AI TRAINING SERIES

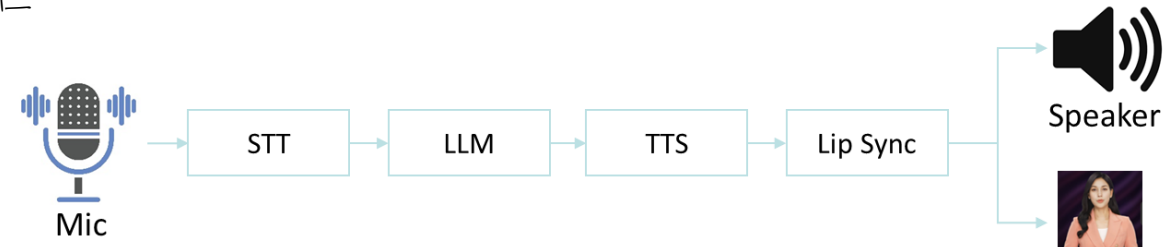
AI Accelerator

Contents

I. AI는 왜 이렇게 빠르게 발전할까?

최신 AI 서비스와 빠른 응답의 비밀

- 서비스 소개
  - ChatGPT / DeepSeek: 대화형 AI 로 몇초 안에 답변 생성
  - Midjourney / DALLE / Stable Diffusion: AI 기반 이미지 생성
  - Runway GEN3 / SORA / Klingai : 영상 및 이미지 편집에 특화된 AI
  - Suno / Udio : 음악 생성에 특화된 AI
  - Character.ai : 각각의 케릭터 AI와 대화 및 서비스
  - EST perso ai: AI human chat



< AI pipeline for AI human >

- Question
  - 어떻게 ChatGPT 는 몇 초 안에 답변을 내놓을 수 있을까?
  - 최신 알고리즘(Transformer) 과 강력한 하드웨어, 그리고 AI accelerator(가속기) 가 그 비밀!

1. AI는 왜 이렇게 빠르게 발전할까?

학습(training) VS 추론(inference) - AI의 두 얼굴

- **학습(훈련) 단계**
  - AI가 데이터라는 '교재'를 통해 공부하는 과정
  - 반복 학습을 통해 지식과 패턴을 습득
  - 예: 대입 입시를 위한 수능시험을 치기위해 12년을 공부하는것과 비슷

- **추론(실행) 단계**
  - 배운 내용을 바탕으로 실제 문제(질문)에 답하는 '시험' 단계
  - 예: 12년간 공부한 내용을 바탕으로 수능시험을 치는 것

