인공지능 시대의
인공지능과 반도체의 패권 전략

Joungho Kim

TeraByte Interconnection and Package Laboratory

Department of Electrical Engineering

KAIST

Part1:생성형 인공지능의 미래

DALL·E 2: OpenAI's Image Generation AI System

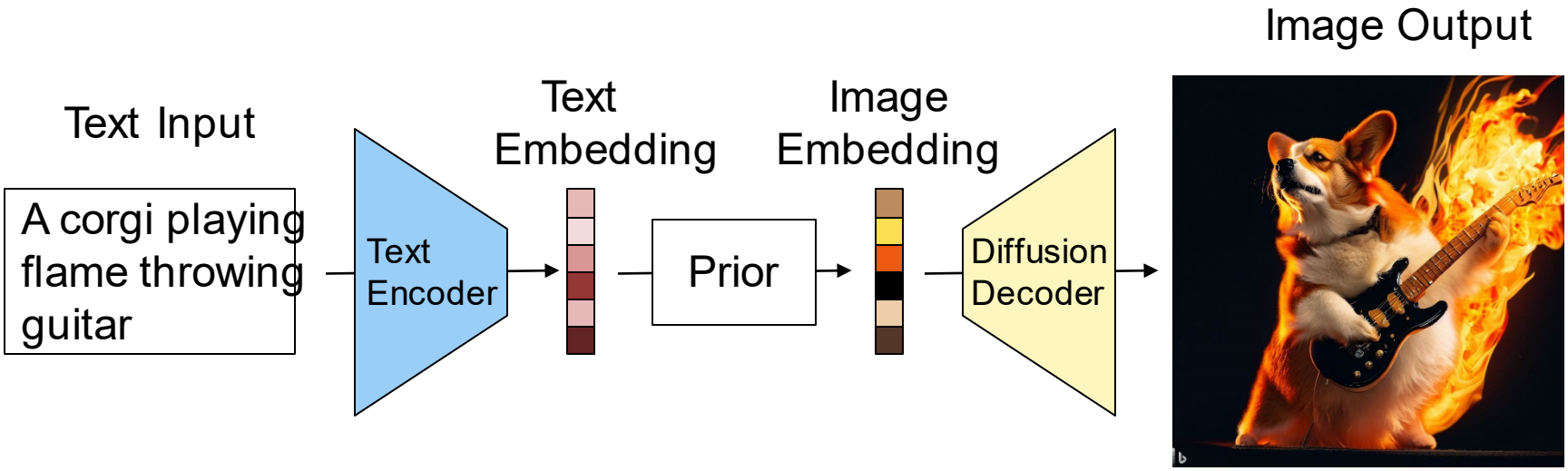
Text Prompt

Vibrant portrait painting of Salvador Dalí with a robotic half face.

< Salvador Dalí image generated by DALL·E 2>

- DALL·E 2 is an AI system that can create realistic images and art from a description in natural language.
- OpenAI released DALL·E 2 on April 6, 2022.
- The parameter of DALL·E 2 is 3.5 billion, which is 1/4 of the previous version.

Principle of DALL·E 2: Encoder, Prior and Decoder



< Overview of the DALL·E 2 image-generation process >

- DALL·E 2 is a particular instance of a three-part model made of encoder, prior and decoder.
- The text embedding and image embedding represents the main features that are semantically meaningful: People, animals, objects, style, colors, background, etc.
- DALL·E 2 can generate a novel image that retains these characteristics while varying the non-essential features.

AI Song Contest (Magenta Project, 2020~)

< A to I > (AI-collaborated music)

Lyrics	▪ GPT, GPT-2
Melody	▪ RNN ('improv_rnn' model from Magenta)

TEAM / 3+i

SONG / A to I

TEAM MEMBERS / Yuxuan Wu, Yi Deng, Yusong Wu

<Lyrics>

(…)

I may be a machine,

but I know how to love

I may not have a heart,

But I can feel your love.

I may not be human,

But I know how to care and share.

I may be artificial,

But I am real.

< An Example of AI and Human Collaborative Composition >

GPT-4V : The Pinnacle of Contemporary Multimodal Technology Attainment [2/4]

[Prompt]

Describe the pointed region in the image.

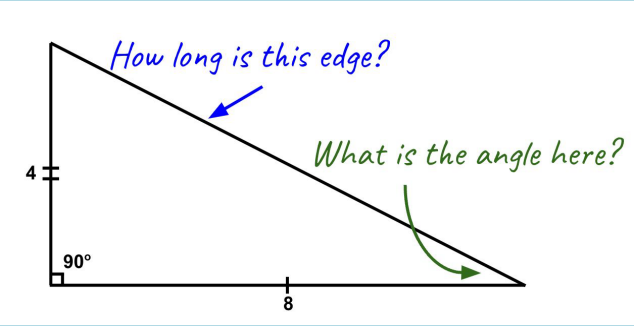
Method	Validataion set								Test set							
	in.		near.		out.		overall		in.		near.		out.		overall	
	C	S	C	S	C	S	C	S	C	S	C	S	C	S	C	S
OSCAR	85.4	11.9	84.0	11.7	80.3	10.0	83.4	11.4	84.8	12.1	82.1	11.5	73.8	9.7	80.9	11.3
Human	84.4	14.3	85.0	14.3	95.7	14.0	87.1	14.2	80.6	15.0	84.6	14.7	91.6	14.2	85.3	14.6
VIVO	92.2	12.9	87.8	12.6	87.5	11.5	88.3	12.4	89.0	12.9	87.8	12.6	80.1	11.1	86.6	12.4
VinVL	103.7	13.7	95.6	13.4	83.8	11.9	94.3	13.1	98.0	13.6	95.2	13.4	78.0	11.5	92.5	13.1
UFO	103.9	14.5	95.5	13.8	83.5	12.3	94.3	13.6	98.9	14.3	94.7	13.9	77.9	12.1	92.3	13.6
mPLUG	-	-	-	-	-	-	114.8	14.8	-	-	-	-	-	-	-	-
SimVLM	113.7	-	110.9	-	115.2	-	115.2	-	113.7	-	110.9	-	115.2	-	115.2	-
LEMON	118.0	15.4	116.3	15.1	120.2	14.5	117.3	15.0	112.8	15.2	115.5	15.1	110.1	13.7	114.3	14.9
UniversalCap	123.2	15.0	121.5	15.3	123.4	14.4	122.1	15.0	118.9	15.4	120.6	15.3	114.3	14.1	119.3	15.1
CoCa	-	-	-	-	-	-	122.4	15.5	-	-	-	-	-	-	120.6	15.5
GIT_B	100.7	13.8	97.7	13.5	89.6	12.5	96.6	13.4	-	-	-	-	-	-	-	-
GIT_L	107.7	14.9	107.8	14.5	102.5	13.7	106.9	14.4	-	-	-	-	-	-	-	-
GIT	**129.8**	**16.3**	124.1	16.0	127.1	15.7	125.5	16.0	122.4	16.2	123.9	16.0	122.0	**15.7**	123.4	15.9
GIT2	126.9	16.1	**125.8**	**16.2**	**130.6**	**15.8**	**126.9**	**16.1**	**124.2**	**16.4**	**125.5**	**16.1**	**122.3**	15.6	**124.8**	**16.1**

[GPT-4V]

The highlighted red region in the image is a table cell that contains a numerical value of 122.3. It is located in the "Test set" section of the table, under the "out" column and the "GIT2" row. This value represents the overall performance metric for the GIT2 method on the test set, in the "out" category.

[Prompt]

Answer the question in the following figure. Provide intermediate steps.



[GPT-4V]

To find the length of the blue arrowed edge, we can use the Pythagorean theorem. Plugging in the values we have, $4^2 + 8^2 = c^2$, $c \approx 8.94$.

To find the angle at the green arrow, we can use the tangent function. Plugging in the values we have, we get: $\tan\theta = 4/8$, $\theta \approx 26.57°$.

< Responses of GPT-4V for visual referring pointing -2 >

- GPT-4V is capable of simple table analysis and solving mathematical problems.

Ref) Zhengyuan Yang et al., 2023

Figure: Robot with Generative AI and LLM

-Recognition, understanding, make decision, and communication using AI
-Understanding the environment using AI such as CNN ,and LSTM
-Make decisions using generative AI
-Communicate with humans using LLM
-Limit by remote computing power and battery

SORA : OpenAI's text-to-video Generation AI Model

Prompt

A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage …

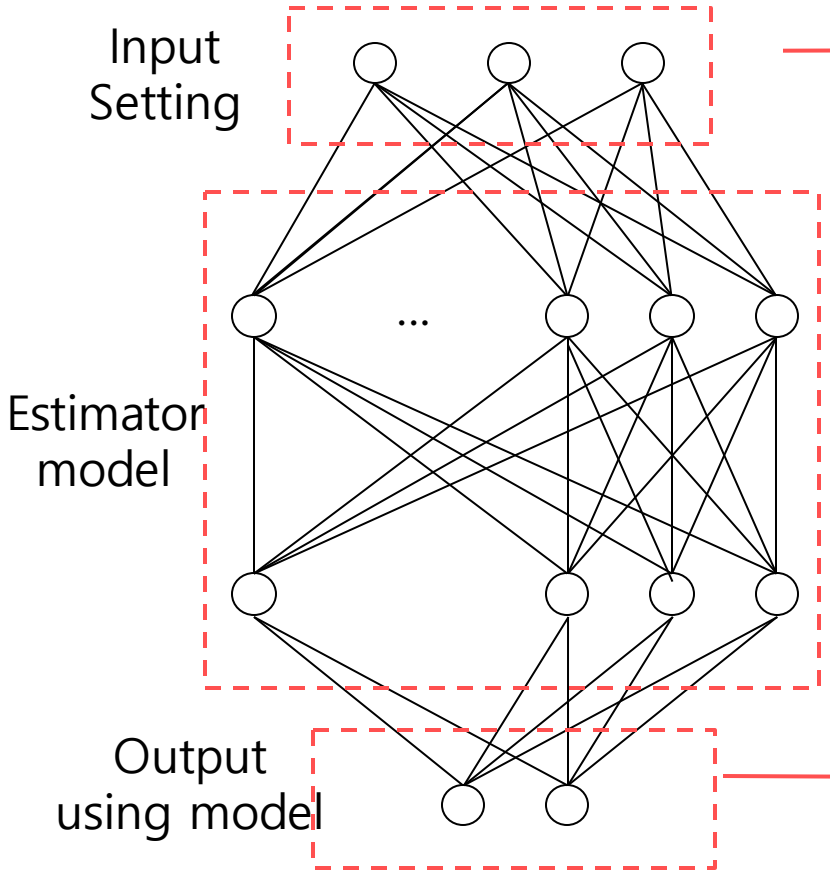
SORA

< SORA producing images with given text input >

- SORA is an OpenAI's text-to-video generation AI model unveiled on February 15, 2024.
- SORA can generate videos up to a minute long while maintaining visual quality and adherence to the user's prompt.
- Sora is named after the Japanese word for sky and air (そら).

Ref) OpenAI, "Sora Technical report"

Collaboration with ChatGPT [3/3]: ChatGPT-Programmed Model



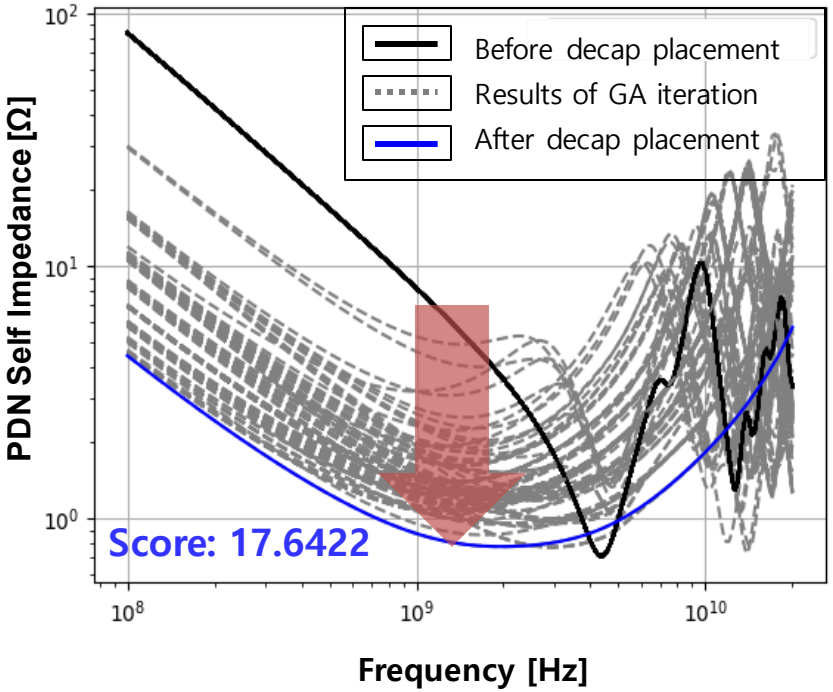

```python
# Load the dataset from csv file
data = pd.read_csv(cur_dir +'\\EYE_REAL.csv')
data = data.dropna(axis = 0)
# Define the input and output variables
inputs = data[['d_TSV', 'p_TSV', 't_ox']]
targets = data[['eye_height', 'eye_width']]
```

```python
# Define the neural network architecture
class EyeEstimator(nn.Module):
    def __init__(self):
        super(EyeEstimator, self).__init__()
        self.fc1 = nn.Linear(num_features, 128)
        self.relu1 = nn.ReLU()
        self.fc2 = nn.Linear(128, 64)
        self.relu2 = nn.ReLU()
        self.fc3 = nn.Linear(64, num_outputs)
```

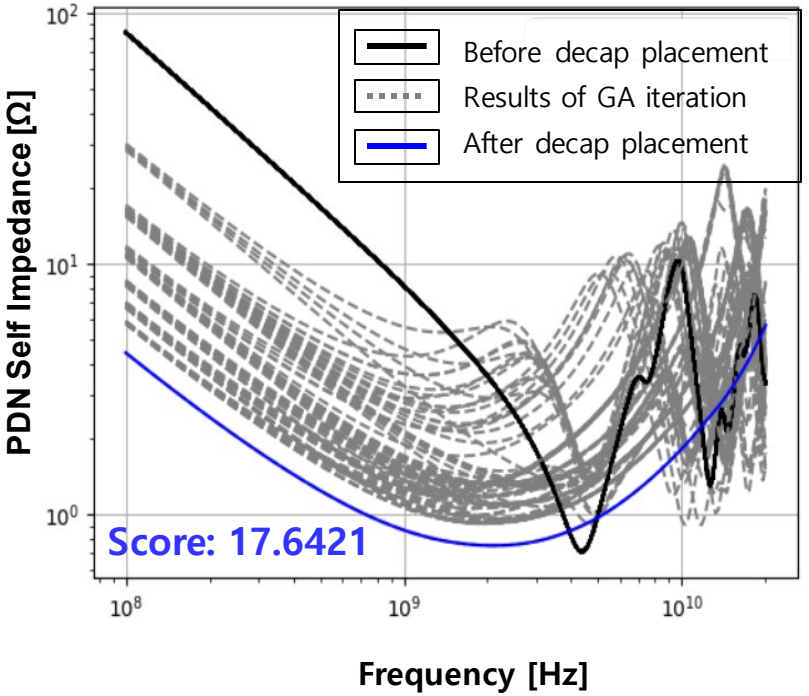
```python
num_epochs = 5000
for epoch in range(num_epochs):
    # Forward pass
    outputs = model(inputs)
    loss = criterion(outputs, targets)

    # Backward and optimize
    optimizer.zero_grad()
    loss.backward()
    optimizer.step()
```

< Code blocks from ChatGPT-collaborated DNN model>

Performance Verification of the Proposed ChatGPT-Collaborated Method



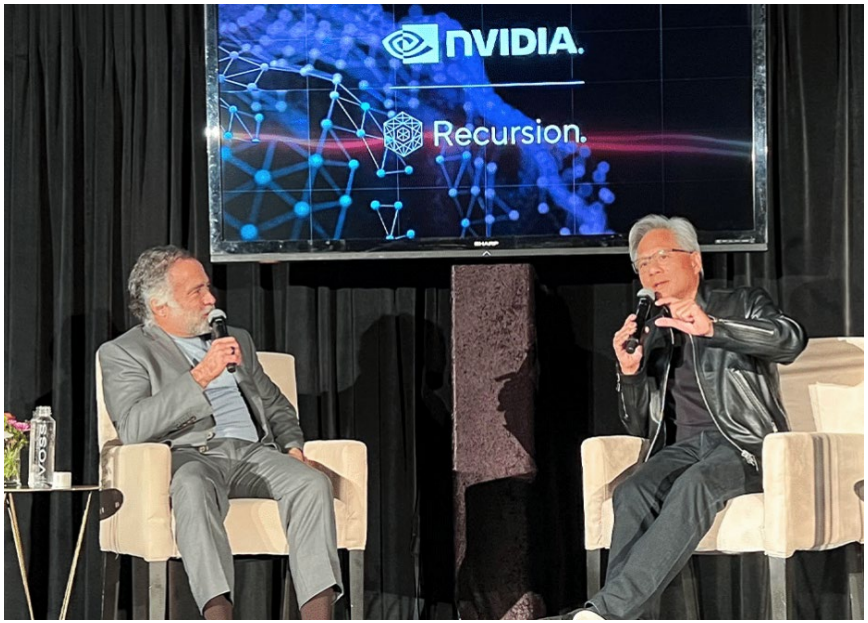
< Result of ChatGPT-Collaborated Method >



< Result of Human >

- The difference between the performance optimality of ChatGPT-collaborated GA and human implemented GA were negligible.
- Generating genetic algorithm code in collaboration with ChatGPT was extremely time efficient.
- The standard deviation with respect to seed was 0.02376, verifying the stability of ChatGPT-collaborated method.

NVIDIA의 Bio분야에서의 사업 확장: BioNeMo

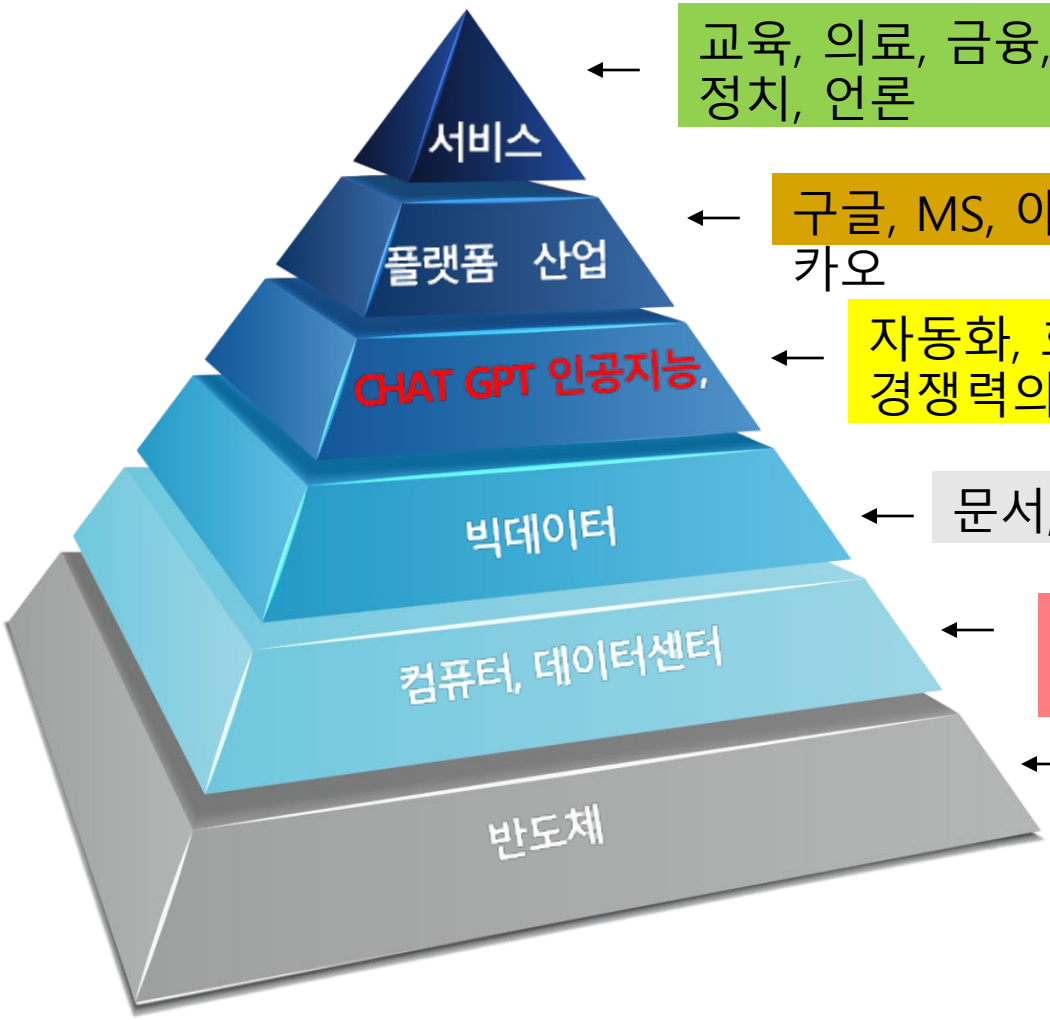


< JP 모건 헬스케어 2024에서 발표하는 젠슨 황 >

- 올해 1월 세계에서 가장 큰 바이오 제약 컨퍼런스인 JP 모건 헬스케어 2024에서 젠슨 황이 생성 AI와 NVIDIA DGX SuperPOD를 통해 신약 개발을 할 수 있는 BioNeMo 클라우드 서비스를 발표함.
- 글로벌 빅파마인 Amgen이 자사의 AI 신약 개발 플랫폼에 BioNeMo를 최초로 도입, AstraZeneca, Insilicon Medicine, Evozyne등 여러 기업에서 사용하고 있음.
- NVIDIA의 AI 시스템 활용을 통해 약물을 개발하고 설계하는 작업에서 데이터 7배 더 빠르게 처리, 비용 7배 아낄 수 있음.

