, Ph.D

Dec 2024

bs.liang@mediatek.com bsliang@gmail.com

The computing power required to train AI models has increased by 100 million times in ten years



100 Yotta FLOP
10^{26} FLOP

1 Yotta FLOP
10^{24} FLOP

1 ZettaFLOP
10^{21} FLOP

1 ExaFLOP
10^{18} FLOP

Source: Stanford HAI, "AI Index 2024", https://aiindex.stanford.edu/report/

Computing power demand brings semiconductor boom

Computing demand increases 100 million times

GPU Computing grows 1000 times

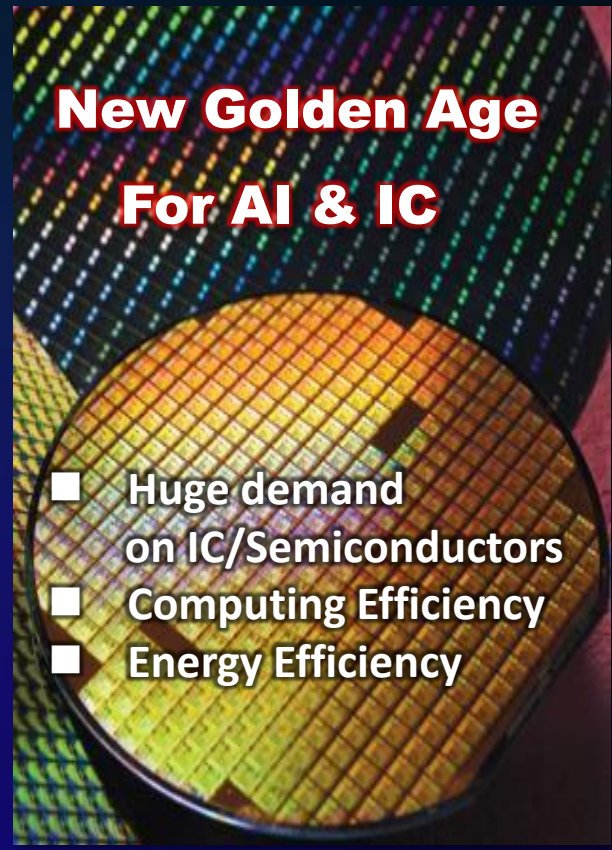
CPU Computing grows 100 times
(Speed of Moore's Law)

100,000 Difference

1,000,000 Difference

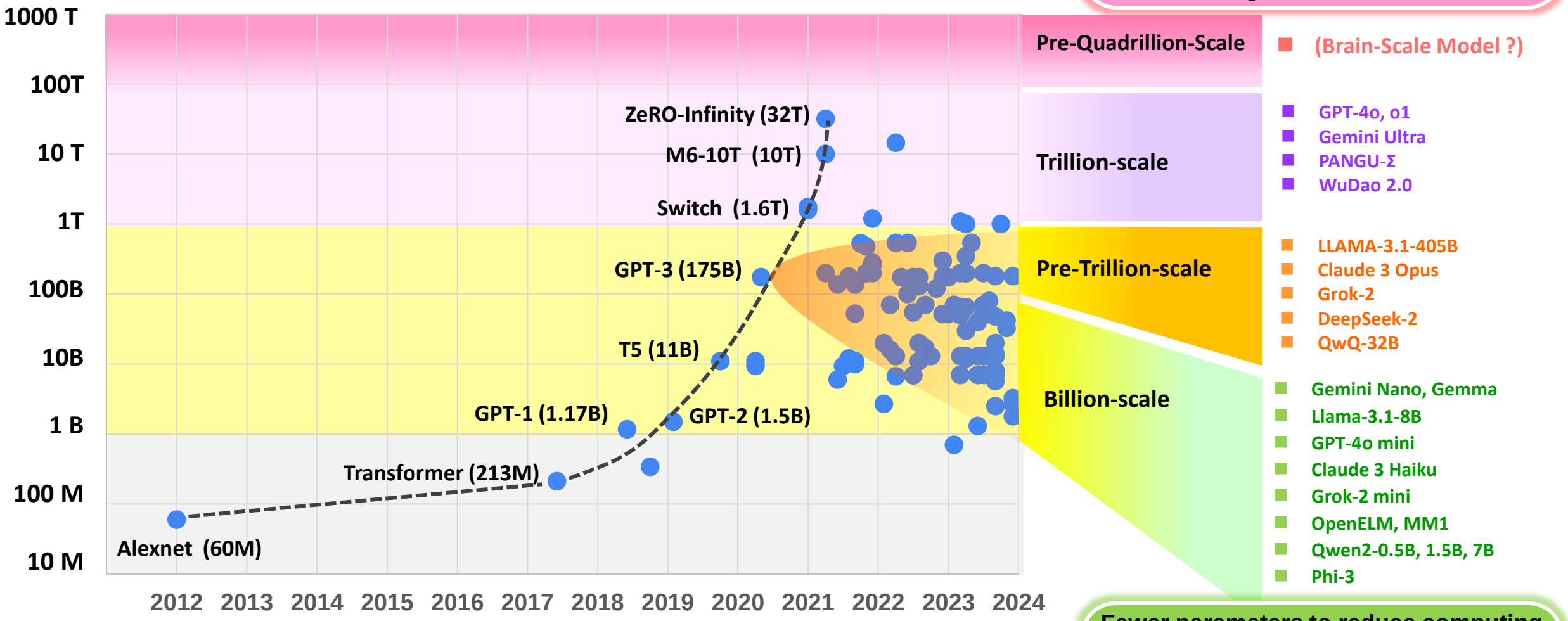
Large number of computing chips to provide computing power

➔ semiconductor boom



Trends of LLM Models

Trends of AI Models in a Decade

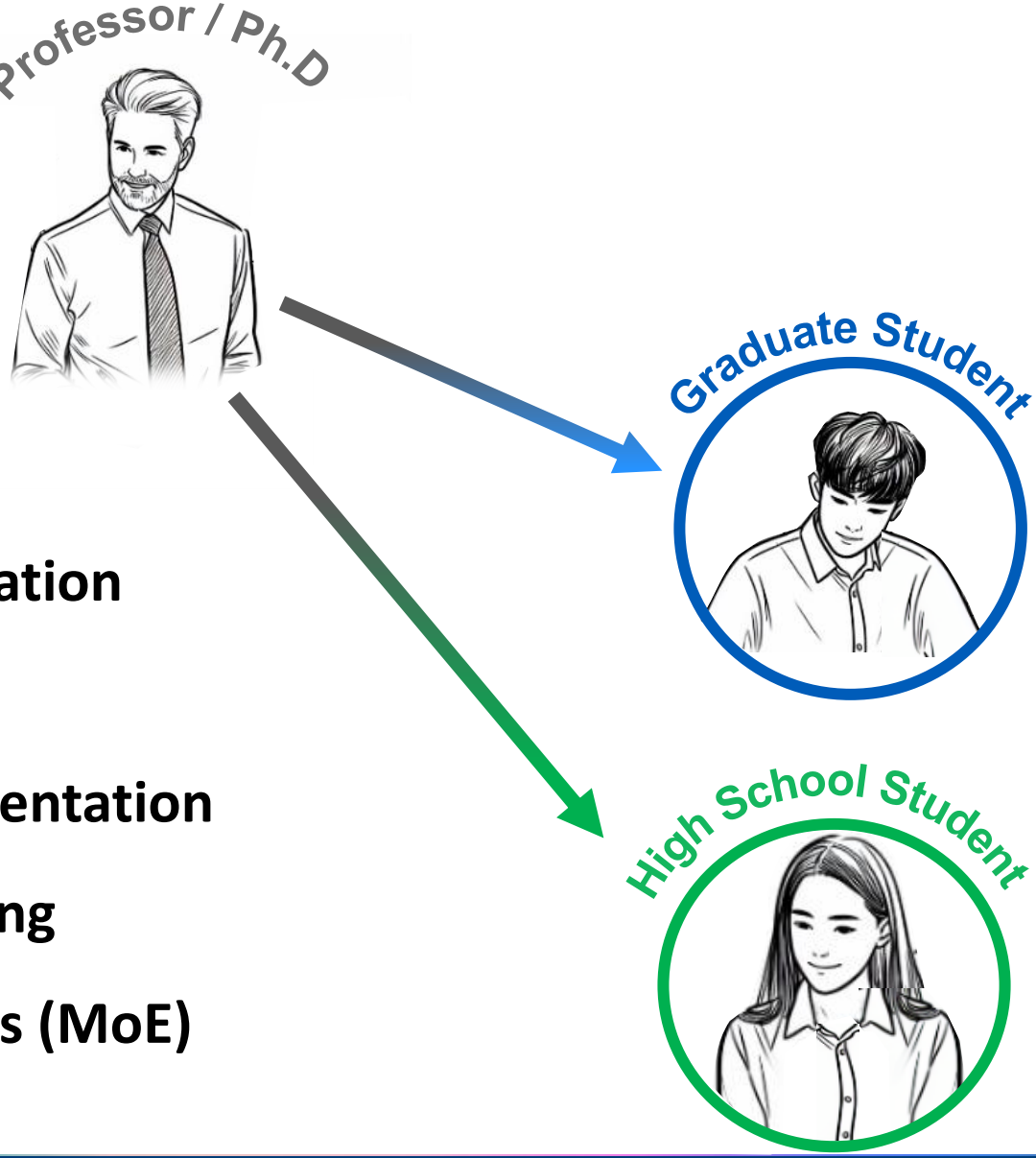


*Numbers in brackets () are AI model parameter size released by their companies or organizations.
For AI models without numbers, the range of their parameter size are estimated.

Trends of LLM Models for Inference

Full-Scale LLM

100s Bn ~ Tn Parameters



Middle-size LLM

10s~100s Billion Parameters

Small-size LLM

Billion Parameters

- Knowledge Distillation
- Model Reduction
  - Number Representation
  - Sparsity / Pruning
- Mixture of Experts (MoE)

Trends of LLM Models for Inference

Training

Inference

Full-Scale LLM

100s Bn ~ Tn Parameters

- Knowledge Distillation
- Model Reduction
  - Number Representation
  - Sparsity / Pruning
- Mixture of Experts (MoE)

Middle-size LLM

10s~100s Billion Parameters

Small-size LLM

Billion Parameters

Inference on Cloud Servers

Inference on Edge Devices

Trends of LLM Models for Inference

Training

Inference

Full-Scale LLM

100s Bn ~ Tn Parameters

- Knowledge Distillation
- Model Reduction
  - Number Representation
  - Sparsity / Pruning
- Mixture of Experts (MoE)

How to improve performance?

