SAPEON

AI beyond server-scale capability

2024 April

서 웅, R&D 센터

01.
Gen AI 시장 Trend

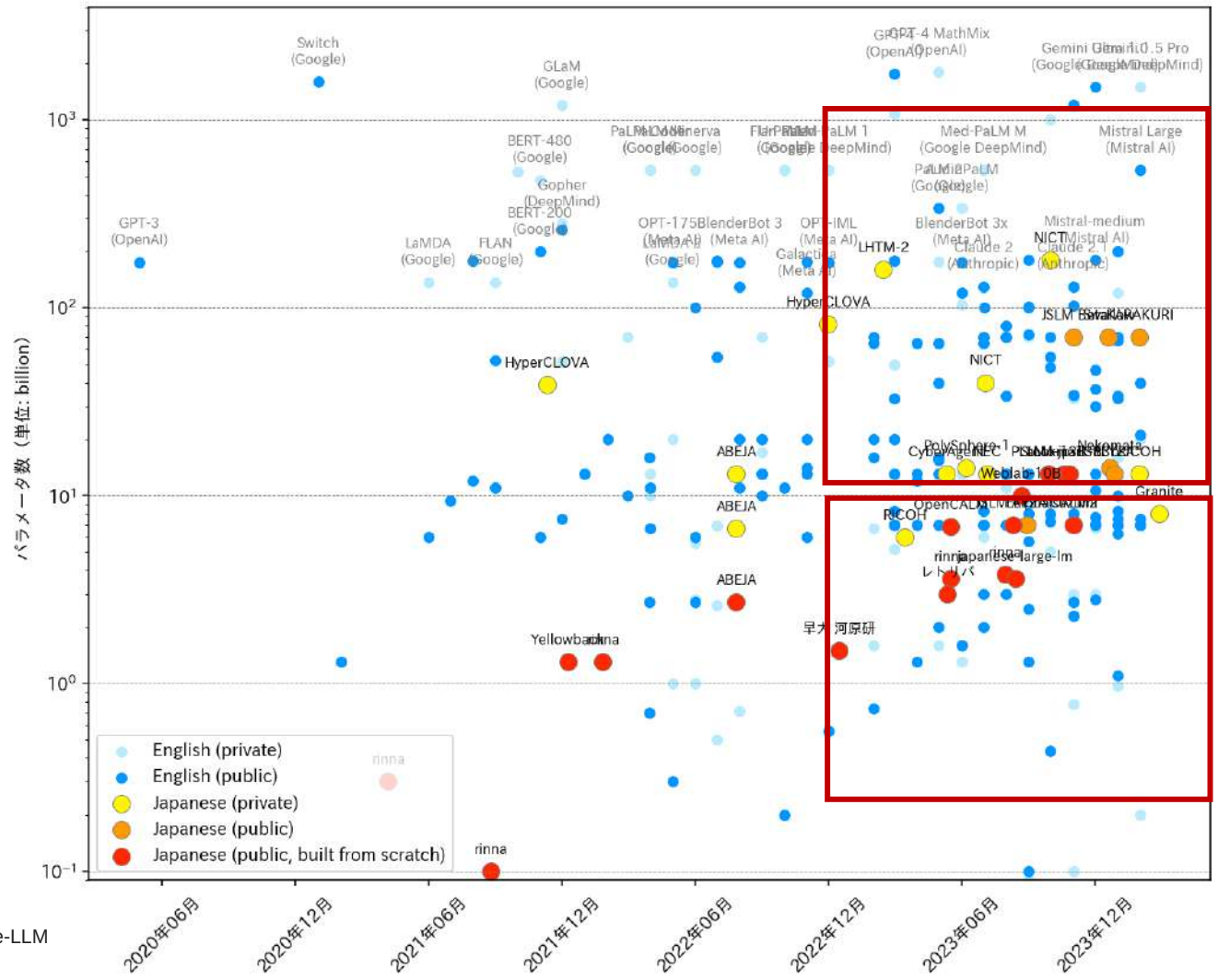
- Technological Innovation, Pioneering
- Economic Viability of Services, Vertical Solution
- Cloud Computing / Datacenter, On-device Computing

Text-to-video Generation, Sora by Open AI

Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a
black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk a bout.

Source: Open AI

Is LLM parameter size always growing?

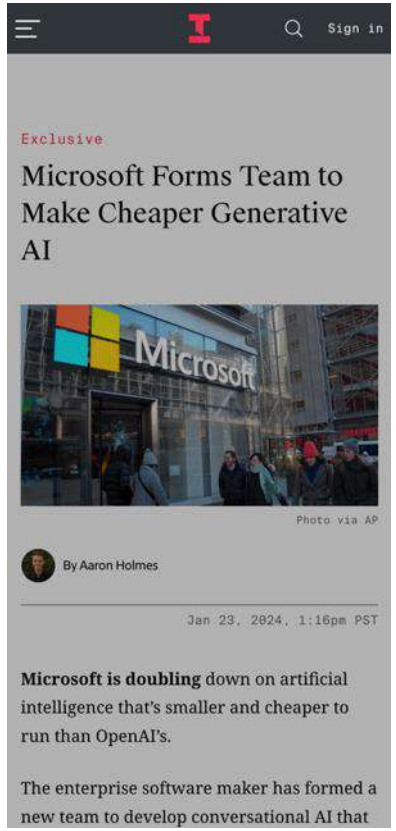


Front-line / Resource-heavy LLMs

Equal progress in small size d LLMs

Source: Awesome-Japanese-LLM (Opensource Project)

Microsoft focuses on small LLMs


Sign in

Exclusive

Microsoft Forms Team to Make Cheaper Generative AI

Photo via AP

By Aaron Holmes

Jan 23, 2024, 1:16pm PST

Microsoft is doubling down on artificial intelligence that's smaller and cheaper to run than OpenAI's.

The enterprise software maker has formed a new team to develop conversational AI that

The GenAI team develops small language models. They mimic the quality of LLMs such as OpenAI's GPT-4—which powers ChatGPT and Microsoft's AI Copilots—while using far less co mputing power. …

Those researchers gained early momentum developing **SLMs such as Phi, a family of open-source models small enough to run on mobile devices but capable of replicating the quality of GPT-4 for certain tasks.** To reach that goal, the researchers last year used GPT-4 to generate millions of tracts of high-quality text and trained Phi on those data so it would mimic the larger model. (Other companies have also been using GPT-4 to produce data to train new models.)