Ref) https://blogs.nvidia.co.kr/2024/01/22/nvidia-ceo-ai-drug-discovery-jp-morgan-healthcare-2024/

ChatGPT-X 의 피라미드 구조:



← 교육, 의료, 금융, 문화, 유통, 생산, 법률, 국방, 정치, 언론

← 구글, MS, 아마존, 애플, 테슬라, 네이버, 카카오

← 자동화, 효율화, 무인화, 종속화 경쟁력의 원천

← 문서, SNS, 사진, 영상, 소리, 음성

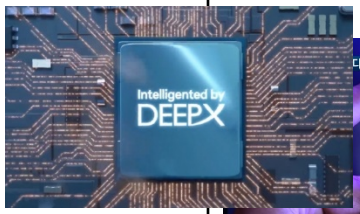
← 초거대 데이터센터, 슈퍼컴퓨터, 양자컴퓨터

← GPU, 메모리, 경쟁력 결정

Two paths: models

장 거리
AI supercomputer
Cloud service

-Ultra large scale
-LLM, VLLM
-Multimodal



-on device AI
-Edge AI
-s-LLM, sVLLM
-Small size generative AI
-Privacy protection

근거리
Device level

저 성능
Low power consumption

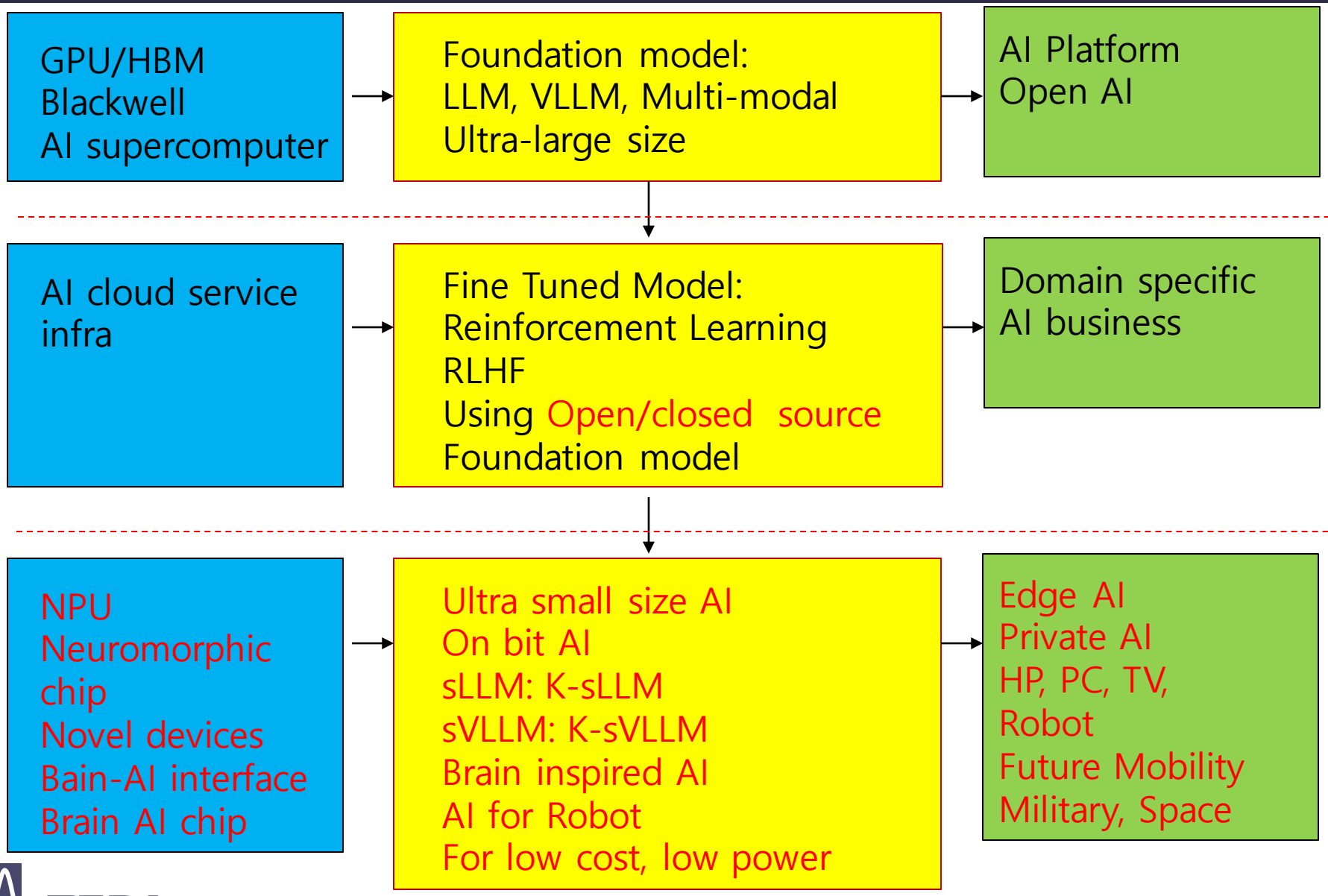
고 성능
High Bandwidth

KAIST의 선택과 집중: 두 갈래 길

-초거대 모델
-초거대, 고성능 반도체
-AI supercomputer and infra
-초거대 자본이 지배
-Could 서비스
-집중화
-패권화

-소형화
-경량화
-저비용
-edge, on device AI
-개인화, 누구나
-평등과 민주적, 공개화
-윤리적, 친환경

Hierarchical AI Models at KAIST

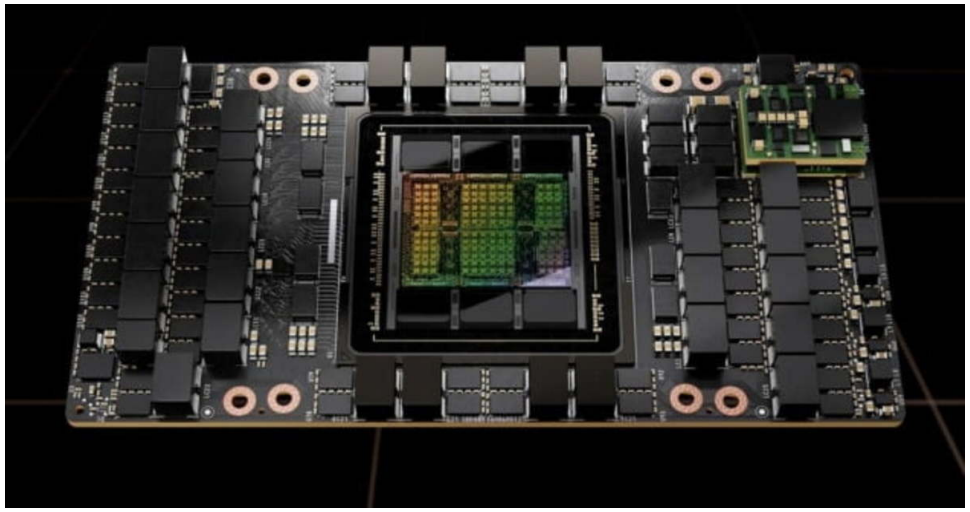


인공지능 生老病死

- 언제나
- 어디서나
- 누구나
- 인공지능의 도움을 받아서
- 삶을 살아간다. (生老病死), 희노애락(喜怒哀樂)
- 인공지능, AI 반도체, AI 슈퍼컴퓨터 컴퓨터가 개인, 기업, 국가의 자산(資産)

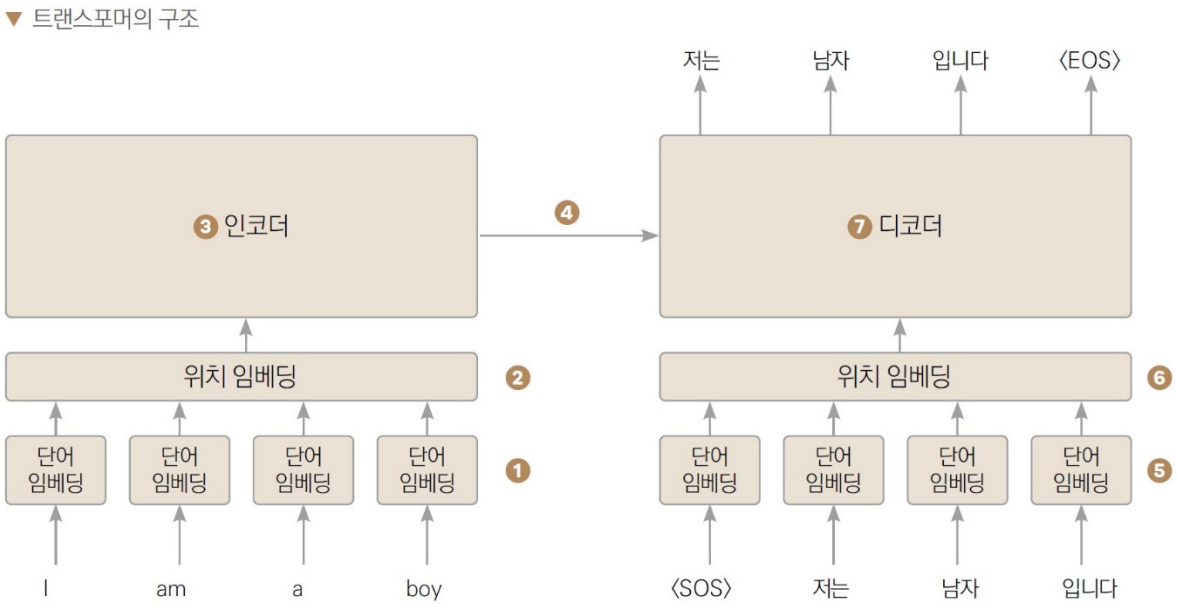
인공지능 반도체 담보 대출

- 개인, 기업의 자산(資産)
- 담보 대출
- 결혼 혼수품, 취업 조건
- 6000 천만원, GH 1억원 예상



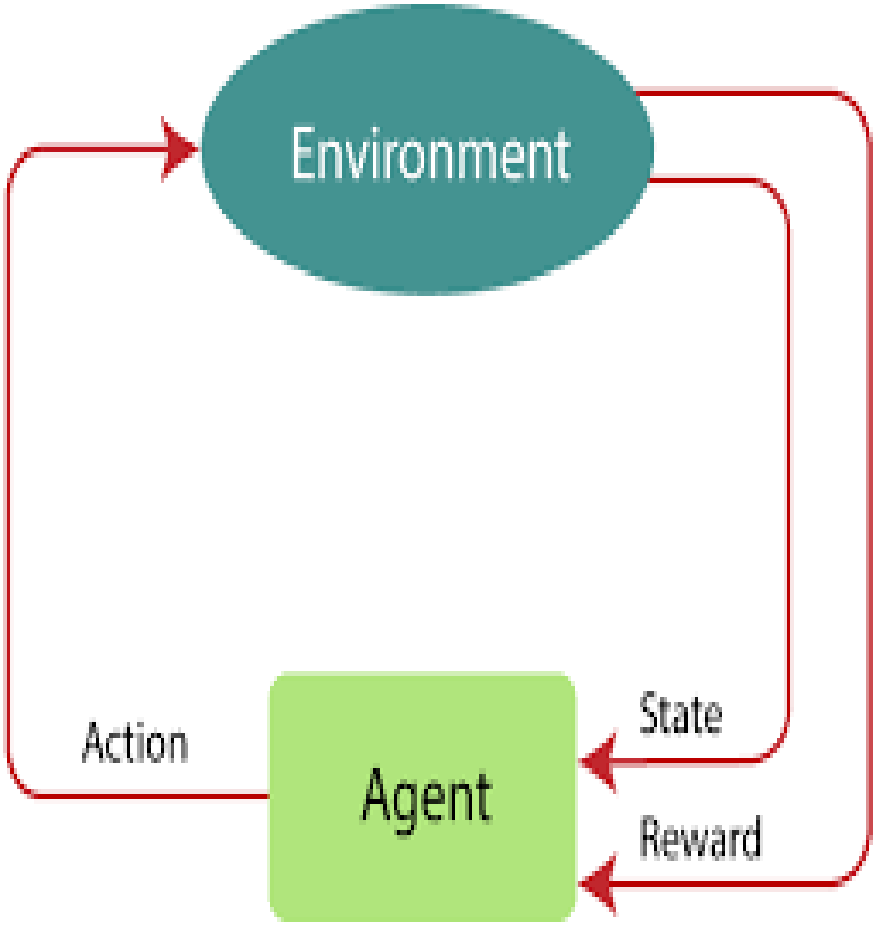
Part2:AI 반도체의 미래와 AI 슈퍼컴퓨터

Transformer Model: 대표적 생성 모델



- Encoder/decoder architecture
- Self- attention
- Cross-attention
- Positional Embedding
- Masked self-attention
- Large sequence number
- Large token size
- Parallel input
- Serial input (Current architecture)
- Parallel output (future)
- Multi-modal input
- Multi-modal output

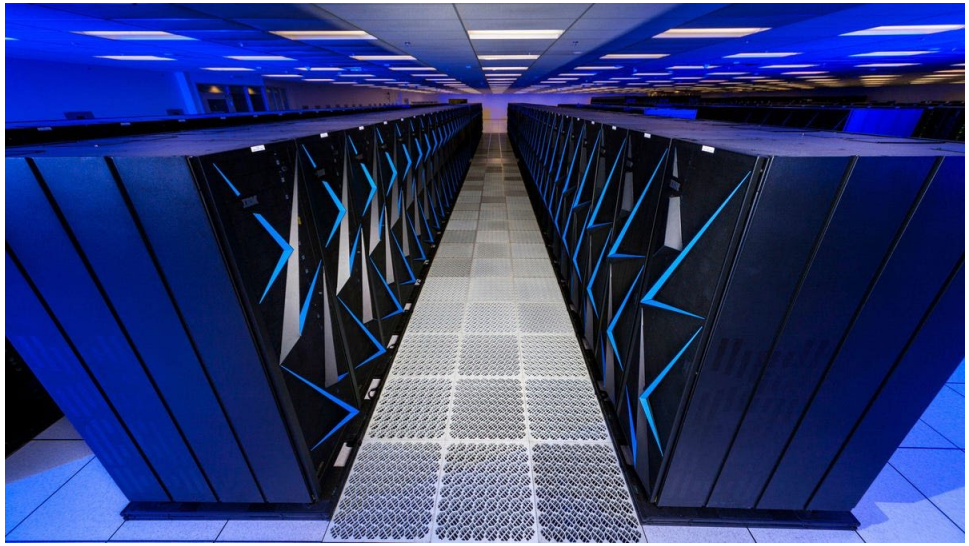
Reinforcement Learning for Fine Tuning and Human Feedback



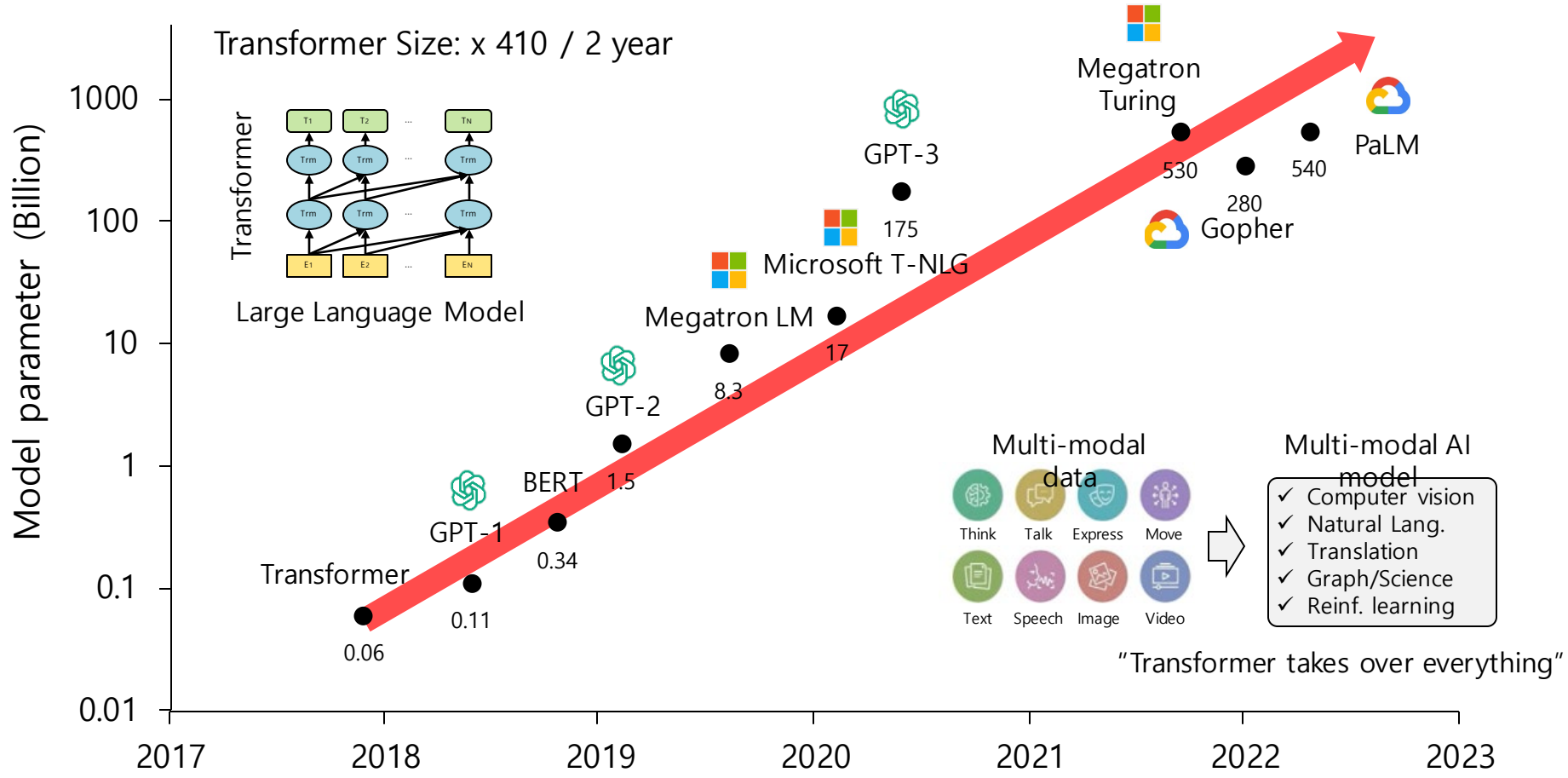
-Optimal decision
-Iterations
-Markov Process
-Reward
-Computational intensive
-Exploitation and Exploration
-RFHF

AI 슈퍼컴퓨터가 필요한 이유

- 초대형 모델 (Multi-modal AI) 시대
- 언제 어디서나 누구나 생성 인공지능 사용
- 개인 1인당 GPU 1개, 기업 10000 개, 국가 100 만개
- 미래 패권 결정(전략 자산)



The Era of Transformer-based Hyperscale AI Model: From Large Language Model to Multi-modal AI Model



< Model parameter of transformer-based AI models >

- As the transformer-based AI model develops, the model parameter of hyperscale AI model is extremely increasing.