Prompt	Guidance from teachers, to know the key points & role (Tips, Hints)
In-context Learning	Provide background information to extent capability (Reference manual, Cheat Sheets)
RAG	Prepare FAQ and standard procedure to answer various problems (FAQ、SOP)
RLHF RLAIF	Hire a coach (Human or AI) to practice and to improve perform. (Mock exams, practice games)
AI Agents	Memory, Planning and Tool Skills Multi-Agents Collaboration (Domain expert, Task Force)
Chain of Thought	Self-reflection, Awareness, Reasoning to Improving Perform. (Self-review, Reasoning)

RLHF= Reinforcement Learning from Human Feedback
RLAIF = Reinforcement Learning by AI Feedback
RAG = Retrieval-Augmented Generation

Trends of LLM Models for Inference

Training

Inference

Full-Scale LLM

100s Bn ~ Tn Parameters

Prompt | RAG | AI Agent | In-context Learning | RLHF RLAIF | Chain of Thought

Middle-size LLM

10s~100s Billion Parameters

Prompt | RAG | AI Agent | In-context Learning | RLHF RLAIF | Chain of Thought

Small-size LLM

Billion Parameters

Prompt | RAG | AI Agent | In-context Learning | RLHF RLAIF | Chain of Thought

Inference on Cloud Servers

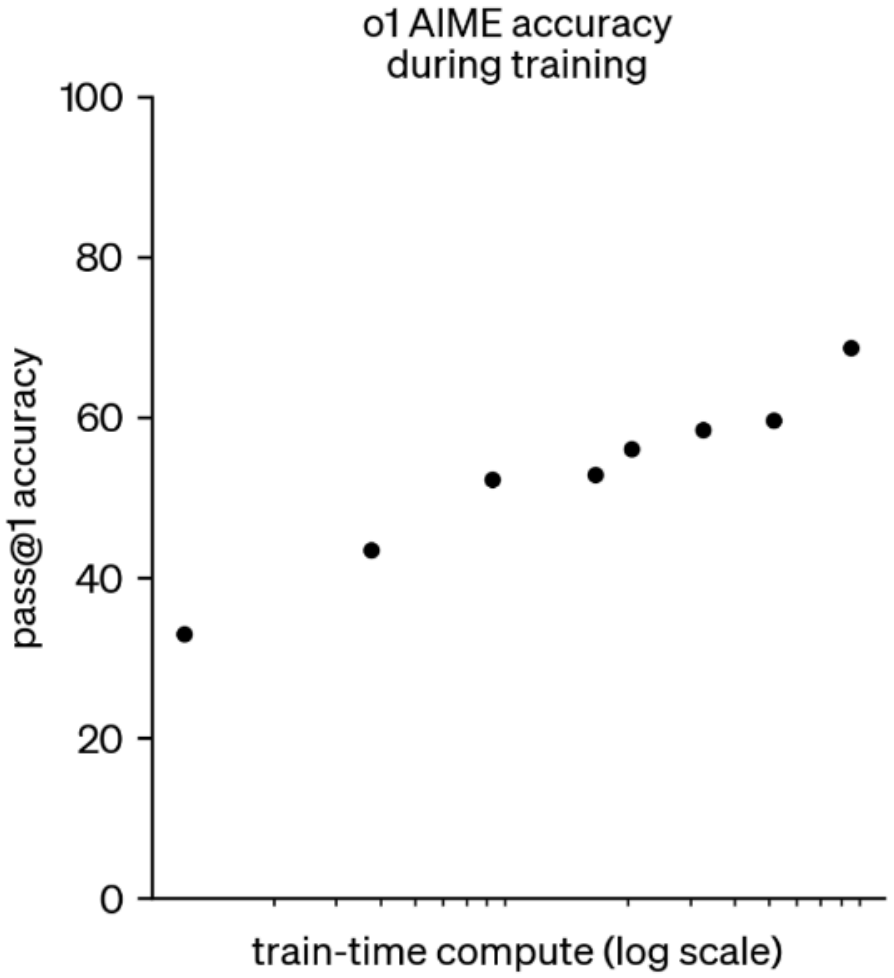
Inference on Edge Devices

- Knowledge Distillation
- Model Reduction
  - Number Representation
  - Sparsity / Pruning
- Mixture of Experts (MoE)

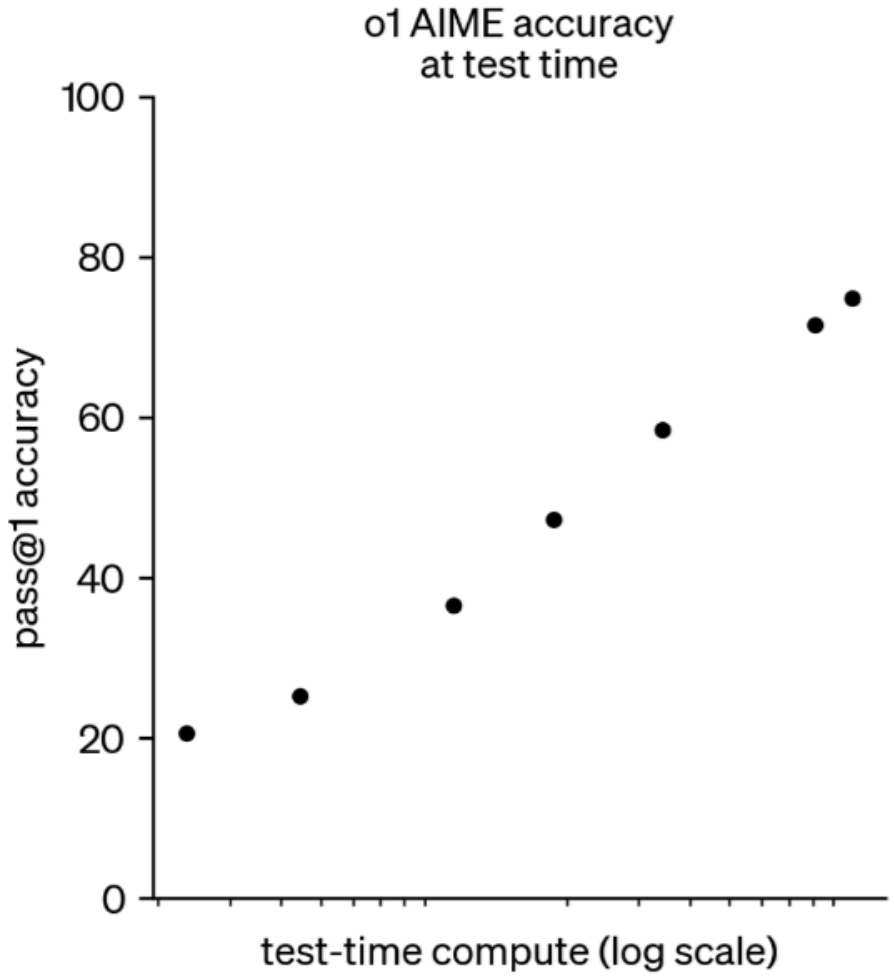
RLHF= Reinforcement Learning from Human Feedback
RLAIF = Reinforcement Learning by AI Feedback
RAG = Retrieval-Augmented Generation

Two Paradigms of Scaling to Increase LLM Performance

Training Time Compute



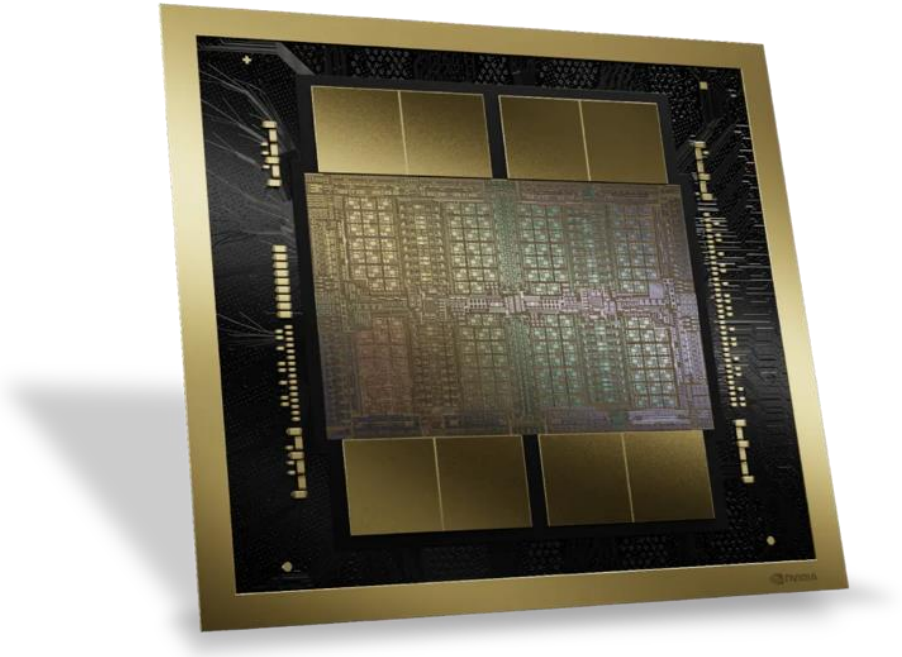
Test-time Compute (Inference)



Source: OpenAI, 2024/9/12, "Learning to Reason with LLMs", https://openai.com/index/learning-to-reason-with-llms/

AI Computing : Data Center vs Edge Devices

NVIDIA® B200

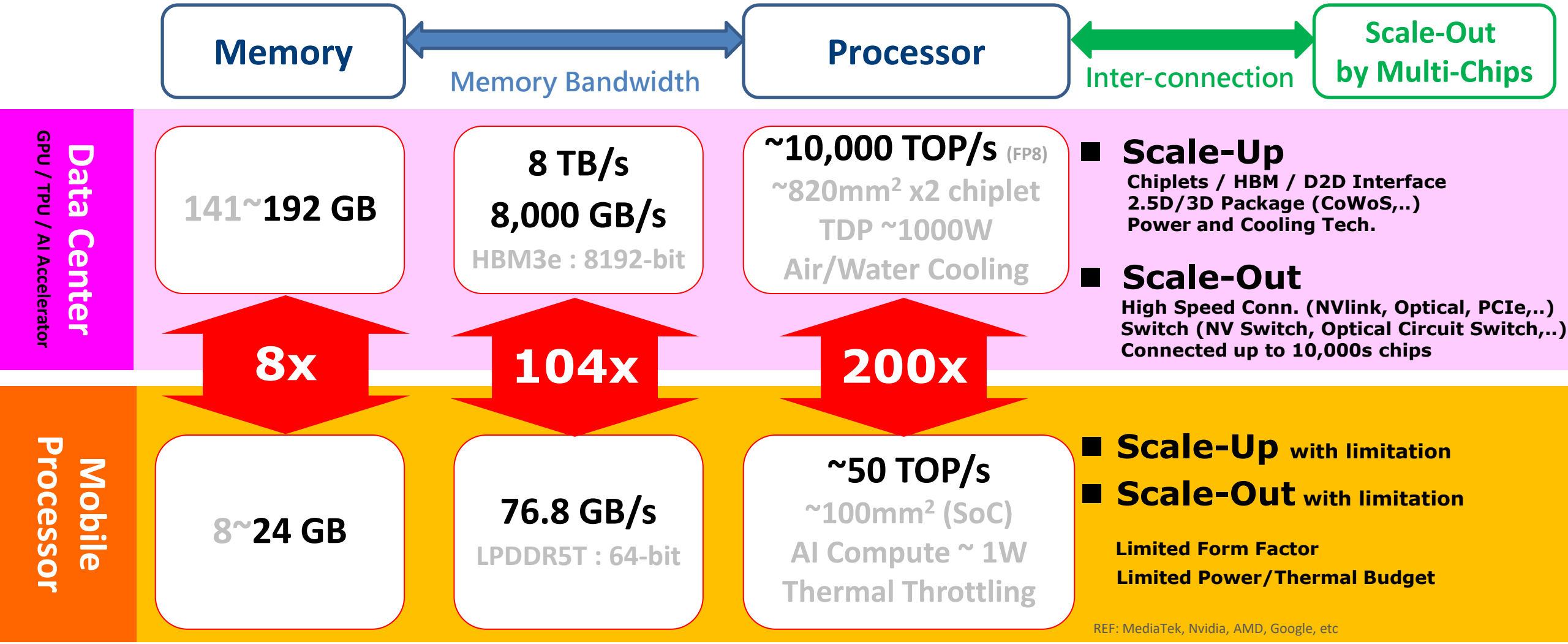


Die Size > 1600mm² (2 chiplet)
TDP ~ 1000W
10,000 TOP/s (FP8 , with Sparsity)
20,000 TOP/s (FP4, with Sparsity)

