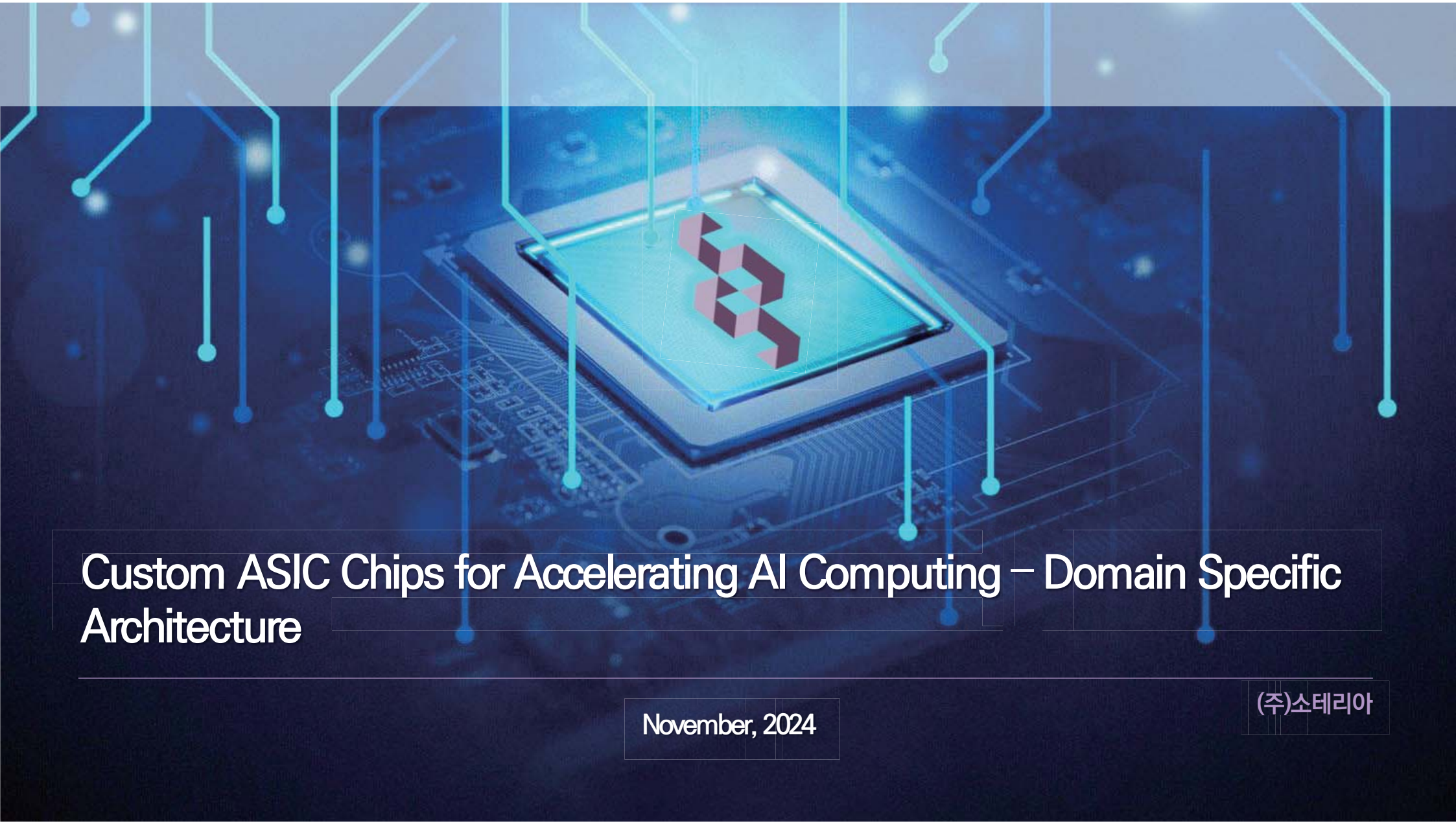
ACK 2024
논문집 Annual Conference of KIPS 2024

초청강연

Custom ASIC Chips for Accelerating AI Computing-Domain Specific Architecture

김종만 대표
(소테리아)