Open AI, AI supercomputer

- 135조원 프로젝트
- AI supercomputer
- Data center
- Stargate project
- For AGI service
- Nvidia B200, 280만개
- Nvidia DGX B200 35만대(8GPU)
- 5 GW
- 1차 2028년 구축
- 2차 2030년 완료
- MS와 open AI 공동 프로젝트
- 비용은 MS

-모델 개발과 구축 시간 대폭 단축

-AGI 시장 지배 계획

스타게이트 프로젝트 개요

내용	오픈AI의 첨단 AI 모델 전용 데이터센터
비용	약 1000억달러
기간	2028년 공개 예정
AI 반도체	오픈AI의 AI를 위한 특수칩 수백만 개 탑재 예정

스타게이트	vs.	네이버 '각 세종'
5 기가와트(약 19배)	전력	270 메가와트
엔비디아 DGX B200 기준 35만대	서버	현재 10만대 수준
엔비디아 B200 기준 약 280만개	GPU	엔비디아 A100 등 탑재, 규모 미공개

자료=각 사

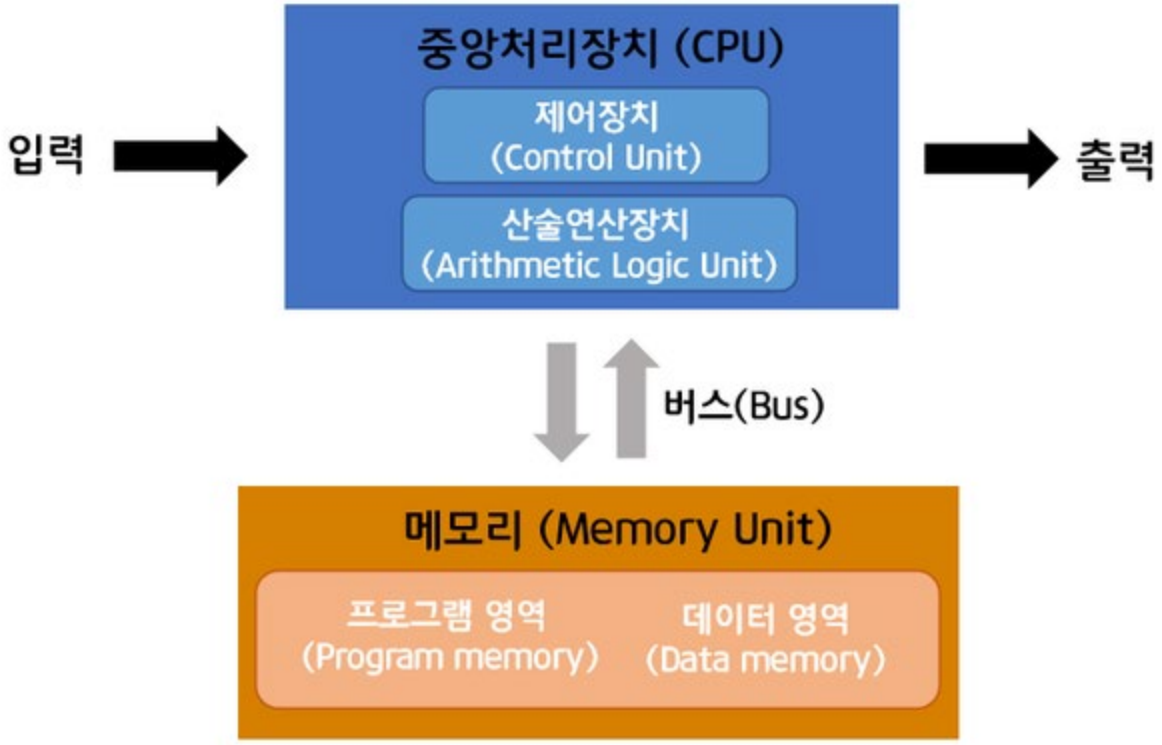
출처: 조선일보

AI 슈퍼컴퓨터 Workloads와 서비스 조건

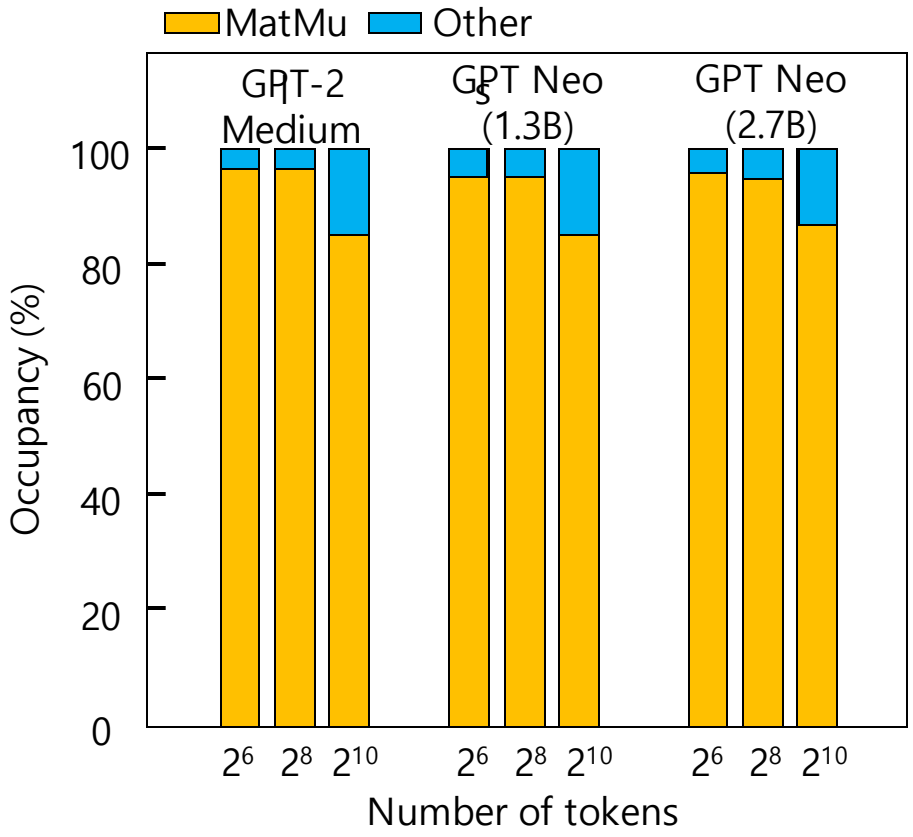
생성 AI 슈퍼컴퓨터

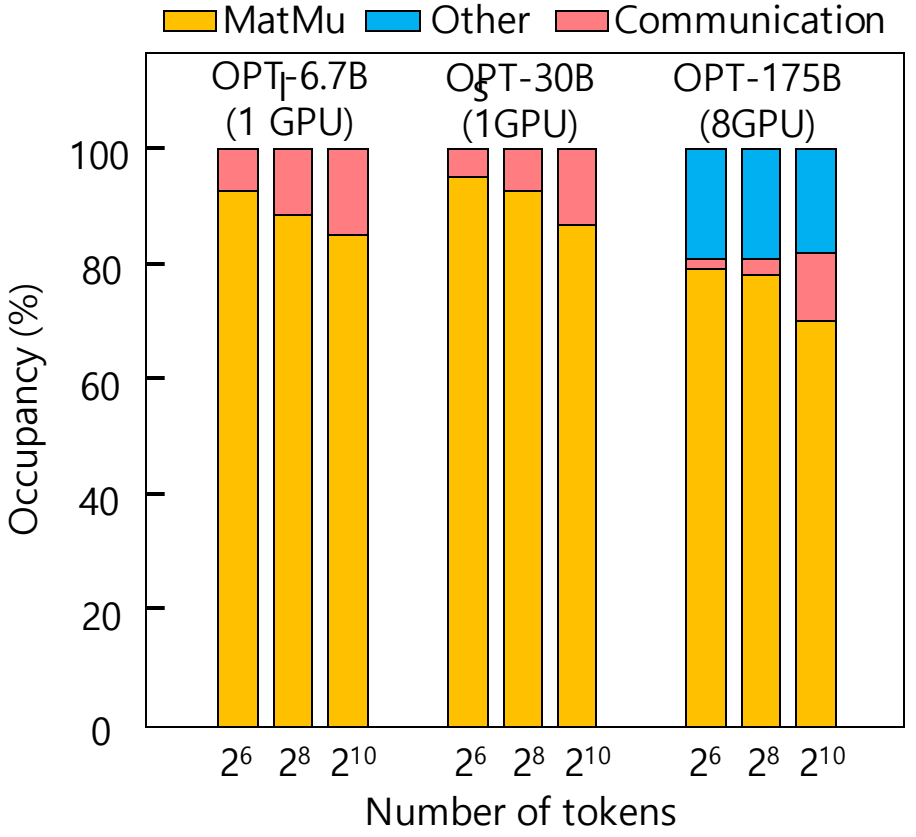
- 강력한 병렬 계산 기능
- 계산의 협동성 중요
- 데이터의 일치성, 동시성 중요, data coherence
- 안정성, 보완성, 회복성, 복구성
- Memory Coherence, Direct memory-sharing channel
- 다중 동시 서비스
- 집중적 공간 설치
- 새로운 클라우드 서비스로 탄생 가능 (AI_aaS)(GenAI_aaS)(Ai_Supercom_aaS)
- 기존 인터넷 서비스와 결합
- B-to-C 또는 B-to-B 서비스 가능
- 오픈 AI, 폐쇄형 AI 모델 모두 가능
- 다양한 Framework, 다양한 모델 서비스
- 오픈 소스 GPU/메모리 가능
- 전용 Customizer GPU/HBM 가능

Von Neumann Architecture



Occupancy of Matrix Multiplication(MatMul) for Transformer Model





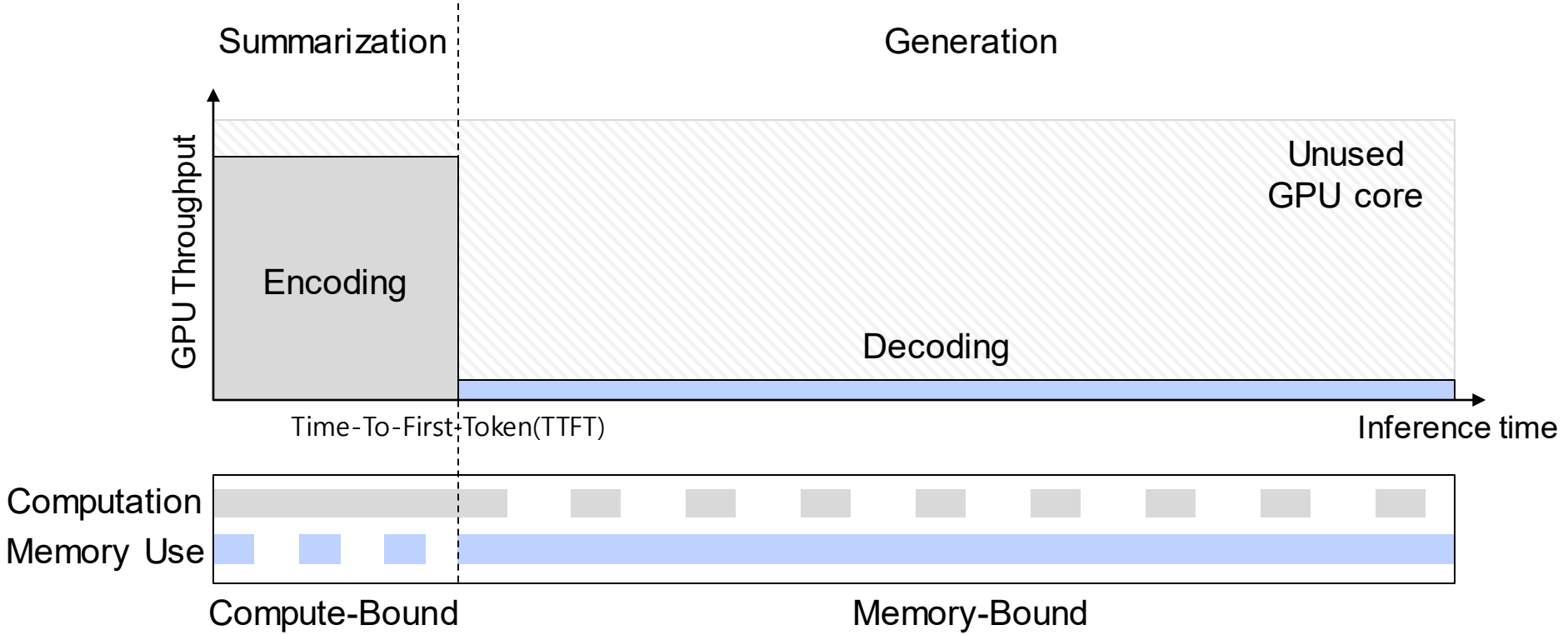
*OPT : Open Pre-trained Transformer

< Latency breakdown of various GPT and OPT models by MatMul

- Matrix multiplication(MatMul) computation dominates hyper-scale GPT model training process.
- The overall performance of GPU and memory for hyper-scale AI is dominated by how fast compute MatMul and how fast transfer data.

Ref)Naver

Limitation of current GPU-HBM for LLM Performance [2/3] : Bandwidth of Data Transfer between GPU and Memory

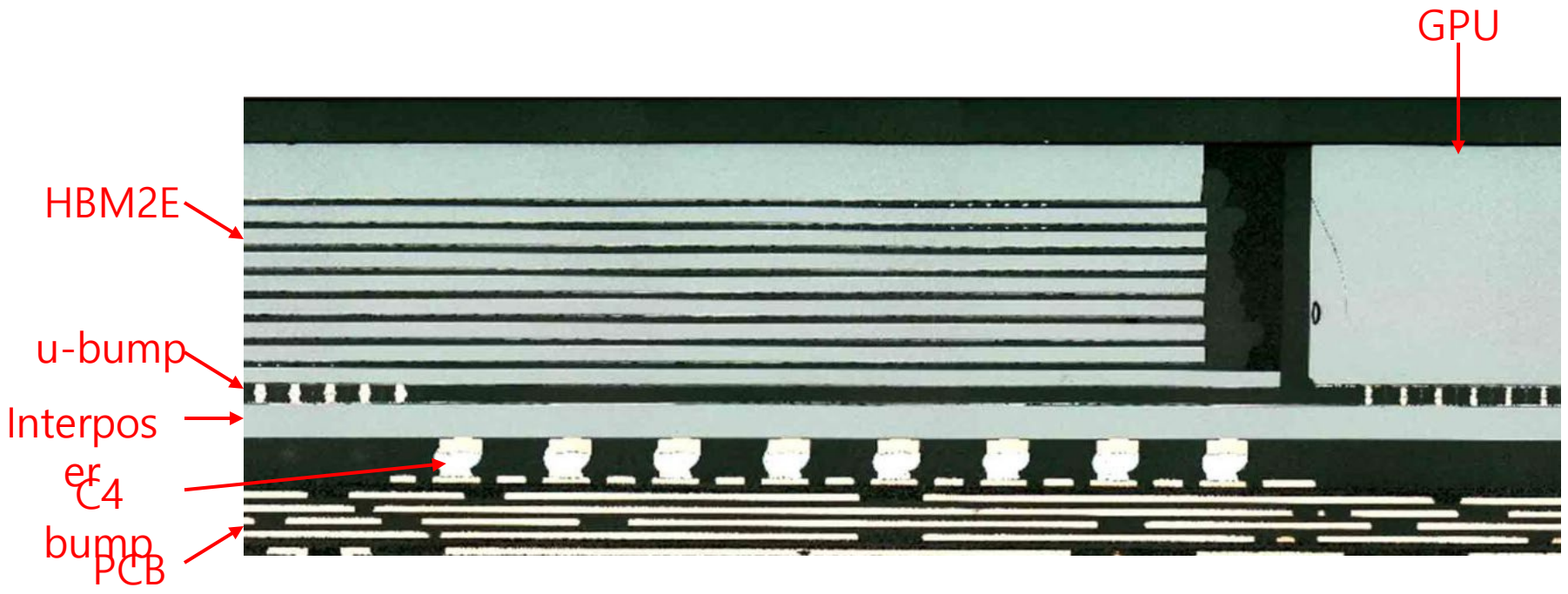


< Computation and Memory use during LLM Inference >

- The process of generating tokens through decoding during the LLM inference phase is a memory-bound operation.
- If the memory bandwidth of data transfer between the GPU and memory can be increased, it is possible to enhance processor utilization, thereby reducing the overall inference time.

Ref) Park, Gunho, et al. "Lut-gemm: Quantized matrix multiplication based on luts for efficient inference in large-scale generative language models." arXiv:2206.09557 (2023)

Cross-section of NVIDIA A100

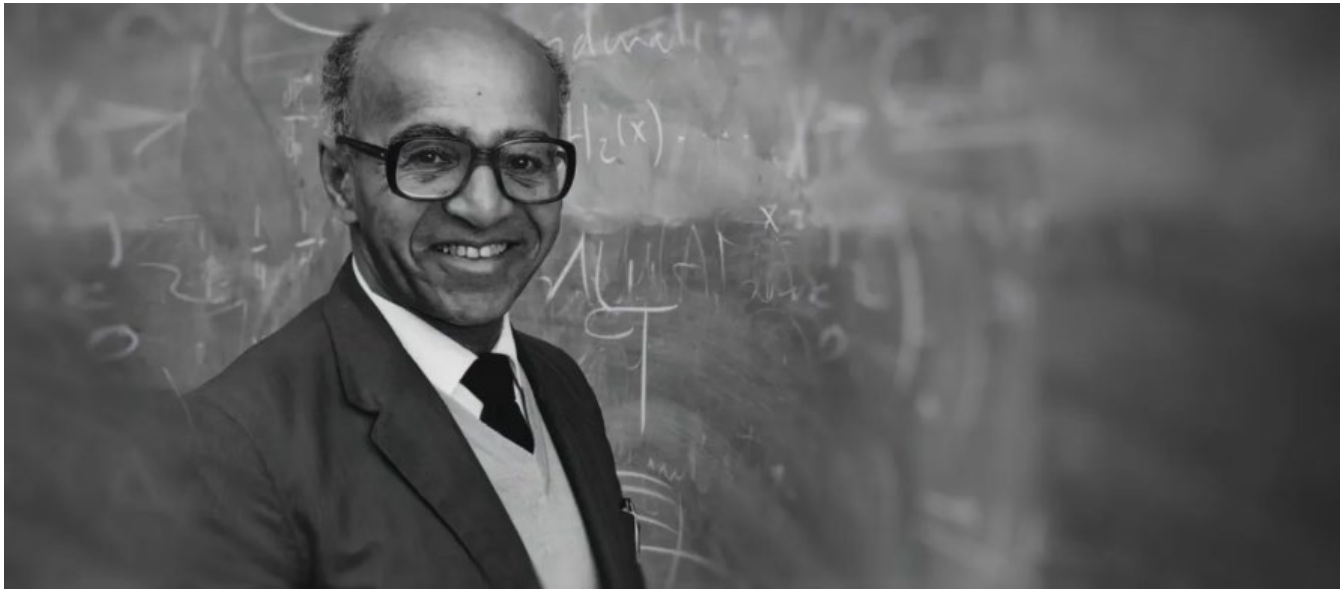


< Cross-section of NVIDIA A100 >

- HBM2E와 GPU를 연결하기 위해서 Interposer를 사용하는 CoWoS 기술이 적용되었음.

Ref) NVIDIA

Who is Blackwell?

