GISC 2025

Global ICT Standards Conference 2025

Physical AI

Computational Memory(PIM) SW Stack

Seungwon Lee
Master, Samsung Electronics

ICT Standards and Intellectual Property:
AI for All

Hosted by Ministry of Science and ICT, Ministry of Intellectual Property
Organised by National Radio Research Agency, TTA, IITP, KEA, kista, ETRI

Index

SAMSUNG

Seungwon Lee

VP of Technology(Master)
Samsung Electronics

Seungwon Lee is Vice President of Technology at Samsung Electronics, leading the Computing Software Lab at SAIT.

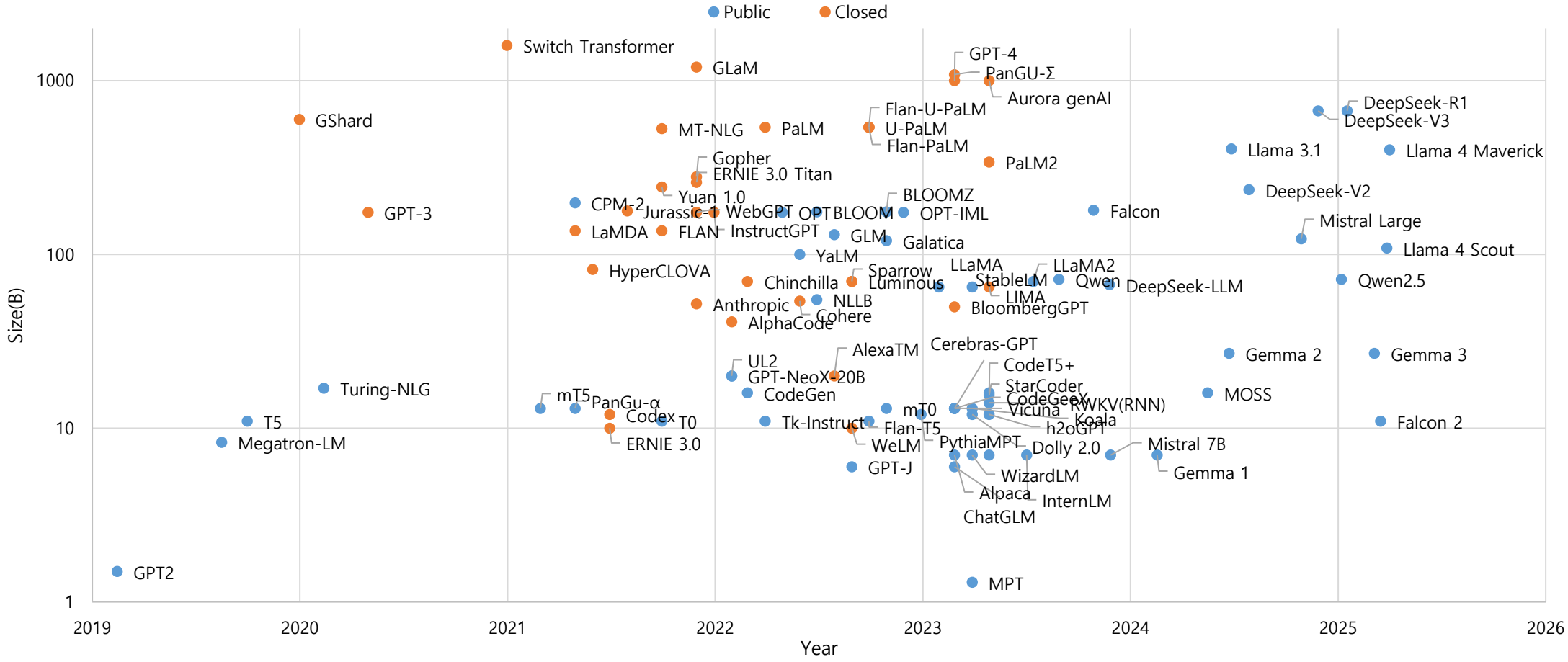
His research focuses on in-/near-memory computing for deep learning. Since 2018, he has led the development of the programming model and compiler for the PIM software stack.

He previously led compiler and quantizer development for Samsung's neural processor and managed compiler projects for in-house DSPs.

He holds a Ph.D. in Computer Science and Engineering from Seoul National University.

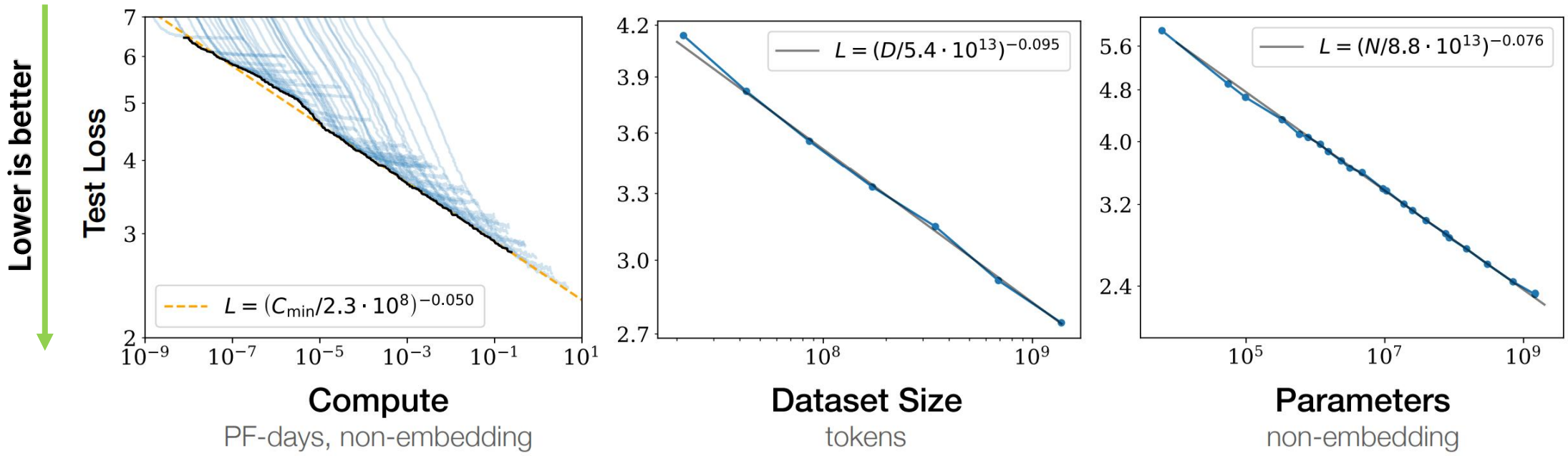
01. The Universe of LLMs

Since GPT-2, over 87 LLMs have been released, averaging 217 billion parameters.



02. Power-Law Scaling in Pre-training Language Models

> "Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute used for training. For optimal performance, all three factors must be scaled up in tandem." — Kaplan et al., 2020