Die Size ~100mm²
TDP ~10W

~50 TOP/s

AI Computing : Data Center vs Edge Devices



Energy Efficiency 5~10x

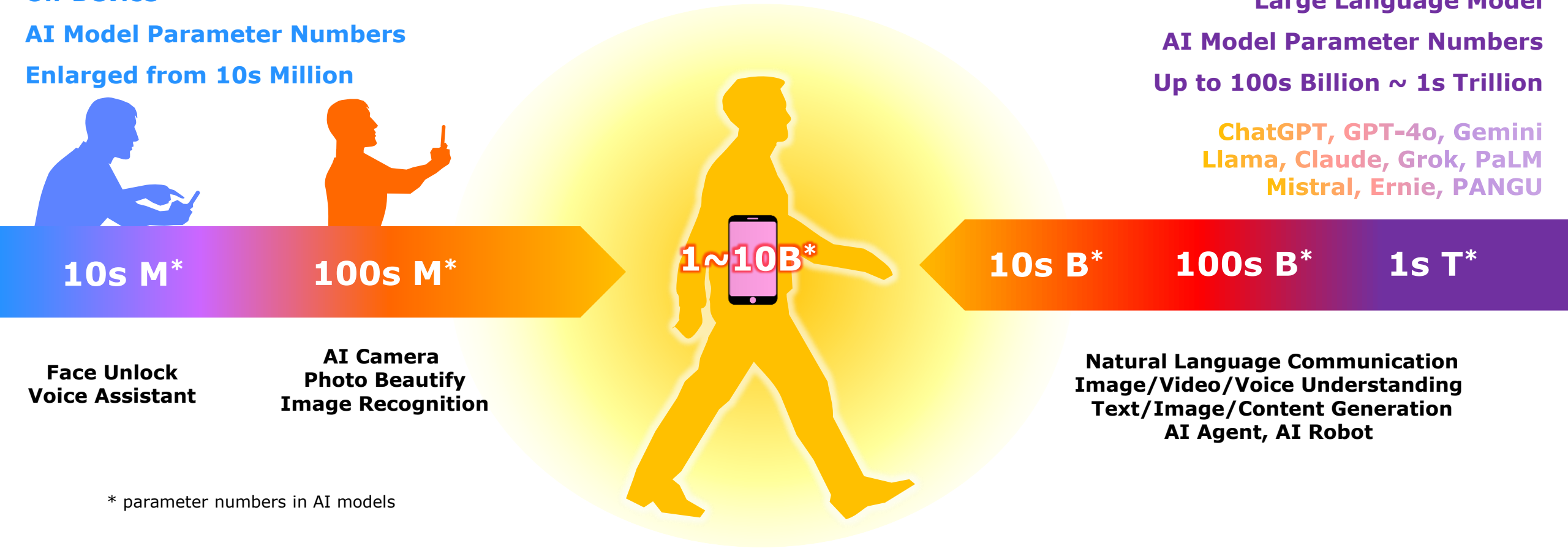
AI Computing in Edge Devices

AI Inference on Mobile Processors for Large Language Models (LLMs)

On-Device
AI Model Parameter Numbers
Enlarged from 10s Million

Large Language Model
AI Model Parameter Numbers
Up to 100s Billion ~ 1s Trillion

ChatGPT, GPT-4o, Gemini
Llama, Claude, Grok, PaLM
Mistral, Ernie, PANGU



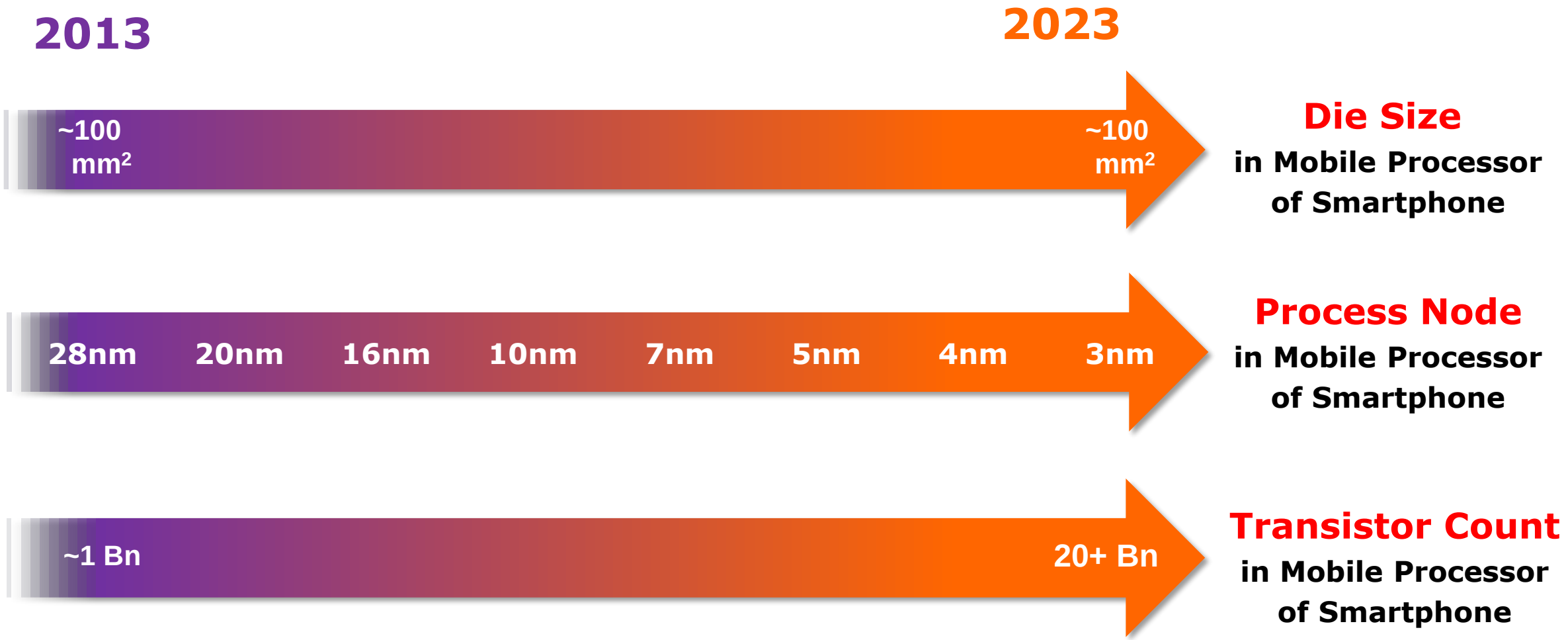
Face Unlock
Voice Assistant

AI Camera
Photo Beautify
Image Recognition

Natural Language Communication
Image/Video/Voice Understanding
Text/Image/Content Generation
AI Agent, AI Robot

* parameter numbers in AI models

Trends of Mobile Processor in a Decade

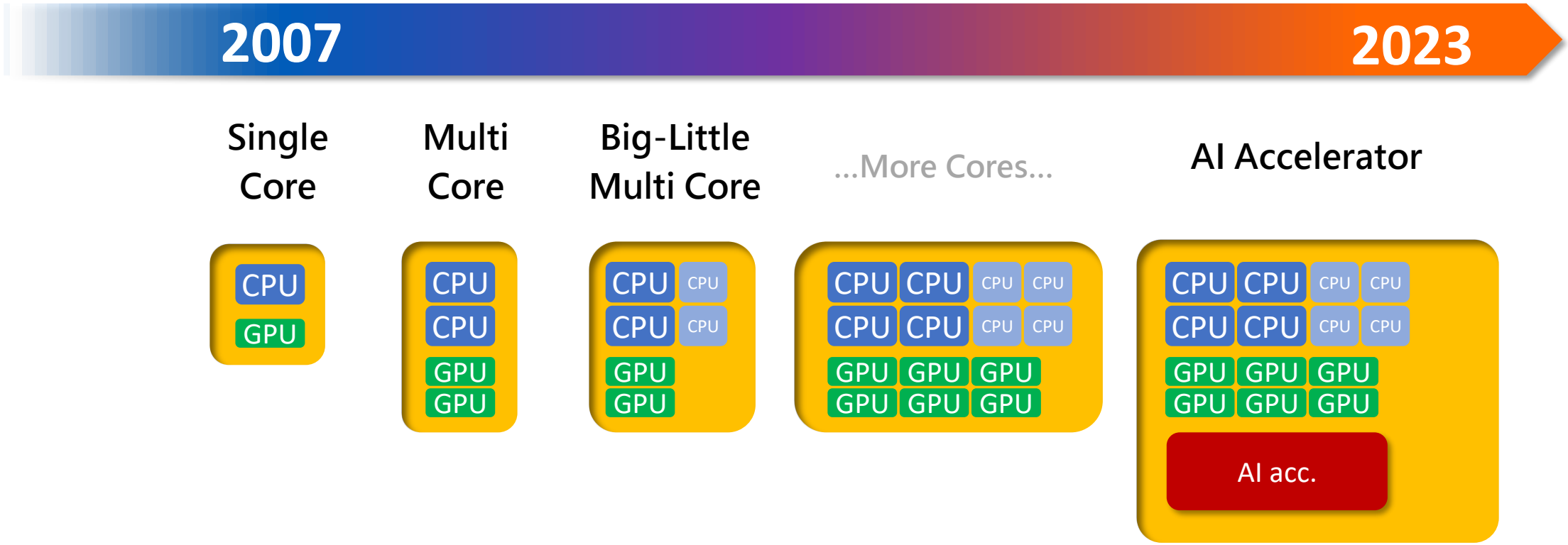


Die Size in Mobile Processor of Smartphone

Process Node in Mobile Processor of Smartphone

Transistor Count in Mobile Processor of Smartphone

Trends of Computing Architecture in Mobile Processor



- CPU Center Processor Unit
- GPU Graphics Processor Unit
- AI acc. AI Accelerator

*The number of CPU, GPU, and AI accelerator cores is for reference. Each application processor will have different architecture and number depending on the actual situation.

Model Size for different Parameter and Number Representation

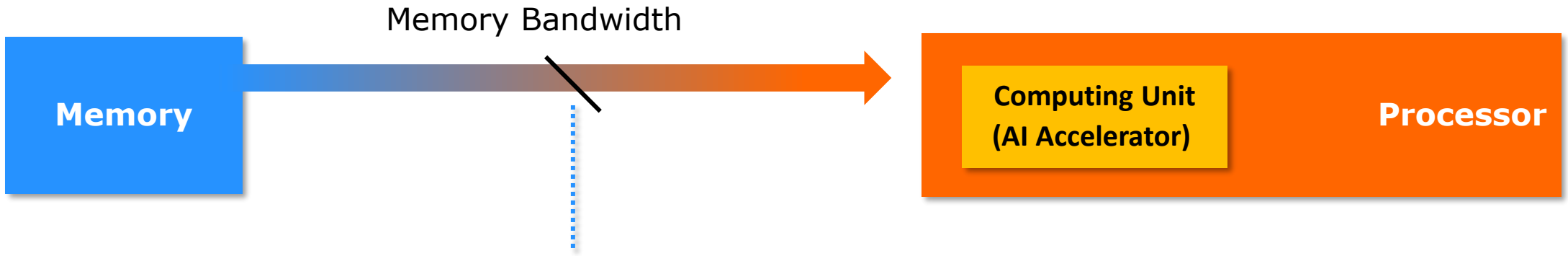
AI Training **AI Inference**

Model Parameters	FP32 **32bit** (4Byte)	FP16, BF16 **16bit** (2Byte)	INT8, FP8 **8bit** (1Byte)	INT4, FP4 **4bit** (0.5Byte)
70 Bn	280 GB	140 GB	70 GB	35 GB
13 Bn	52 GB	26 GB	13 GB	7.5 GB
7 Bn	28 GB	14 GB	7 GB	3.5 GB
3.25 Bn	13 GB	6.5 GB	3.25 GB	1.625 GB
1.8 Bn	7.2 GB	3.6 GB	1.8 GB	0.9 GB