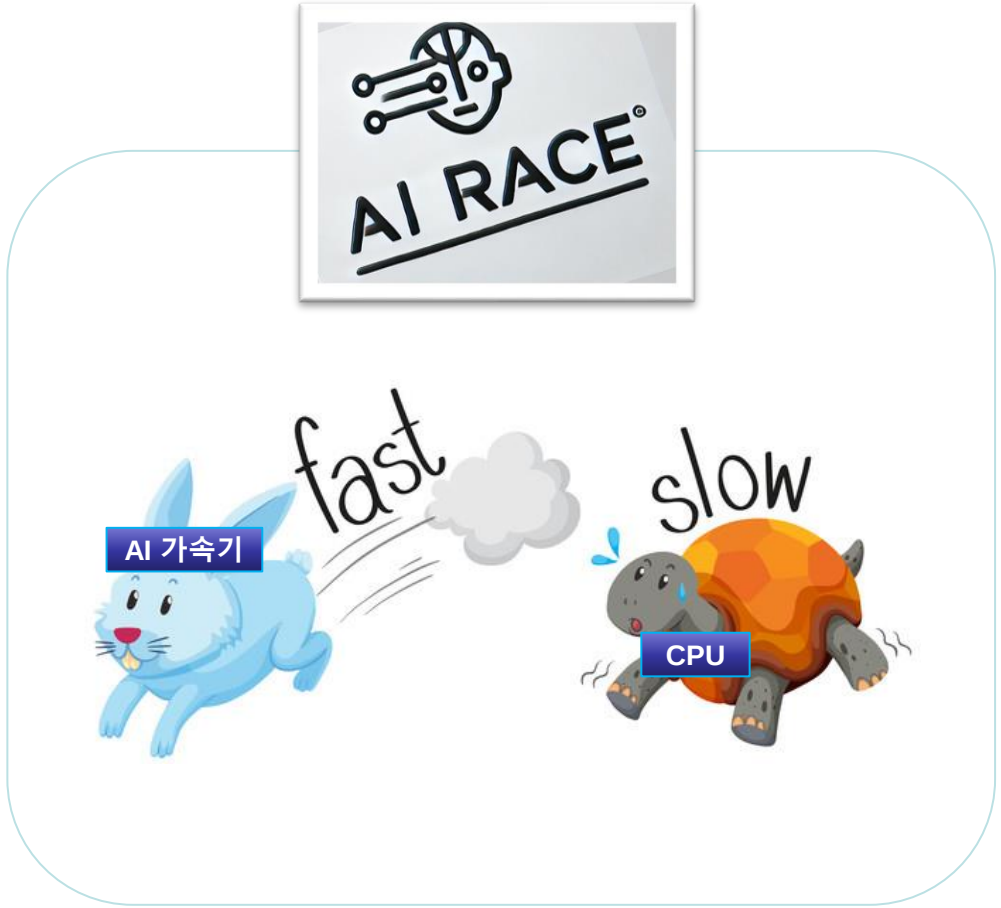
항목	학습 (Training)	추론 (Inference)
연산량	매우 큼 (고차원 행렬 연산, 역전파 연산)	비교적 적음 (순전파만 수행)
HW 가속기	고성능 GPU, TPU, NPU 필요	CPU, 저전력 GPU, NPU도 가능
메모리 요구량	매우 큼 (모델 및 배치 데이터 저장 필요)	상대적으로 적음 (한 개 입력 데이터 처리 가능)
전력 소비	높음 (고성능 연산 지속)	낮음 (경량화 가능)

I. AI는 왜 이렇게 빠르게 발전할까?

연산량 폭발! AI 가속기가 왜 필요한가?

- **LLM 발전과 함께 발생된 연산량 폭발 현상**
  - AI모델이 복잡해 질수록 계산해야 할 데이터와 연산이 급증
  - 일반 CPU로 처리하면 "느림"과 "비효율" 문제 발생
  - 예: GPT-4 는 1,760억개의 파라미터를 갖고 있고 이를 CPU 로만 실행하면 질문 하나 답하는데 몇분 소요됨

- **AI 가속기 등장**
  - GPU, TPU, NPU 등의 특수한 HW가 대량 연산을 병렬 처리로 빠르게 수행하여 실시간 처리를 가능하게 함
  - 예: AI 가속기를 사용하면 GPT-4 는 즉시 응답 가능