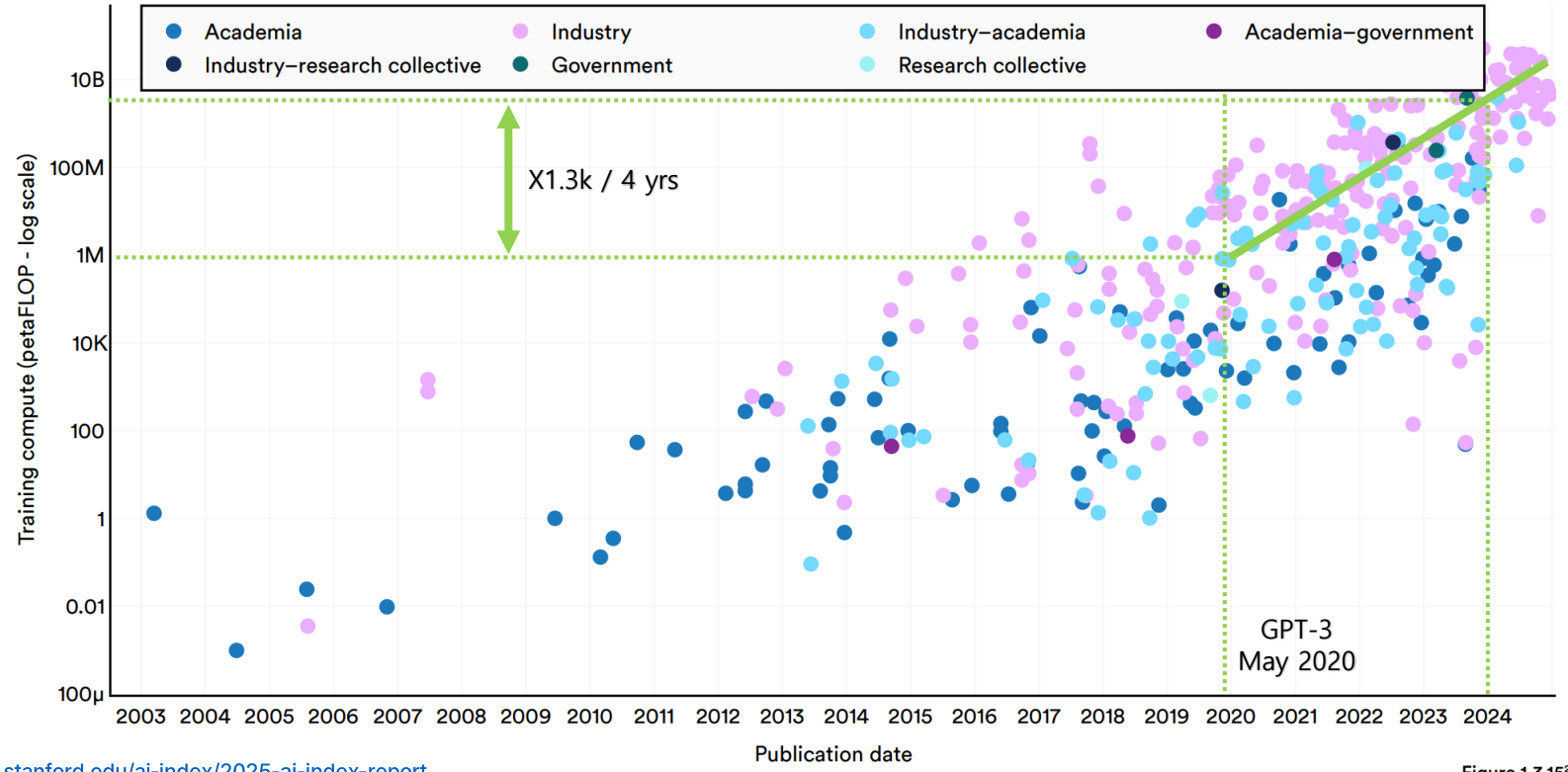
OpenAI, "Scaling Laws for Neural Language Models," arXiv 2020: https://arxiv.org/pdf/2001.08361

03. The Compute Boom and Industry Leadership

- Training compute for top models doubles every 4–5 months
- Recent industry-driven models account for 90% (ruled by scaling laws)

Training compute of notable AI models by sector, 2003–24

Source: Epoch AI, 2025 | Chart: 2025 AI Index report

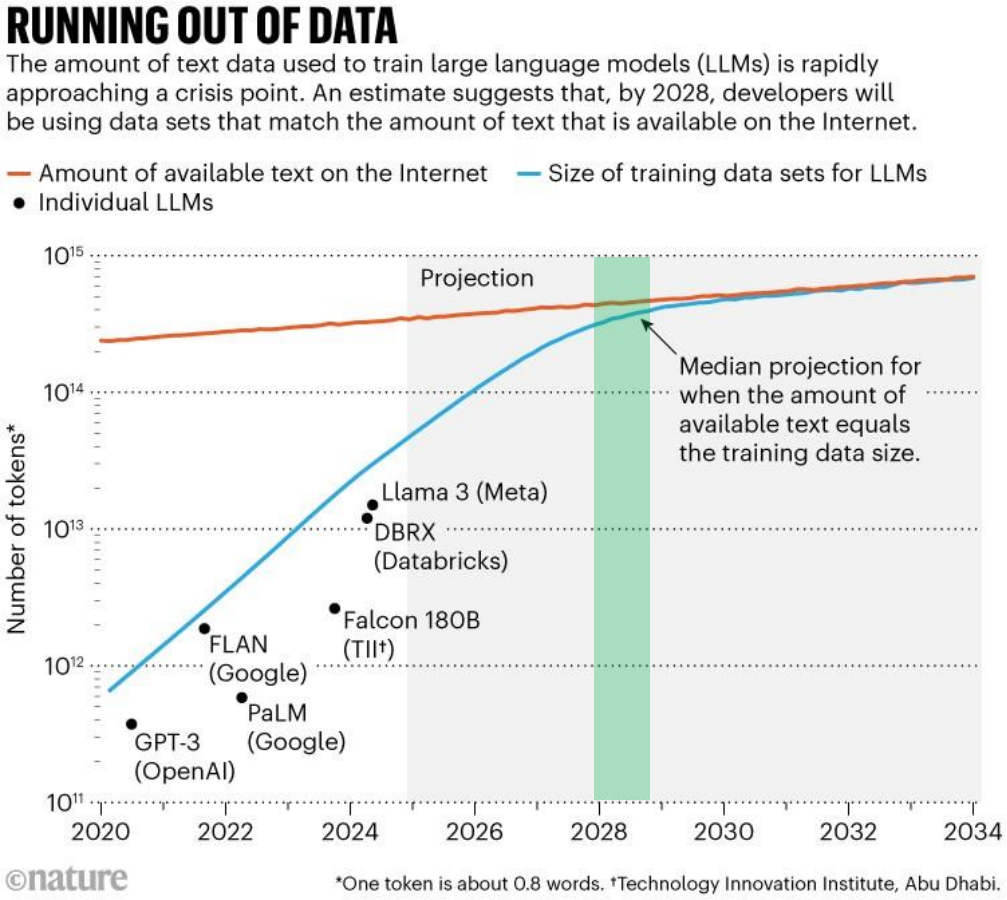


The 2025 AI Index Report: https://hai.stanford.edu/ai-index/2025-ai-index-report

Figure 1.3.15[23]

04. Training Data is Running Out

"We've achieved peak data and there will be no more.," NeurIPS, 13th Dec. 2024.




Feature

THE AI REVOLUTION IS RUNNING OUT OF DATA. WHAT CAN RESEARCHERS DO?

AI developers are rapidly picking the Internet clean to train large language models such as those behind ChatGPT. Here's how they are trying to get around the problem. By Nicola Jones

The Internet is a vast ocean of human knowledge, but it isn't infinite. And artificial intelligence (AI) researchers have nearly sucked it dry.

The past decade of explosive improvement in AI has been driven in large part by making neural networks bigger and training them on ever-more data. This scaling has proved surprisingly effective at making large language models (LLMs) – such as those that power the chatbot ChatGPT – both more capable of replicating conversational language and of developing emergent properties such as reasoning. But some specialists say that we are now approaching the limits of scaling. That's in part because of the ballooning energy requirements for computing. But it's also because LLM developers are running out of the conventional data sets used to train their models.

A prominent study[1] made headlines this year by putting a number on this problem: researchers at Epoch AI, a virtual research institute, projected that, by around 2028, the typical size of data set used to train an AI model will reach the same size as the total estimated stock of public online text. In other words, AI is likely to run out of training data in about four years' time (see 'Running out of data'). At the same time, data owners – such as newspaper publishers – are starting to crack down on how their content can be used, tightening access even more. That's causing a crisis in the size of the 'data commons', says Shayne Longpre, an AI researcher at the Massachusetts Institute of Technology in Cambridge who leads the Data Provenance Initiative, a grass-roots organization that conducts audits of AI data sets.

The imminent bottleneck in training data could be starting to pinch. "I strongly suspect that's already happening," says Longpre.

"I DON'T THINK ANYONE IS PANICKING AT THE LARGE AI COMPANIES."

Although specialists say there's a chance that these restrictions might slow down the rapid improvement in AI systems, developers are finding workarounds. "I don't think anyone is panicking at the large AI companies," says Pablo Villalobos, a Madrid-based researcher at Epoch AI and lead author of the study forecasting a 2028 data crash. "Or at least they don't e-mail me if they are."