Single digital Billion-Scale Parameter LLM

DRAM Size : >128GB 128~24GB 32~8GB 8~4GB 4~2GB <2GB

Bandwidth Utilization Ratio

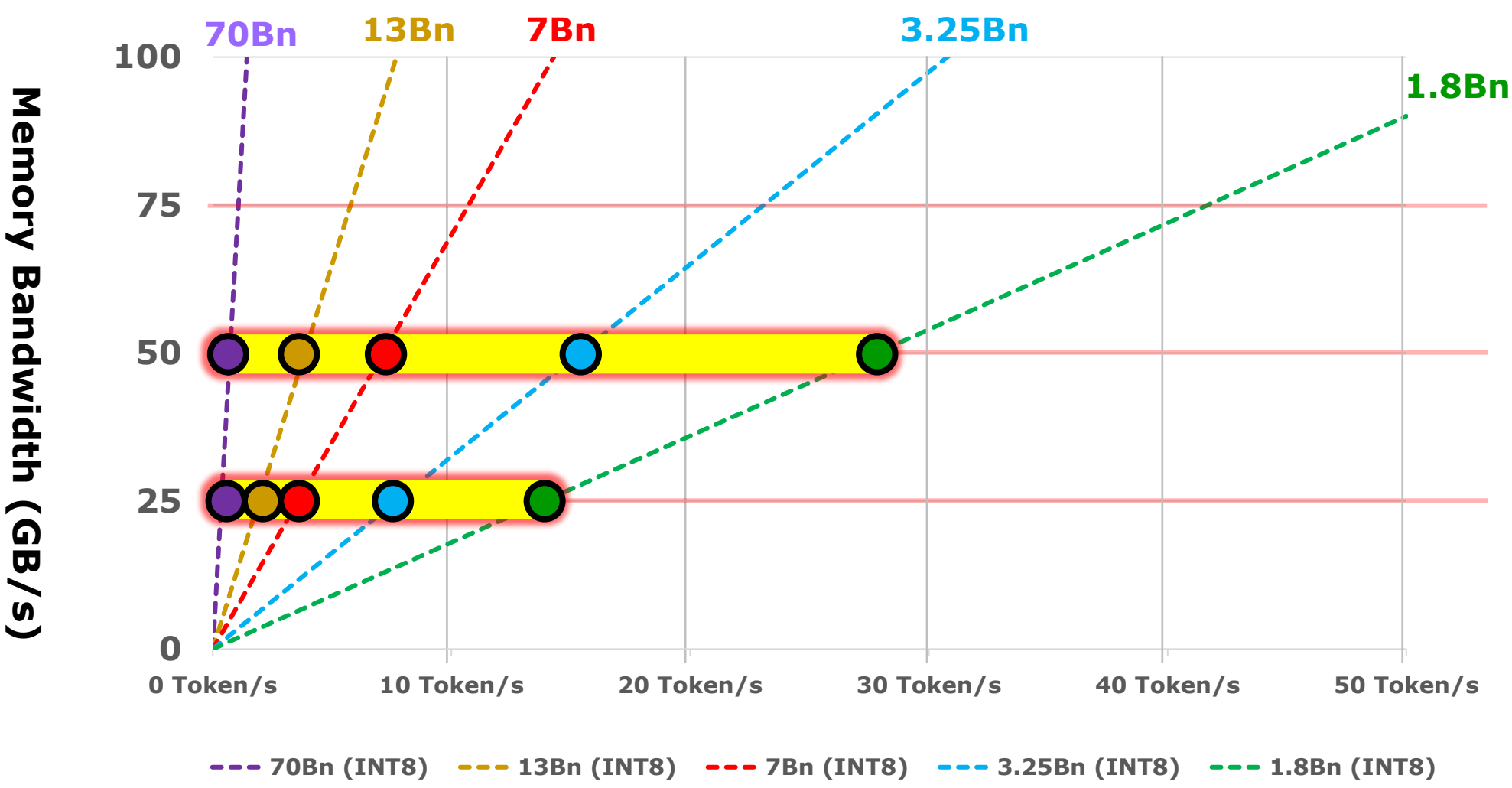


MBU

Model Bandwidth Utilization

$$\frac{\text{Real memory bandwidth for LLM}}{\text{Peak memory bandwidth}}$$

Token Speed under Memory Bandwidth Limit (LLM @ INT8/FP8)



Memory:
LPDDR5T-9600
(9.6Gbps, 64bit)
76.8 GB/s

100%
Model Bandwidth Utilization

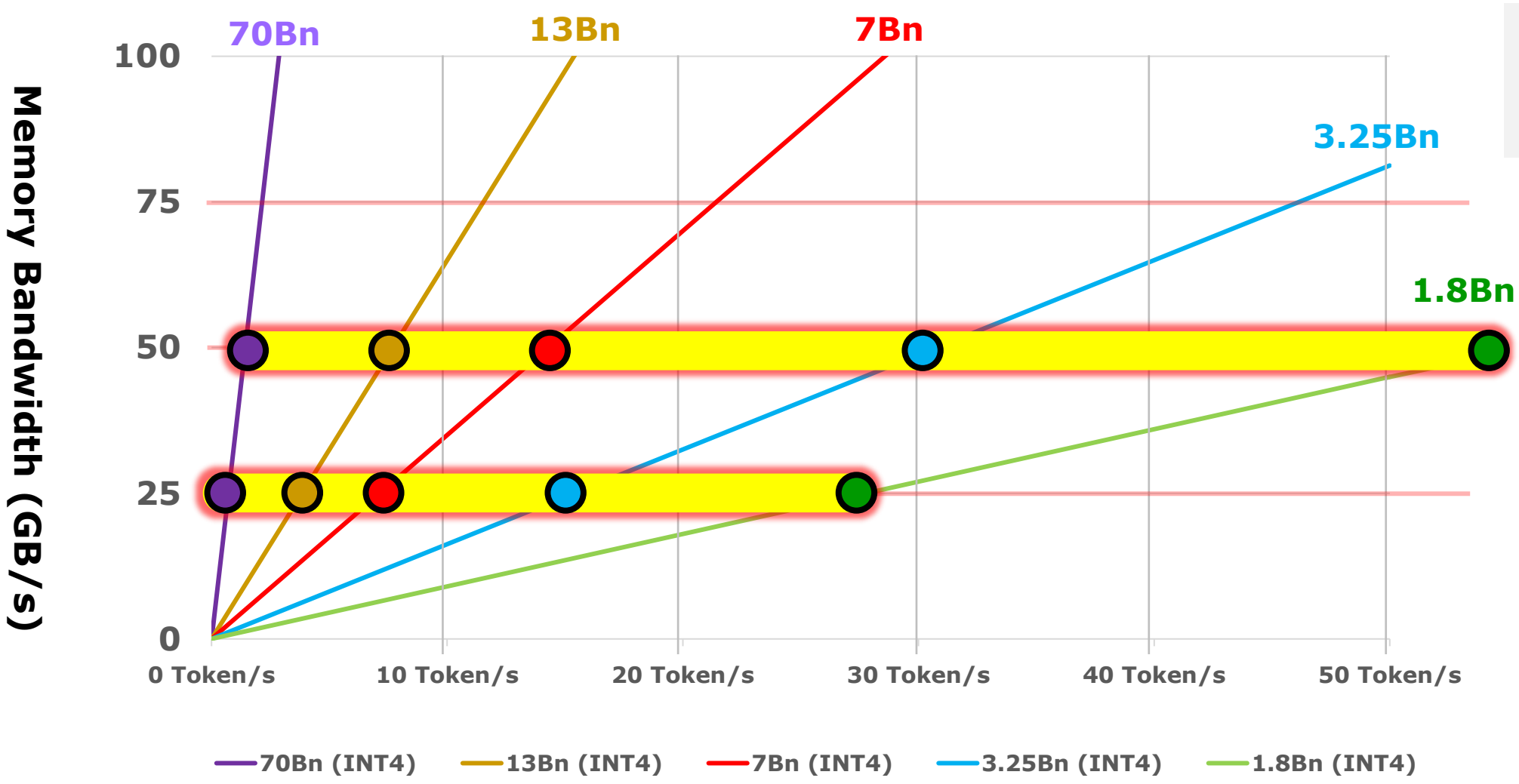
65%
Model Bandwidth Utilization

33%
Model Bandwidth Utilization

* Trivial LLM porting without optimize model for bandwidth

- Only consider Bandwidth of "Memory to Processor"

Token Speed under Memory Bandwidth Limit (LLM @ INT4/FP4)



Memory:
LPDDR5T-9600
(9.6Gbps, 64bit)
76.8 GB/s

100%
Model Bandwidth Utilization

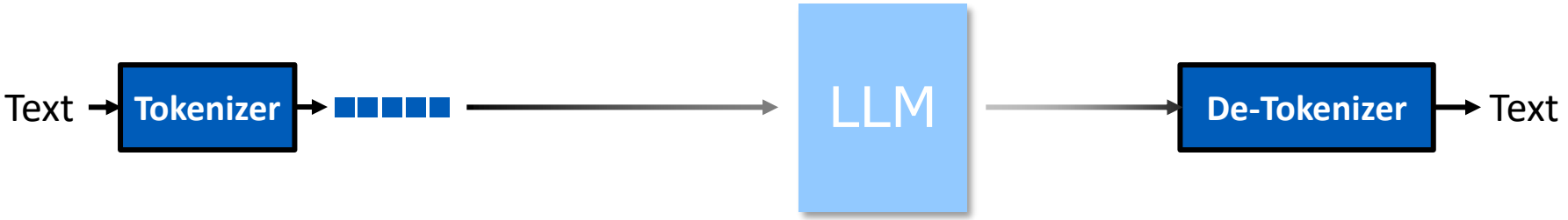
65%
Model Bandwidth Utilization

33%
Model Bandwidth Utilization

* Trivial LLM porting without optimize model for bandwidth

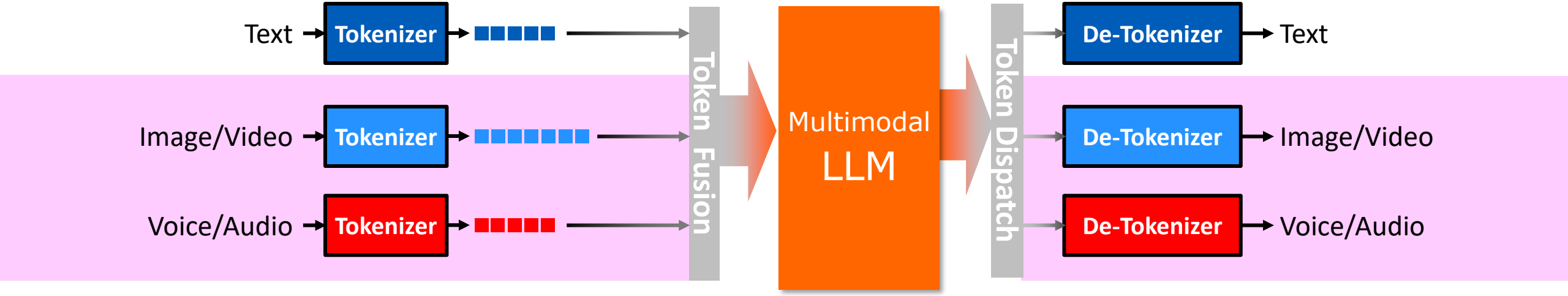
- Only consider Bandwidth of "Memory to Processor"