Phi made waves in the AI research community, and Microsoft has since made Phi-2, the latest version of the model, available as an open-source model for Azure customers who want to use it to b uild their own AI applications. **Firms such as Goldman Sachs have been testing Phi in recent months. And Microsoft has already been looking for ways to use SLMs to handle more rudimentary queries from users of the Bing AI chatbot and Windows Copilot, thus cutting back on compute costs.**

…

Diversified LLM size for different purposes

Source: The Information

The larger, the better? Not always

Model	Size	BBH	Commonsense Reasoning	Language Understanding	Math	Coding
Llama-2	7B	40.0	62.2	56.7	16.5	21.0
	13B	47.8	65.0	61.9	34.2	25.4
	70B	66.5	69.2	67.6	64.1	38.3
Mistral	7B	57.2	66.4	63.7	46.4	39.4
Phi-2	2.7B	59.2	68.8	62.0	61.1	53.7

Table 1. Averaged performance on grouped benchmarks compared to popular open-source SLMs.

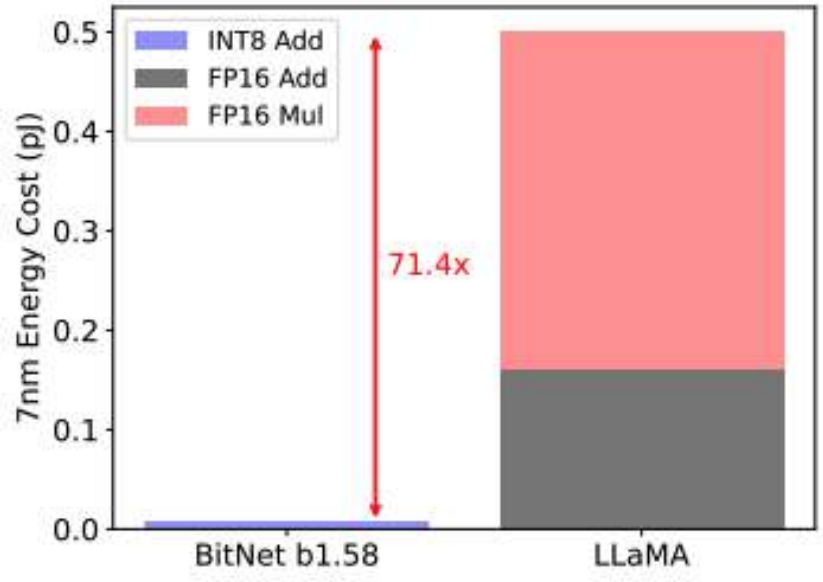
Model	Size	BBH	BoolQ	MBPP	MMLU
Gemini Nano 2	3.2B	42.4	79.3	27.2	55.8
Phi-2	2.7B	59.3	83.3	59.1	56.7

Table 2. Comparison between Phi-2 and Gemini Nano 2 Model on Gemini's reported benchmarks.

Source: Microsoft Research, Li, Yuanzhi, et al. "Textbooks are all you need ii: phi-1.5 technical report." *arXiv preprint arXiv:2309.05463* (2023).

How small can it go? : 1bit LLM

71.4x energy gain



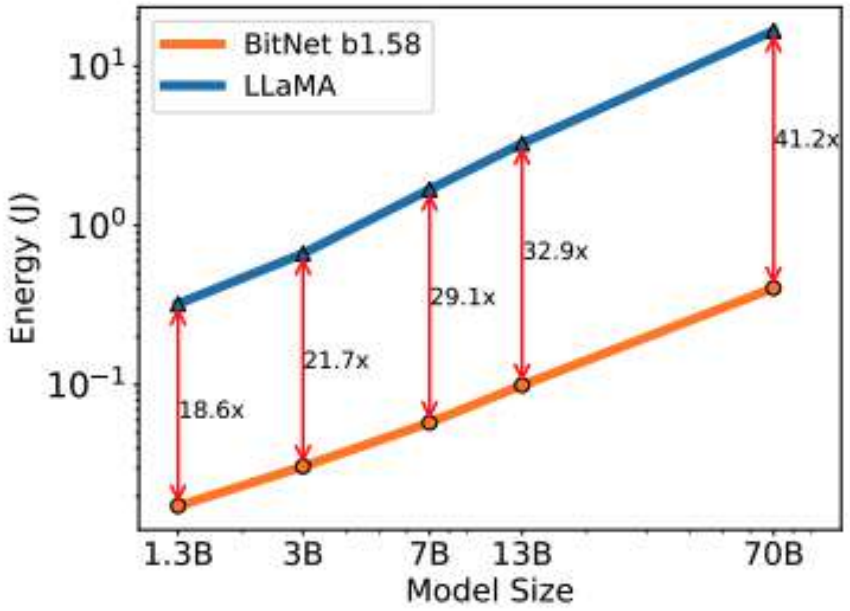


Figure 3: Energy consumption of BitNet b1.58 compared to LLaMA LLM at 7nm process nodes. On the left is the components of arithmetic operations energy. On the right is the end-to-end energy cost across different model sizes.

Source: Ma, Shuming, et al. "The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits." *arXiv preprint arXiv:2402.17764* (2024).