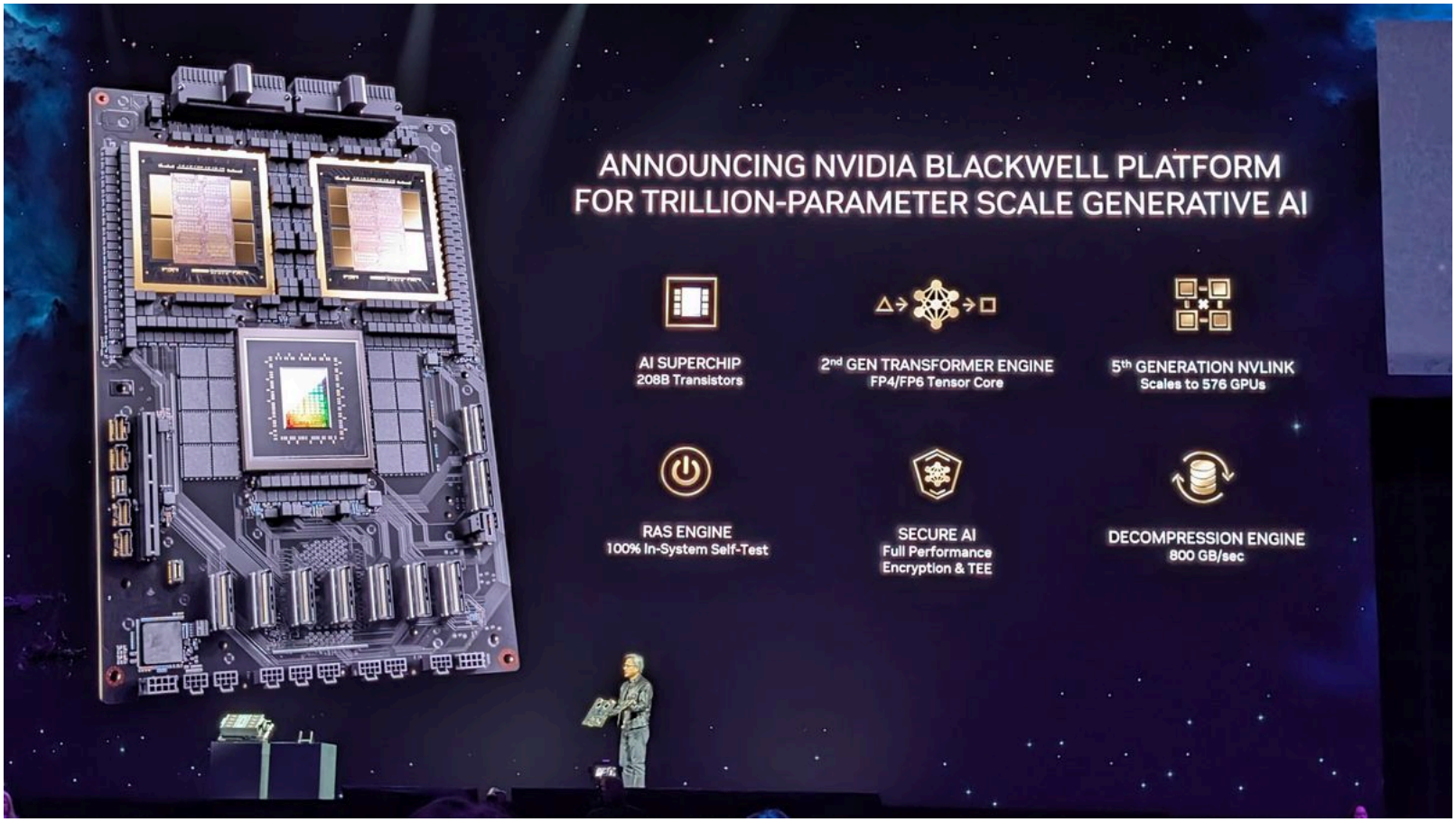


< David Harold Blackwell >

- David Harold Blackwell(1919 – 2010)은 통계, 게임 이론 및 정보 이론을 전문으로 하는 미국의 통계학자이자 수학자이며, Rao-Blackwell 정리를 정립한 인물임.
- Rao-Blackwell 정리는 두 확률 변수가 조건부 독립일 때, 조건부 분포의 엔트로피가 두 변수의 엔트로피의 조합과 같다는 정리로, **데이터 압축 및 머신러닝 등의 분야에서 중요하게 다뤄 짐**.
- 아울러, 그는 미국 국립과학원(National Academy of Sciences) 회원이 된 최초의 아프리카계 미국인입니다.
- 위의 업적을 기려, NVIDIA 에서 차세대 GPU 를 Blackwell 로 명명한 것으로 사료됨.

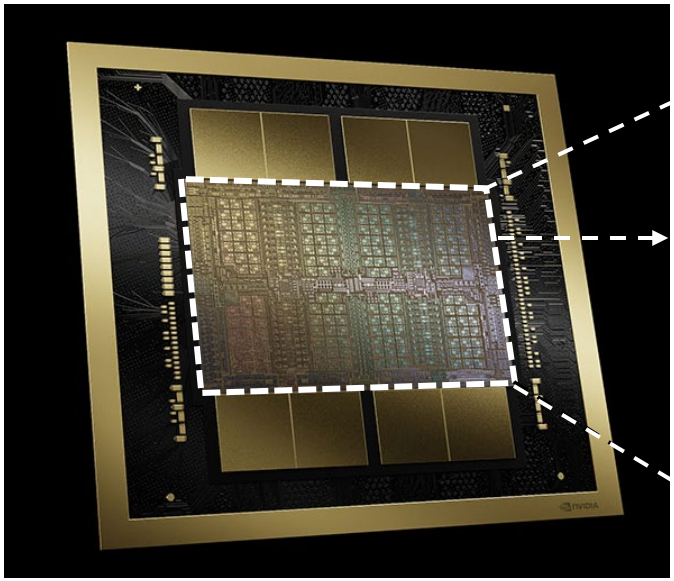
- Blackwell 이전에 개발된 Grace CPU 와 Hopper GPU 는 미국의 컴퓨터 과학자이며 프로그래밍 언어 코볼의 개발을 주도한 Grace Hopper 해군 제독의 이름에서 따옴.

Blackwell Architecture by Nvidia



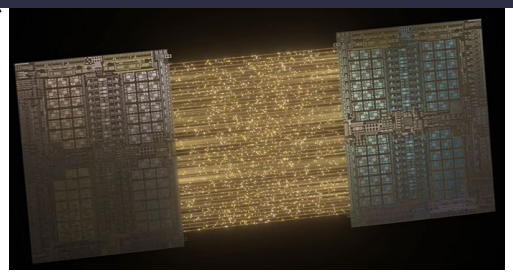
- 2026 년 양산 예상
- Nvidia CPU (Grace), Nvidia GP(2 Blackwell GPU), 8 HBM3e
- NVlink
- AI Supercomputer, 72GPU

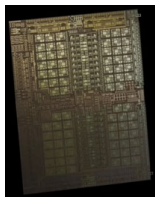
Comparison between Blackwell and Hopper - Chip Level



B200

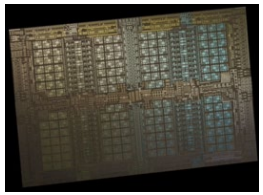
< NVIDIA B200 module >





Largest chip as possible

- 1040 억개 트렌지스터
- TSMC 4NP 공정



Two largest dies possible -unified as one GPU

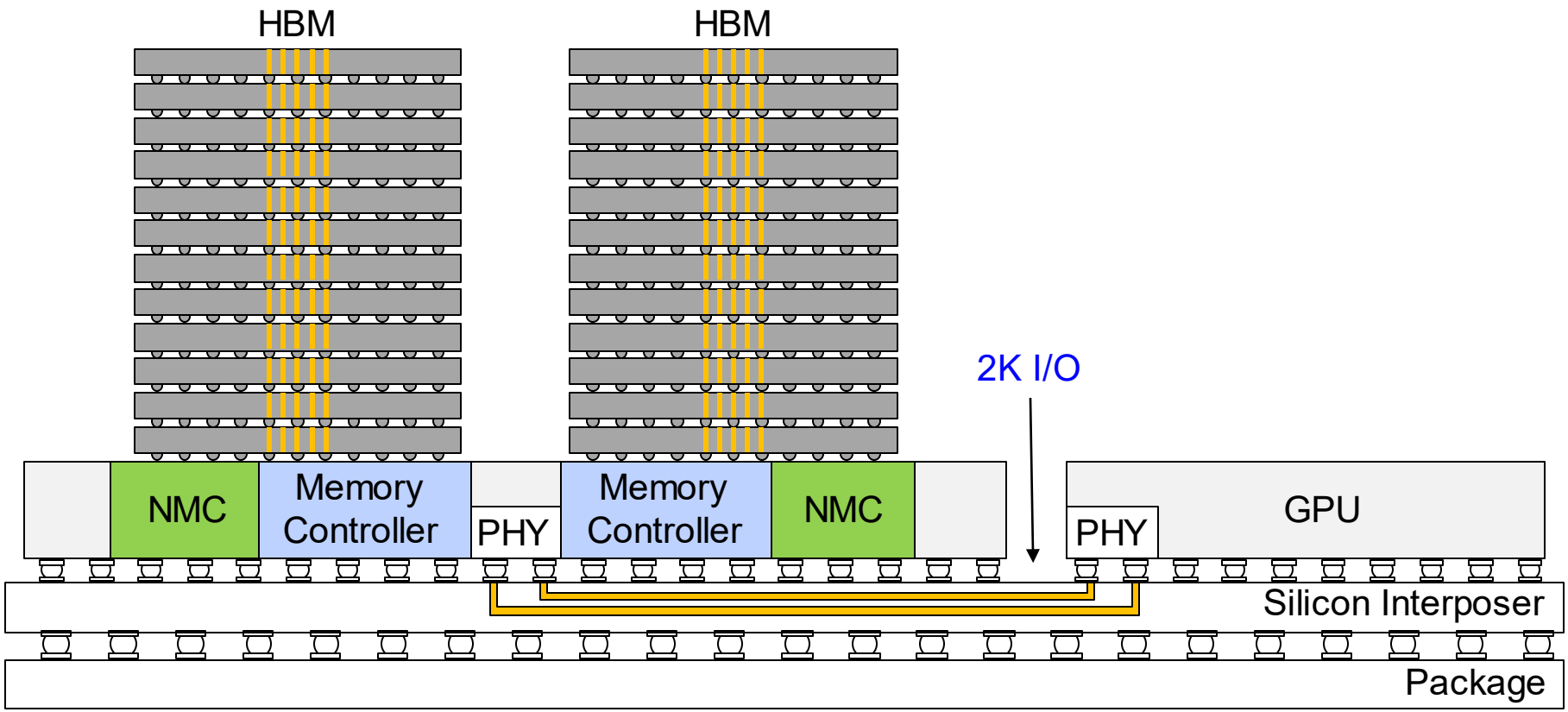
- 2080 억개 트렌지스터
- TSMC 4NP 공정
- Full cache coherence

- B200 은 2개의 GPU를 합친 구조를 가지며, 2000억개 이상의 트랜지스터로 구성됨.
- 현재 물리적으로 가능한 가장 큰 사이즈의 칩을 cache coherence 를 보장하며 하나의 unified GPU 로 구현한 구조임.
- 두 칩 사이는 10 TB/sec 의 NV-HBI (NVIDIA-High Bandwidth Interface)로 연결되어 하나의 Die 로 동작하게 됨.

Ref) NVIDIA

Ref) NVIDIA, GTC 2024

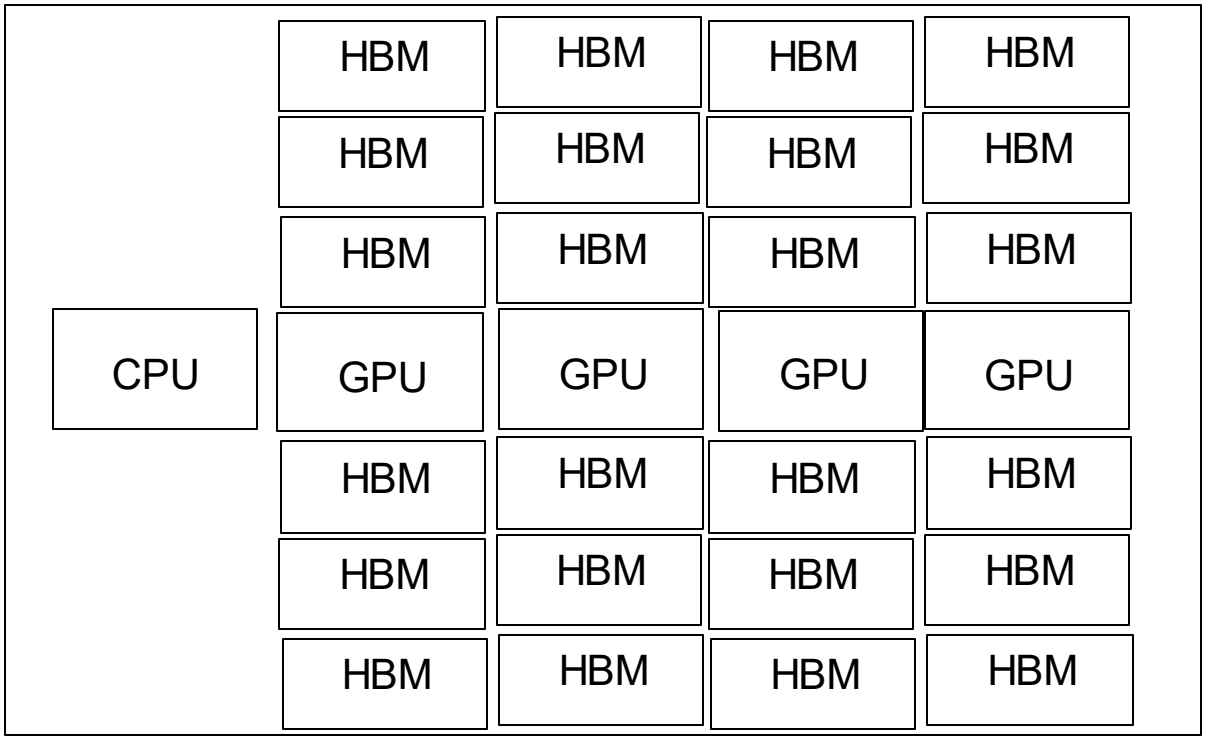
Proposal of Twin HBM-NMC Architecture



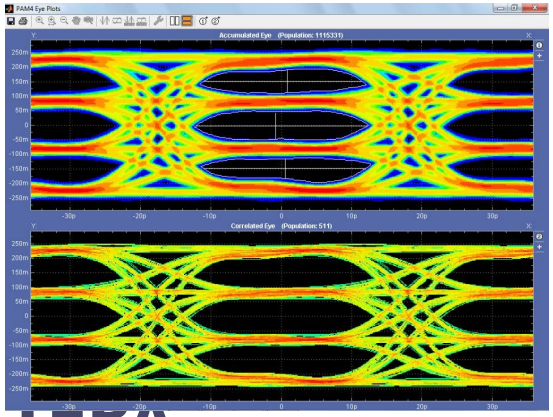
< Overview of Twin HBM with NMC architecture >

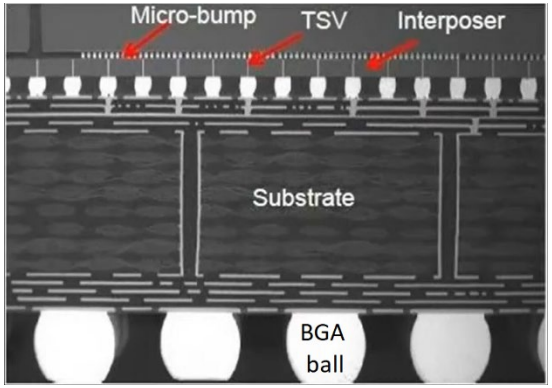
- Two memory stacks are located on top of a logic die.
- The logic die contains an NMC and a Memory Controller.
- The logic die and the GPU are connected through 2048 interposer channels.

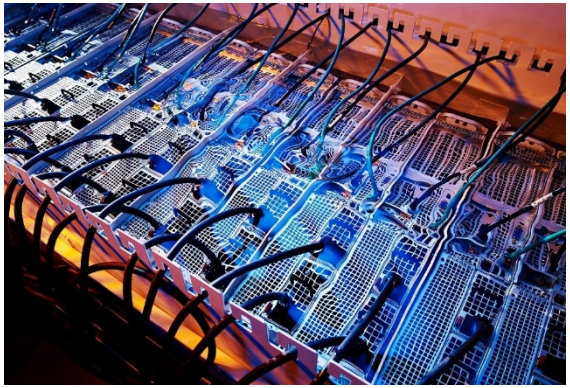
Next Generation HBM n.0



1) Multi-array HBM
2) >10K I/O
3) Si interposer
4) Passive and active embedded Si interposer
5) Immersion Cooling
6) 100Bps
7) PAM-N
8) CXL
9) C-to-C interface
10) GPU-to-GPU interface

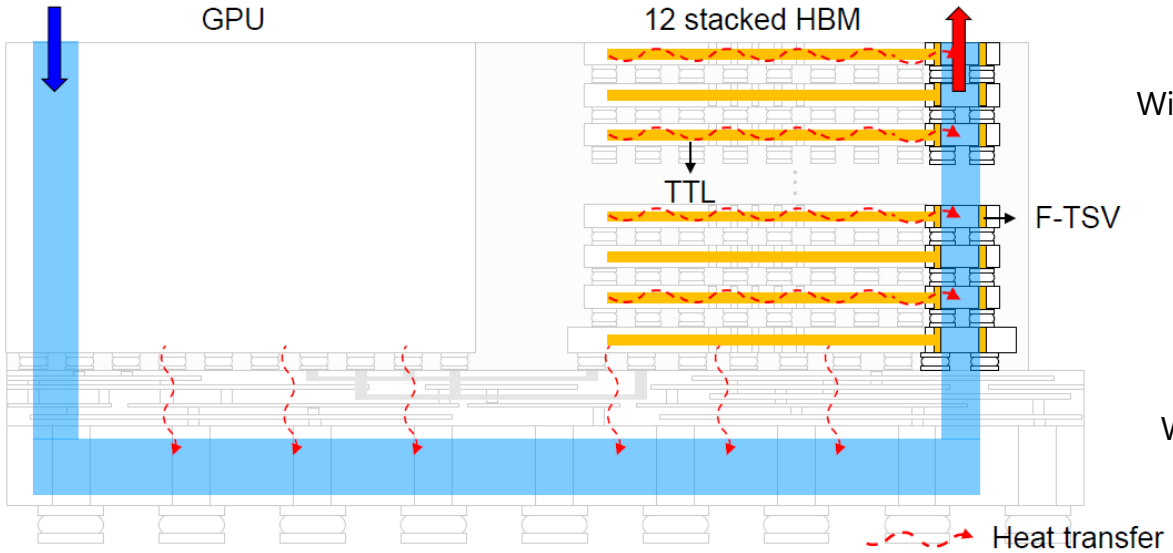




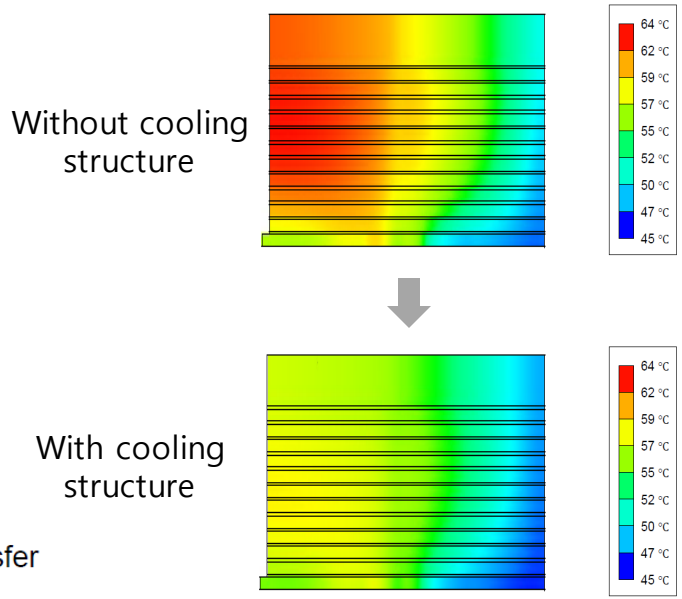


HBM Packaging Based Research at Teralab
: Embedded Cooling Structure for Next-Generation HBM Module