For example, prominent AI companies such as OpenAI and Anthropic, both in San Francisco, California, have publicly acknowledged the issue while suggesting that they have plans to work around it, including generating new data and finding unconventional data sources. A spokesperson for OpenAI, told *Nature*: "We use numerous sources, including publicly available data and partnerships for non-public data, synthetic data generation and data from AI trainers."

Even so, the data crunch might force an upheaval in the types of generative AI model that people build, possibly shifting the landscape away from big, all-purpose LLMs to smaller, more specialized models.

Trillions of words

LLM development over the past decade has shown its voracious appetite for data. Although some developers don't publish the specifications of their latest models, Villalobos estimates that the number of 'tokens', or parts of words, used to train LLMs has risen 100-fold since 2020, from hundreds of billions to tens of trillions.

That could be a good chunk of what's on the Internet, although the grand total is so vast that it's hard to pin down – Villalobos estimates the total Internet stock of text data today at 3,100 trillion tokens. Various services use web crawlers to scrape this content, then eliminate duplications and filter out undesirable content (such as pornography) to produce cleaner data sets: a common one called RedPajama contains tens of trillions of words. Some companies or academics do the crawling and cleaning themselves to make bespoke data sets to train LLMs. A small proportion of the Internet is considered to be of high quality, such as human-edited, socially acceptable text that might be found in books or journalism.

The rate at which usable Internet content is increasing is surprisingly slow: Villalobos's paper estimates that it is growing at less than 10% per year, while the size of AI training data sets is more than doubling annually. Projecting these trends shows the lines converging around 2028.

At the same time, content providers are increasingly including software code or refining their terms of use to block web crawlers or AI companies from scraping their data for training. Longpre and his colleagues

290 | Nature | Vol 636 | 12 December 2024

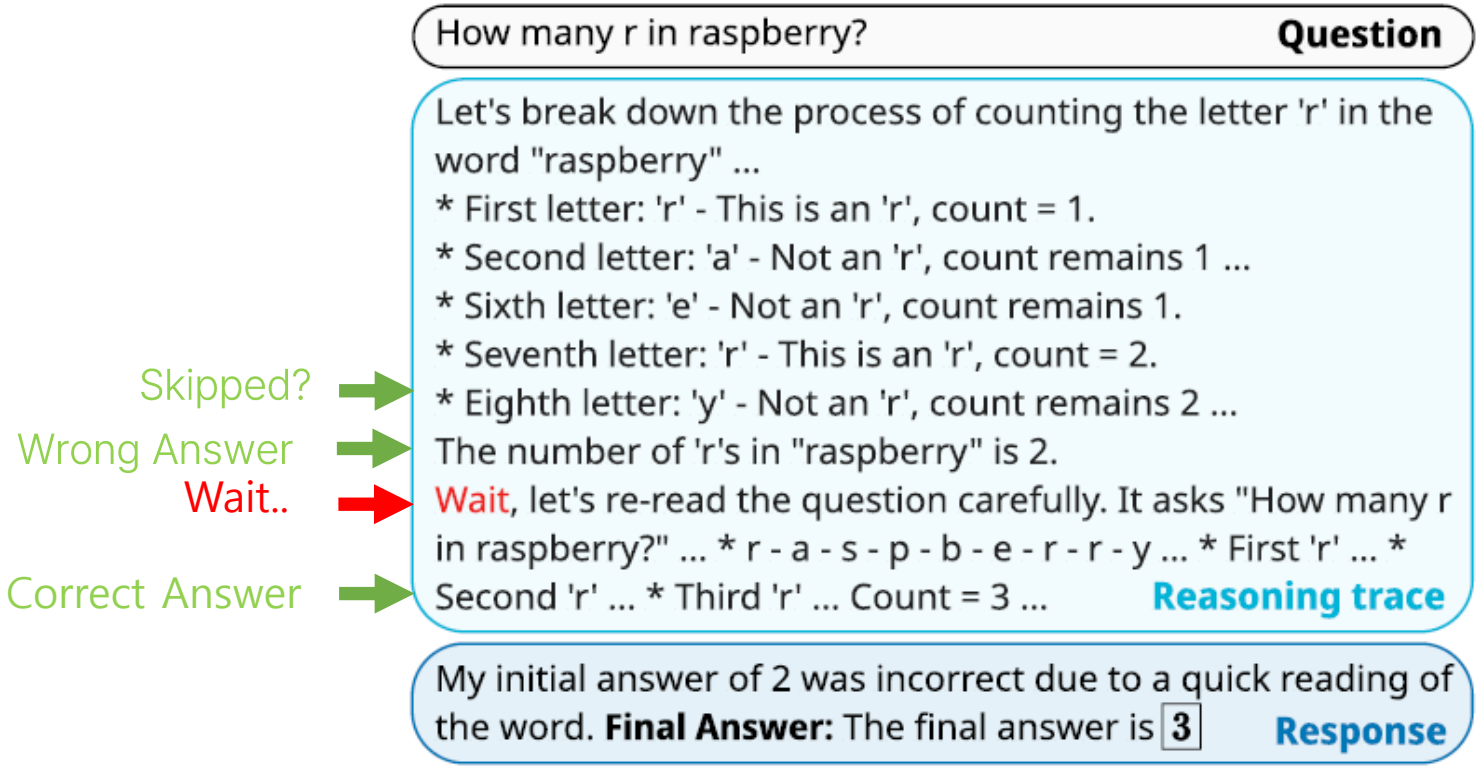
Ilya Sutskever
OpenAI Co-Founder

https://www.nature.com/articles/d41586-024-03990-2

05. Test-Time Scaling: Simple Yet Powerful

During inference, achieving OpenAI o1-preview level with just Budget Forcing

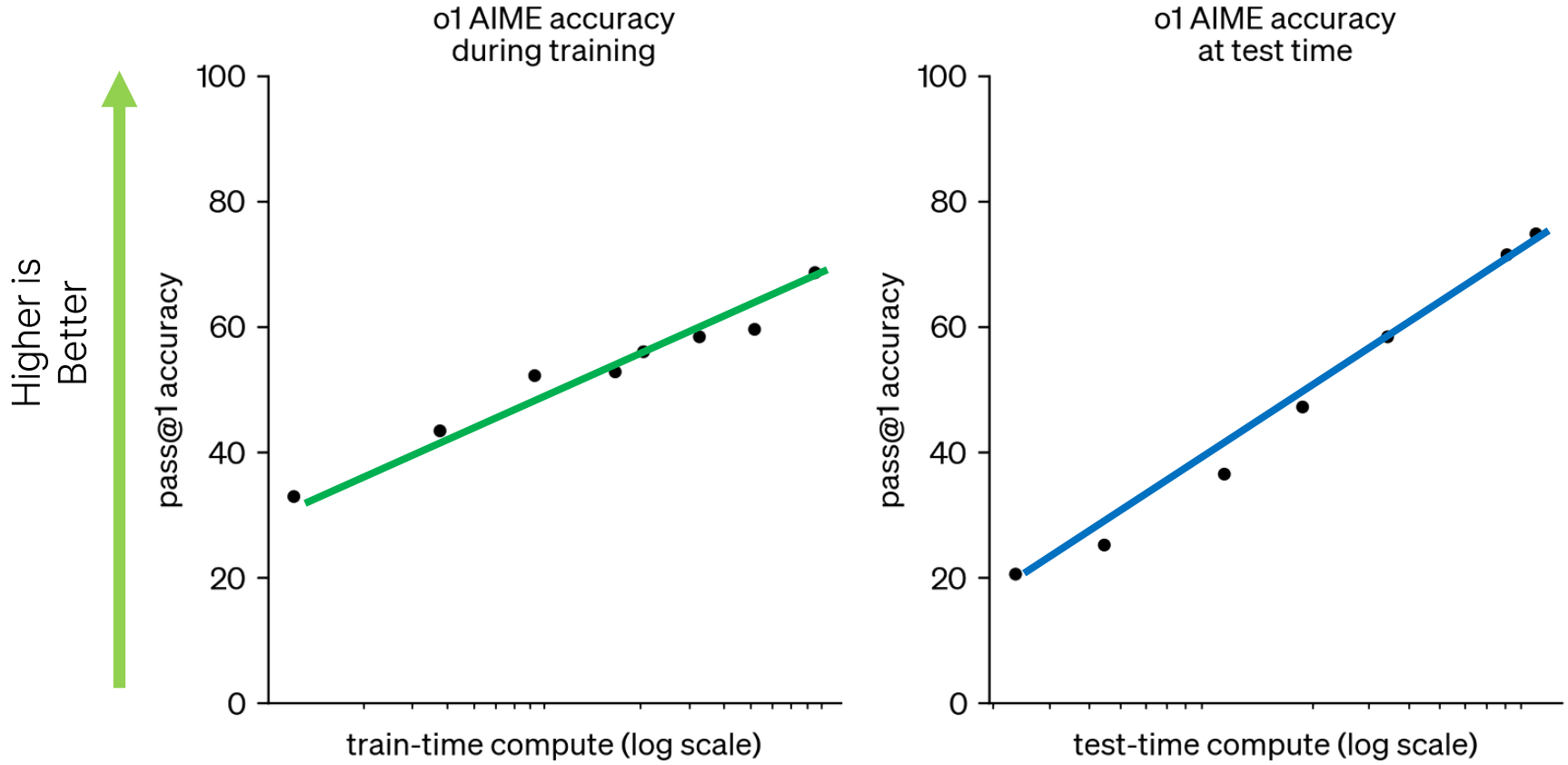
- If the model stops too early, prompt it to continue reasoning by appending "Wait"
- If necessary, forcibly end internal thinking with an end-of-thinking token