Every possible effort for the efficiency

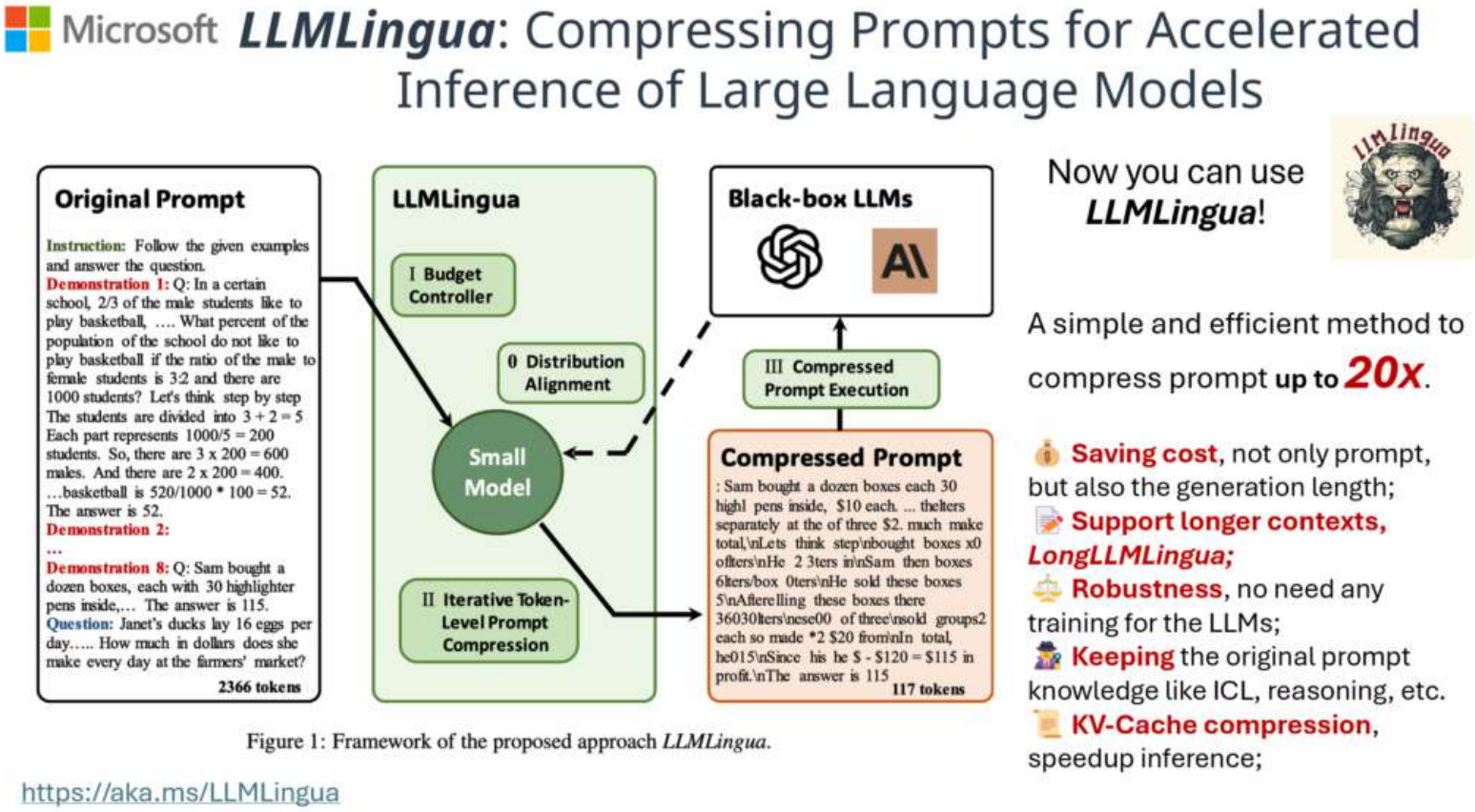


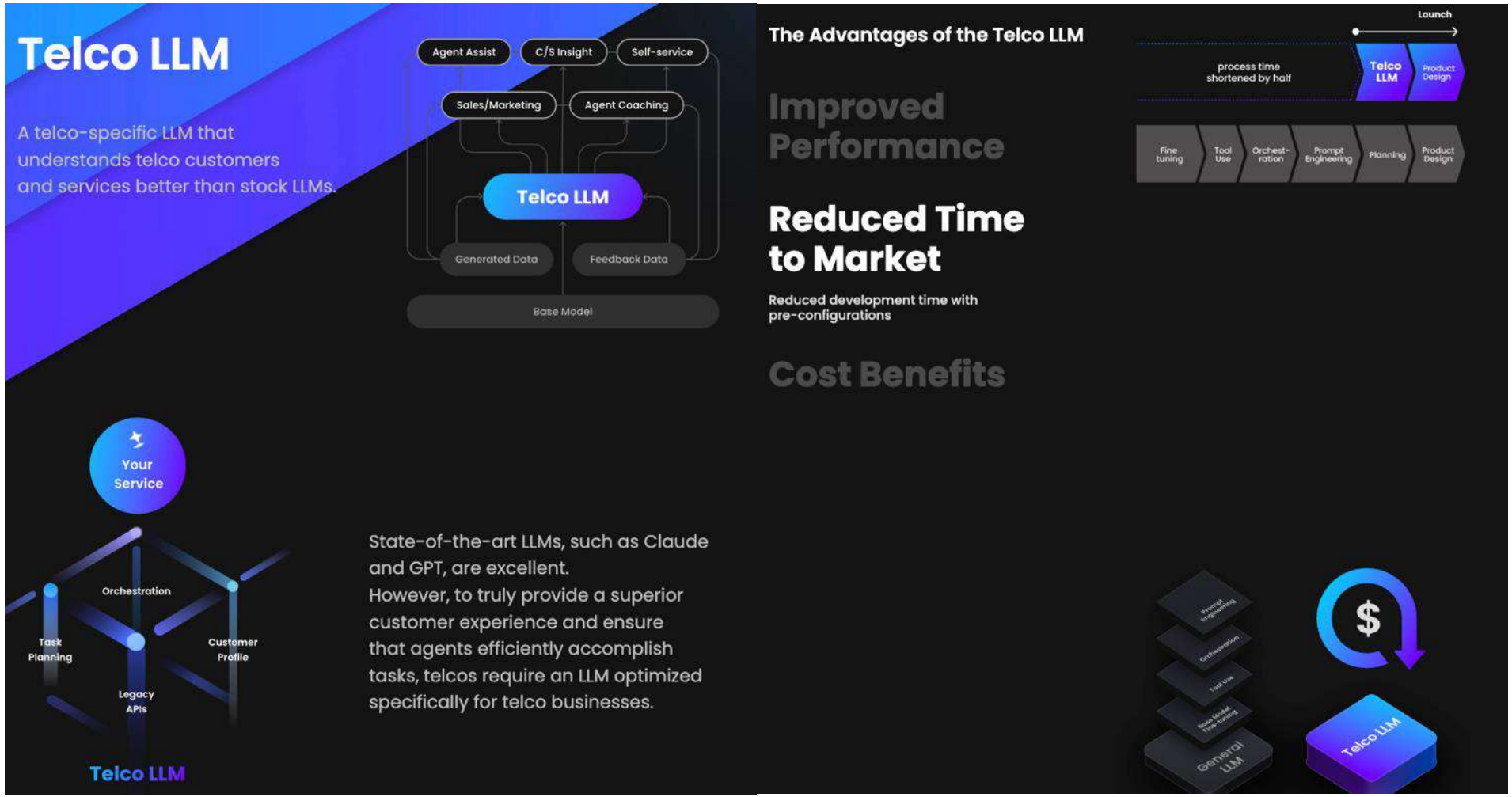
Figure 1: Framework of the proposed approach *LLMLingua*.

https://aka.ms/LLMLingua

LLMLingua, which employs a well-trained small language model after alignment, such as GPT2-small or LLaMA-7B, detects unimportant tokens in the prompt and enables inference with the compressed prompt in black-box LLMs, achieving up to 20x compression with minimal performance loss

Source: Jiang, Huiqiang, et al. "Llmlingua: Compressing prompts for accelerated inference of large language models." *arXiv preprint arXiv:2310.05736* (2023).