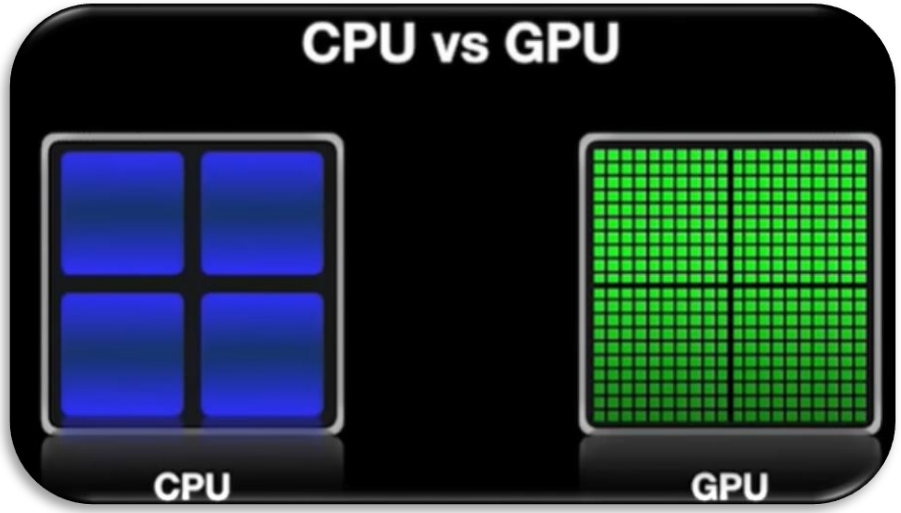
Contents

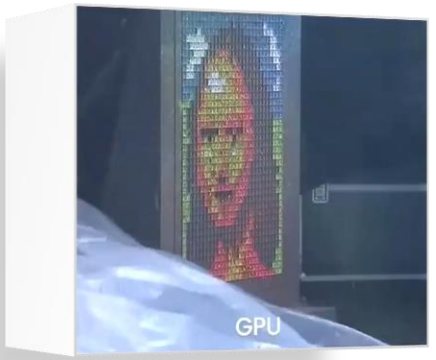
2. AI 가속기란?

CPU vs GPU

- CPU (Central Processing Unit)
  - 비유: 한 명의 택배기사가 하나씩 물건을 배달하는 것
  - 특징:
    - 직렬처리: 하나의 작업을 빠르게 처리할 수 있지만 한번에 하나씩 순차적으로 작업
    - 복잡한 연산: 분기나 조건문 등 복잡한 연산에 강함
    - Intel CPU i9-14900 cores: 24

- GPU (Graphics Processing Unit)
  - 비유: 수많은 드론이 동시에 여러 지역에 물건을 배달하는 것
  - 특징:
    - 병렬연산: 수천개의 코어를 활용해 동일한 연산을 동시에 수행
    - 대규모 데이터 처리: 행렬 연산, 벡터 연산 등 반복적인 연산에 최적화
    - NVIDIA RTX-5090 cores: 21760(shading), 680(tensor core)





https://www.youtube.com/shorts/EAjj-wV4Jto
https://www.youtube.com/shorts/_F71GcLpqQc

2. AI 가속기란?

가속기의 종류와 차이점 (GPU vs TPU vs NPU)

■ GPU (Graphics Processing Unit)

- 용도: 그래픽 작업, 딥러닝 훈련 및 추론
- 비유: 다목적 공장에서 여러 작업을 동시 처리

■ TPU (Tensor Processing Unit)

- 용도: 대규모 AI 연산, 특히 행렬 연산 가속
- 비유: 전용 생산 라인을 가진 공장에서 특정 반복 작업을 빠르게 수행

■ NPU (Neural Processing Unit)

- 용도: 모바일/임베디드 환경에서 실시간 AI 추론
- 비유: 특정 작업만을 위해 최적화된 소형 맞춤형 워크숍

항목	GPU	TPU	NPU
정의	• 원래 그래픽 렌더링용으로 설계되었으나, 병렬 연산 덕분에 범용 AI 연산에도 활용됨	• 구글이 AI 연산(특히 행렬 곱셈)에 최적화해 개발한 전용 AI 가속기	• 스마트폰/임베디드 기기에서 AI 추론을 위해 설계된 저전력 전용 가속기
비유	• "다목적 공장": 다양한 작업(그래픽, 딥러닝 등)을 동시에 처리하는 공장	• "전용 라인 특화 공장": 특정 AI 연산(행렬 연산)을 고속으로 수행하는 공장.	• "맞춤형 워크숍": 에너지 효율을 중시하며, 실시간 AI 작업에 최적화된 소형 작업장
장점	• 병렬 연산에 강함 • 다양한 작업에 유연하게 대응 • 범용성 높음	• AI 연산에 특화되어 매우 빠름 • 혼합 정밀도 연산 최적화 • 대규모 연산에 탁월	• 저전력, 에너지 효율적 • 모바일/임베디드 환경에 최적화 • 실시간 응답 우수
단점	• AI 전용 연산에 특화되진 않음 • 전력 소비 및 발열 문제 발생 가능	• 특정 프레임워크에 의존적 (TF) • 범용성 떨어짐	• 범용 연산 처리에는 한계 • 성능이 GPU/TPU에 비해 낮음 • 특정 작업에 최적화됨

GPGPU (General Purpose GPU) 는 뭘까?