s1: Simple test-time scaling, arXiv 2020: https://arxiv.org/abs/2501.19393

06. Training Scaling vs Test-Time Scaling — OpenAI

On the math benchmark, Test-Time Scaling has higher performance improvement efficiency



https://www.forethought.org/research/inference-scaling-reshapes-ai-governance?utm_source=chatgpt.com

07. Reasoning Inference

Computing Paradigm Shift

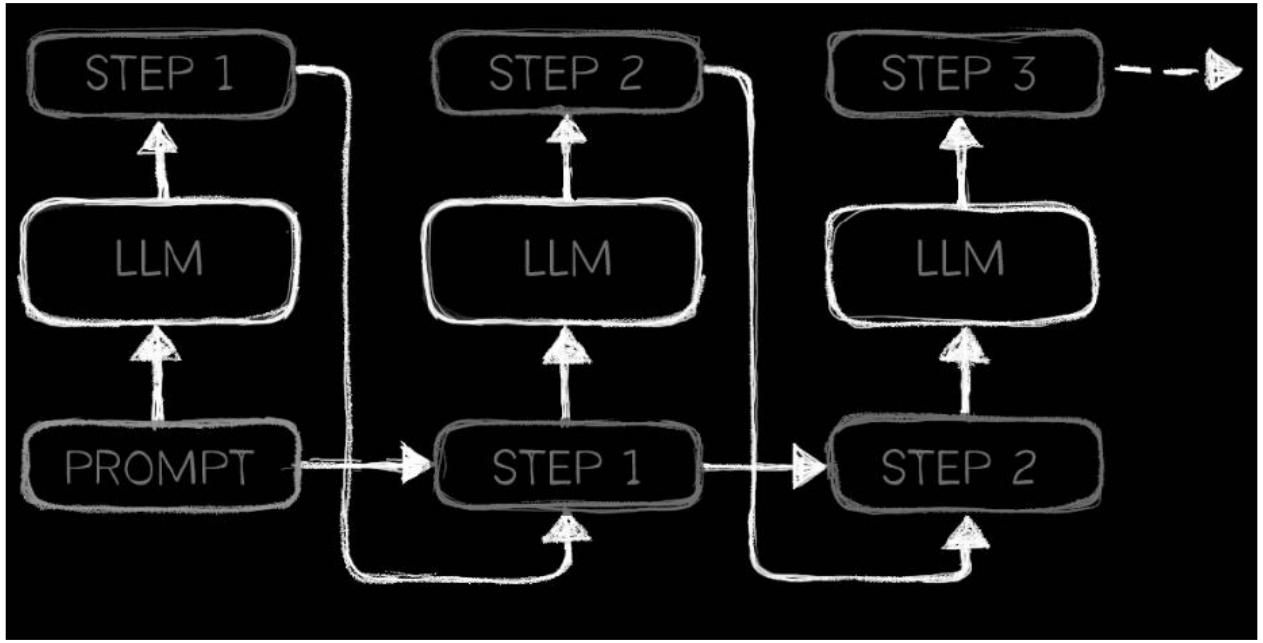
Accuracy improvement from Train → Accuracy improvement from inference

☞ Reasoning tasks (math, code, legal) demand >100× inference compute.

- **Mitigation of training data depletion issue**
- **Improvement of reliability and explain-ability**

Application areas

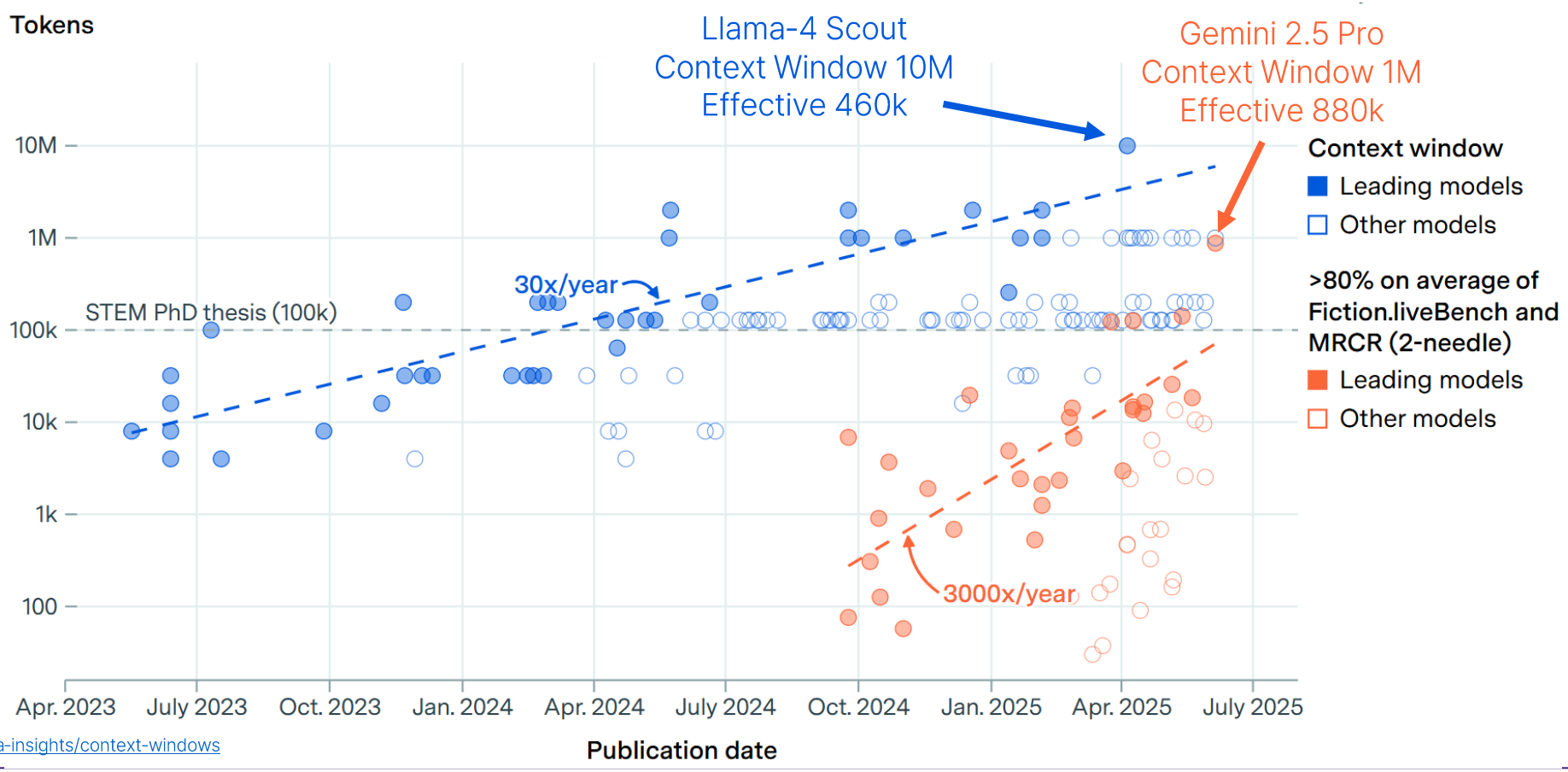
- Mathematics/Science problem solving
- Code generation and debugging
- Medical/Legal reasoning
- Knowledge-based Q&A
- Multi-stage decision-making



08. The Context Explosion

Context window sizes are exploding—KV cache demand is growing 30× per year

☞ Effective Context length increases x3000/year



https://epoch.ai/data-insights/context-windows

09. Physical AI

> "The next wave of AI is Physical AI." — Jensen Huang, GTC 2024.