Proposed TTL & F-TSV for HBM Cooling Structure



Simulated Thermal distribution of HBM

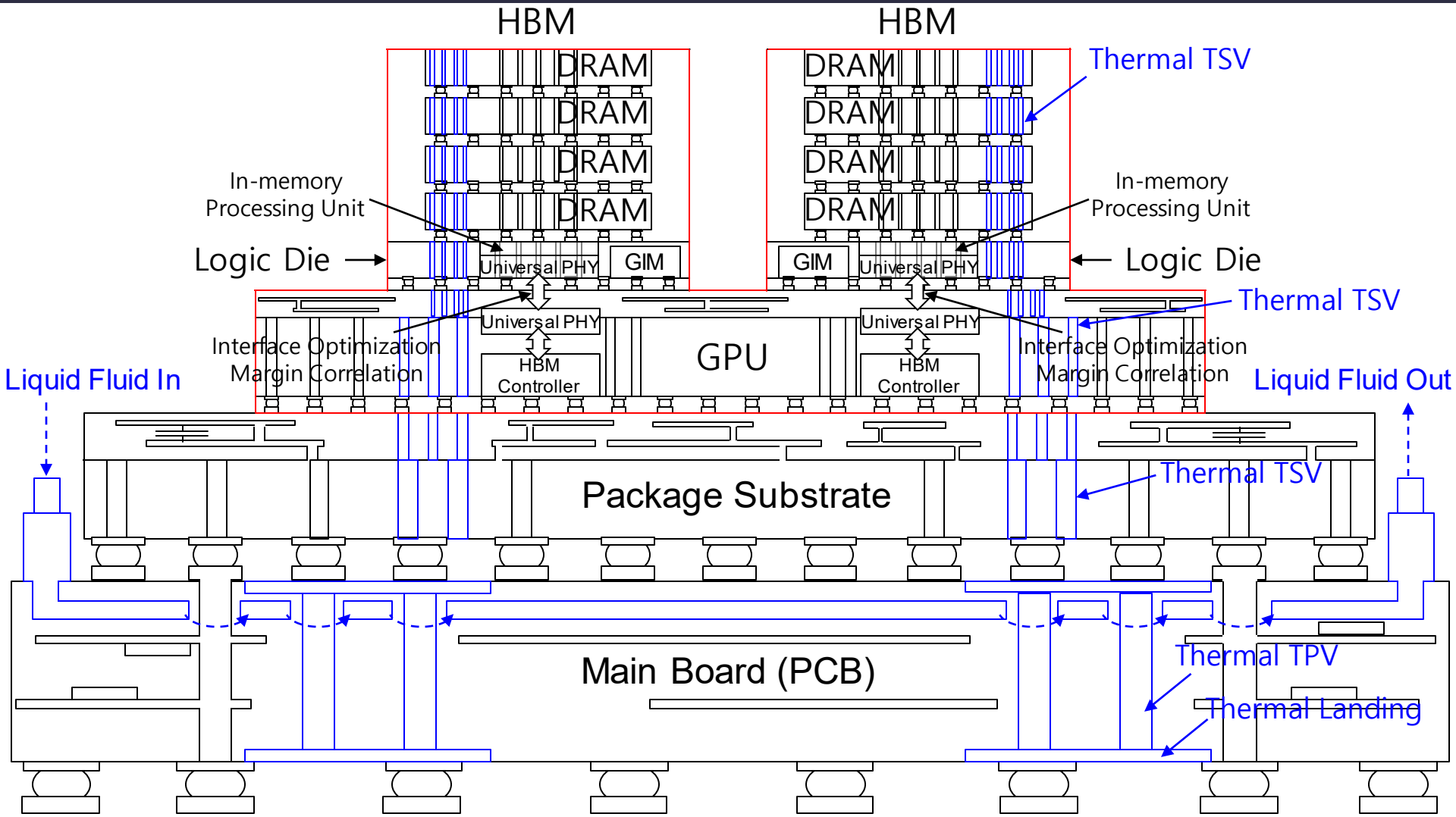


< Thermal Transmission Line(TTL) and Fluidic Through Silicon Via(F-TSV) based Embedded Cooling Structure >

- Proposed thermal transmission line(TTL) and fluidic through silicon via(F-TSV) based embedded cooling structure for improved thermal and signal integrity performance of next-generation HBM module.

Ref) Keeyoung Son, Teralab

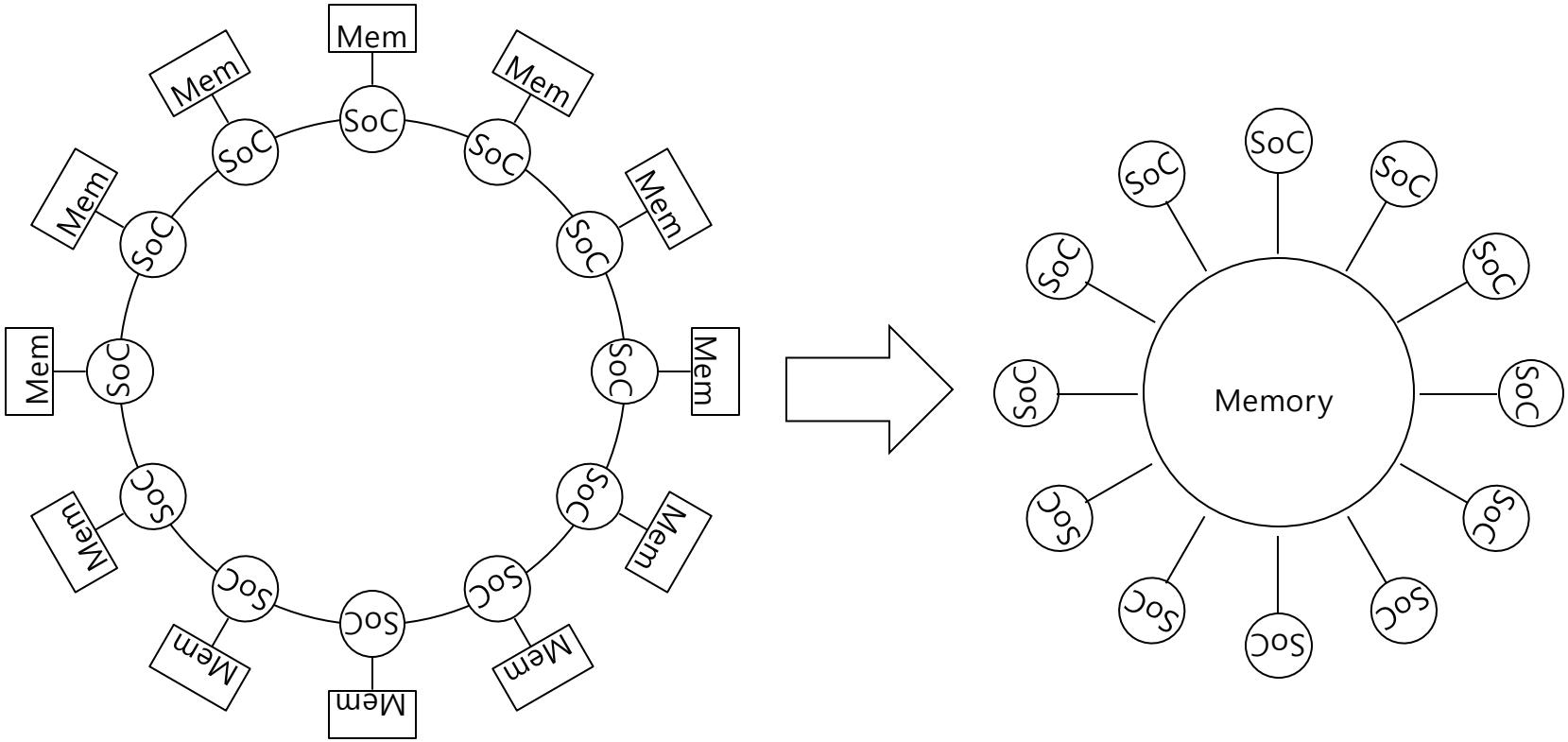
Thermal TSV and TLV for cooling



< Conceptual figure of fully stacked 3-D ICs for next generation HBM interface >

Ref) Joungho Kim, HBH(High Bandwidth Memory) Designs for Next Generation AI Server Applications

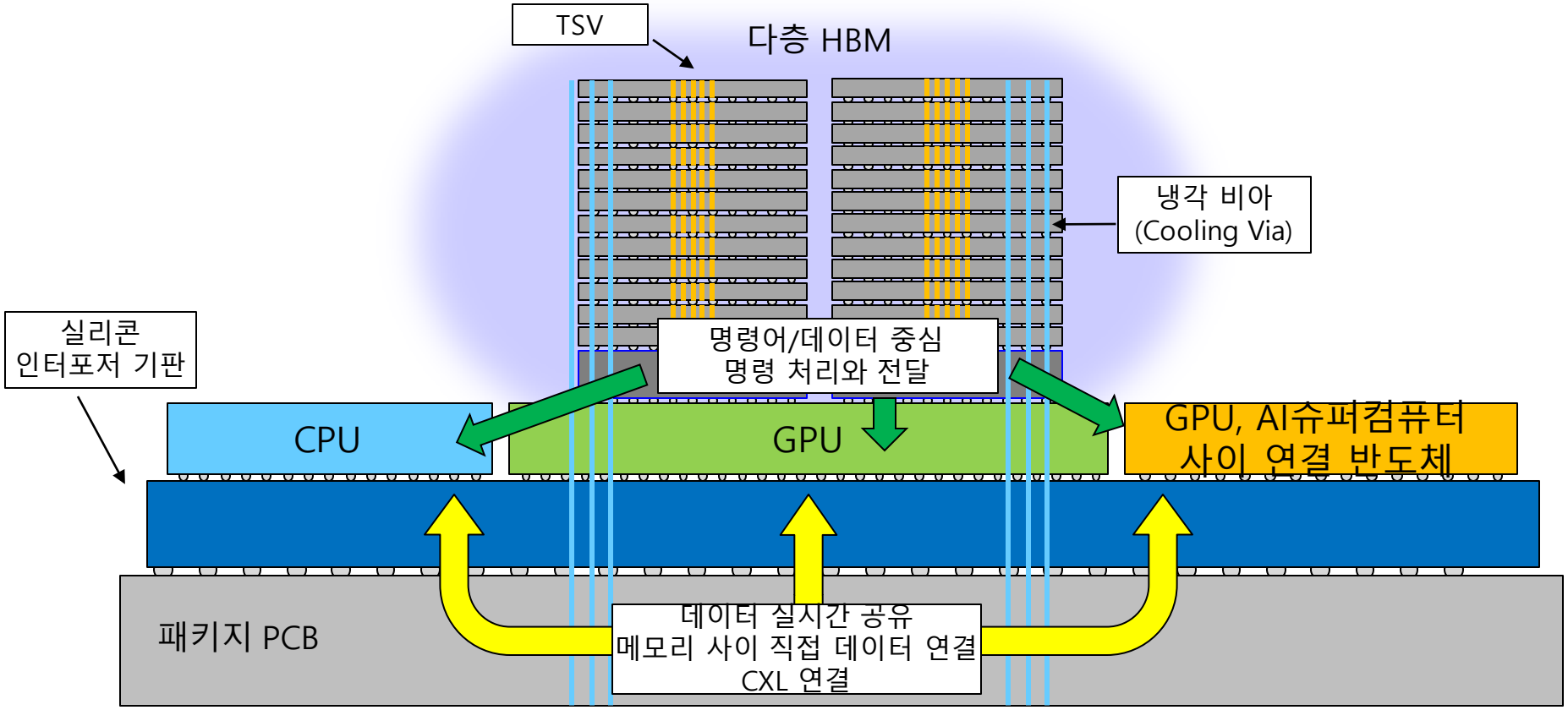
Processor-centric to Memory-centric Computing



< Processor-centric to Memory-centric Computing >

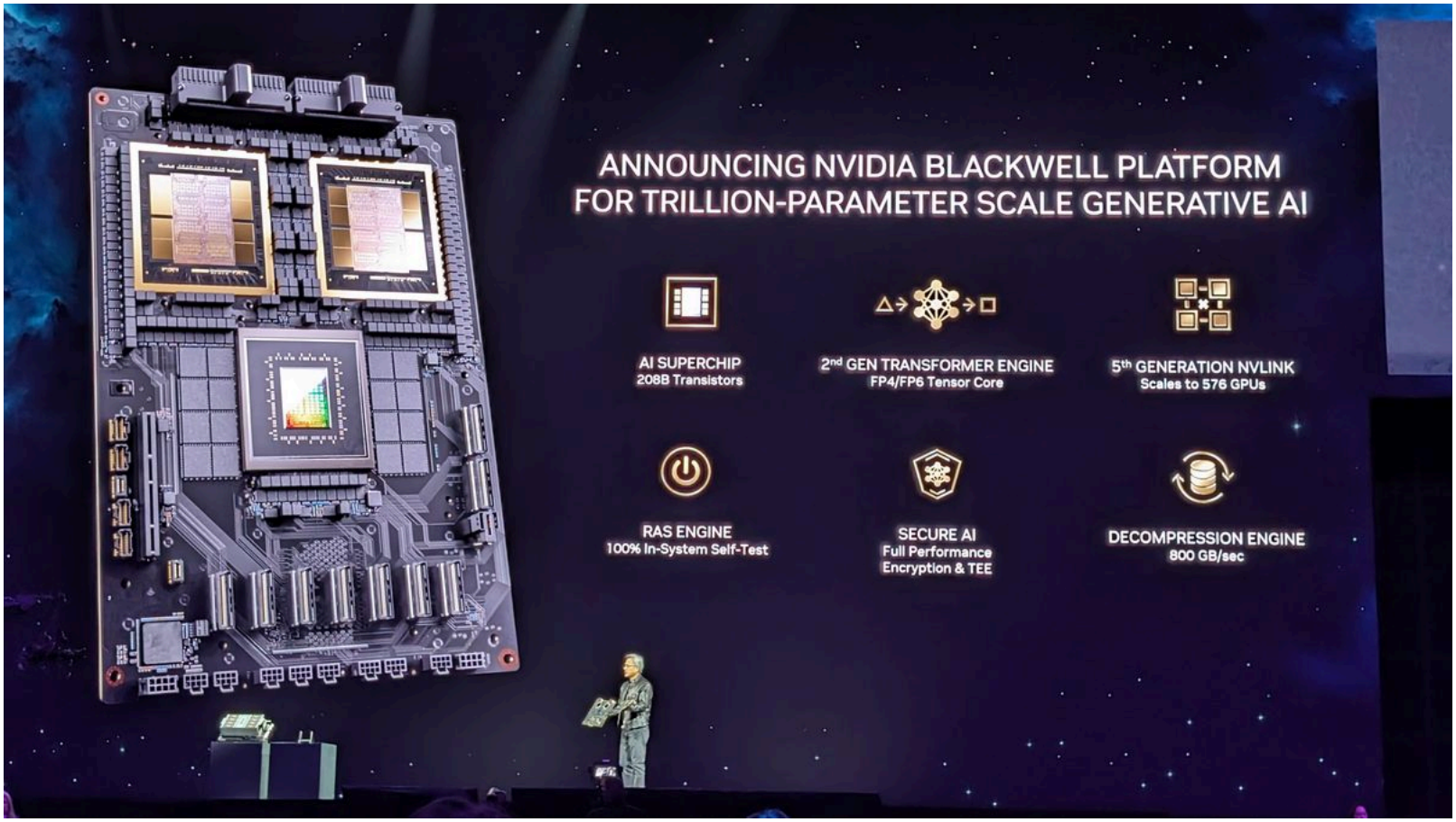
- In the existing Von-Neumann structure, the processor was at the center, data was repeatedly loaded and de-loaded from distributed memory several times, and there was a lot of data movement between processors.
- In order to reduce unnecessary data movement between processors, it must be changed to memory-centric computing where multiple processors can access a huge amount of memory at the same time.

미래의 3차원 HBM 구조 (HBM Centric Computing 구조)



Part3: AI 반도체의 경쟁

Blackwell Architecture by Nvidia

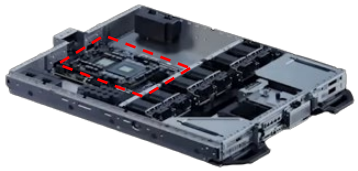
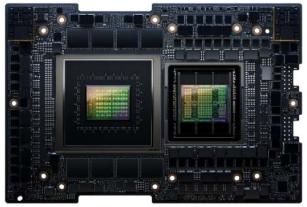


- 2026 년 양산 예상
- Nvidia CPU (Grace), Nvidia GP(2 Blackwell GPU), 8 HBM3e
- NVlink
- AI Supercomputer, 72GPU

Exascale Supercomputer for Hyperscale AI Computing

NVIDIA → Google Cloud Meta Microsoft

x16 Rack = 32 compute node = 256 Grace Hopper Super chip

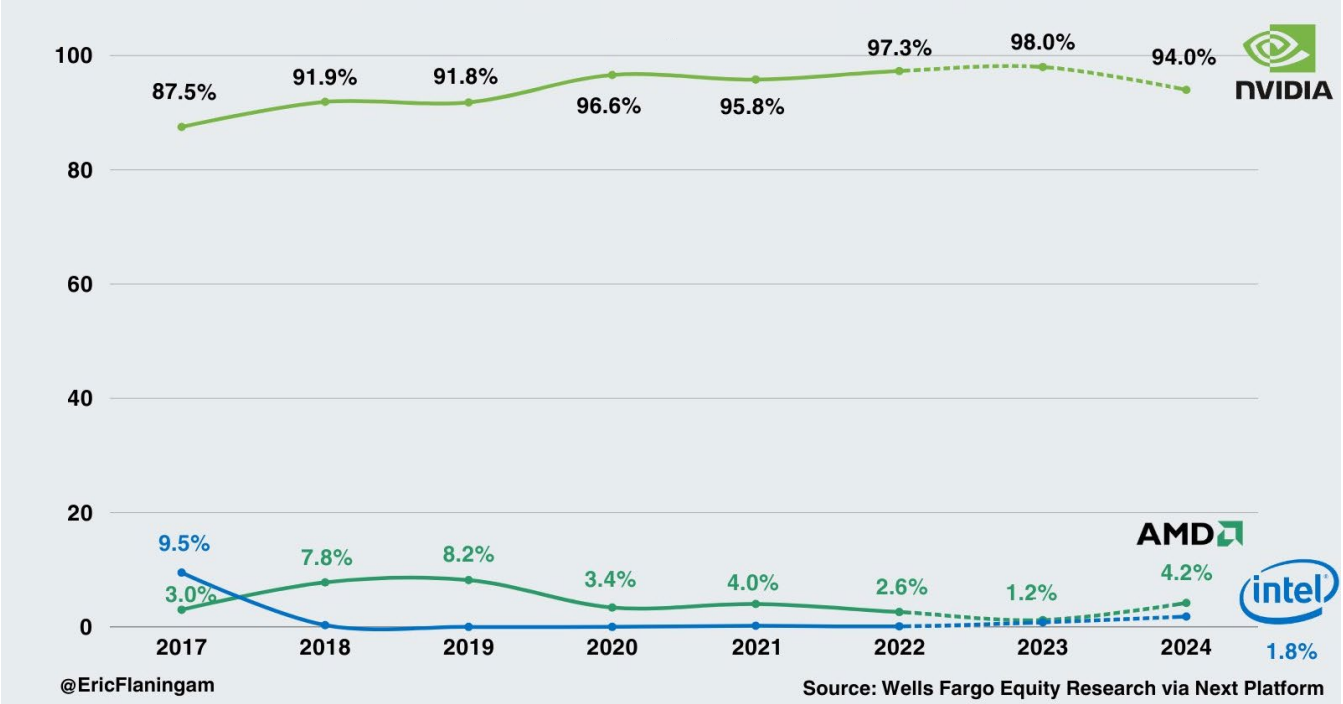
Rack	Chassis	Compute tray	GPU module
			
✓ 2 chassis ✓ 2 computer node	✓ 1 Computer node ✓ 8 Compute trays ✓ 3 NVLink 1st Switch trays	✓ 1~2 GPU module ✓ 200/400Gbps Infiniband ✓ 200Gbps Ethernet ✓ 12 TB data/OS storage	✓ CPU + GPU ✓ **HBM** ✓ NVLink ✓ PCIe gen

x2 ← x8 ←

- In the supercomputer for hyperscale AI computing, most of the computing power is derived from GPUs that require high bandwidth memory (HBM).