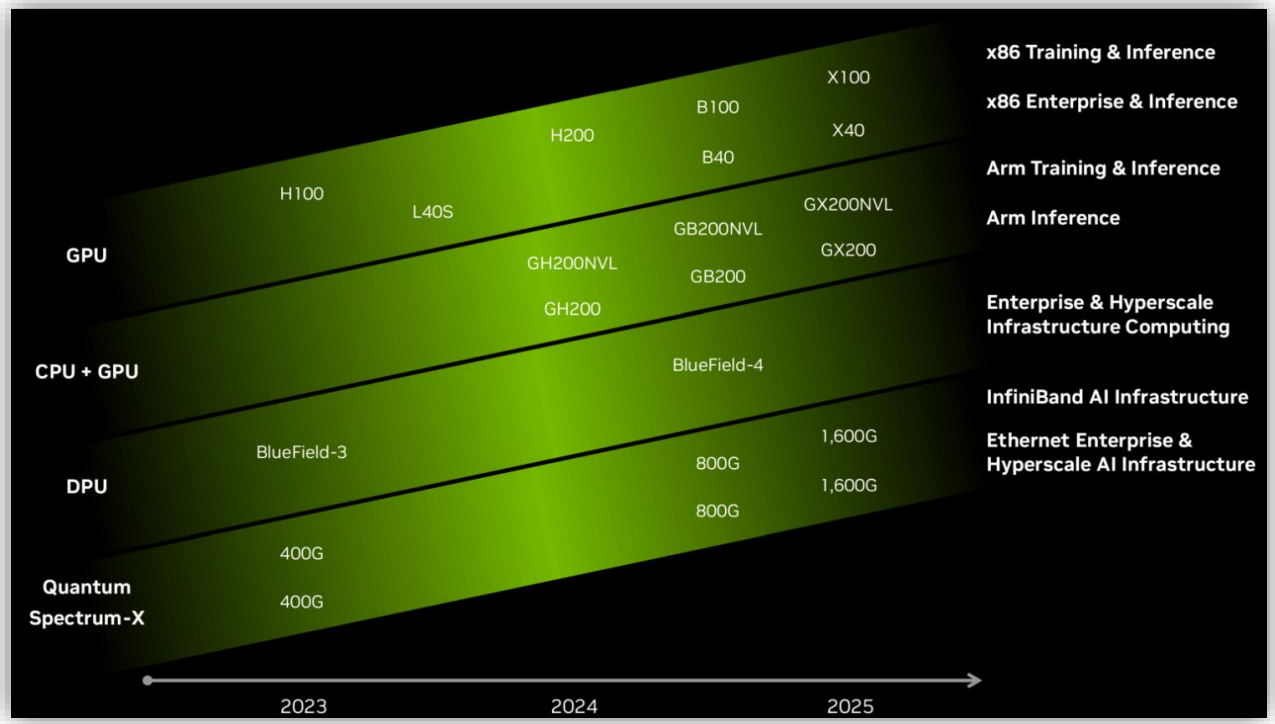
과학 계산, 딥러닝, 물리 시뮬레이션 등 그래픽 이외의 연산에 활용하는 기술

Contents

3. Big Tech 제조사별 AI 가속기 Trend

NVIDIA (내가 제일 잘나가)

- 데이터 센터용 GPU (시장점유율: 90% 이상)
  - 주요 제품: H200 (Hopper - 3,958 TOPS, 141GB mem, 700W, ~$50,000)
- Client 용 GPU (시장점유율: 70~80%)
  - 주요 제품: RTX-5090 (Blackwell - 3,352 TOPS, 32GB mem, 1000W, $1,999)
- Edge Device 용 Solution
  - 주요 제품: Jetson (Orin / Xavier)