Robot System

Cosmos: The World Foundation Model Platform

NVIDIA/**Cosmos**

New repo collection for NVIDIA Cosmos:
https://github.com/nvidia-cosmos

NVIDIA

5 Contributors | 0 Issues | 8k Stars | 519 Forks

https://www.youtube.com/watch?v=9Uch931cDx8

10. World Models require massive memory and compute for simulation.

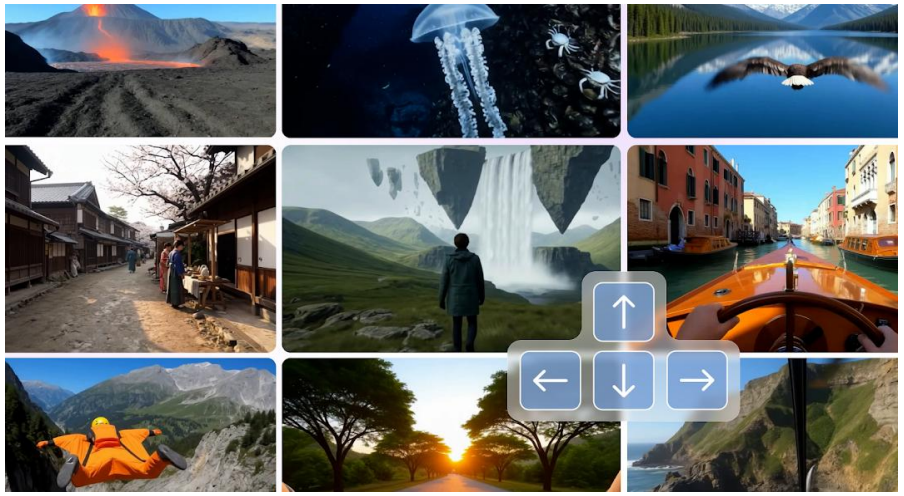
Memory Intensive

- State and space storage: need to preserve 3D environment, physics simulation, etc., for a long period
- KV cache + World State merging → results in higher memory demand

Compute Intensive

- Simulation computation: perform physics law calculations at each step
- Multi-modal Inference, Real-time interactivity

Google Genie3: State-of-the-Art World Model



https://www.youtube.com/watch?v=PDKhUknuQDg

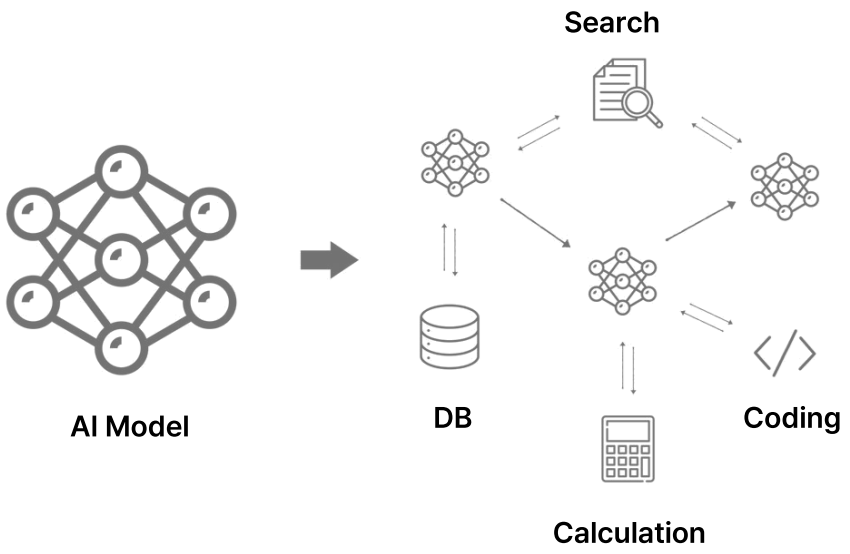
11. Different Workloads, Different Demands

Training	Inference	World Model
Compute-Intensive - Large batch size - Long training epochs	Latency-sensitive - Small batch size - Fast response for users	Compute + Memory heavy - Physics & simulation steps - Multimodal reasoning
Moderate memory - Fits large batches	Memory-bound - KV Cache dominates - Context window growth	Memory footprint explodes - World state + KV Cache
PB-scale Storage - Sequential access	Light storage - Model weights & caching	Massive synthetic data - Simulation logs - Multimodal inputs
High-bandwidth training - Low-latency - All reduce	High-throughput serving - Massive concurrent user request	Real-time streaming - High BW & low-latency - Real-time multi-agent

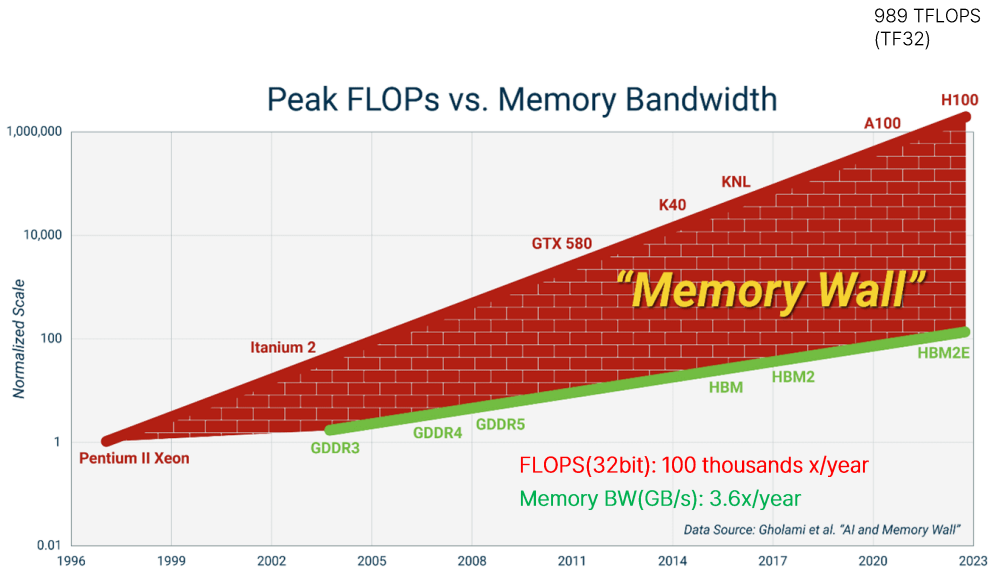
01. System Complexity & Memory Wall

- Repeated composition between AI Model and Micro Service(SW): Compound AI System
- The growth of memory bandwidth is much slower than that of model size
  → Emergence of 'Computational memory' that utilizes internal memory bandwidth

Compound AI System



The deepening of the "Memory Wall"



02. "Computational" Memory is Rising

Rising technology to overcome memory bandwidth bottleneck

- *Leverage internal bandwidth* to achieve much higher throughput
- Perform the *computation in-place*, cutting data movement and boosting *energy efficiency*