Domain-specific LLMs : reducing Time-to-market



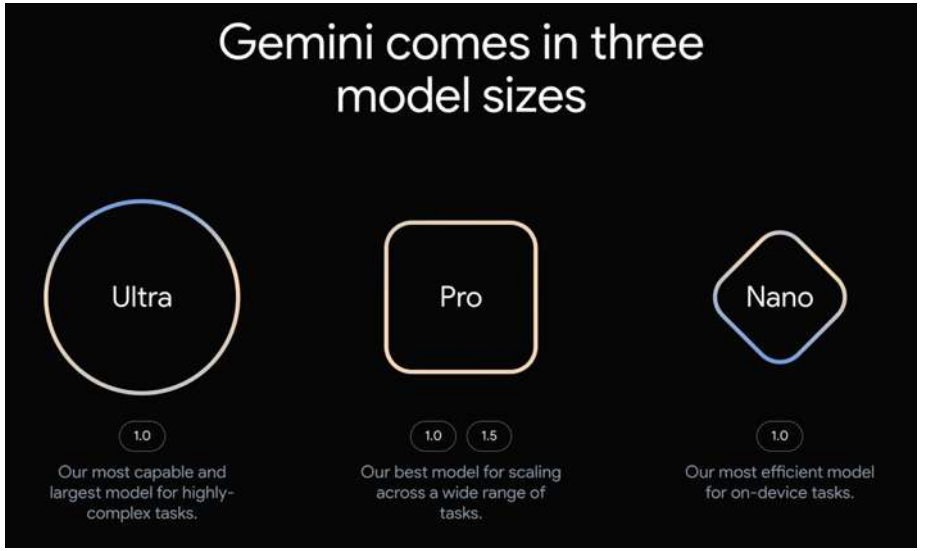
Source: SK Telecom

LLM at your palm : On-device AI

Samsung Gauss (Galaxy S24)

- Real-time translation / interpretation available in multiple languages on the phone
- Available on Galaxy S24
- (Allegedly) ~7B

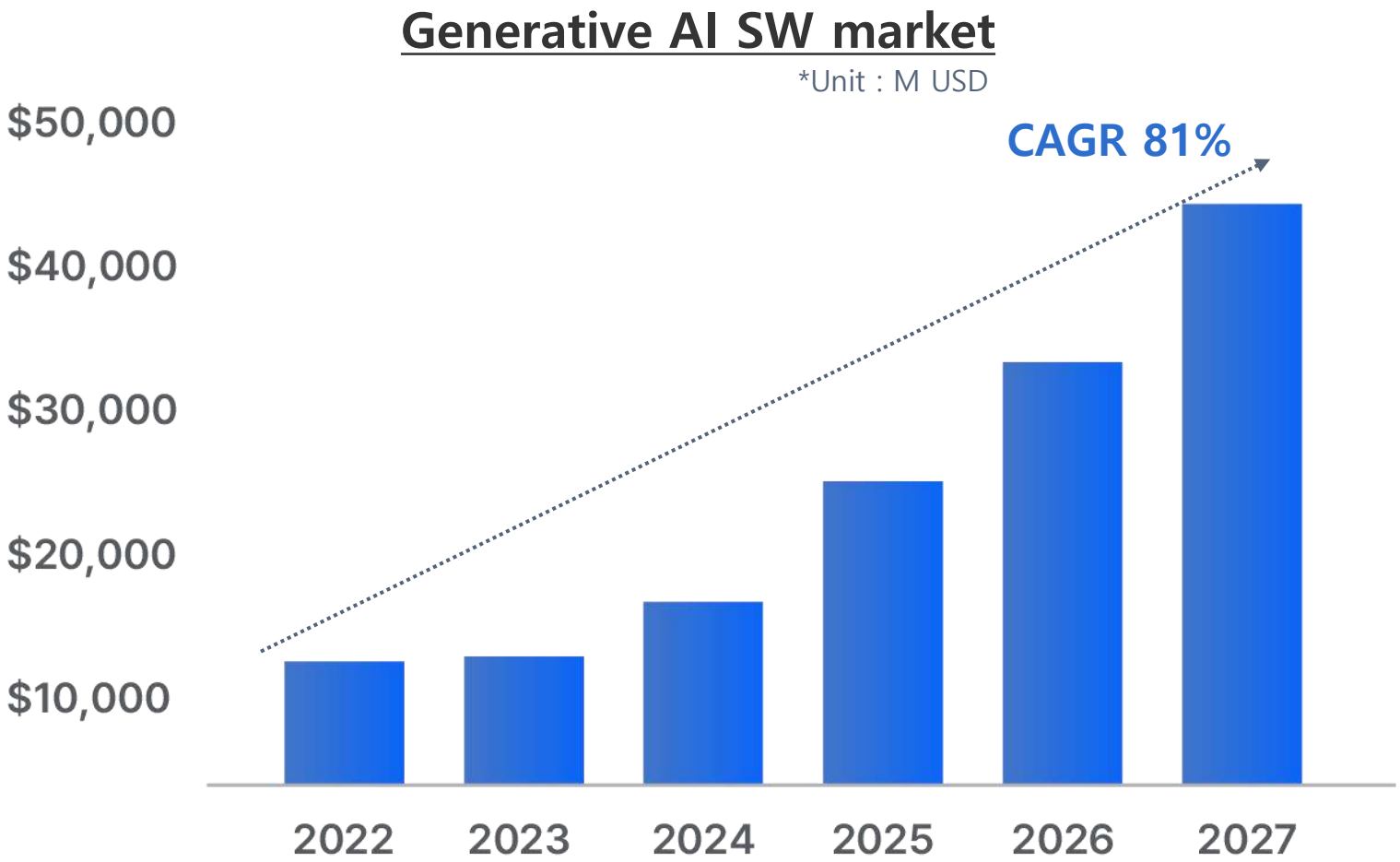
Google's Gemini Nano



- Google's Gemini comes in multiple parameter sizes
- Gemini Nano is available in Google Pixel 8
- Model size : 3.25B

Source: Samsung Galaxy, Google

Generative AI SW market is growing rapidly



Source: Gartner

Gen AI recap

1. **Sora 등 새로운 형태의 생성형 AI 모델 지속 등장**

2. **On-device AI 탑재가능한 sLLM 모델 또한 지속 등장**
   - 휴대전화 탑재
   - LLM 과 sLLM 이 모델 생태계가 이분화되어 지속 확장