Token Process of LLM



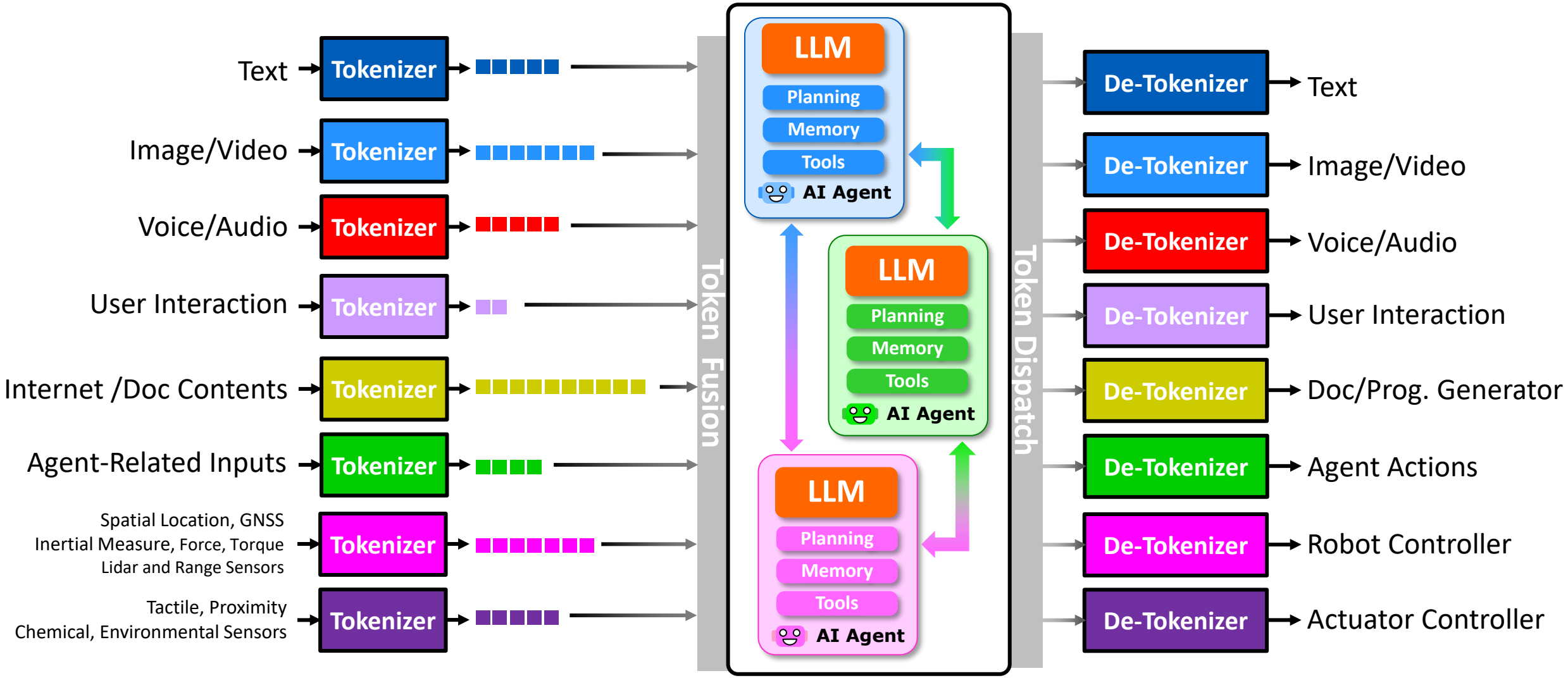
Mobile Devices
w/ text-based LLM

Token Process of Multimodal LLM



Mobile Devices
w/ Multimodal Gen AI

Token Process for AI Robotics with Multi AI Agents



Token Speed

How Fast can Humans Read? → related to how fast should LLM generate text

Read word-by-word

Skim Reading looking *only* for the general or main ideas
Scan Reading looking *only* for a specific information

Read Normally 250 words/min

(Adults Silent-reading Speed 238 words/min)

The IEEE International 3D Systems Integration Conference (3DIC) will be held at the Hotel Metropolitan Sendai and Sendai Kokusai Hotel in Sendai, Japan in September 25-27, 2024. This conference combines the previous ASET and IEEE EDS Japan sponsored International 3D System Integration Conference, held in Tokyo in 2007 & 2008 and the Fraunhofer and IEEE CPMT sponsored International 3D System Integration Workshop held in 2003 & 2007 in Munich.
So much has changed since the first combined conference in San Francisco in 2009, 3D integrated circuits have moved from academic curiosity to sc commercial reality. The underlying technologies continually improve evolve, incorporating new methods and resources. Today's hot topics i chiplets, photonic interconnect, micro-die handling, and exotic substrates. This 2024 conference will focus on the research and science of 3DICs. It will cover relevant 3D/Chiplet topics from manufacturing, processes, materials, and equipment to circuit designs, design methodology, and applications.

Read Proficiently 1000 words/min

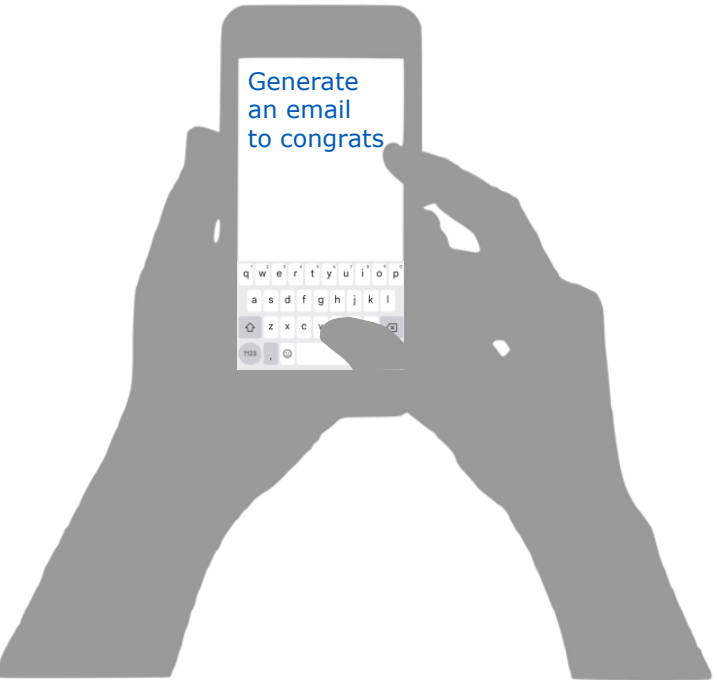
(Skimming speeds of 1000+ words/min)

The IEEE International 3D Systems Integration Conference (3DIC) will be held at the Hotel Metropolitan Sendai and Sendai Kokusai Hotel in Sendai, Japan in September 25-27, 2024. This conference combines the previous ASET and IEEE EDS Japan sponsored International 3D System Integration Conference, held in Tokyo in 2007 & 2008 and the Fraunhofer and IEEE CPMT sponsored International 3D System Integration Workshop held in 2003 & 2007 in Munich.
So much has changed since the first combined conference in San Franc 2009, 3D integrated circuits have moved from academic curiosity to solid commercial reality. The underlying technologies continually improve and evolve, incorporating new methods and resources. Today's hot topics include chiplets, photonic interconnect, micro-die handling, and exotic substrates. This 2024 conference will focus on the research and science of 3DICs. It will cover relevant 3D/Chiplet topics from manufacturing, processes, materials, and equipment to circuit designs, design methodology, and applications.

LLM in Mobile Devices

LLM (Large Language Model)

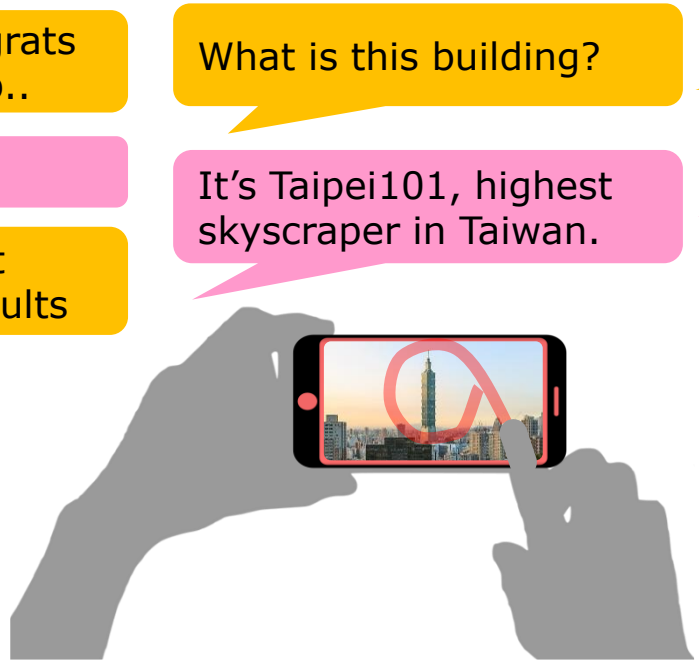
- Typing Text



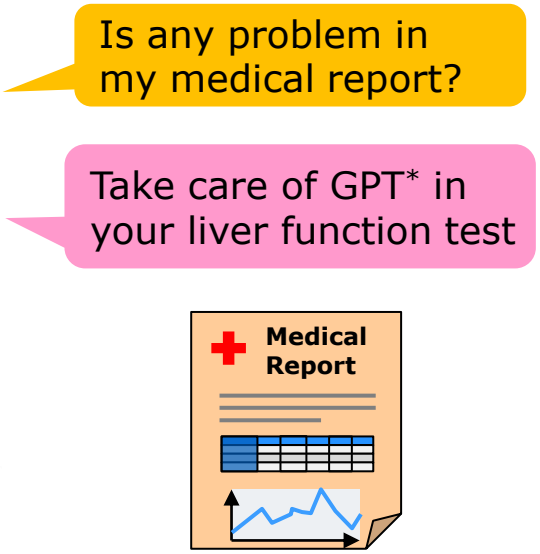
Multimodal LLM

- Voice
- Nature Language Communication

- Camera
- Visual indicator



- Image
- Doc, Table



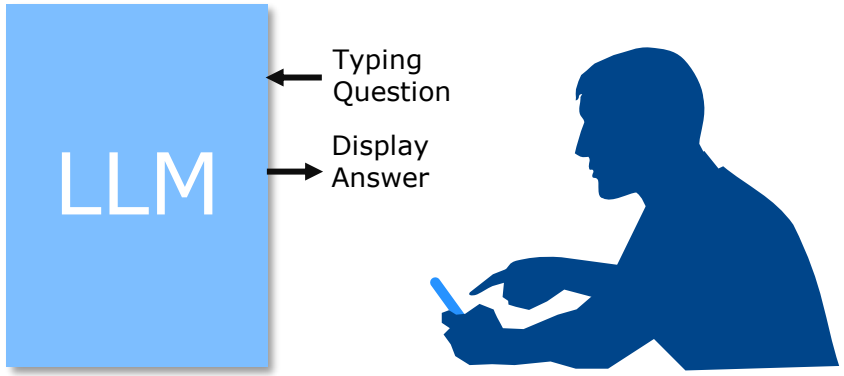
* GPT (glutamyl pyrubic transaminase) in liver function test