Custom ASIC Chips for Accelerating AI Computing – Domain Specific Architecture

November, 2024

(주)소테리아

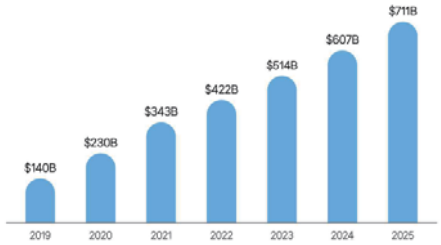
CONTENT

01 Prologue | (1) AI 반도체 패러다임 변화와 급격한 시장 수요 증가에 따른 급성장 시장

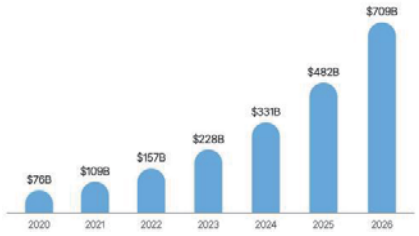
ChatGPT와 Nvidia 제품들로 인공지능 시장 확대

새로운 인공지능 시나리오는 더 많은 특수 칩 요구 : AI 임계치(tolerances)

- 24년 현재, ChatGPT와 구글 등 다른 회사들에서 출시한 인공지능 솔루션들은 이미 전세계 인류의 생활에 깊숙이 침투
- 이러한 대규모 인공지능 모델을 연산하기 위해서는 GPU와 CPU 및 메모리의 엄청난 대규모로 구성된 Mega Data Center 들이 필요
- 특히 병렬계산에 특화된 Nvidia의 GPU 같은 경우 H100이라는 제품이 4만 USD 달러 이상에 판매 되고 있으나 공급이 부족하여 시장에서 구할 수 없는 지경
- 현재 시장에서 주로 사용하고 있는 언어모델 인공지능 모델에 향후 이미지 및 동영상 인공지능 모델이 도입될 경우 필요한 HW 연산 능력치는 기하급수적으로 증가
- AI 임계치에 맞는 상용 시스템 개발과 운용을 위해서는 고도화된 지능형 제품의 급증 및 데이터의 폭발적 증가가 필연적이며, ChatGPT 등의 실제 예를 살펴보아도 대규모 데이터 처리와 전송에 에너지 소모와 탄소 배출량이 문제가 되고 있어 더 효율적인 컴퓨팅이 필요하다는 사안이 중요



〈 Gartner, '22 〉



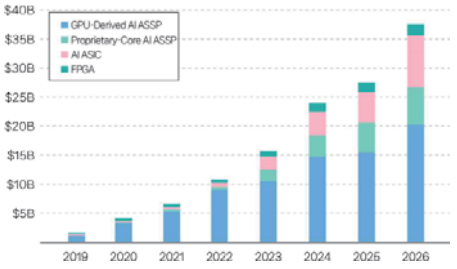
〈 Statista, '22 〉

〈 인공지능 반도체 활용 전망 (각 통계 출처를 바탕으로 KISTEP 재구성, 2023.12) 〉

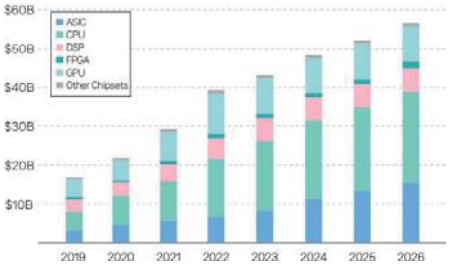
향후 연평균 30%를 상회하는 폭발적인 시장 성장 전망

응용분야에서는 데이터센터 대비 엣지컴퓨팅 분야의 확대 전망

- 유력 매체에 따르면 인공지능 반도체 시장 규모는 가파른 증가가 전망되며, 미래 주요 산업으로 성장할 것으로 전망(매체에 따라 현수준·예측치는 다소 상이하나, 대체로 큰 폭의 성장을 전망)
- 인공지능 반도체의 부상은 시스템반도체 산업에서 새로운 기회를 창출하며, 인공지능 활용 제품·서비스의 성장에 따라 관련 시장이 성숙 중
- 데이터센터(클라우드컴퓨팅) 분야는 시장 규모 확대와 함께 GPU의 강세가 이어질 것으로 전망되며, '23년 이후로 ASIC의 비중도 점차 확대할 것으로 전망
- 엣지컴퓨팅 분야 역시 지속적인 시장 규모 확대가 전망되며, 전력 소모 등 GPU의 한계로 '22~'23년 매출 정점에 도달한 후 ASIC·FPGA 등으로 대체될 것으로 예상