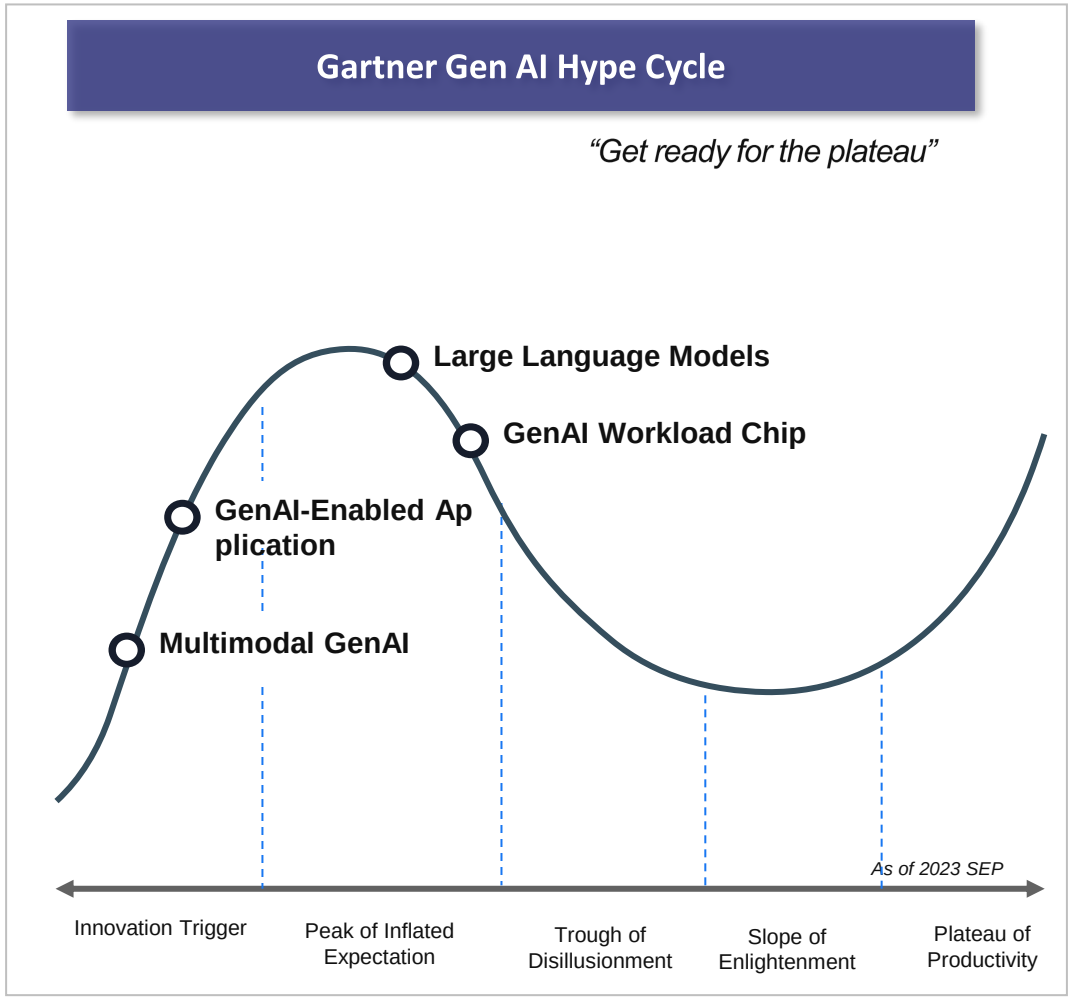
3. **상용단계에서 비용효율화를 고려한 LLM 디자인 / 적용 방법 확대**
   - 1 bit LLM, 프롬프트 압축 등

4. **Telco LLM 등 Domain-specific LLM 대두**
   - Application specific 모델로, 불필요한 모델 구조 및 Parameter 최소화
   - 모델 적시 및 즉각 적용 가능

02.

Gen AI 와
AI Semiconductor

As Gen AI is getting mature, HW market is getting there, too



Source: Gartner, SAPEON Analysis

2022-2024 Gen AI 관련 HW Market Dynamics

- 2022.03 NVIDIA
  - **H100 GPU 출시**
    - 1,979 TFLOPS, HBM3 80GB
- 2022.11 OpenAI
  - **ChatGPT 서비스 개시**
- 2023.08 Google Cloud
  - **TPUv5e 발표 및 Nvidia와 파트너십 확대**
    - 394 TOPS, HBM2e 16GB
- 2023.11 Microsoft / NVIDIA
  - **Maia 100 발표 및 Nvidia와 파트너십 확대**
  - **H200 GPU 발표**
    - 1,979 TFLOPS, HBM3E 141GB
- 2023.12 AMD
  - **AMD MI300X GPU 출시**
    - 2,615 TFLOPS, HBM3 192GB
- 2024.02 OpenAI / NVIDIA
  - **Sam Altman, AI Chip Project 언급**
  - **Hyperscaler 위한 Custom Chip 사업 준비**
- 2024.03 NVIDIA
  - **B100 GPU 발표**

Now, the inference – production at scale matters

NVIDIA

TC

Nvidia launches NIM to make it smoother to deploy AI models into production

Frederic Lardinois

@fredericl / 7:00 AM GMT+9 • March 19, 2024

Comment

Image Credits: Haje Kamps / TechCrunch

At its GTC conference, Nvidia today announced Nvidia NIM, a new software platform designed to streamline the deployment of custom and pre-trained AI models into production environments. NIM takes the software work

SAMSUNG

Kye Hyun Kyung • Following
Samsung Electronics (삼성전자) CEO of Samsung Semiconductor
3d •

What is the practical future of #AI? Here at #SamsungSemiconductor, we are increasingly focused on what is known as Artificial General Intelligence (#AGI): AIs with capabilities greater or equal to humans which can learn on their own without being trained on human data first.

To help pave the road for AGI, I'm excited to announce the establishment of the Samsung Semiconductor AGI Computing Lab located in both the United States and South Korea, for which we have already begun recruitment efforts. These specialized research labs, overseen by my colleague Dong Hyuk Woo, endeavor to create an entirely new type of semiconductor: one specifically designed to meet the incredible processing demands of future AGIs.