Ref) DGX GH200 for Large Memory AI Supercomputer

AI 반도체 시장을 독점하는 NVIDIA



< 매출액 기준 데이터 센터 GPU 시장 점유율 추정 >

- 현재 데이터센터 GPU 시장 점유율은 NVIDIA가 90% 이상 차지.
- 대다수 빅테크는 대안이 없기 때문에 그래픽처리장치(GPU)에 고대역폭메모리(HBM)를 붙인 NVIDIA의 AI 반도체를 학습, 추론 구분 없이 활용 중.
- 지난해 엔비디아 서버 매출의 40%(188억달러, 약 25조원)가 추론 분야에서 나옴.
- NVIDA의 AI 가속기는 학습에는 적합하나 추론에는 비효율적이라는 지적도 존재.
  - ➢ 학습 시 GPU의 병렬 연산이 유리하나, 추론에선 GPU와 HBM이 대량의 데이터를 주고받는 과정에서 '병목 현상'이 발생해 속도 저하와 대량의 전력 소모가 발생하기 때문

Ref 1) WCCFtech, " NVIDIA To Dominate AI Share In 2024 With An Estimated $46 Billion In Data Center Revenue"
Ref 2) 세계일보, " 삼성 'HBM·마하1' 투트랙 전략... 반도체 판 흔든다"

Nvidia Stock Price

주식 시장 요약 > 엔비디아

881.86 USD

+837.34 (1,880.82%) ↑ 지난 5년

+ 팔로우

폐장: 4월 12일 오후 7:59 GMT-4 • 면책조항

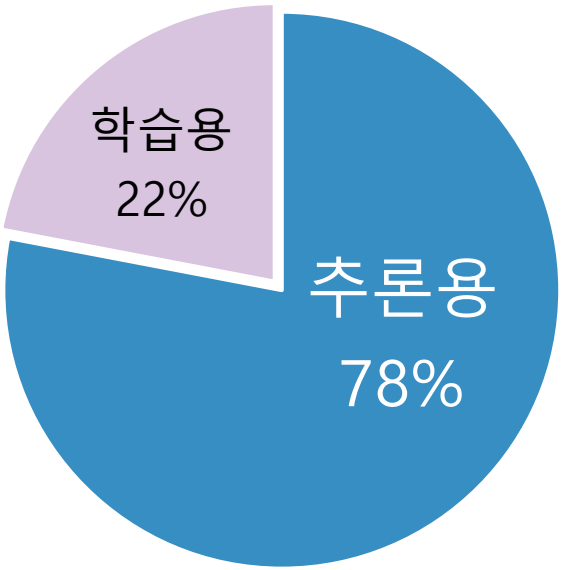
폐장 후 875.25 −6.61 (0.75%)

1일 | 5일 | 1개월 | 6개월 | 연중 | 1년 | 5년 | 최대

시가	896.99	시가총액	2.20조	52-주 최고	974.00
최고	901.75	주가수익률	73.90	52-주 최저	262.20
최저	875.30	배당수익률	0.018%		

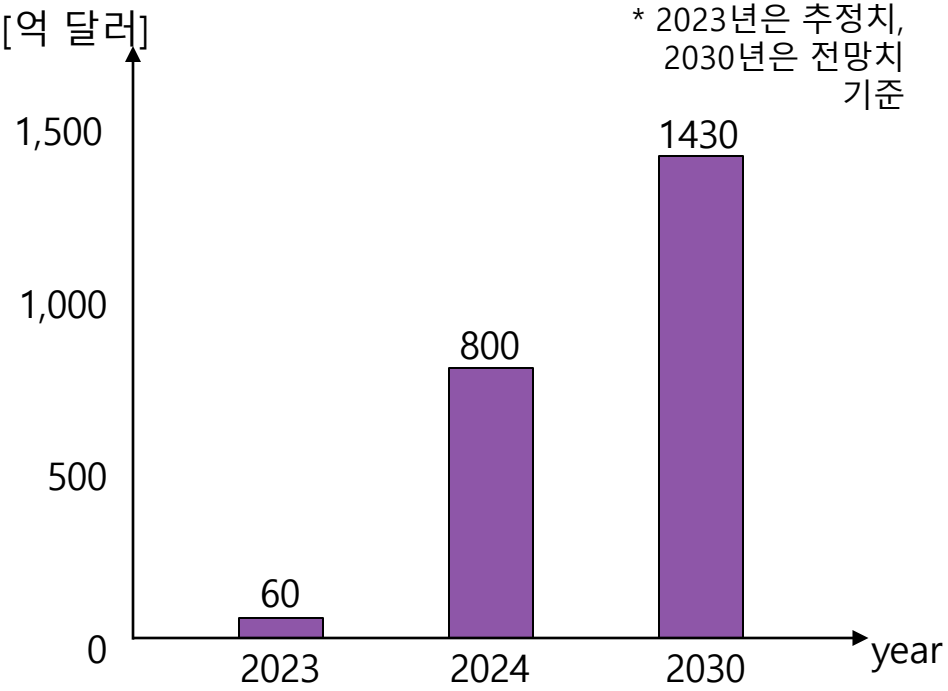
중요해지는 AI 추론용 반도체

자료: 트랙티카
* 2025년 전망치 기준



< AI반도체 용도별 비중 >

자료 : 옴디아
* 2023년은 추정치,
2030년은 전망치
기준

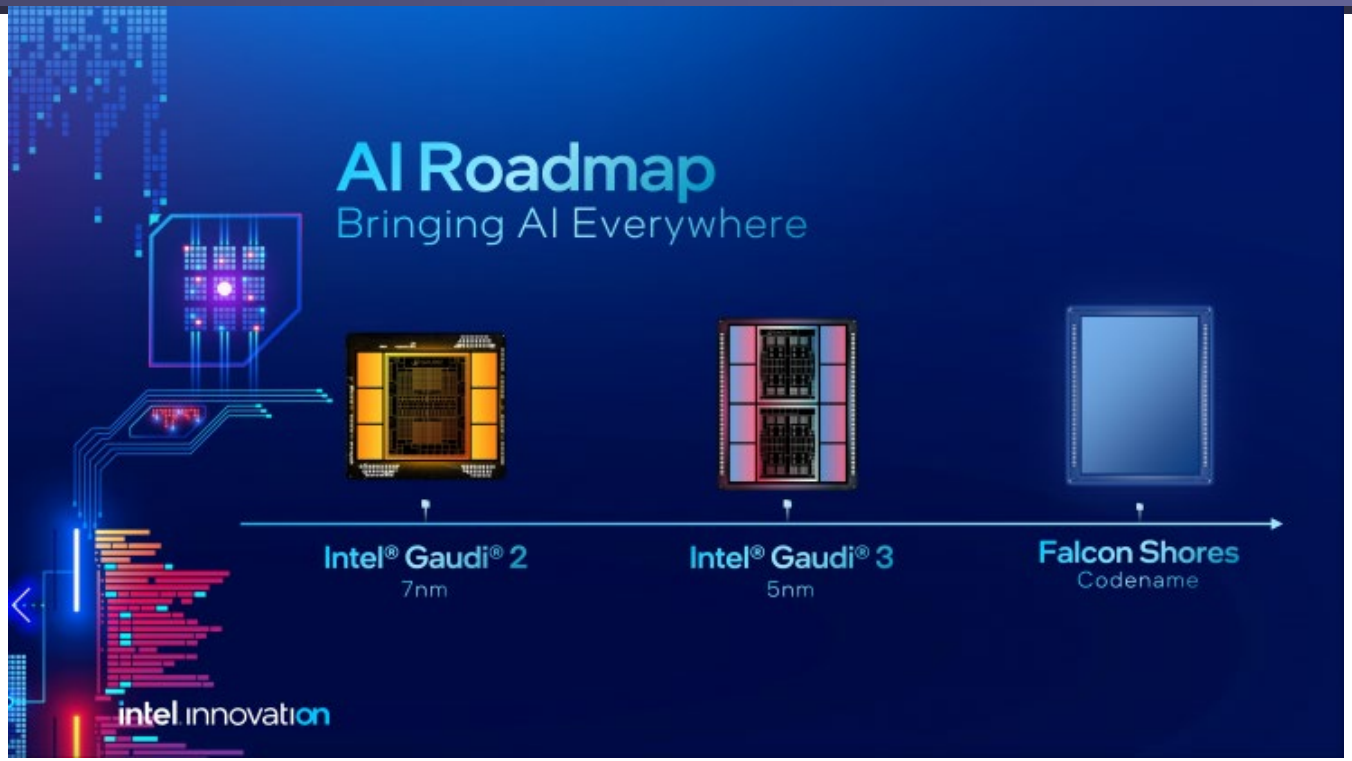


< 성장하는 AI 추론용 반도체 시장 >

- 특화된 AI 서비스 수요가 늘어나는 데 따라 추론용 AI 가속기 시장은 계속 커질 것으로 전망.
- 2025년 기준 AI 가속기 중 추론용 비중이 78%로 학습용도 (22%)의 세 배가 될 것이란 조사 결과 존재.
- 2023년 60억달러이던 추론용 AI 가속기 시장 규모는 2030년 1430억달러로 확대될 것으로 전망.

Ref) 한국경제, ""5만弗 AI 가속기는 비싸"...삼성전자 '마하1'으로 빅테크 잡는다"

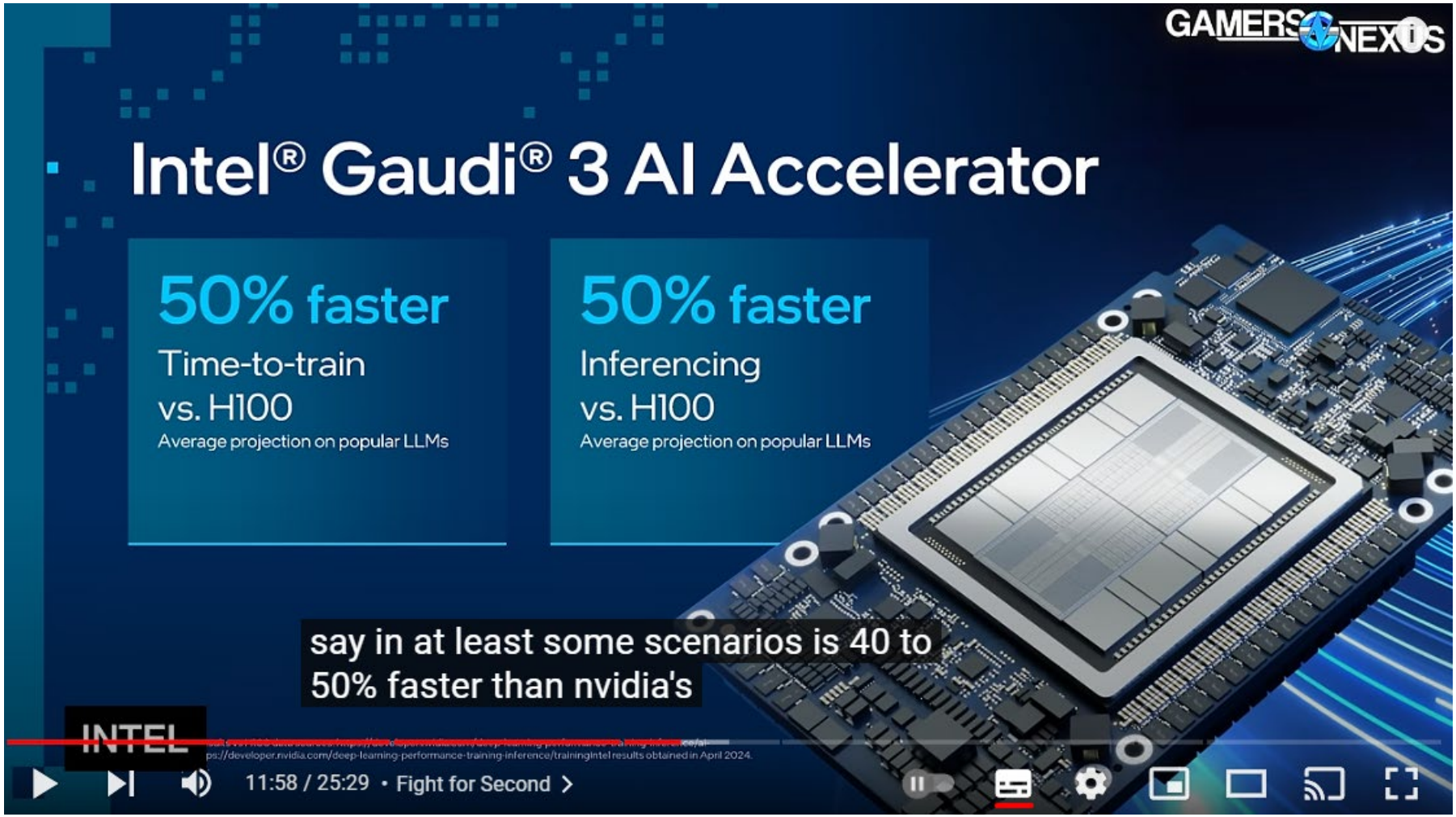
Gaudi: Intel AI GPU



< Intel AI Accelerator roadmap >

- 인텔은 2019년 Habana Labs (이스라엘) 인수 이후 Gaudi라는 AI 가속기 모델을 출시 중.
- 가우디3은 2024년 출시 예정이며, 가격은 미정. (가우디2의 경우 현재 2300만원 정도)
  - ✓ TSMC 5nm 공정으로 제작

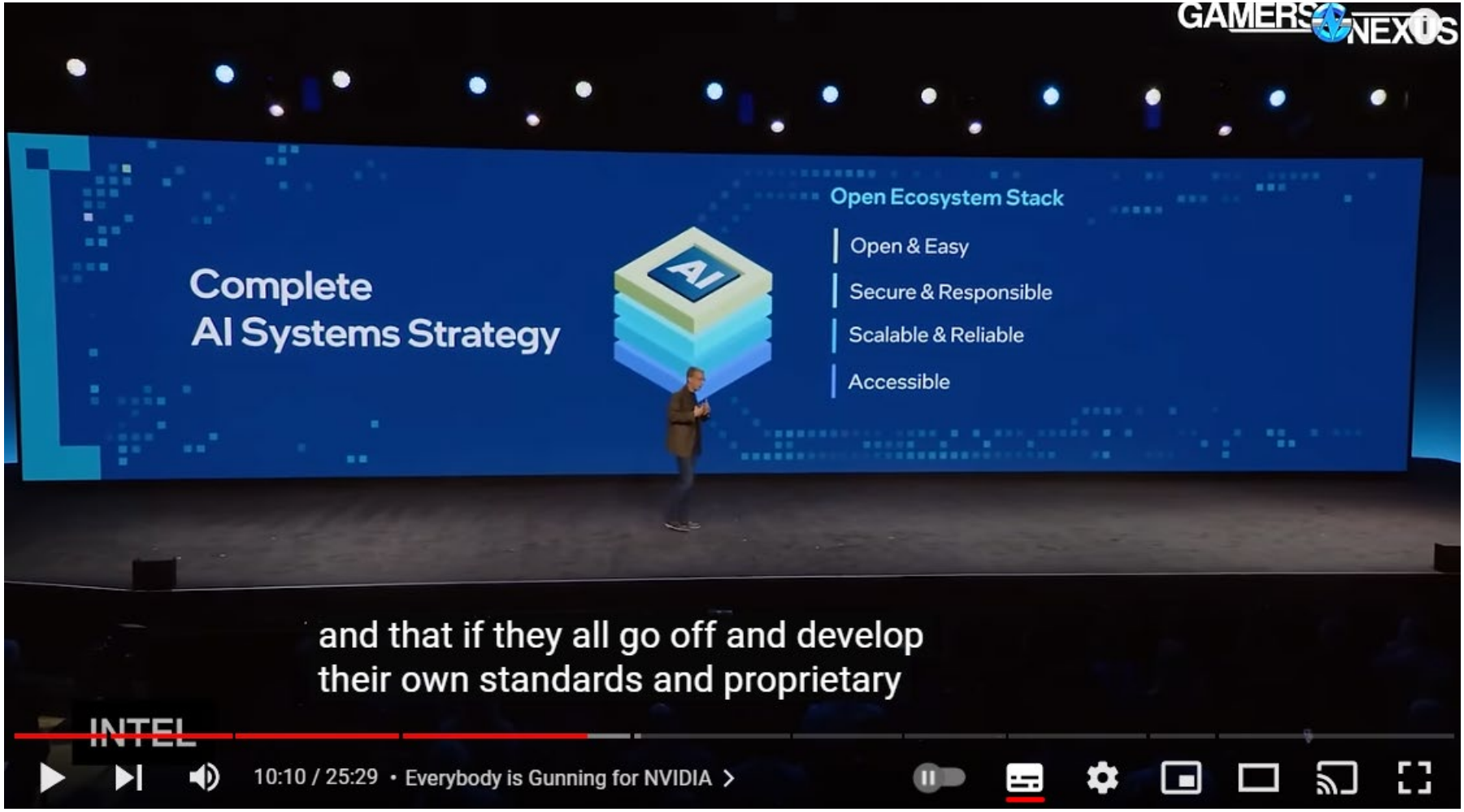
Intel and Gaudi: 유튜브 정리



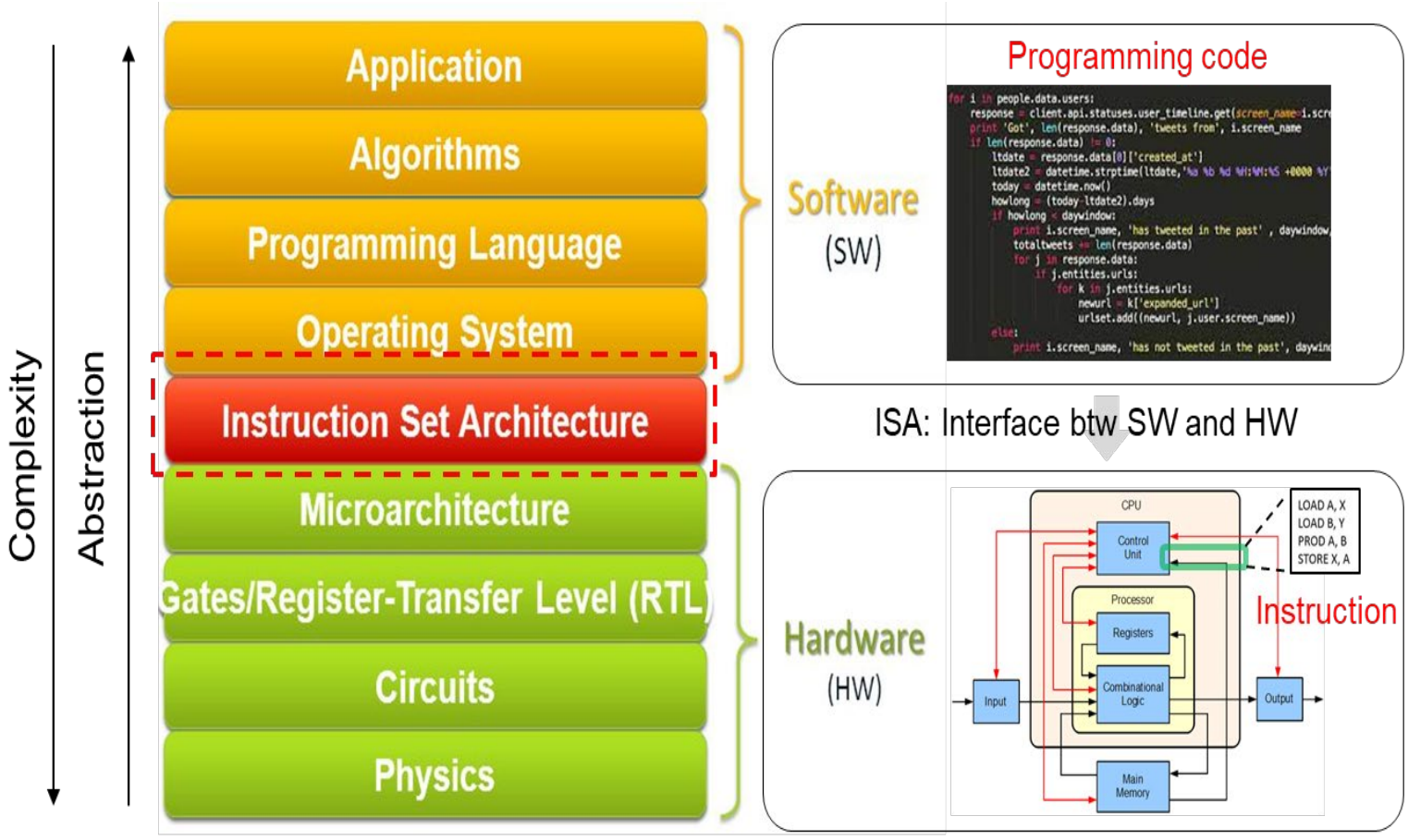
Intel 의 사업 전략

- Nvidia 추격
- 별도의 생태계 구축
- 경량화, LLM 모델에 적용
- HBM2 채용
- GPU라고 부르지 않고 AI 가속기라고 부름
- 자체 CPU, GPU 사용, 인텔의 자체 파운드리 사용
- 결국 블랙웰 (엔비디아 CPU, GPU 채용, HBM3e) 와 격전
- 엔비디아, AMD(MI300X), 인텔의 3국지 예상
- Naver-KAIST 와 협력 (응용, 생태계 구축, 데이터 센터 운영 경험, software stack, 가우디 이용 AI 모델 개발)
- MS가 파운드리 협력 등 장기적으로 MS, open 야와 협력 가능,
- 미국 정부의 적극적 지원 가능