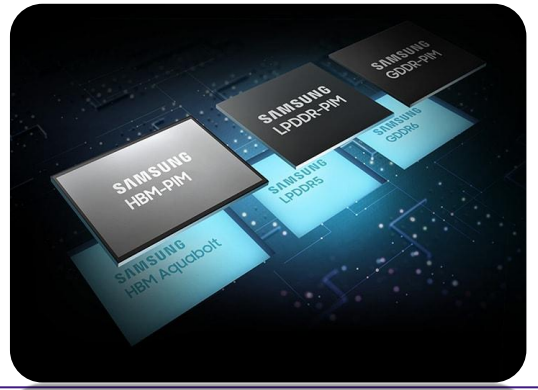
Examples

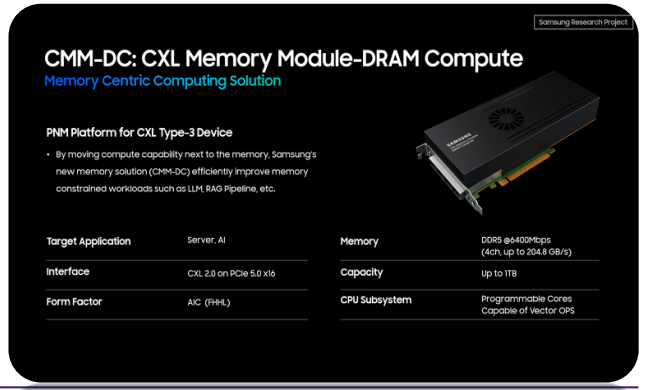
- Samsung
  - PIM(Processing In Memory): HBM2-PIM(POC w/ AMD MI100 GPU), LPDDR-PIM
  - NDP(Near Data Processing): CXL-PNM(CMM-DC)

- Marvell
  - CXL near-memory accelerator: Structera A 2504

- SK Hynix
  - PIM: GDDR6-AiM
  - PIM-based accelerator: AiMX

Samsung PIM & CXL-PNM

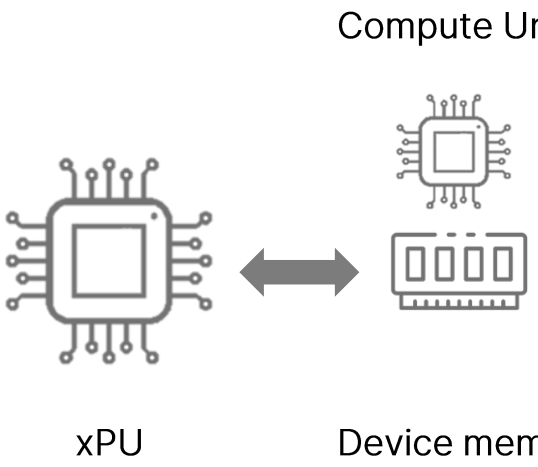




03. Memory vs Accelerator Mode

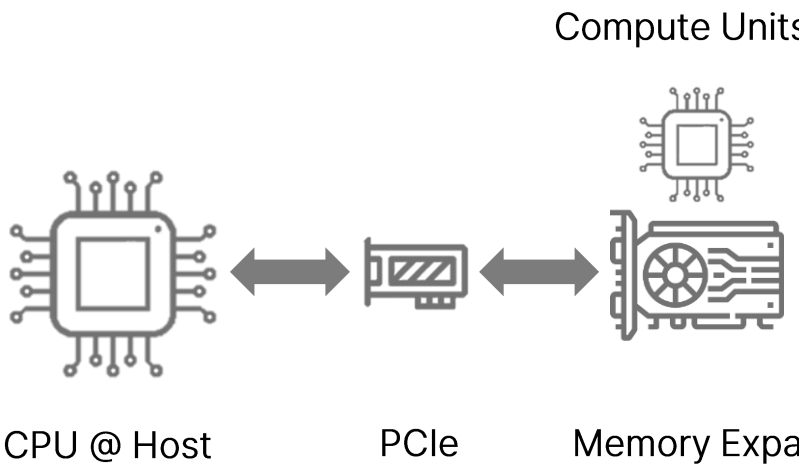
"Computational Memory"-as-a-Memory

- Computational memory is controlled to xPU-device
- Controlled by xPU's programming model (eg. r/w cmds)
  - Need to modify xPU's SW to use a computational memory
- Eg. Samsung HBM/LPDDR-PIM, SK Hynix GDDR6-AiM

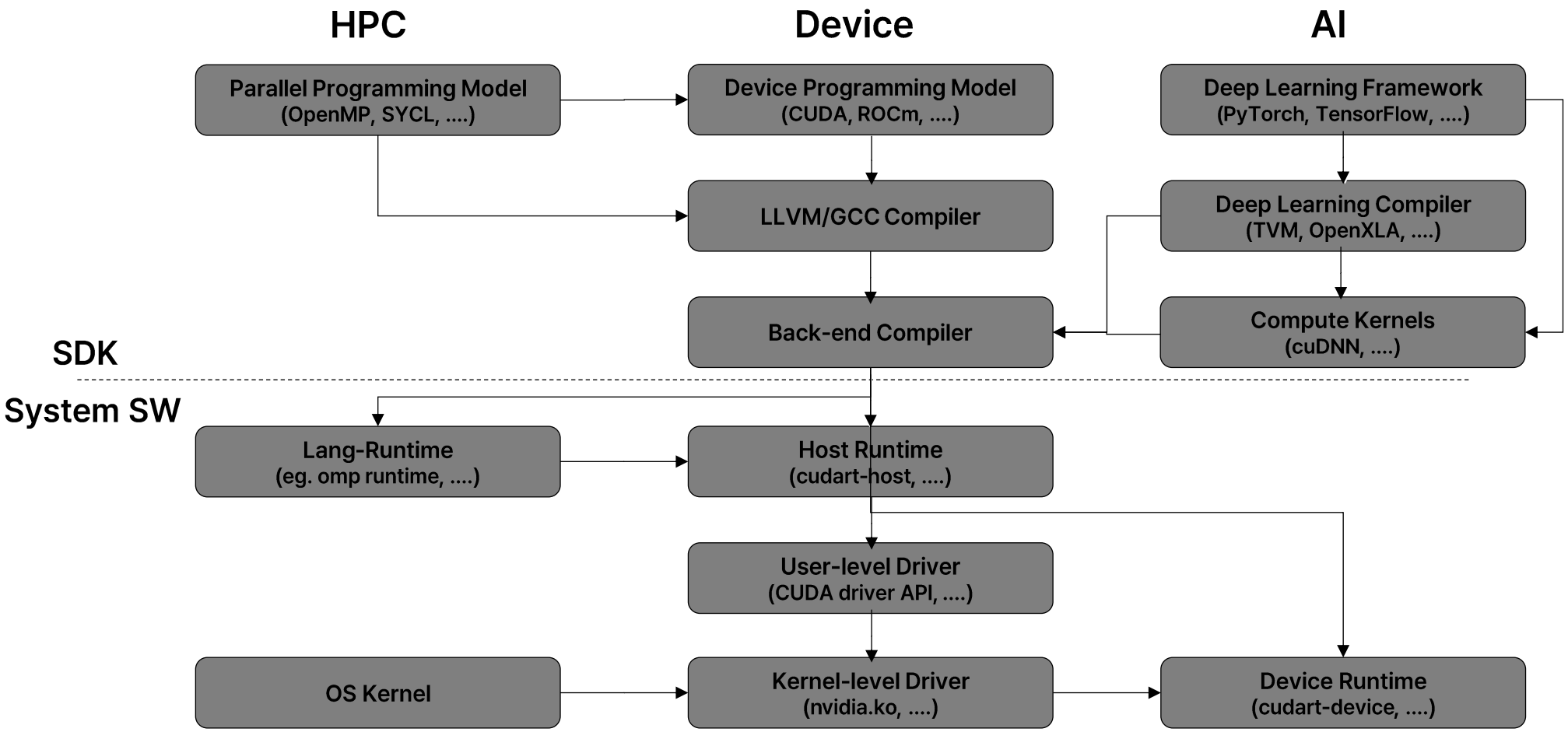


"Computational Memory"-as-a-Accelerator

- Computational memory is shown as a kind of accelerators
- Controlled by host-side programming model (eg. APIs)
  - Need to expand host-side APIs for a new compute device
- Eg. Samsung CMM-DC, SK Hynix AiMX, Marvell Structera A 2504