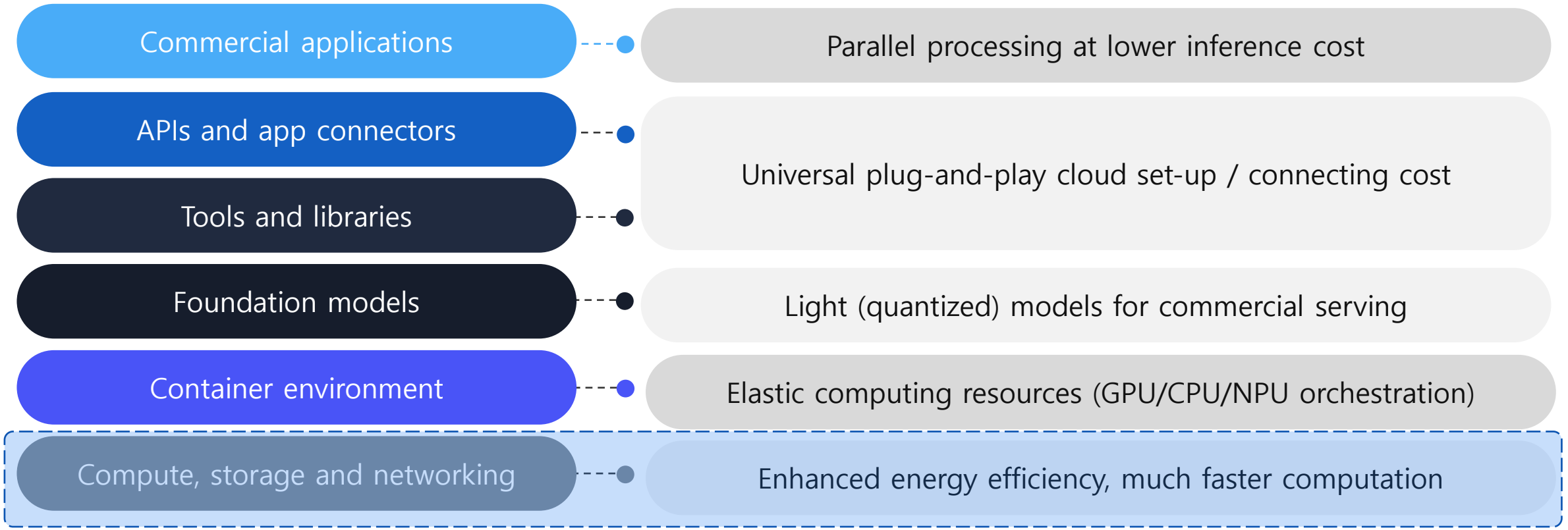
Initially, the AGI Computing Lab will focus on developing chips for Large Language Models (#LLM), with a focus on inference and service applications. To develop chips that will dramatically reduce the power necessary to run LLMs, we are revisiting every aspect of chip architecture, including memory design, light-weight model optimization, high-speed interconnect, advanced packaging, and more. Our plan is to continuously release new versions of our AGI Computing Lab chip designs, an iterative model that will provide stronger performance and support for increasingly larger models at a fraction of the power and cost.

Through the creation of the AGI Computing Lab, I am confident that we will be better positioned to solve the complex system-level challenges inherent in AGI, while also contributing affordable and sustainable methods for the future generation of advanced AI/ML models.

모델 트레이닝 보다는, pre-trained 된 모델의 Scalable 한 추론에 포커스

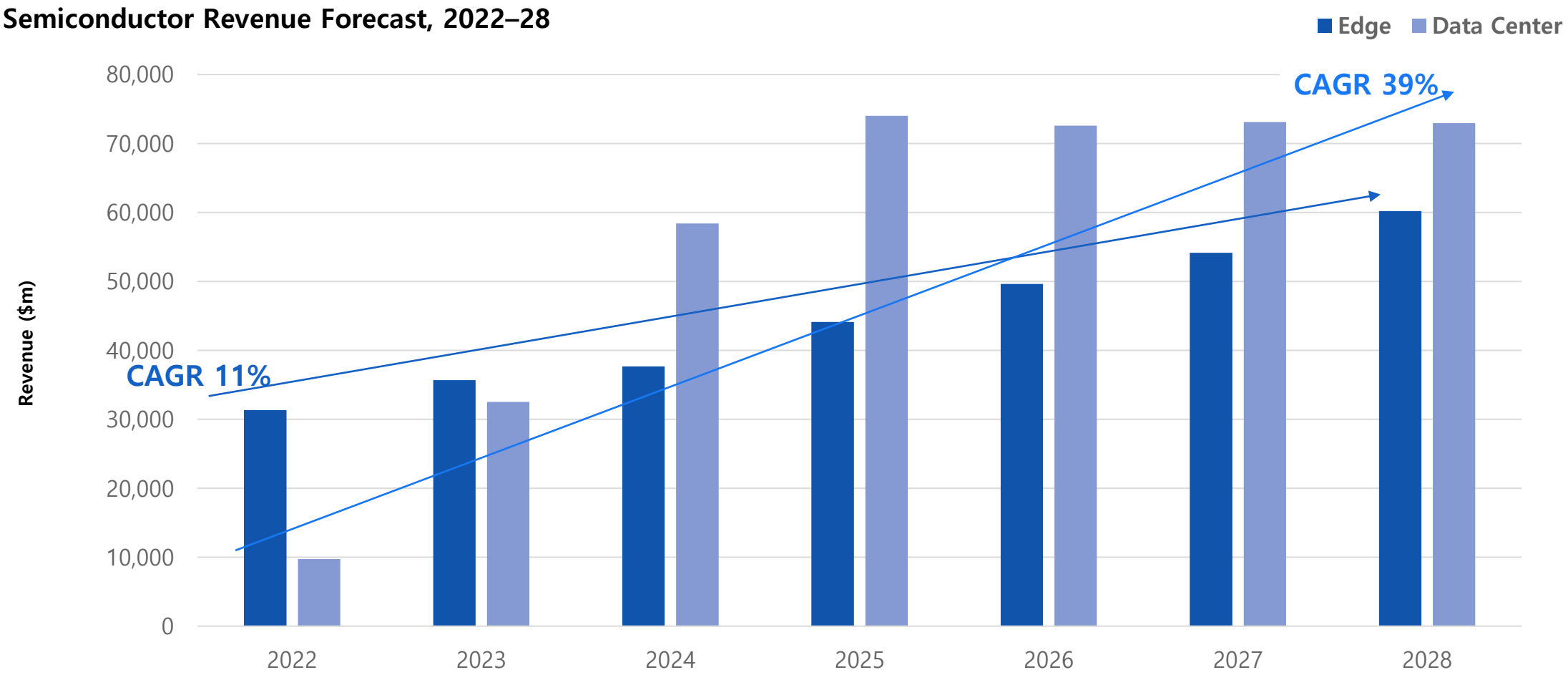
Source: Techcrunch, Linkedin post

Gen AI vertical value chain



Source: Omdia, SAPEON Analysis

AI Semiconductor market is growing: Edge and Datacenter alike



Source: Omdia

Gen AI x AI Semiconductor

	sLLM	LLM 등 Gen AI
Model size	**Relatively small :** below 70B	**거대 Parameter size, Large :** 70B+

	On-device / Edge향 NPU	Datacenter 향 NPU
Form factor	**Chip / Card** for Edge box / server	**PCIe, OAM :** for datacenter server systems (rack-scale)
Power	**Low Power** : ~5W (Cellphone) -35 W	**High Power** : 200 – 700 W / card
Network	**Not required** : Low latency, Secure privacy	**Required** : High-speed to avoid potential latency

모델 크기, 사용환경 등에 따라 Generative AI 모델에 적합한 NPU가 상이할 수 있음

Source: SAPEON Analysis

OCP Accelerator Module (OAM)