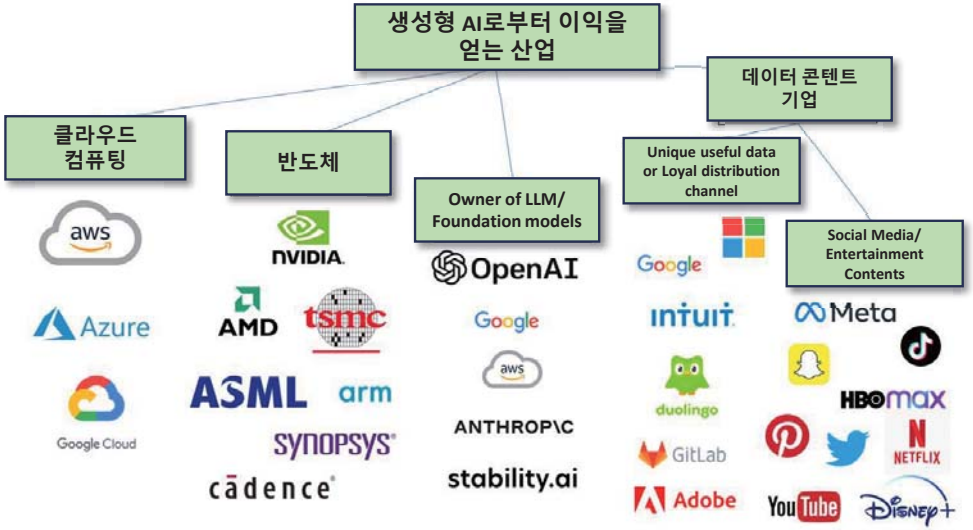
〈 (데이터센터용)연산유닛별 점유율 전망 〉



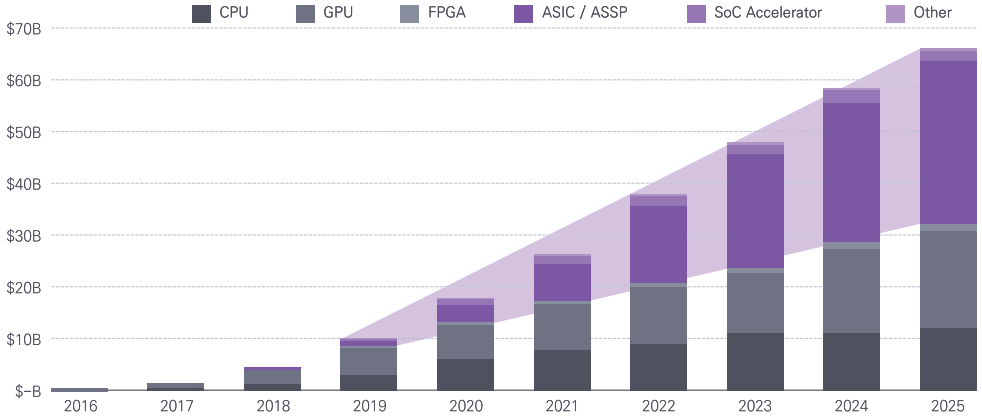
〈 (엣지컴퓨팅용)연산유닛별 점유율 전망 〉

〈 데이터센터·엣지컴퓨팅용 인공지능 반도체 시장 전망 (2023 인공지능 반도체, KISTEP, 2023.12.) 〉

01 Prologue | (2) 반도체와 AI 융합으로 전산업군에 비즈니스 기회

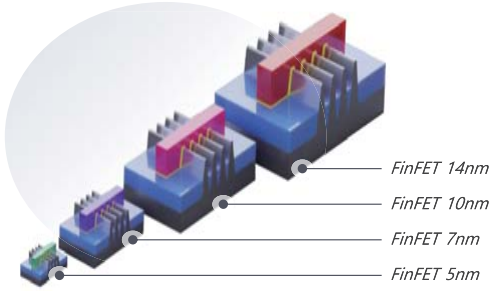


- Worldwide Deep Learning Chipset 매출



- Artificial Neuron / Synapse
- Event-driven / Asynchronous
- Analog/Digital/Mixed Spiking Signals
- Neuromorphic Processor