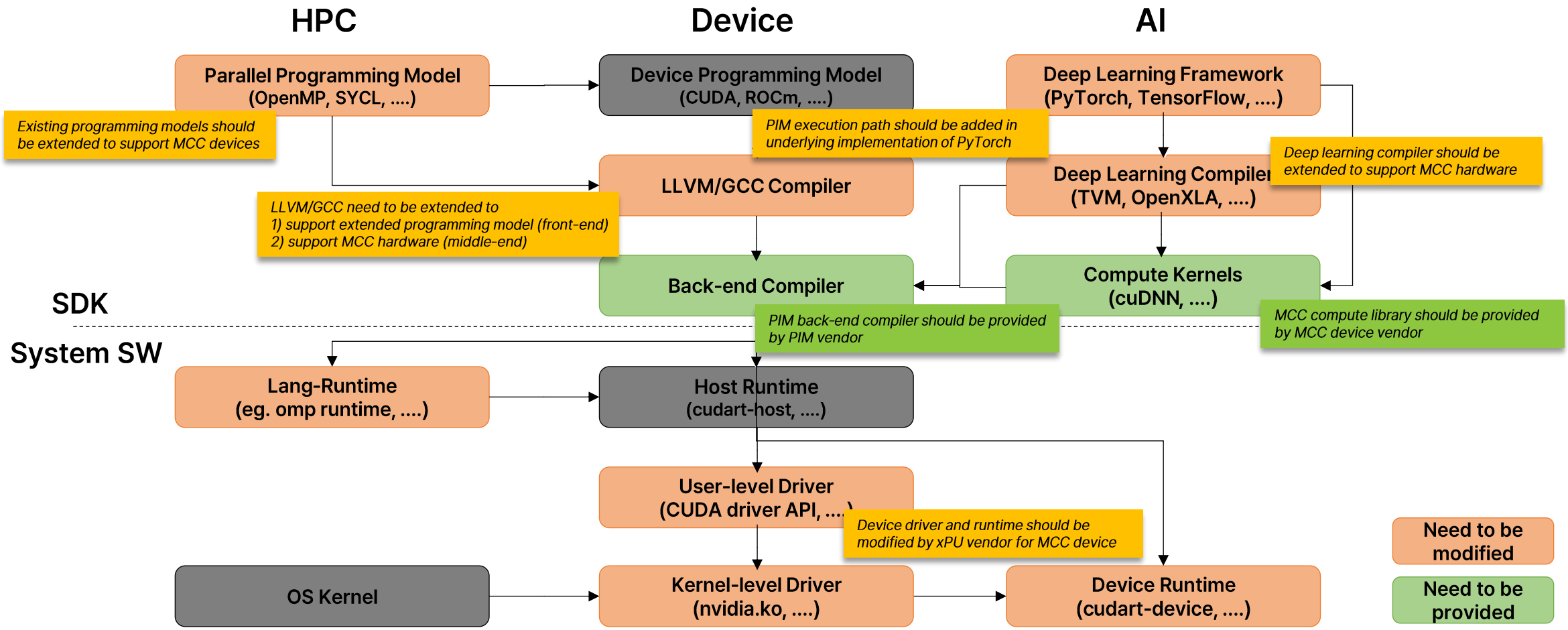


04. Legacy SW Stack for AI and HPC



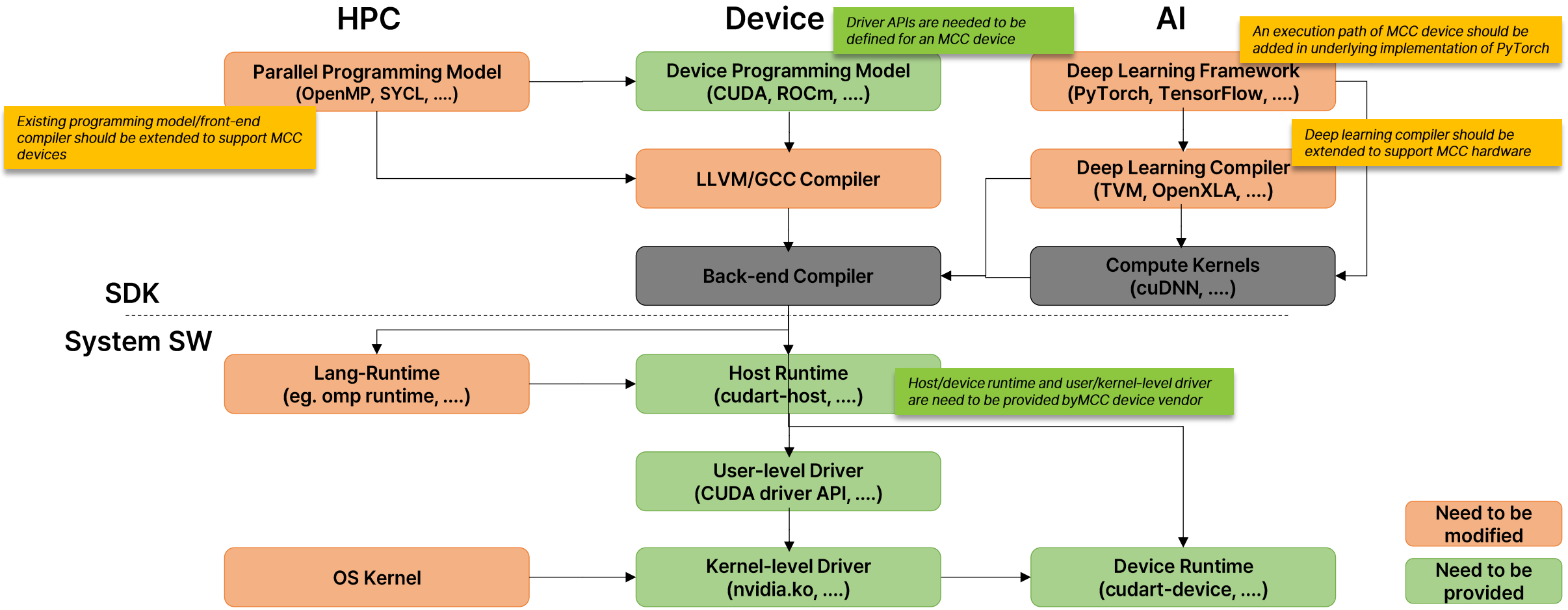
05. SW Stack: “Computational Memory”-as-a-Memory

xPU venders for PIM need to modify their drivers and runtimes



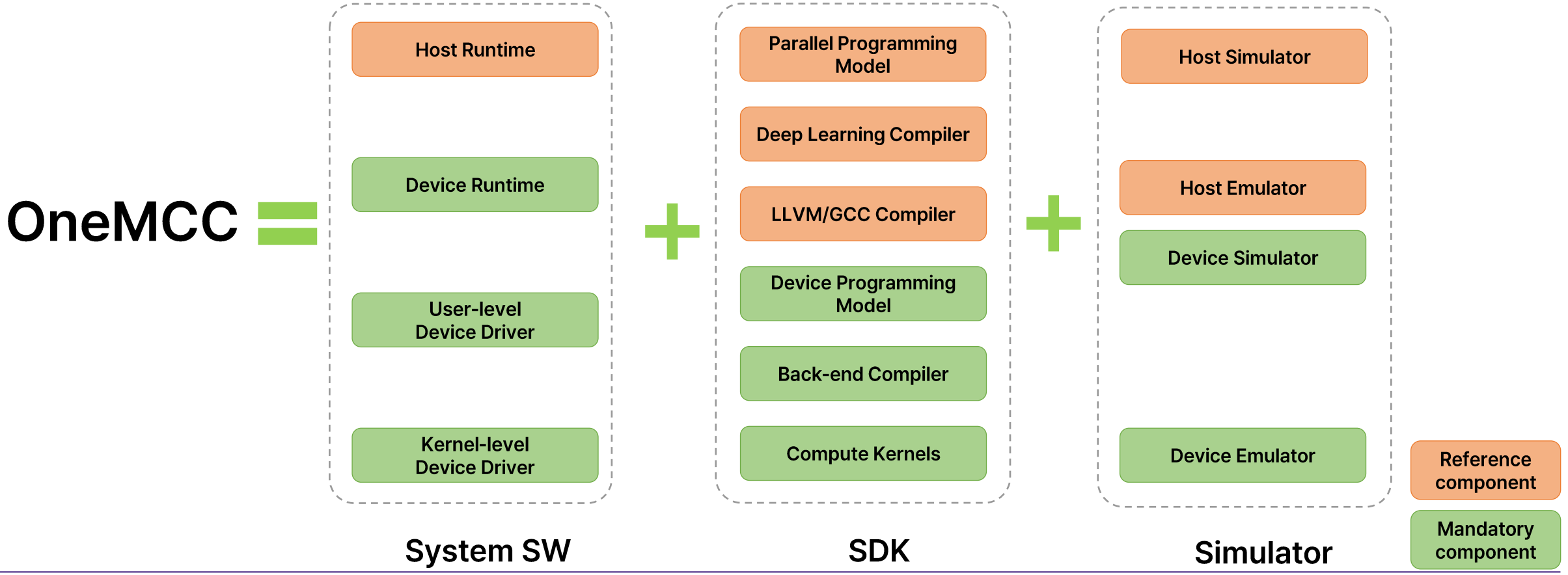
06. SW Stack: “Computational Memory”-as-a-Accelerator