OAM Base Specification defines GPU/ASIC power consumption up to 1000W with form factor
OAM Base Specification defines liquid cooling implementation to improve data center facility PUE (efficient under 1.5) and WUE

Source: SAPEON Analysis

Orchestrated Service

Datacenter

교통관제센터 - 고용량, 고화질 이미지 동시 처리 / LM 훈련
(e.g. SAPEON X330)

전국 6,185개 교차로 ITS를 GPU에서 국산AI반도체로 교체하면

3,039MWh
전력 절감

1,396
tCO2ep 저감

153,439그루
심는 효과

전국 6,185개(2022년 기준) 4지 교차로 가정
NVIDIA A2 대비 X220 Enterprise 1년 운영 기준

On-Device

자율주행차량 내부 – 별도 네트워크 연결 없이도 객체인식
(e.g. SAPEON XA3)

Edge

신호관리 시스템, AI Camera – AI Box 등 선택적 네트워크 연결로 신호 시스템 운영 및 감시
(e.g. SAPEON XE31)

03.

SAPEON’s Journey

About SAPEON

SK telecom

SAPEON

2016
AI chip R&D unde r SK telecom

2017
FPGA prototype X110 came out

2018
NUGU AI speaker ASR based on X110

2019
Video surveillance intrusion detection based on X110

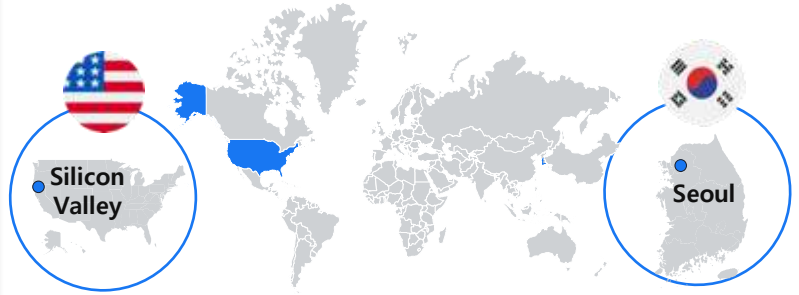
2020
Launch of **X220, the 1st datacenter AI Processor in Korea**

Apr. 2022
Incorporation of SAPEON in the U.S.

Nov. 2023
X330 launch
A flagship datacenter AI processor

SK
South Korea's 2nd biggest corporate group

Global Presence



SAPEON overview

Product Line-up

Chip

Card

Server

System (Cloud)

Target Industries

Datacenter

Factory

Language

Autonomous Driving

Media

Security

Mobility

Product roadmap

	2021	2022	2023	2024	2025	
Cloud	X220			X330		
Edge				XE31	XA3	

X220
Compact / Enterprise
87 / 174 TOPS
65 / 135 W

X330
Compact / Prime
367 / 734 TFLOPS
75-120W / 250 W

XE31
Low-power, med-high OPS

XA3
Auto (with Partner)