Types of Tokens

Simple LLM

Single Person

One Simple LLM Task

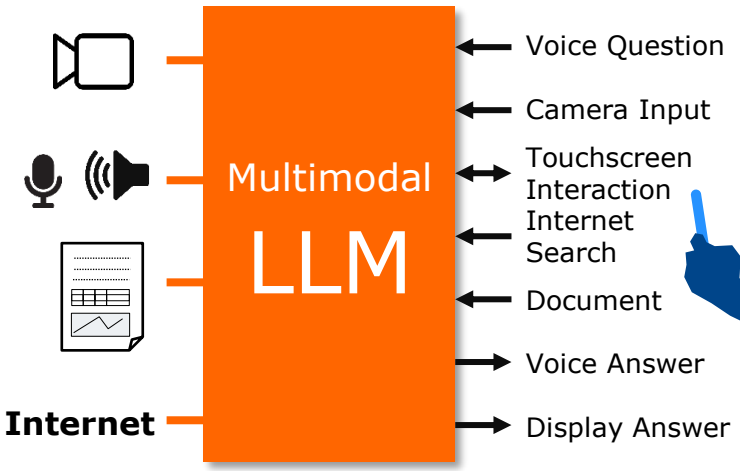


Only Tokens for Text

Read Normally
250 words/min
~8.33 Token/sec

Read Proficiently
1000 words/min
~33.3 Token/sec

Multimodal Interaction

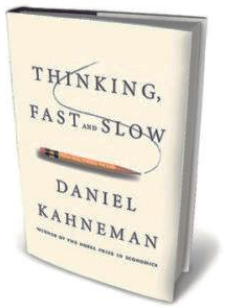


More Various Tokens Concurrently

- **Tokens for Text**
- **Tokens for Image/Video**
- **Tokens for Voice /Audio**
- **Tokens for User Interactive**
- **Tokens for Internet Contents**
- **Tokens for Document Contents**

LLM Collaboration with Mobile OS and Cloud

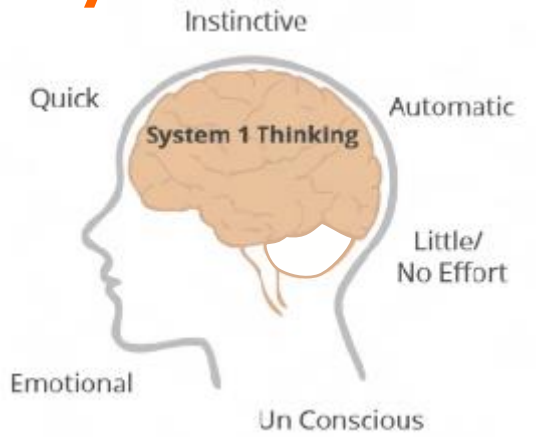
Current LLM: from “System 1” to “System 2”



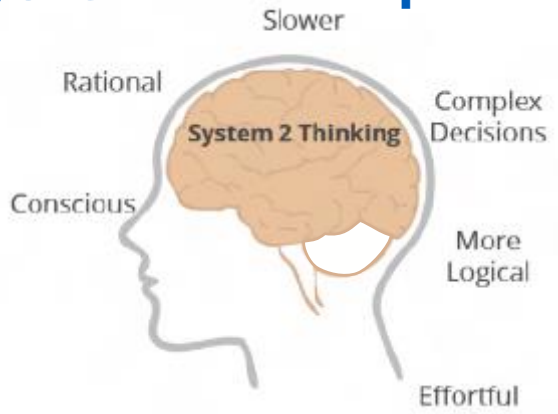
Cerebrum

Current LLM Status

System 1 : Instinctive



System 2 : Deep Thinking



Inspired from System 1 and 2 :

Can we collaborate both Device-side and Cloud-side for System 1 and 2 ?

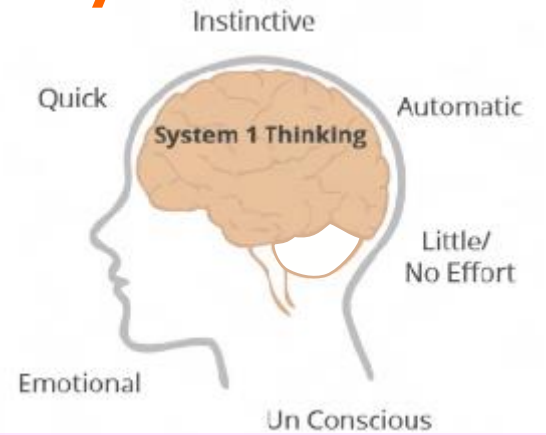
LLM – Thinking Fast, Deep Thinking and Swift Response

LLM on Device

LLM in Cloud

Cerebrum

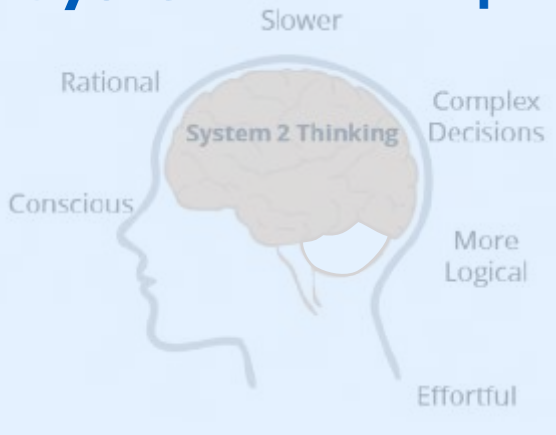
System 1 : Instinctive



Thinking Fast

Fast

System 2 : Deep Thinking



Deep Thinking

~~Slower~~ Fast

Cerebellum

Brainstem

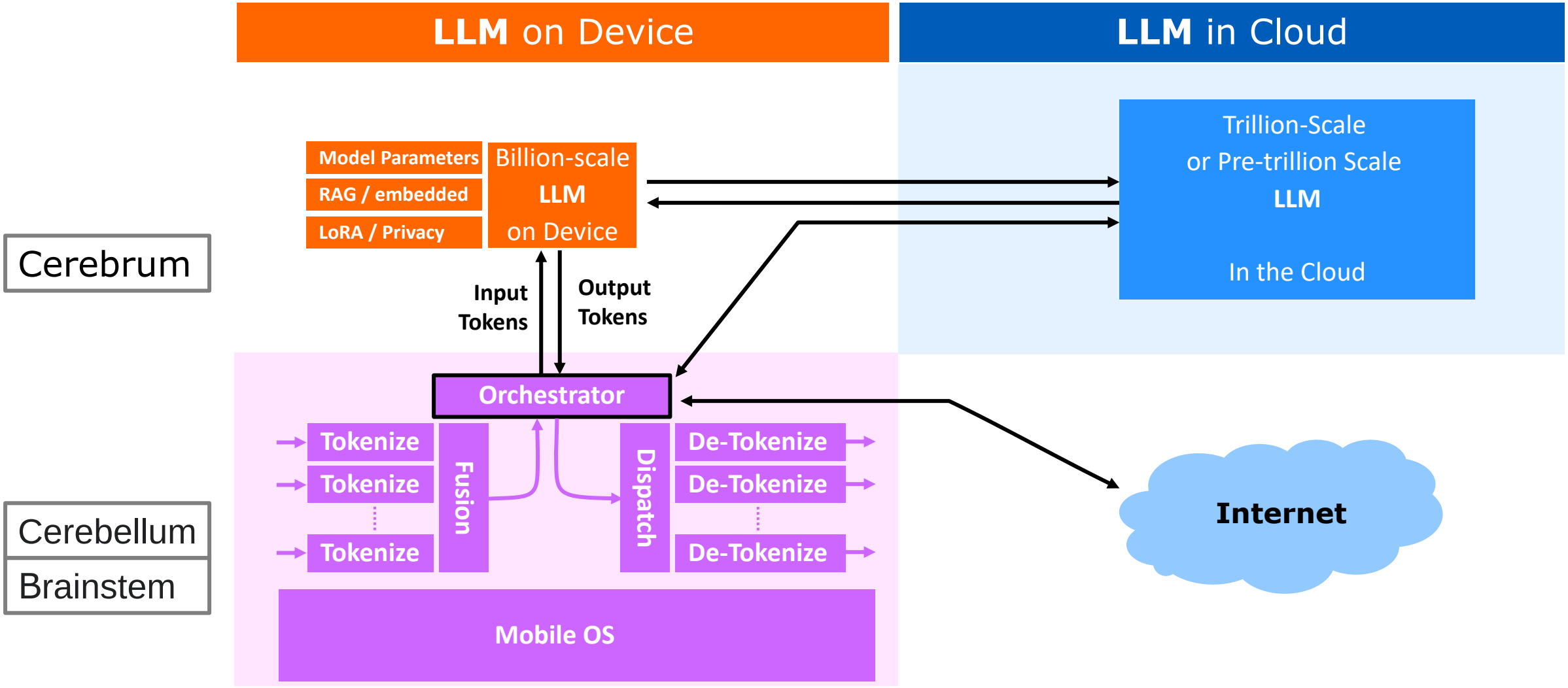
Quick Response
Multimodal Signals
Multi-Tasks Concurrently

Swift Response

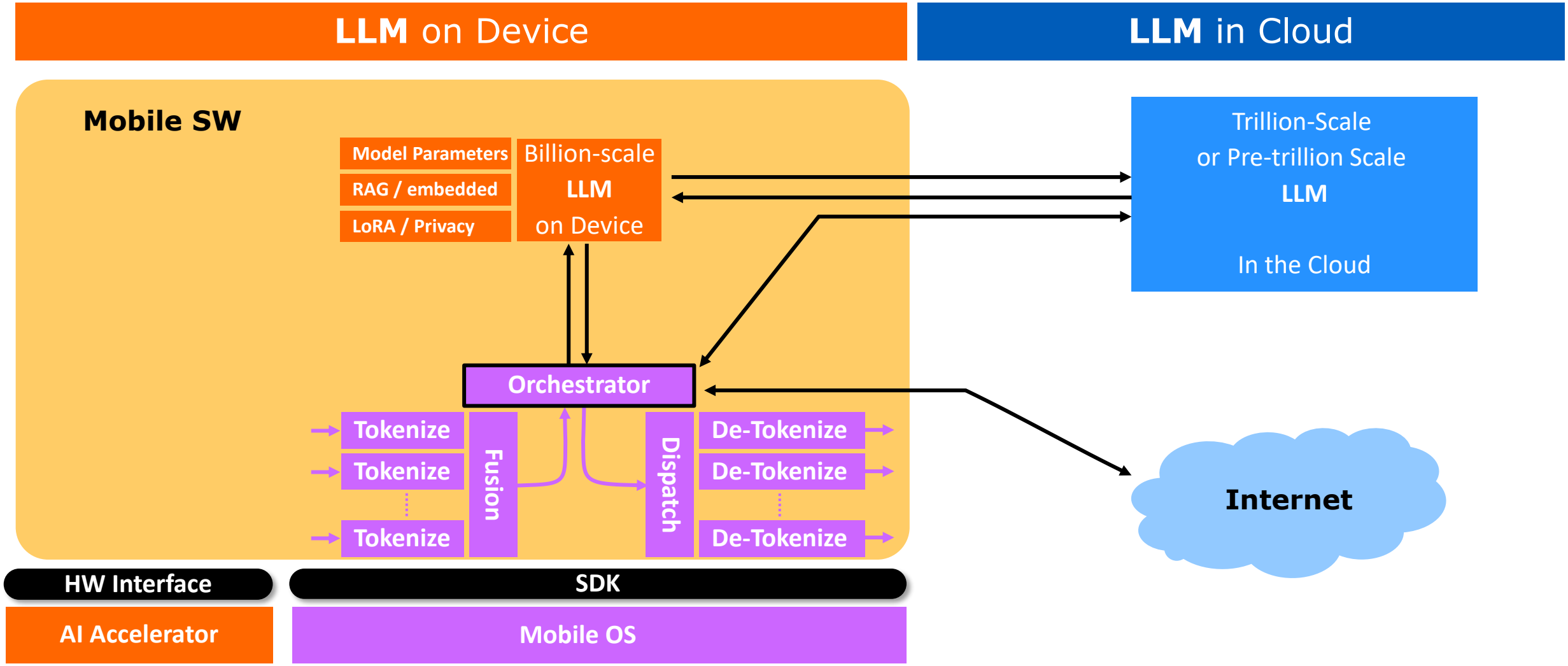
provide sufficient computing resources to speed-up