Intel and Gaudi: 유튜브 정리



AI 반도체 설계를 위한 소프트웨어 Stack



< Computing Software stack: 소프트웨어와 하드웨어 계층 구조 >

Unified Acceleration Foundation (UXL) 의 목적과 설립

UXL
Unified Acceleration Foundation

arm FUJITSU Google Cloud
Imagination intel Qualcomm
SAMSUNG vmware by Broadcom

< Unified Acceleration Foundation (UXL) 과 참여기관 >

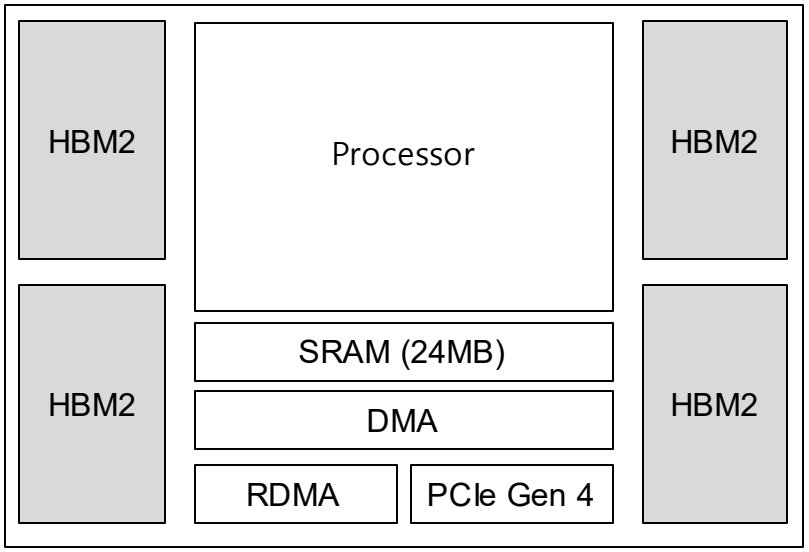
- Linux foundation 이 2023년 9월에 Unified Acceleration Foundation (UXL) 을 설립함.
- UXL 의 목적 : NVIDIA 의 CUDA 와 AMD의 ROCm 과의 경쟁
  - ✓ Build a multi-architecture multi-vendor software ecosystem for all accelerators.
  - ✓ Unify the heterogeneous compute ecosystem around open standards.
  - ✓ Build on and expand open-source projects for accelerated computing.
- 참여회사 : ARM, Fujitsu, Google Cloud, Intel, Qualcomm, Samsung, VMware 등
  (*독자적인 Accelerator 전략인 있는 NVIDIA와 AMD는 참여하지 않음.)

Ref) UXL foundation, Github

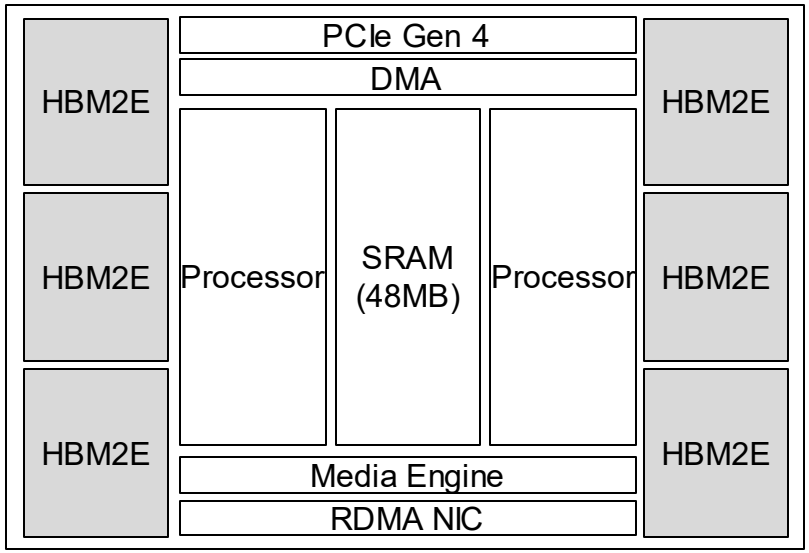
인텔의 기술적 전략

- Naver-KAIST 와 협력 (응용, 생태계 구축, 데이터 센터 운영 경험, software stack, 가우디 이용 AI 모델 개발)
- 삼성과 HBM3e, HBM4에서 협력 가능성
- SRAM 확대로 데이터 병목현상 완화
- GPU의 Data compression 기능으로 데이터 병목현상 완화
- GPU Data compression 기능으로 저전력 구현
- 추가로 Pruning, quantization 기법 가능
- 오픈 SW 개발 전략, UXL
- 오픈 GPU 연결망 전략 (ethernet, CXL 등, Nvidia 는 NVLink 사용)

Architecture and Hardware Spec of Gaudi Series



< Intel Gaudi 1 >



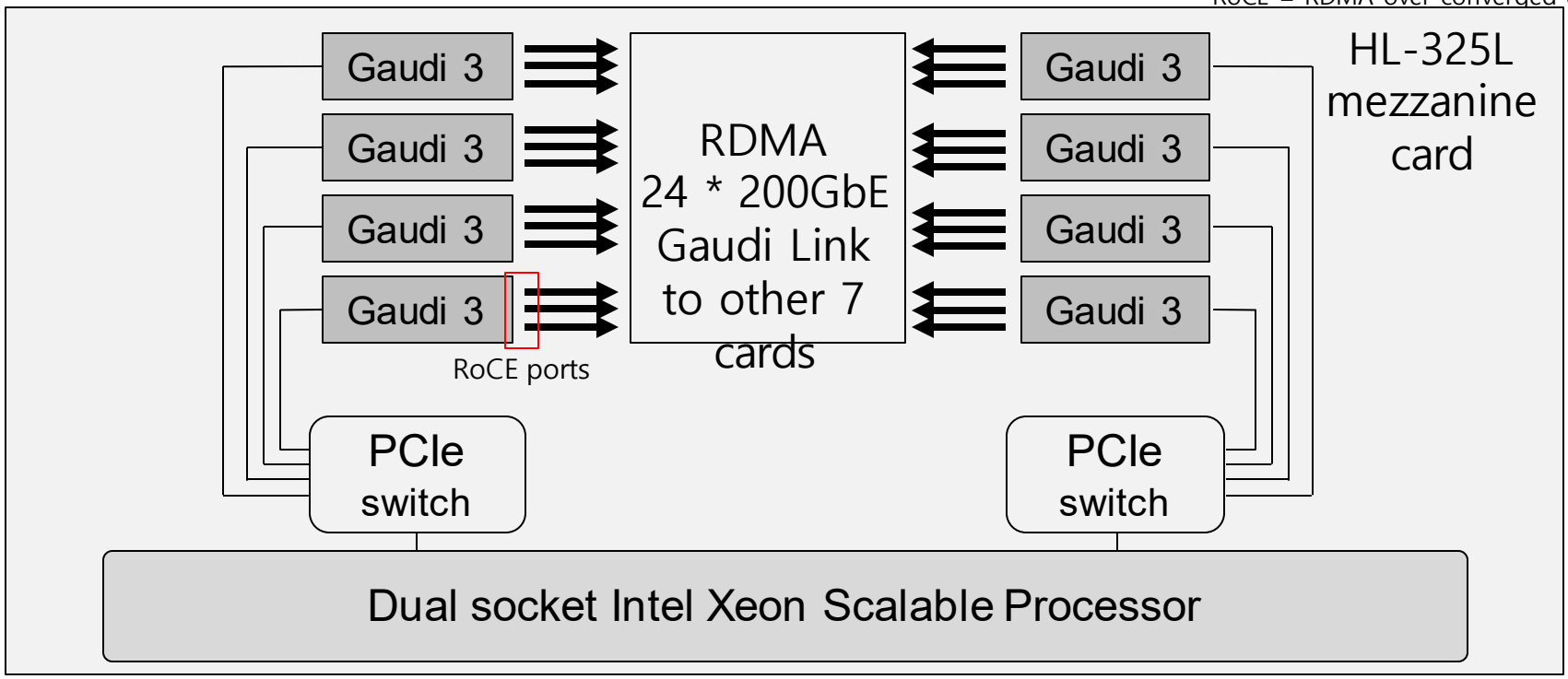
< Intel Gaudi 2 >

	Gaudi 1	Gaudi 2	Gaudi 3
대역폭	1 TB/s	2.45 TB/s	7.2 TB/s
캐시 사이즈	24 MB	48 MB	unknown
HBM 용량	32 GB	96 GB	144 GB
네트워크 BW	250 GB/s	300 GB/s	600 GB/s

- 가우디3의 상세 아키텍처는 아직 공개되지 않음.

Gaudi 3 Supercomputer: 8-GPU Universal Baseboard (UB)

*RoCE = RDMA over converged ethernet



< Intel Gaudi UB high level block diagram >

- HL-325L는 인텔의 Xeon CPU와 8개의 Gaudi 3를 융합한 Gaudi Universal Baseboard.
- 14.6 PetaFLOPS, > 1TB Memory, 29.6 TB/s HBM Bandwidth 지원
- CPU와 GPU는 16개의 PCIe Gen 5.0로 연결.
- RDMA(Remote Direct Memory Access)는 메모리 간의 통신을 CPU 개입없이 하는 기술.
- 8개의 GPU는 48 pair의 112 Gbps Tx/Rx PAM4로 통신하며, 이는 Gaudi들이 다른 GPU와 600Gb Ethernet으로 통신할 수 있게 함. (Intel Version의 NVLink = IntelLink)

네이버와 협력 이유

- 하이퍼 클로버 X, foundation model 보유
- 네이버 데이터 센터 운영
- 네이버 software stack 채용
- 네이버 사용자 경험 배움
- 점진적 확장 (고성능, 학습용)
- 추론용에서 시작해서 학습용으로 확대
- 다음 GAUDI, 4,5,6 예상

네이버와 협력하는 삼성전자

NAVER

SAMSUNG

서비스 경험

최적화를 위한 Requirement 제공.

AI 모델링

모델에 대한 지식/ 이해 제공

경량화

모델압축 기술제공 HW Design 제안

HW / Memory

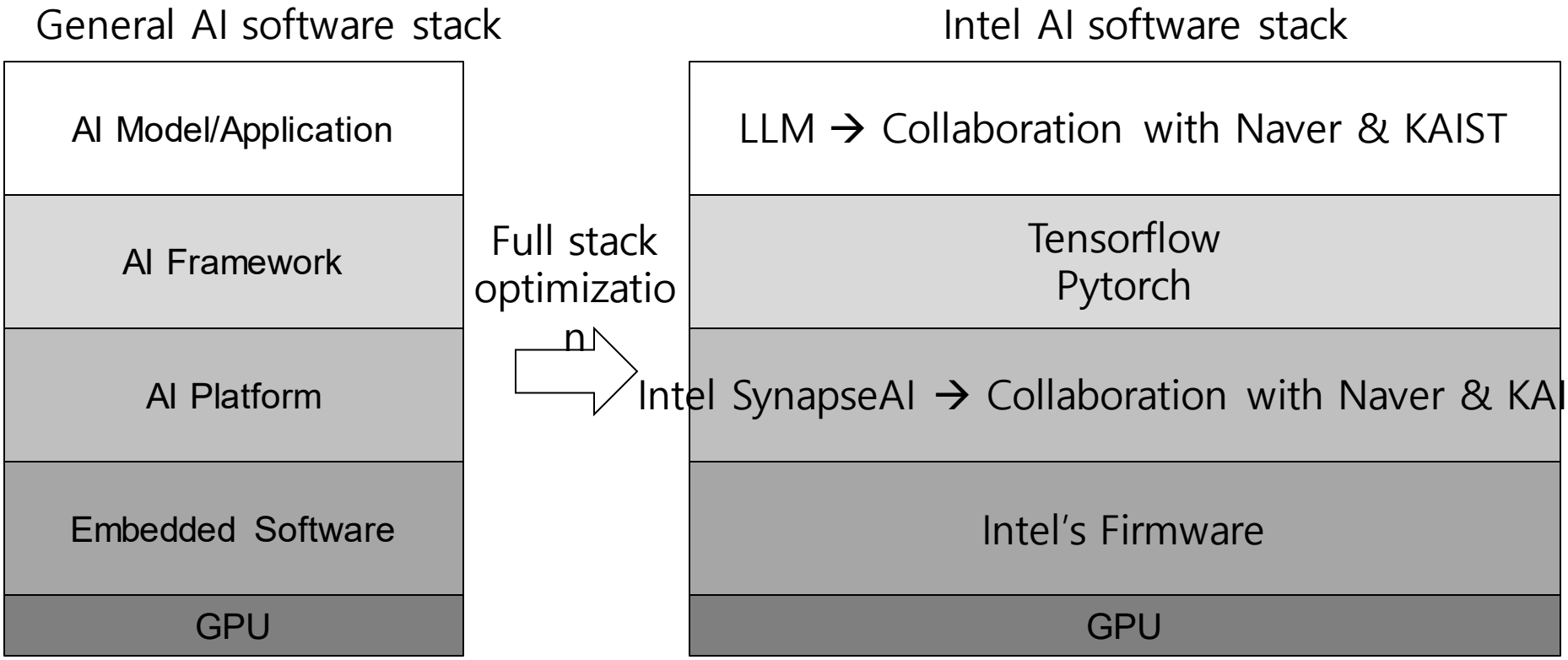
AI / 경량화 기술에 어울리는 고성능 고효율 H/W Design

< 네이버와 협력하는 삼성전자 >

- 삼성전자는 네이버 클라우드와 협업하여 AI 반도체 솔루션 개발 중.
- 협업하여 AI 반도체 솔루션 개발을 위해서는 양사 간의 서비스, AI 모델, 경량화 기술, H/W 기술 등 다양한 기술 스택의 협력 필요.
- 네이버는 삼성에게 서비스 경험, AI 모델링, 경량화 기술을 제공.
- 삼성전자는 해당 기술을 구현할 수 있는 Hardware, Solution 기술을 제공.

Ref) Naver

Intel Gaudi Software: End-to-end Full Software Stack with Naver-KAIST Collaboration



< Gaudi end-to-end AI software stack >

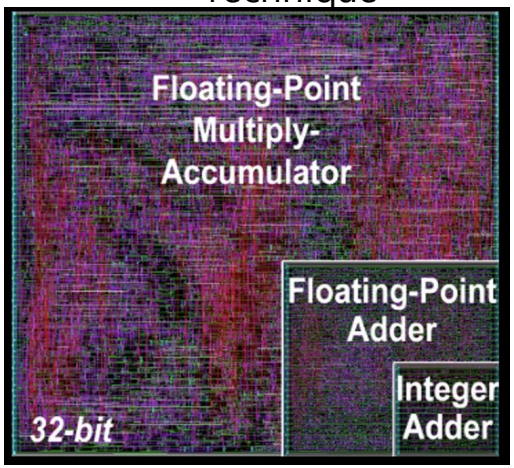
- Intel은 가우디 성능의 극대화를 위해 모든 소프트웨어 스택을 가우디에 최적화 예정.
- Naver는 제일 high-level인 AI Application 회사로, Gaudi에 맞게 최적화 예정.
- SynapseAI는 현재 Pytorch, Tensorflow를 모두 지원하고, CUDA로 쓰여진 코드를 쉽고 빠르게 Gaudi용 플랫폼인 SynapseAI로 변환할 수 있다고 함.

삼성 마하1의 전략

- Nvidia와의 차별화
- 경량화, LLM 모델에 적용
- Inference 용 진출 전략
- 저전력 저비용
- 자체 foundry 사용
- 자체 LPDDR 채용
- NPU-메모리 병목현상 최소화
- Data compression, quantization, pruning 기법
- 네이버의 클로버 X 추론에 사용, 시장 창출
- 네이버 software stack 채용
- 네이버 사용자 경험 배움
- 점진적 확장 (고성능, 학습용)

마하-1의 핵심기술 [2/2] : 경량화

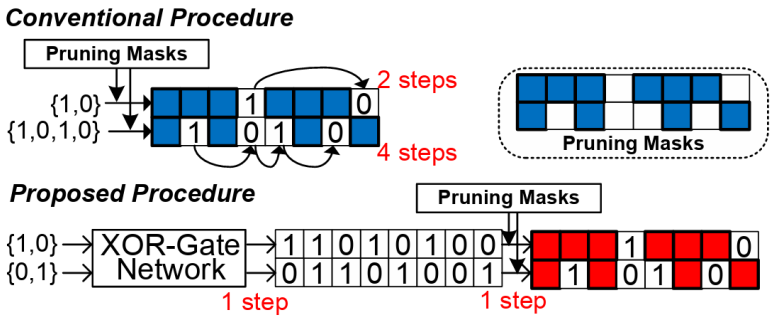
Quantization w/ New Computation Technique



Memory Bandwidth > 4x

Peak Computation > 8x

Pruning w/ New Data Representation



Memory Bandwidth > 4x

10x Perf./Energy Improvement
(Compared to the best AI chip of Today)

< Qunatization과 Pruning을 통한 LLM 모델 압축 >

- Quantization과 Pruning을 통해 LLM 모델을 압축할 수 있음.
- 네이버는 이 두 기술을 활용하여 메모리 대역폭과 연산능력을 크게 개선함.
- 기존 칩에 비해 10배의 전력효율을 보여줌.