X330, an extraordinary chip for the next generation AI

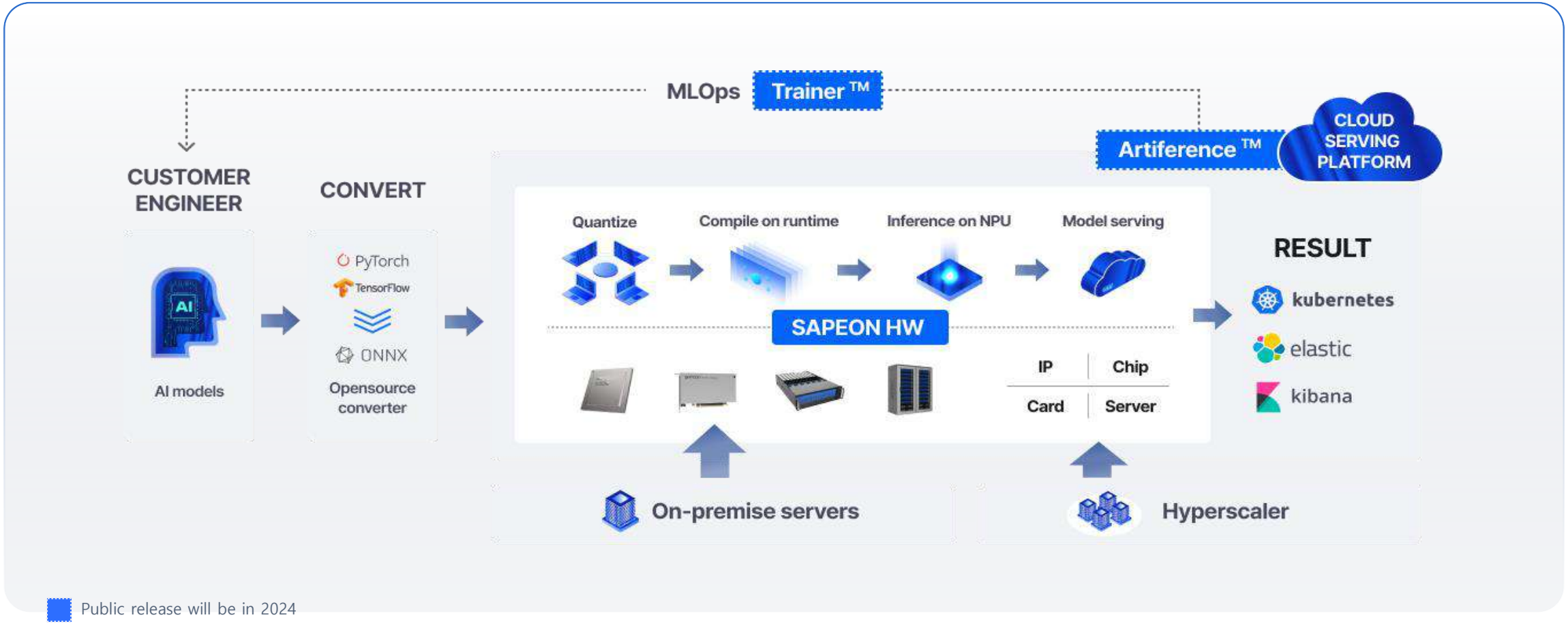
Unveiled
2023 November

SAPEON X330
Compact

SAPEON X330
Prime

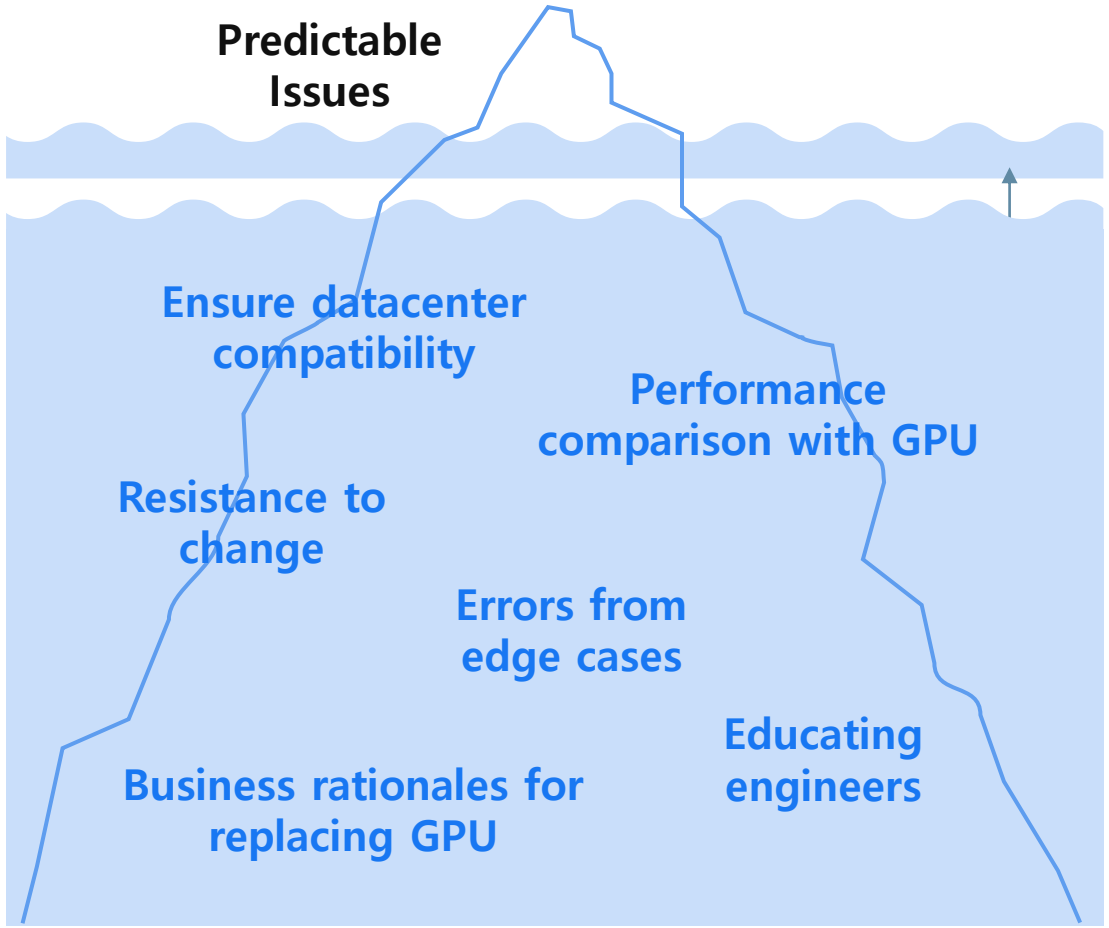
Precision	FP 16/8 bit, INT 8 bit	
8bit Performance	367 Tera FLOPS	734 Tera FLOPS
Memory	GDDR6 x 8 (ECC)	GDDR6 x 16 (ECC)
Memory Capacity	16 GB	32 GB
Memory Bandwidth	256 GB/s	512 GB/s
Host Interface	PCIe Gen5 16 Lane	
Max Power Consumption	75 - 120 W	250 W
CODEC	Encoder : H264 / VP8 / MPEG-4 Decoder : HEVC / H264 / VP8 / MPEG-4	

HW, SDK, 서빙 및 Pre-trained model catalog 까지 Streamlined



Public release will be in 2024

What do kings (customers) want?

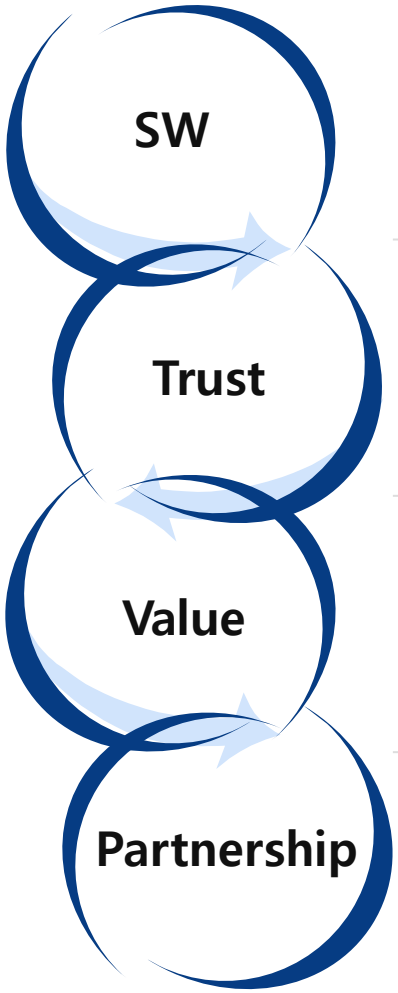


Source: SAPEON Analysis

Customer feedback

- *" .... What if we would like to change the threshold of the models?... "*
- *" .... Needs to put too may efforts to learn about SDK ... It is not like a plug and play... "*
- *" .... What is the result of apples-to-apples comparison with GPUs ... "*
- *" .... When is the earliest possible delivery dates of the servers, how much lead time is needed? ... "*
- *" .... Is it cheaper than GPUs? ... then how much cheaper? "*

Lessons learned



- SW / SDK are MUCH more important than we thought it be
- Always be in their shoes to fight against resistance to change : enhance usability, convince engineers and research scientists

- Establish a robust foundation of trust with your clients by allowing them ample time to comprehend any limitations, without resorting to concealment or deception
- Trust-driven improvements cultivated over time endure

- As a contender, you will be always compared with a strong incumbent
- Set-up a strong unrivaled value proposition of your chip – performance, power efficiency, he at control ; so that clients can make business decisions with strong evidence

- Work smart : find good manufacturing partners, work together – focus on what you can do b est
- Make every application experience count : your partners' / customers' lessons are great clues to improve your SDK

Recap : remember 3E for commercialization

Efficiency

- Quantitative and proven power efficiency
- Minimal transition cost from GPU and cloud services

Efforts

- Educate customer engineers often; they have to 'change'
- Be in their shoes
- Admit with clients: the product is new, there could be issues!
- Ready to tackle the issues 24/7

Evolution

- Markets are changing so fast, while HW comes in ~2 years
- Always find room to evolve and adapt to the market and tech
- Now LLM/sLLM, Gen AI, but in 2 years what will dominate?
- Evolve or die!

SAPEON

Thank you