➔ Cloud Computing

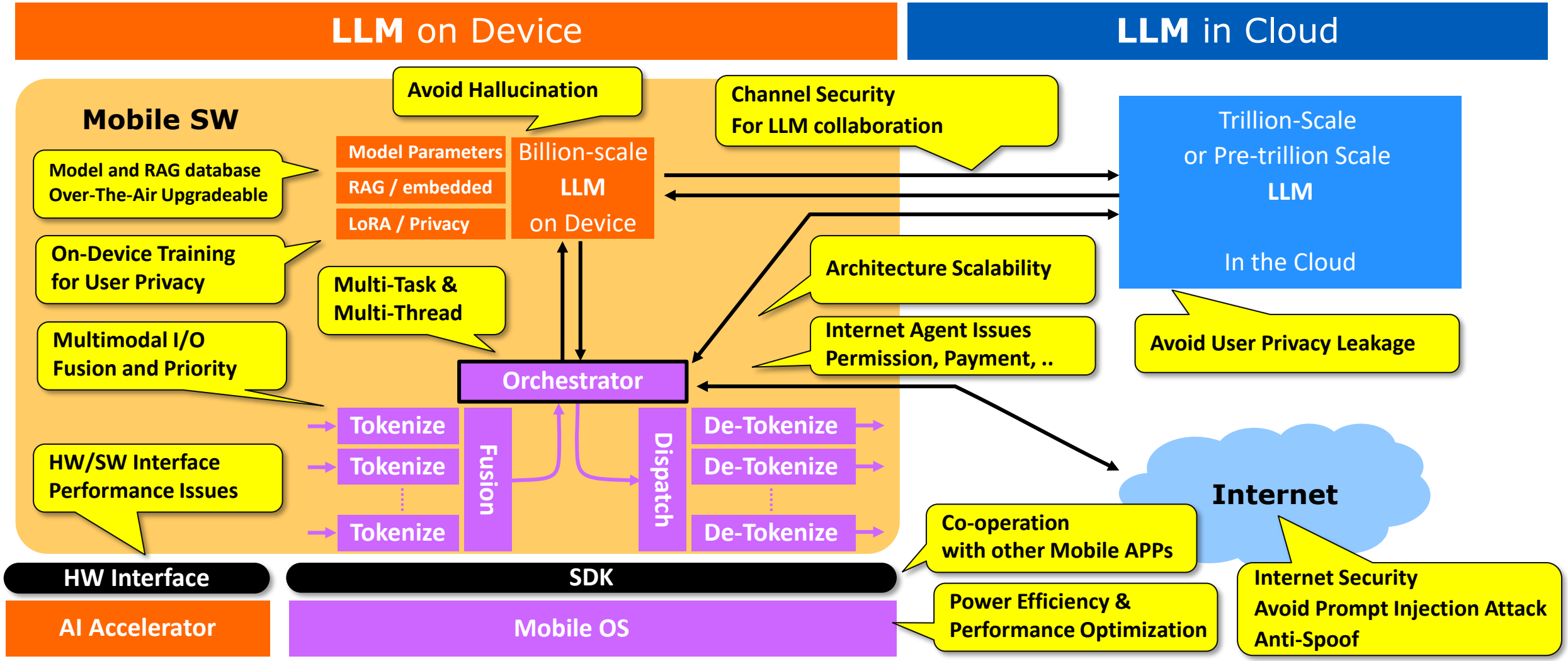
LLM on Device - Collaborate with Cloud and Mobile OS



LLM as Mobile Software on Mobile Processor

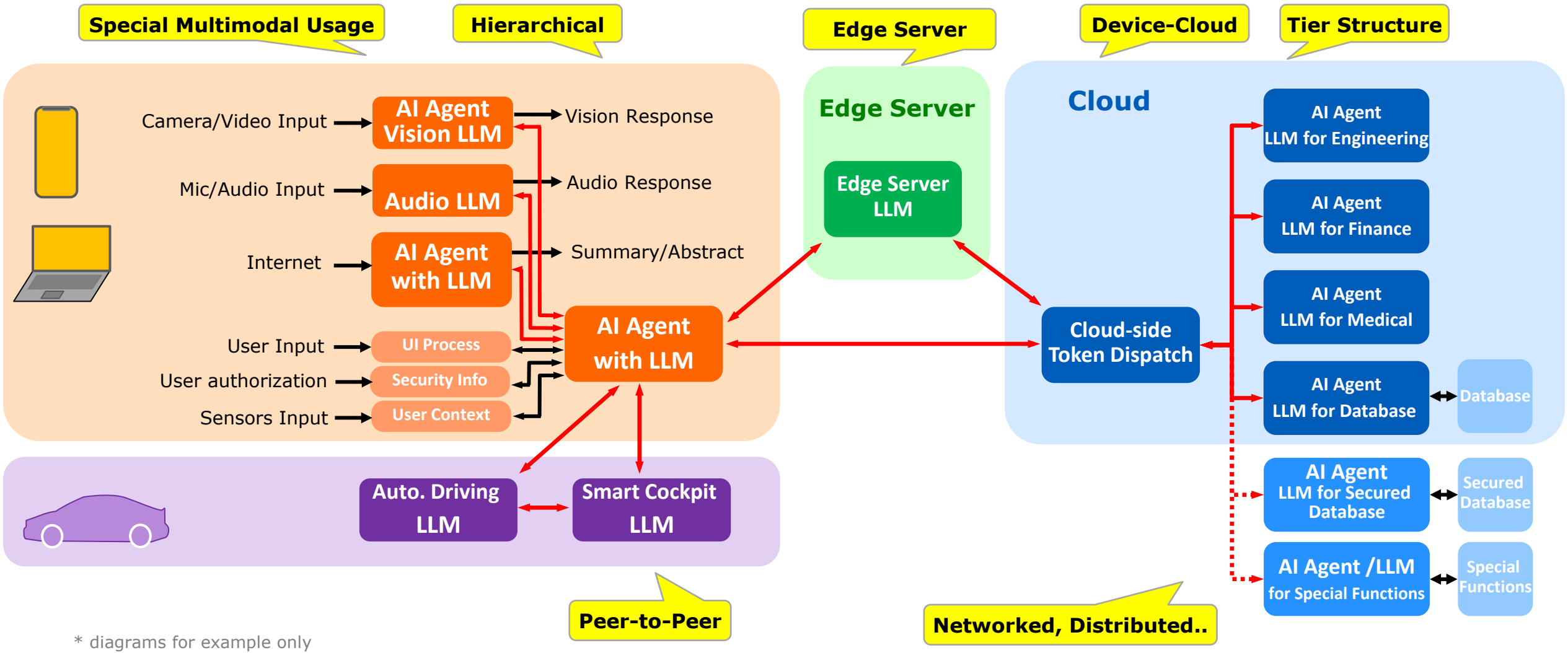


LLM on Mobile Processor : Issues



Trends of AI Computing for Data Center and Edge Devices

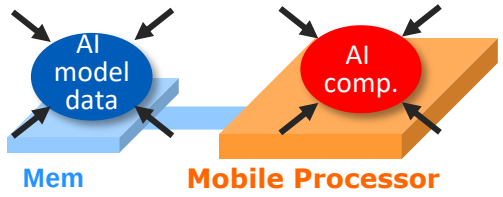
Collaboration among AI Models



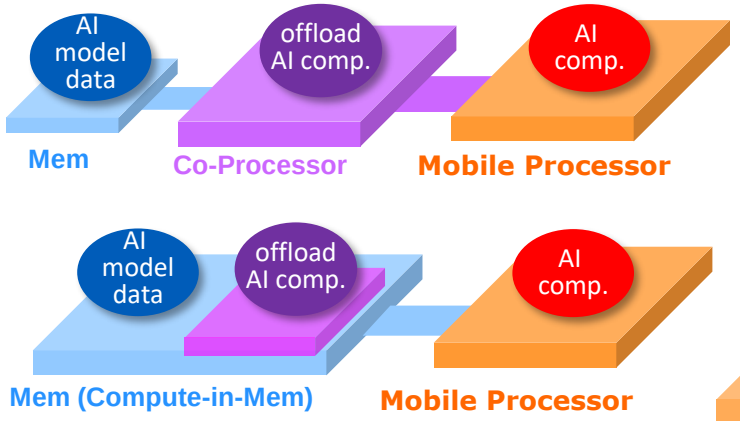
* diagrams for example only

Trends of AI Computing on Edge Devices

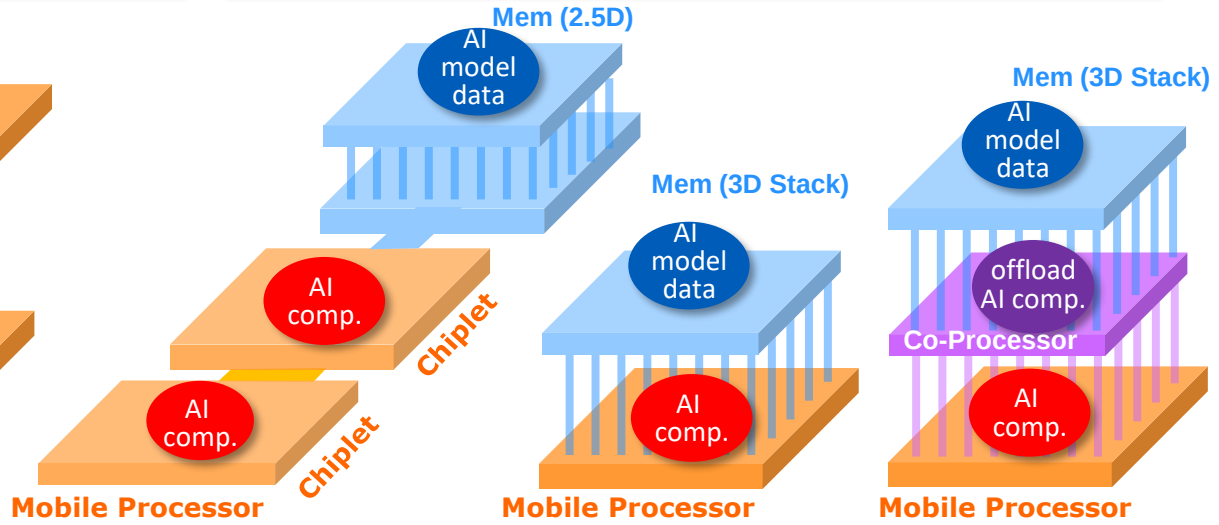
Reduce
Number Representation
Pruning/Sparsity
Smaller model by KD



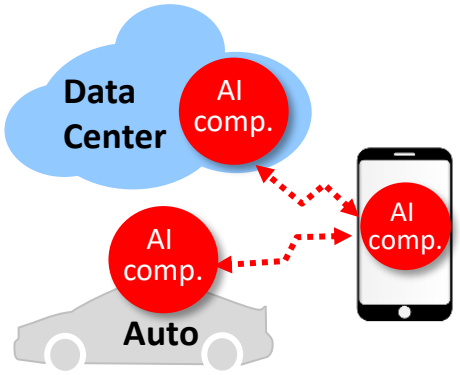
Offload
Co-processor
Compute in Memory



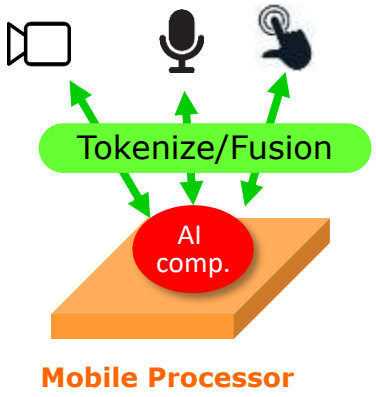
Scale-up, Scale-out
Chiplets
2.5D/3D Integration