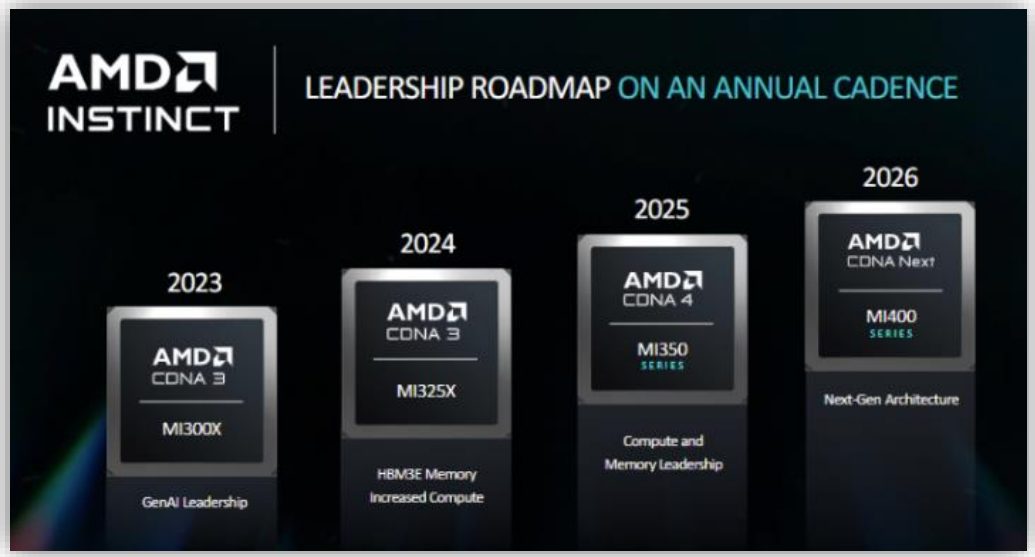


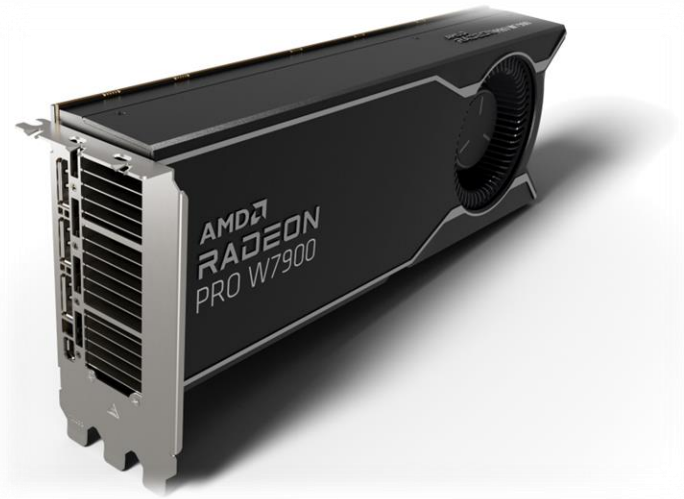
3. Big Tech 제조사별 AI 가속기 Trend

https://www.amd.com/content/dam/amd/en/documents/corporate/events/advancing-ai-2024-distribution-deck.pdf

AMD (끊임없는 도전)

- 데이터 센터용 GPU
  - 주요 제품: MI325X (2,600 TOPS, 128GB mem, 1000W, ~$20,000)
- Client 용 GPU
  - 주요 제품: Radeon RX W7900 XTX (240 TOPS, 24GB mem, 800W, ~$1,000)
- Edge Device 용 Solution
  - 주요 제품: Ryzen Embedded





3. Big Tech 제조사별 AI 가속기 Trend

Intel (조용한 추격자)

- 데이터 센터용 GPU
  - 주요 제품: Gaudi®3 (1,835 TOPS, 128GB mem, 900W, ~$13,000)
- Client 용 GPU
  - 주요 제품: Arc B580 (233 TOPS, 12GB mem, 190W, ~$249)
- Edge Device 용 Solution
  - 주요 제품: 200HX Arrow Lake - AI PC including NPU (13 TOPs)

3. Big Tech 제조사별 AI 가속기 Trend

Google / Amazon / MSFT (Cloud 3대장 - NVIDIA GPU 너무 비싸! 내가 만들어 쓸꺼야!)