- 초저전압 다크 실리콘 기술




Planar


FinFET


GAAFET Gate All Around

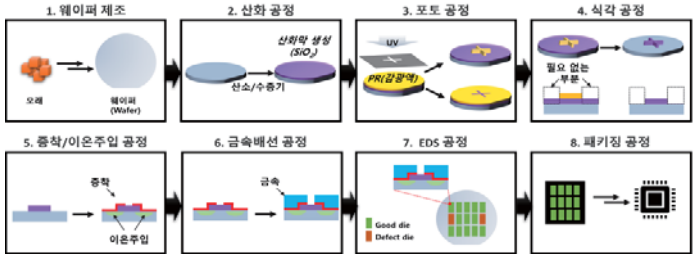

MBCFET Multi Bridge Channel

01 Background | (3) 반도체 산업 분야 및 Value Chain

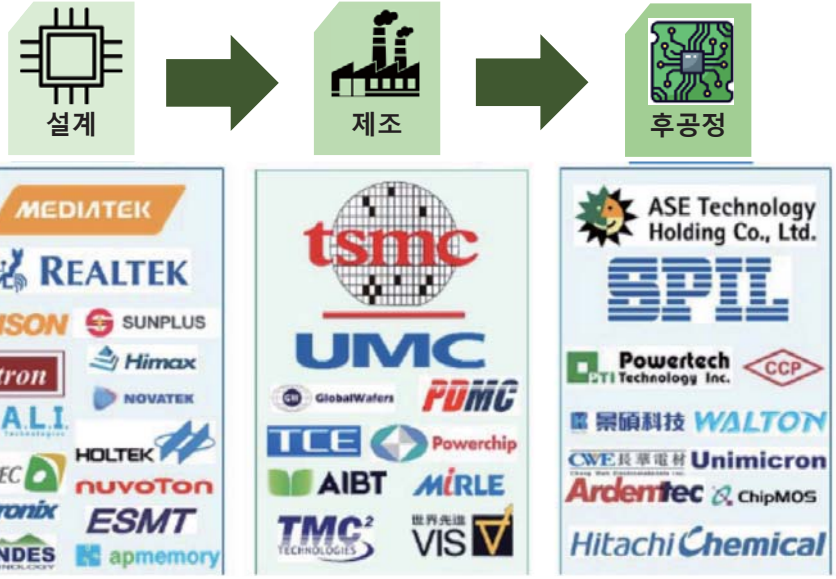
SOTERIA

반도체 Value Chain: 설계 (Design) → 생산 (Manufacturing) → 조립&검사 (Packaging&Test)