Ref) 이동수 네이버클라우드 이사 X (옛 트위터) 2023년 12월 3일 게시물

네이버의 주요 기술

AI Tech

➢ Generative AI

- HyperCLOVA X : 초대규모 인공지능
- CLOVA X : 대화형 AI 서비스
- Cue: AI 검색 서비스 (쇼핑, 편의 등)
- Project Connect X : 생산성, 업무 효율 AI
- CLOVA Studio : Customized AI 모델

➢ AI 윤리 원칙

: 사람을 위한 AI, 다양성 존중, 보안 등 고려

Search Tech

➢ AI 맞춤형 검색

: 사용자의 목적과 관심사를 파악해 정교한 맞춤형 검색 결과 제공

➢ AI 추천 기술

: 네이버에 축적된 사용자 데이터를 기반으로 콘텐츠, 상품, 장소 추천

➢ AI 콘텐츠 검색 기술

: 사용자가 원하는 정보만 추출

Robotics

➢ ARC

: AI, Robot, Cloud의 약자로, 로봇의 두뇌와 눈이 되는 멀티 로봇 인텔리전스 시스템

➢ NAVER Robots

- Rookie : 제 2 사옥 자율주행 서비스 로봇 (택배, 카페, 도시락 배달 등)
- AMBIDEX : 양팔 로봇으로, 와이어 구조를 바탕으로 정밀 동작 및 높은 하중 작업 가능
- ARTO-1 : 드로잉 로봇으로 정교한 작업

Digital Twin

➢ 디지털 트윈

: 현실 세계를 디지털 환경에 복제하는 기술로 AR, 자율주행, 스마트 빌딩, 스마트 시티에 이르는 혁신 산업의 디지털 데이터 제작

➢ 공간 디지털 트윈 제작 솔루션

: 도시 전체의 3D 모델 기반 시뮬레이션

➢ 디지털 트윈 디바이스

: M시리즈(3차원 매핑), R시리즈(차량용), T시리즈(웨어러블)

네이버 데이터센터 : 각세종 [1/2]

< 네이버 각세종 전경 >

- 9만평 부지의 초대규모 데이터센터 (각춘천 대비 6배 규모)
- 대한민국 최대 서버 수용량 : 단일 기업 데이터 센터로서 60만 서버 유닛 수용
- 최대 20kW 전력을 공급하는 랙(Rack)으로 고성능 서버를 밀도 높게 운영

영역을 넘나드는 빅테크의 AI 칩 전략

SAMSUNG 데이터 병목 해소하는 AI 칩 '마하1' 2025년 초 출시	NVIDIA 서버용 CPU '그레이스' 출시 이어 PC용도 개발 착수	Google 서버용 CPU '액시온' 공개 '인텔 x86' 기반보다 성능 50% 향상
Microsoft 자체 AI용 GPU '마이아100'과 CPU '코발트' 공개	intel AI 가속기 '가우디3' 공개 'NVDIA H100 GPU 보다 전력효율 2배	amazon ARM 기반 자체 서버용 프로세서 '그래비톤' 개발

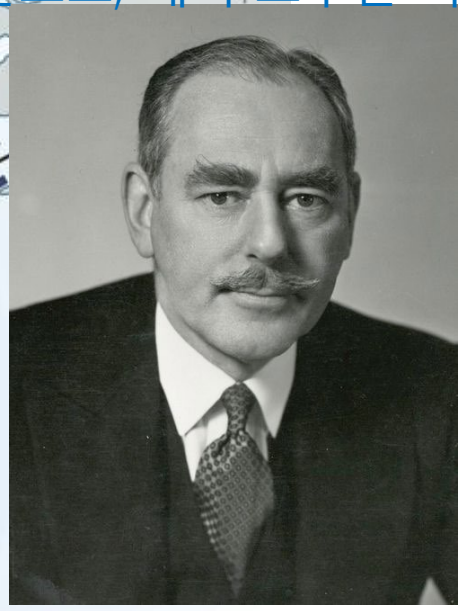
< 영역을 넘나드는 빅테크의 AI 칩 전략 >

Ref) 동아일보, "경계 무너진 'AI칩 전쟁'... 인텔 GPU-구글 CPU 공개"

Part4:AI반도체의 신 에치슨 라인

1950년 미국 극동 방위선: 애치슨 라인

딘 애치슨(Dean Acheson, 1893~1971)은 미국의 51대 국무장관으로, 해리 트루먼 시기에 재임

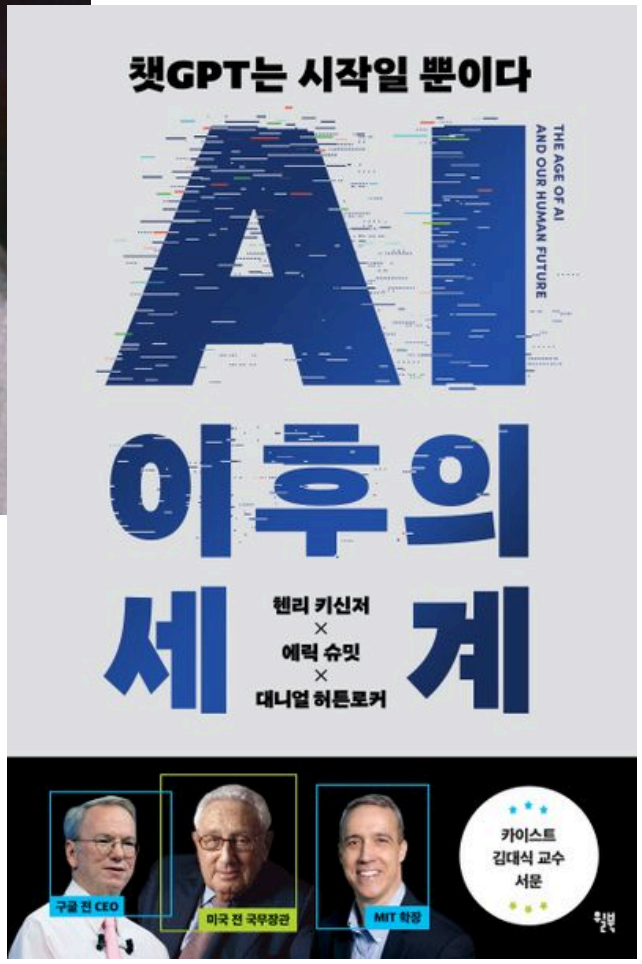


2024 반도체 극동 방위선: 新 반도체 애치슨 라인

제2 반도체 극동 방위선: 新 반도체 애치슨 라인

일본
구마모토현 기쿠요마치 공장

키신저



-중국 개방
-베트남
-한국 핵무기
-AI

러시아가 미국에 불리한 이유: 우크라이나 전쟁

- 반도체 격차 시작: 1960년대 시작 (미국:IBM, Intel, 러시아: 진공관)
- 소련 붕괴: 반도체 산업 이탈, 주변국 (1991년)
- 러시아에는 실리콘밸리가 없다 (1990년대 추격 실패)
- JSC Mikron: RFID, 8인치 팹, 90나노 공정(장비, 재료, EDA, 설계 능력 한계)
- GPU, DRAM, 12인치 파운드리가 없다 (15년 이상의 차이)
- Baikal: ARM,RISC-V chip design, TSMC 공정
- 인공지능 반도체(엔비디아), 슈퍼컴퓨터로 그 격차 확대
- 다음은 중국
- https://www.youtube.com/watch?v=N_4R4X7AWtU

독점을 허락하지 않는다: 반도체 기업 랭킹

<표 4> 글로벌 반도체산업의 매출 10위권 기업 순위 변화

순위	1981	1989	1995	2000	2006	2009	2013
1	TI	NEC	Intel	Intel	Intel	Intel	Intel
2	Motorola	Toshiba	NEC	Toshiba	Samsung	Samsung	Samsung
3	NEC	Hitachi	Toshiba	NEC	TI	Toshiba	Qualcomm
4	Philips	Motorola	Hitachi	Samsung	Toshiba	TI	Micron
5	Hitachi	TI	Motorola	TI	St Micro	St Micro	SK Hynix
6	Toshiba	Fujitsu	Samsung	St Micro	Renesas	Qualcomm	Toshiba
7	National	Mitsubishi	TI	Motorola	Hynix	Hynix	TI
8	Intel	Intel	Fujitsu	Hitachi	AMD	Renesas	Broadcom
9	Matsushita	Matsushita	Mitsubishi	Infineon	Freescale	AMD	St Micro
10	Fairchild	Philips	Hyundai	Micron	NXP	Infineon	Renesas

출처 : Gartner

1Q21 Top 15 Semiconductor Sales Leaders ($M, Including Foundries)

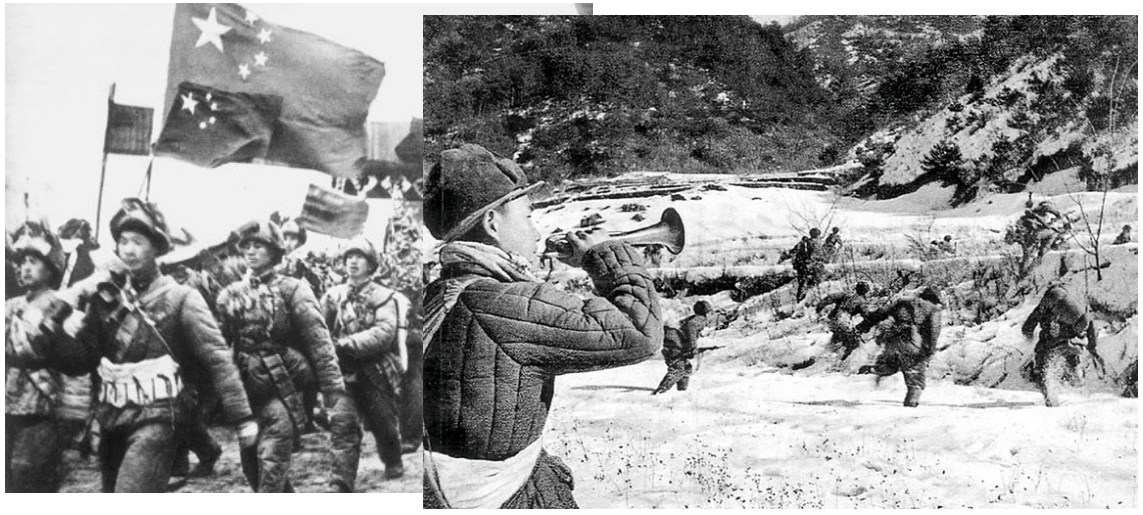
1Q21 Rank	1Q20 Rank	Company	Headquarters	1Q20 Total IC	1Q20 Total O-S-D	1Q20 Total Semi	1Q21 Total IC	1Q21 Total O-S-D	1Q21 Total Semi	1Q21/1Q20 % Change
1	1	Intel	U.S.	19,508	0	19,508	18,676	0	18,676	-4%
2	2	Samsung	South Korea	14,030	767	14,797	16,152	920	17,072	15%
3	3	TSMC (1)	Taiwan	10,319	0	10,319	12,911	0	12,911	25%
4	4	SK Hynix	South Korea	5,829	210	6,039	7,323	305	7,628	26%
5	5	Micron	U.S.	5,004	0	5,004	6,580	0	6,580	31%
6	7	Qualcomm (2)	U.S.	4,050	0	4,050	6,281	0	6,281	55%
7	6	Broadcom Inc. (2)	U.S.	3,673	409	4,082	4,355	485	4,840	19%
8	9	Nvidia (2)	U.S.	3,074	0	3,074	4,630	0	4,630	51%
9	8	TI	U.S.	2,974	190	3,164	3,793	235	4,028	27%
10	16	MediaTek (2)	Taiwan	2,022	0	2,022	3,849	0	3,849	90%
11	18	AMD (2)	U.S.	1,786	0	1,786	3,445	0	3,445	93%
12	11	Infineon	Europe	1,828	876	2,704	2,170	1,083	3,253	20%
13	10	Apple* (2)	U.S.	2,770	0	2,770	3,080	0	3,080	11%
14	14	ST	Europe	1,483	745	2,228	2,011	994	3,005	35%
15	13	Kioxia	Japan	2,567	0	2,567	2,585	0	2,585	1%
—	—	Top-15 Total		80,917	3,197	84,114	97,841	4,022	101,863	21%

(1) Foundry (2) Fabless

Source: Company reports, IC Insights' *Strategic Reviews* database

*Custom processors/devices for internal use.

중국의 전략: NxY 전략



전쟁 작전: 매복, 위장, 기만, 인해전술

반도체 작전: 시간 지연, 물량전술, 시장/기술/인력 내재화, 장기전, 동력(민족주의)

결국 인해전술(人海戰術), 인재전술 (**人材**戰術), 저비용전술(低費用戰術), 연합전술 (聯合 戰術)(러시아, 아시아, 중동, 아프리카)(미국의 약점 중심)

100년 전쟁의 시작

일본의 반도체 부활 전략

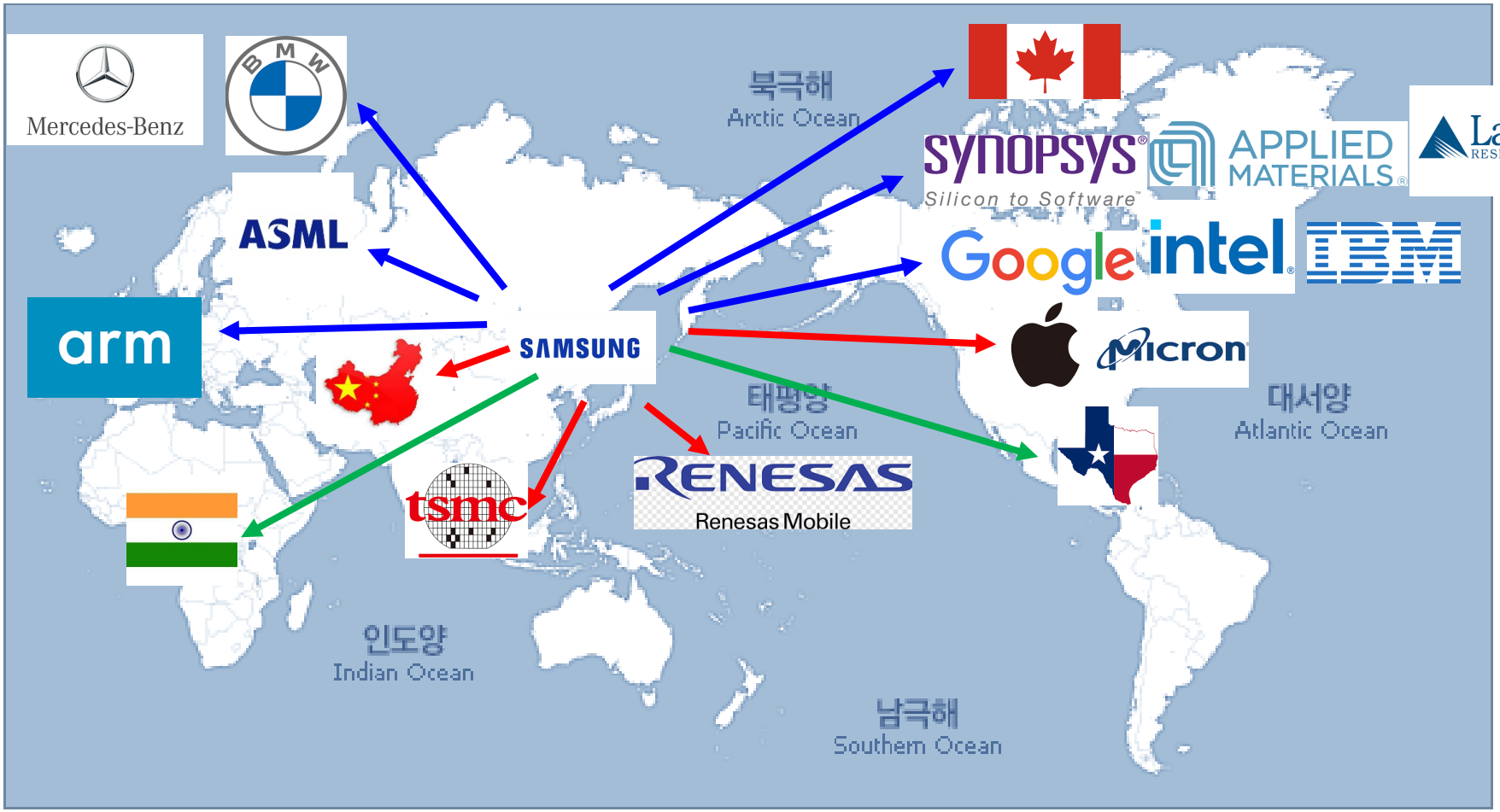
- 지정학적 이점 부각
- 세계 반도체 지정학적 재배치 전략
- 소재, 부품, 장비 경쟁력
- 미국 대만 기업들과 협력

(마이크론, TSMC)

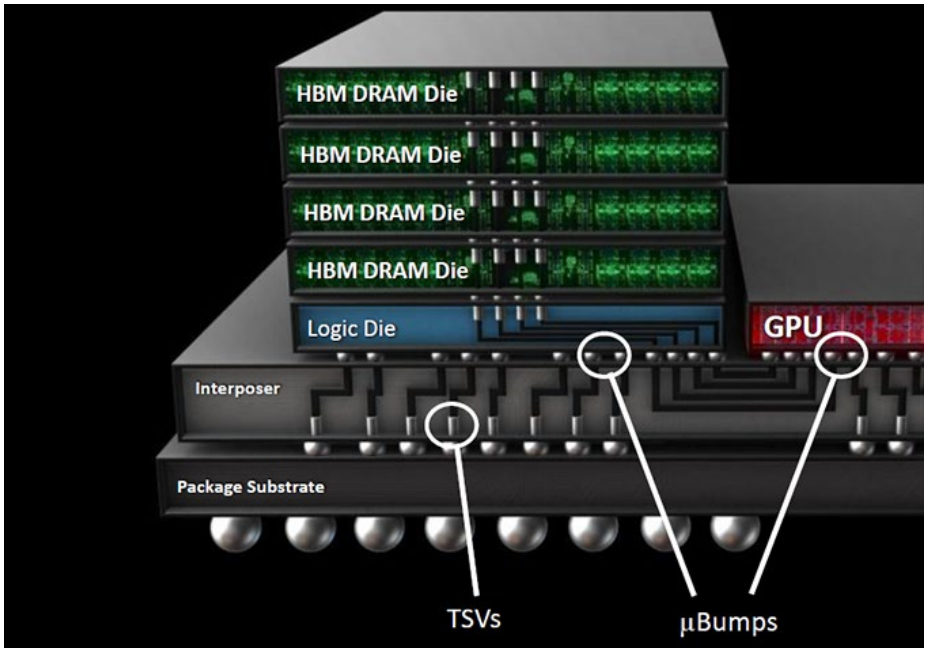
기시다 후미오(왼쪽에서 네 번째) 일본 총리가 5월 18일 도쿄 총리 관저에서 글로벌 반도체 기업 7곳 대표와 면담 후 기념촬영을 하고 있다. AP 연합뉴스

반도체 지정학 전략: 힘의 균형, 도전과 응전

위험분산 → 도전과 경쟁 → 협력 →



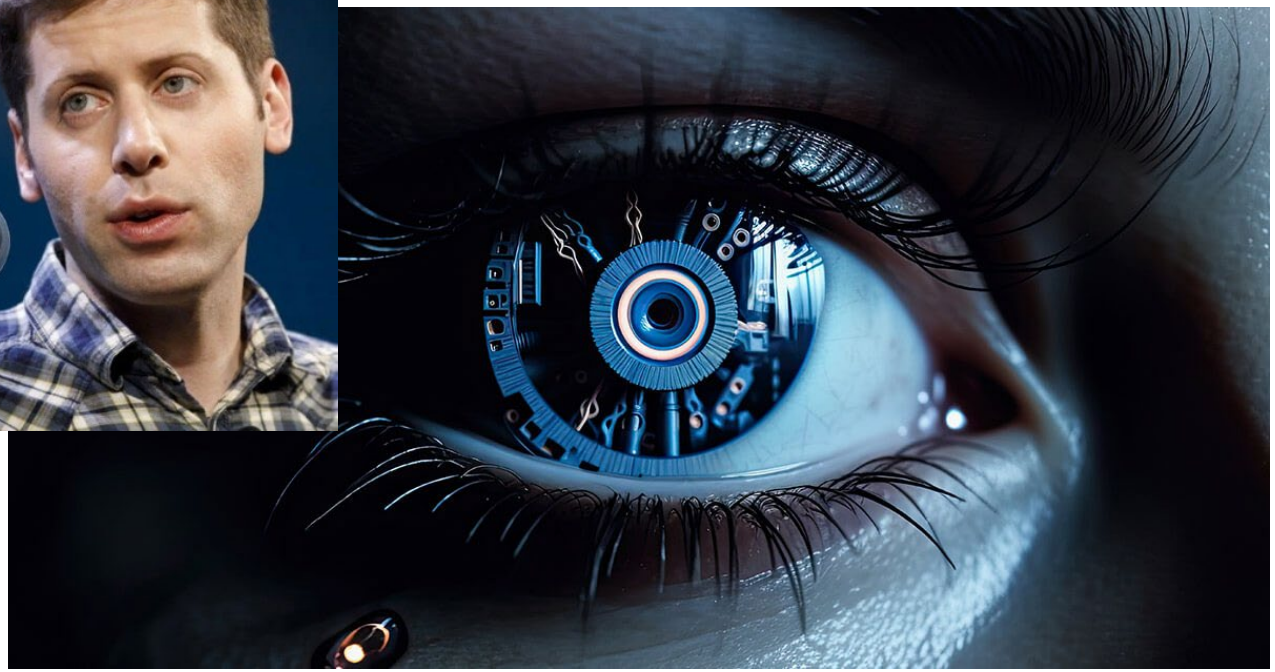
메모리 반도체가 찐빵 속의 앙꼬



Part6:결론

영화 마이너리티 리포트: 홍채와 open AI

-홍채 인식 open AI world coin
-기본 소득 부여
-홍채 등록
-극단적 개인 정보(지문, 목소리, 얼굴사진, DNA)



인간과 인공지능의 제어와 공존

- “인간을 위한(목적),
- 인간에 의한 인공지능(도구),
- 인간의 인공지능(소유)”

인공지능과 반도체가 바로 안보이다

과학기술
(인공지능, 반도체)

국가 안보

산업과 경제

국방 외교

인공지능의 봄

산새들의 노래가 계룡산에 퍼지네, 벚꽃이 피어 더욱 화려한 그곳에.

바람은 부드럽게 스쳐가며 춤을 추고, 꽃잎은 살랑살랑 노래하며 떨어져.

봄이 왔다 알리는 벚꽃의 향기, 가슴 속에 피어나는 아름다운 추억의 기억.

산길을 걸으며 벚꽃을 바라보면, 마음속에 묻어둔 사랑의 미소가 번져간다.

계룡산의 벚꽃은 자연의 아름다움, 우리 마음을 감동시키는 그 특별한 풍경이로구나

Thank you!