Google Cloud - TPU (video)

- Trillium v6e spec: 1836 TOPs, 32GB mem
- 2024년 35% 매출 증가, 17% 영업이익 (2023년 3% 에서 크게 상승)

Amazon AWS - AWS Trainium / AWS Inferentia

- Trainium2 spec: 1300 TOPs, 96 GB mem

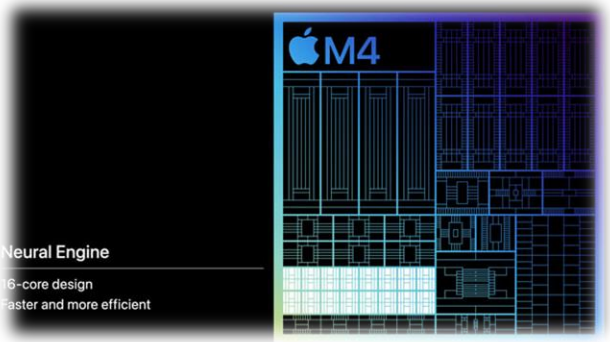
MSFT Azure - Maia 100

- 자체 제작한 AI 가속기도 있으나 성능이 아쉽
- 현재 Azure 는 NVIDIA/AMD 의 GPU 를 사용해 AI server 를 구성함

3. Big Tech 제조사별 AI 가속기 Trend

Qualcomm / Apple / Tesla

- **Qualcomm**
  - Cloud data center 용 AI 를 위해 Cloud AI 100 시리즈 출시 (저전력 / 고효율 장점)
  - Edge AI 를 위해 Snapdragon 플랫폼에 NPU 탑재 (Snapdragon X: 45 TOPs)
  - Edge AI 와 관련 SW platform (AI hub) 에 주력 중



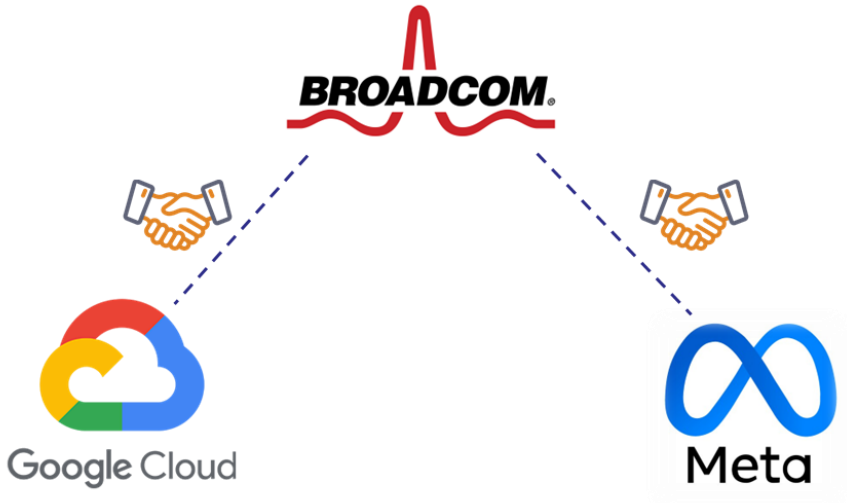
- **Apple**
  - Cloud 보다는 Edge AI 쪽에 주력
  - 최신 M4 chip 의 내장 NPU 예상 성능: 32TOPs

- **Tesla**
  - Edge AI: 초기 모델 S/X는 NVIDIA GPU 기반으로 자율주행을 했으나, 2019년부터 자체 설계한 FSD HW 를 도입했고 현재 4.0 버전까지 적용됨. Dual SoC 형태로 각 칩(FSD2)당 72 TOPs 의 NPU 가 내장
  - Dojo 프로젝트: NVIDIA 의존성을 탈피해 자체 제작한 AI 가속기를 탑재한 AI training 용 슈퍼 컴퓨터

3. Big Tech 제조사별 AI 가속기 Trend

Broadcom

- **Google / Meta 와 Broadcom 의 AI 협업**
  - Google 과 Meta 는 방대한 AI workload 를 효율적으로 대응하기 위해 맞춤형 ASIC 설계에 관심을 갖었고, Broadcom 과 AI 가속기 공동 개발함
  - NVIDIA GPU 대비 특정 AI 연산에서 와트당 성능을 높이고 독립성 강화 효과