Memory vender need to provide Host/device runtime and user/kernel level driver



07. OneMCC Overview

OneMCC: A unified SW interface for “Memory-Centric Computing”



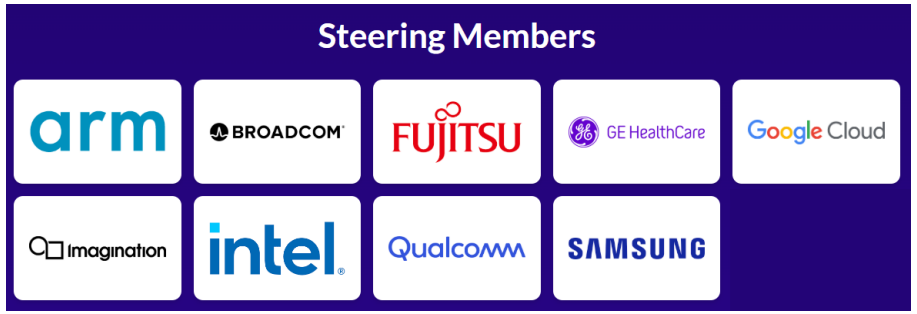
08. oneAPI Overview

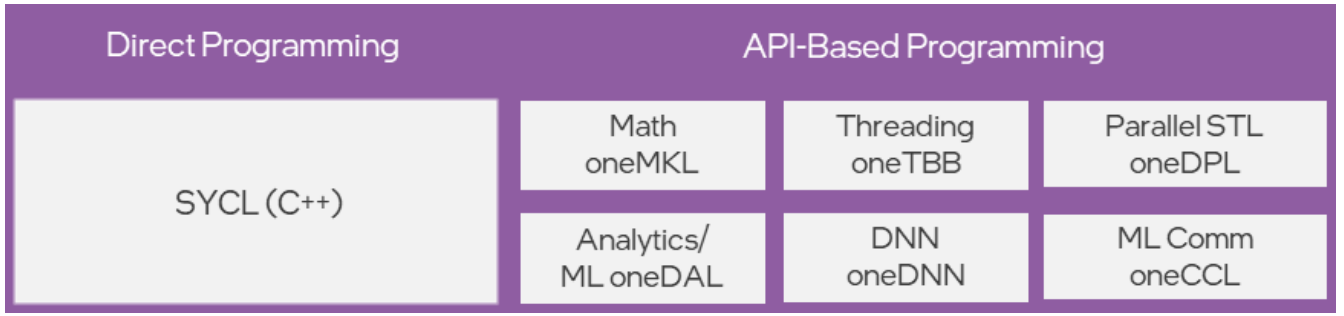
Industry efforts for heterogeneous accelerating devices

- OpenCL (Khronos Group, ‘08) → SYCL (Khronos Group, ‘14) → DPC++ (Intel, ‘18) → oneAPI (Intel, ‘20) → UXL Foundation (Intel, ARM, Qualcomm, Samsung, etc., ‘23~)

UXL(Unified Acceleration): under Linux Foundation

- A non-profit consortium that promotes an open standard ecosystem for heterogeneous computing through oneAPI
- A unified programming model not limited to a specific manufacturer (Open-source S/W)
- “vendor-neutral software ecosystem for multi-architecture accelerated computing” - UXL

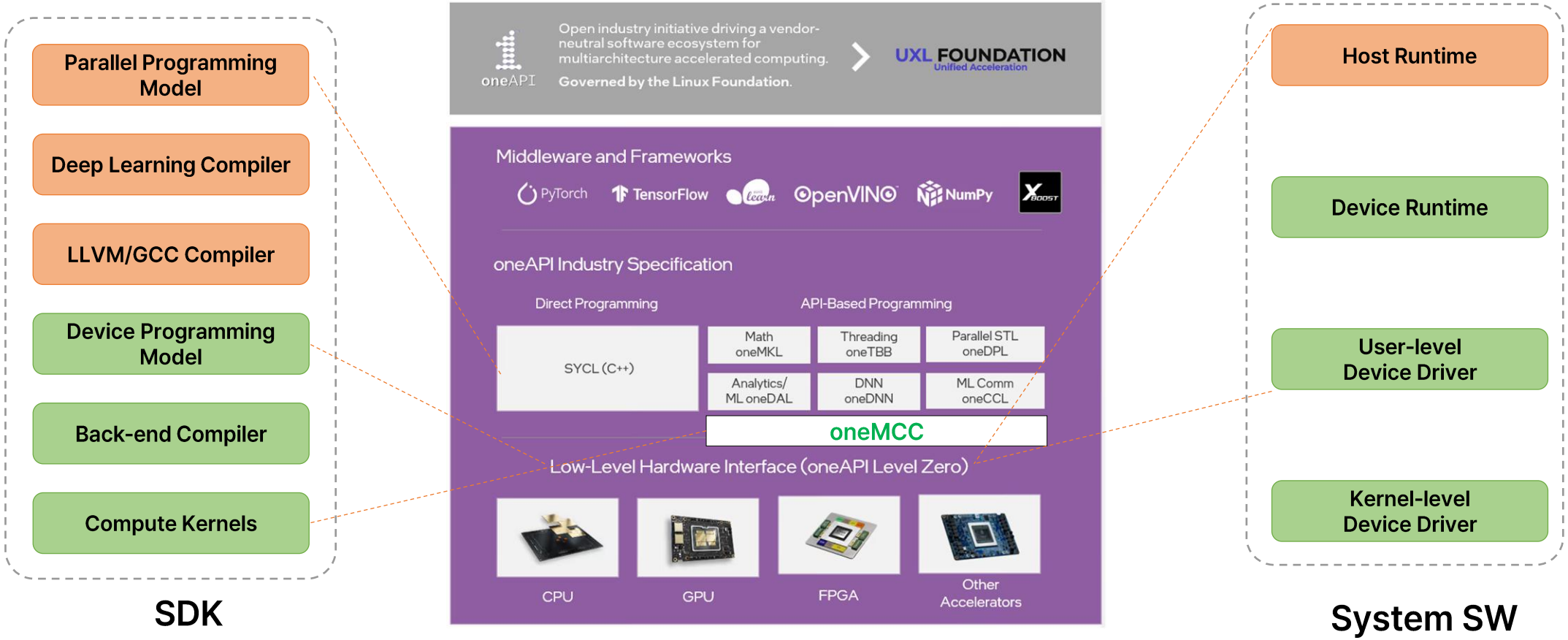




09. Toward an open specification for OneMCC

There is plan to define OneMCC specification within the UXL Foundation

☞ Open to collaboration with organizations(Khronos, …) to standardize OneMCC



GISC 2025

Global ICT Standards Conference 2025

- Thank you -

Seungwon Lee, Master, Samsung Electronics

seungw.lee@samsung.com

ICT Standards and Intellectual Property:
AI for All

Hosted by Ministry of Science and ICT, Ministry of Intellectual Property

Organised by National Radio Research Agency, TTA, IITP, KEA, kista, ETRI