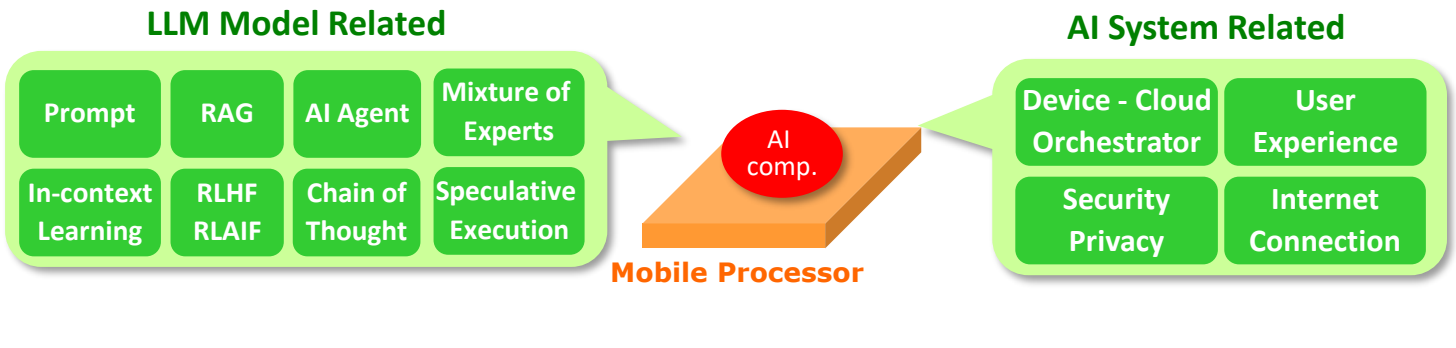
Collaboration



Multimodal
Fusion/Pre-Processing

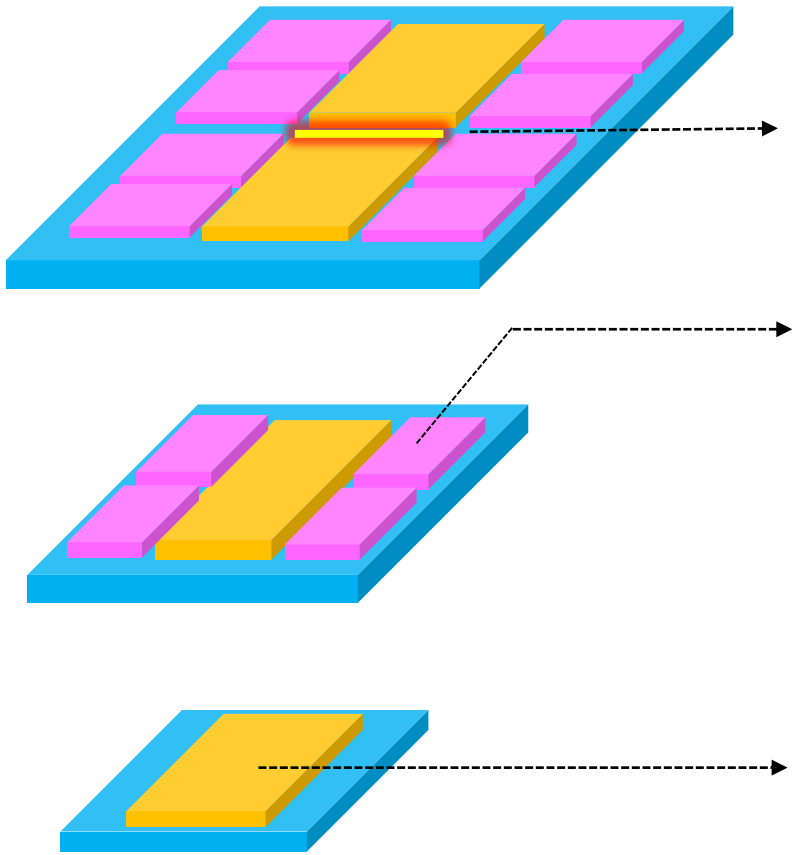


Model/System Architecture Optimization



Trends for Scale-up for AI Computing in Data Center

- Data Flow Architecture
- Compute in Memory
- Compute Near Memory
- Neuromorphic Computing
- Analog Computing
- Approximate Computing
- Quantum Computing / Quantum-Inspired Computing



more 3D IC Tech

3D SoC, 3D IC, Monolithic 3D
3D Heterogeneous Integration
Advanced Package (Wafer-level Package, Panel-level Packaging, ..)
Cooling Tech (chip-level, package-level, system-level,..)

Chiplets

- **Multi-Chiplets**
  - High-Speed Die-to-Die Interface
  - 2.5D/3D Package
  - Different Process for Different Chiplet
  - Cooling Tech for High Power

Package with Memory

- **Wide Data Bus + High-Speed Signal + 3D-Stacked**
  - HBM (High Bandwidth Memory)

Increase **Compute & Bandwidth**

But Need to take care **Package, SerDes Thermal Issues**

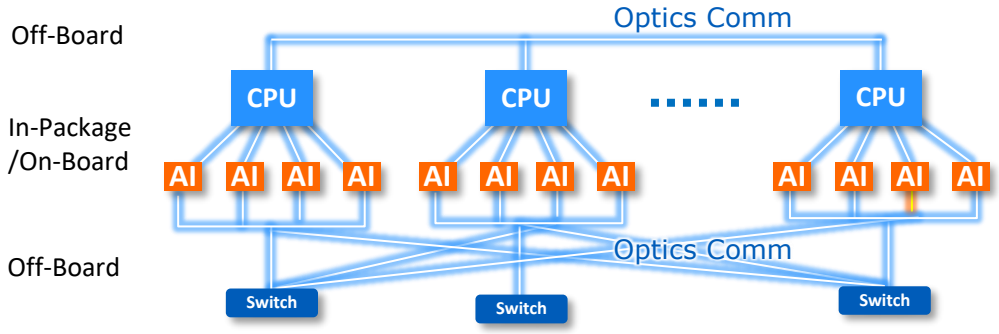
Scale-up on Single Die

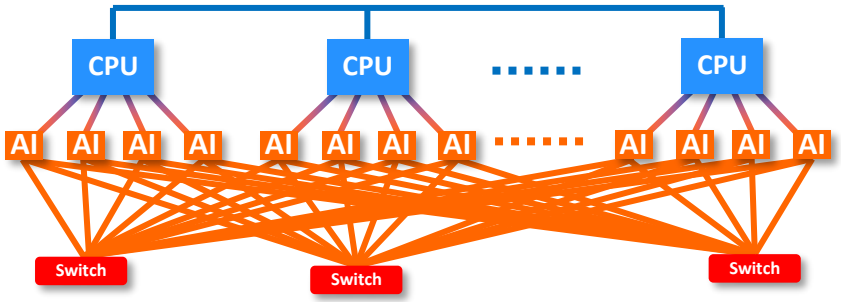
- **Semiconductor Process**
- **More Transistors**
  - Die Size ➔ Reticle Limitation
  - Compute Density ➔ Design Optimization
- **DSA (Domain Specific Architecture)**
  - Optimized for Matrix Multiplication
  - Number Representation
  - Sparsity

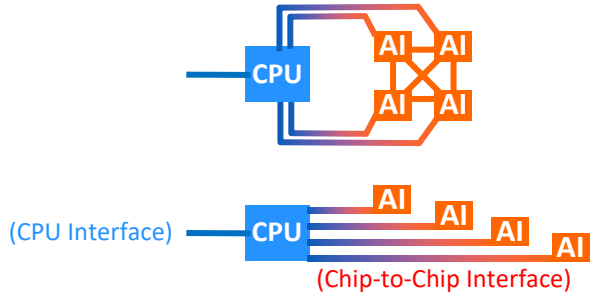
Increase **Compute**

But Require **More Bandwidth** and **System Software Optimization**

Trends for Scale-out for AI Computing in Data Center







* AI Chips can be GPU, TPU or ASIC

3D IC Tech

Optical Connection

- ☐ High Bandwidth with Low Power
- ☐ Co-Packaged Optics (CPO)
- ☐ Packaging, Fiber Cable, Reliability Elec-Optical Signals, SerDes,…

Large System Compute w/Low Power
But Need to Handle
Optical System Issues

High-Speed Chip-to-Chip Connection with Switch

- ☐ High-Speed High-Bandwidth Switch
- ☐ High-Efficiency Data Exchange in GPU (or TPU) subsystem
- ☐ Enable Larger-Scale System for LLM
- ☐ Flexibility for different Parallelism Schemes (Tensor Parallel, Pipeline Parallel, Expert Parallel, Data Parallel)

Increase
Large System Compute

But Problems on
Power, Thermal Issues

High-Speed Chip-to-Chip Connection

- ☐ CPU-GPU, GPU-GPU connection
- ☐ High-bandwidth, Low Latency
- ☐ Unified Memory Allocation

Increase
Compute Efficiency For Small Group GPUs (or TPUs)

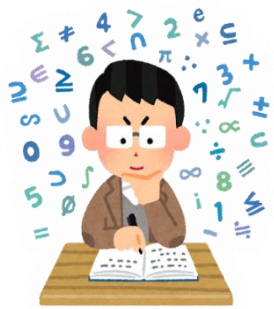
AI Chip Market (Forecasted in an article of *The Economist*)

- GPU or TPU with Software Stack
- Efficiency for AI Model Training
- Pre-Training, Post-Training, Evaluation
- **AI Model Ability & Time-to-Market**

- GPU, TPU or ASIC
- Efficiency for AI Inference
- Fine Tune, RAG, Chain-of-Thought, Agent
- **Inference Capability for Million Users & Energy Efficiency**

- Mobile Processor or ASIC
- Efficiency for AI Inference on devices
- User Interaction, Privacy, Multimodal, Agent
- **Inference with Smooth User Experience, Orchestrator of Device-Cloud Collaboration & Energy Efficiency**

Training (15%) | **Inference** (85%)



15% Training

45% Data Center Inference

40% Edge Device Inference

Data Center (60%) | **Edge Device** (40%)

- **Scale-up and Scale-out**
  to **max** the scale of overall compute and bandwidth for performance
  to **min** the energy of "compute per unit" and "comm per bit" for efficiency
- Heavily rely on the progress of semiconductor and package technologies
- **Affordable for data center companies. Sustainable for long-term operations**

- **Hardware/Software/System co-design**
  to collaborate edge and cloud resources to boost AI performance
  to offer smooth and swift user experience to use AI
- Heavily rely on domain specific design, and semi/package tech.
- **Affordable for end users. Penetrate to various AI applications**

- Ref: The Economist, "A Cambrian moment: AI has propelled chip architecture towards a tighter bond with software", Sep. 16th , 2024. / Pic Source : irasutoya.com

AI Computing on Edge Devices and Data Center

Edge Devices

AI Model

Billion-level Parameter Model

Advantage of Edge Devices

- Always Standby
- High Energy-Efficiency
- Privacy Personalization
- Realtime Response

Wireless Comm

5G

6G

Wi Fi

Data Center

AI Model

Pre-Trillion-level / Trillion-Level Parameter Model

Advantage of Cloud Computing

- Exa-scale Computing
- Scale-Out
- Massive Data
- Electricity