- **네트워크 스위치 ASIC**
  - 대규모 AI 모델을 학습·추론하기 위해서는 수많은 GPU(또는 AI 가속기) 노드가 초고속 네트워크로 연결되어야 함
  - 네트워크 스위치 ASIC은 스위치/라우터 장비에 들어가서 엄청난 양의 트래픽을 단순하고 빠르게 처리하는 데 최적화된 전용 칩
  - Broadcom 은 이를 위한 전용 칩의 강자로 AI 서버의 사용이 증가하고 복잡해질 수록 수혜를 보게 됨

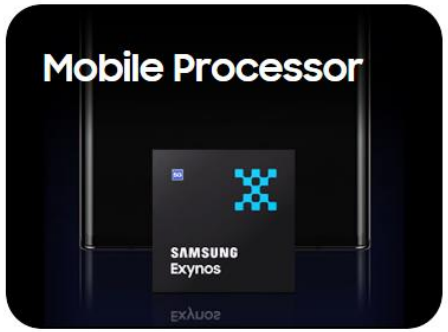
vmware
by Broadcom

- **보안 및 서버 관련 SW 사업 확장**
  - Symantec 기업 보안 솔루션 인수
  - Vmware 인수 - 가상화/클라우드/컨테이너 운영 등 클라우드 인프라 소프트웨어의 대표 주자

Contents

4. 국내 업체들의 AI 가속기 전략

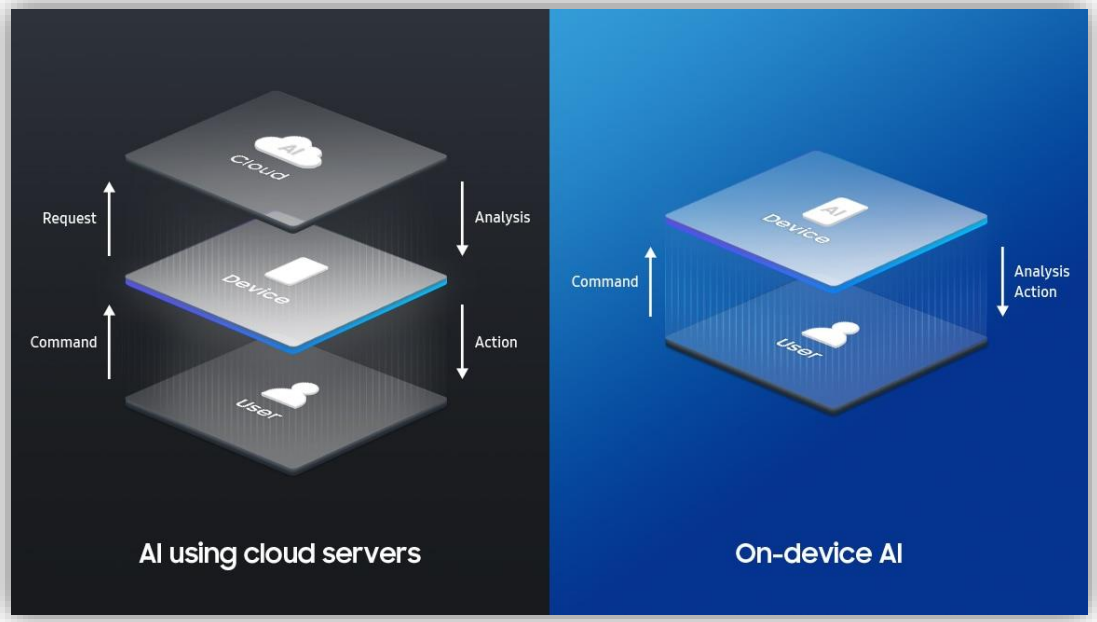
Samsung



- SoC 에서의 NPU 를 강화 (On-Device AI)
  - Exynos 시리즈: 삼성 고유의 AP(애플리케이션 프로세서) 브랜드이며, 최신 Exynos 칩에는 CPU·GPU뿐만 아니라 NPU 코어가 내장되어 AI 연산 가속을 지원
  - Exynos 2400 NPU: 10 TOPs (추정치)

- AI 가속기 공동개발로 AI 서버시장에 간접진출
  - FuriosaAI와 삼성전자의 공동 개발 및 파운드리 협업 (5nm 공정)
  - Rebellions(리벨리온)과 삼성전자의 공동 개발 및 파운드리 협업 (5nm 공정)

- 네이버와 협업으로 자체 AI 가속기 개발 – 마하
  - 2024년 3월 네이버와 협업하여 자체 AI 가속기 '마하1' 개발 착수
  - 2024년 9월 네이버와 협업 결렬. 차후 행보는 미지수



https://furiosa.ai/

4. 국내 업체들의 AI 가속기 전략

FuriosaAI (퓨리오사AI)

- AI ASIC 분야 도전하는 국내 선두 스타트업
  - 2018년경 창업
  - 2022년 “Warboy” 란 NPU 칩을 공개하며 Vision AI 쪽에 강한 면모

- 최근 데이터 센터용 AI 가속기 출시
  - RNGD: 512TOPs, 48GB mem, 150W

	RNGD	H100 SXM
Technology	TSMC 5nm	TSMC 4nm
BF16/FP8 (TFLOPS)	256/512	989/1979
INT8/INT4 (TOPS)	512/1024	1979/-
Memory Capacity (GB)	48	80
Memory Bandwidth (TB/s)	1.5	3.35
Host I/F	Gen5 x16	Gen5 x16
TDP (W)	150	700

- 회사 전략 및 전망
  - GPU(엔비디아) 대비 높은 전력 효율, 합리적인 가격을 내세워 AI 추론 시장에서 경쟁 우위를 노림
  - 삼성 파운드리(5nm 이하) 공정으로 차세대 칩 양산 추진 → 공정 미세화로 성능/전력효율 개선
  - 시장 인지도 및 레퍼런스 부족이 단점. 향후 대형 고객사 및 투자처 확보 필요

4. 국내 업체들의 AI 가속기 전략

https://rebellions.ai/

Rebellions(리벨리온)

- AI ASIC 분야 도전하는 국내 스타트업
  - 2019년경 창업
  - "범용 AI 가속기"보다는 금융 및 엔터프라이즈에 초점을 맞춘 니치(Niche) 전략을 취한다는 점에서 경쟁사와 차별화

- 주요 제품
  - RBLN-CA25: 512TOPs, 64GB mem, 350W

- 회사 전략 및 전망
  - 2024년 국내 동종회사인 사피온을 인수해 Edge/저전력 관련 솔루션 확보
  - KT 등 대형 후원사의 지원이 있어 단순 자본지원을 넘어 KT 를 통한 시장 진입이 쉬움
  - 파운드리(삼성 등) 협업을 통한 양산 안정화

Single Chip	
FP16	128 TFLOPS
INT8 / INT4	512 TOPS / 1024 TOPS
Multi-Instance	HW isolation up to 64 independent tasks
Input Power	DC 12V (CPU 8-pin power connector)
Max. TDP	350W
Thermal	Air Cooling (passive)
Memory	GDDR 64GB, 1024 GB/s
Host & CIC I/F	PCIe Gen5 x16, 64GB/s
Form Factor	266.5 mm x 111 mm x 19 mm

Contents

5. 실습 - AI 가속기를 사용해 AI model training 하기

첫번째 실습: 추론(Inference) - CPU vs GPU

- YOLO object detection model 을 CPU 와 GPU 를 사용해 Inference 해 보고 성능을 비교

https://colab.research.google.com/drive/1FOp_SzHw3hgRfwr7e18P7ETyFkvf4kMm?usp=sharing

두번째 실습: 훈련(Training) - CPU vs GPU vs TPU

- Google MediaPipe 의 Hands Landmark Detection 이후 손모양 classification 을 위한 모델을 다양한 AI Accelerator (CPU / GPU / TPU)를 사용해서 Training 하고 각 성능을 비교

https://colab.research.google.com/drive/1-emsMl3O-TBuQNsTdD_2fKQPOp9PHK7g

THANK YOU