기업 유형: 종합반도체회사 (IDM) / IP → 팹리스 → 디자인 하우스 → 파운드리 → OSAT

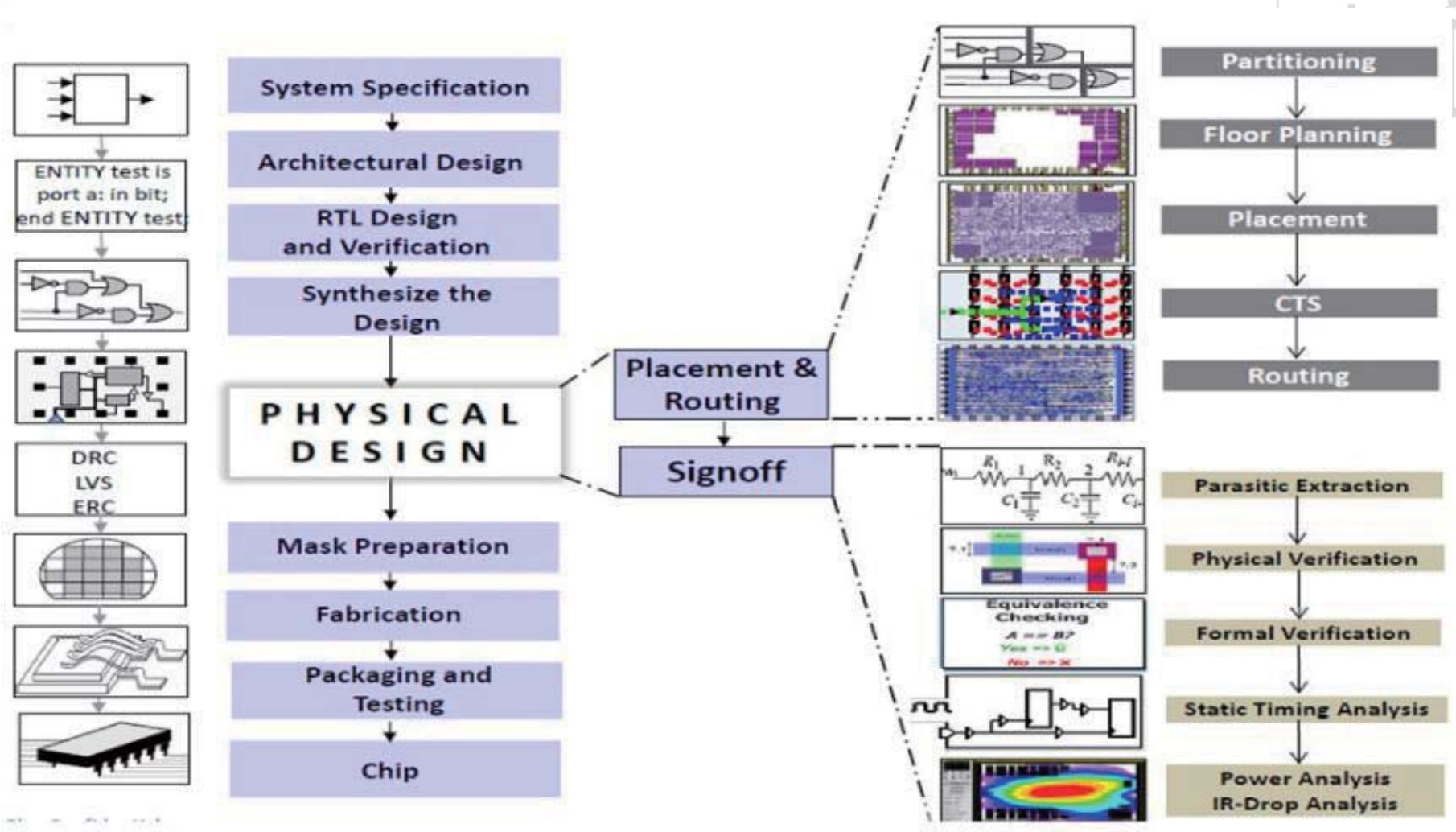


SAMSUNG, intel, SK hynix



IP (Intellectual Property)	팹리스 (Fabless)	디자인하우스 (Design House)	파운드리 (Foundry)	OSAT (Outsourced Semico-nductor Assembly & Test)
- 지적재산권 기업 - 설계 라이선스 판매 - 예: Arm, 시놉시스 arm SYNOPSYS® 팹리스는 통상적으로 메모리 컨트롤러, 인터넷트 및 인터페이스 등의 IP를 직접 제조하지 않는 특성이 있음	- 반도체 생산 공장 (Fab) 없이 반도체 설계/개발 - 예: 엔비디아, 퀄컴 NVIDIA Qualcomm	- 팹리스가 설계한 제품을 파운드리 공정에 맞춰 최적화 서비스를 제공 - 예1: TSMC VCA (GUC, Alchip, ASICLAND 등) - 예2: SSF DSP (가온칩-스, AD테크놀로지 등)	- 팹리스가 설계한 제품을 전문 위탁생산 - 예: TSMC/SSF 전공정: Wafer → Deposition (산화, 증착) → 포토 공정 → 식각 공정 → 이온주입 → CMP → 금속배선 → 패키징 & 테스트 후공정: Back Grinding (후면연삭) → Wafer Dicing (절단) → Wire Bonding → Molding → Package Test (Final Test) → 기타	- 반도체 생상 후공정인 패키징, 조립, 테스트 수행 전문 기업

01 Background | (4) ASIC Design and PD Flow



01 Problem Statement | (5) 시장·산업 이슈 대응과 부합되는 역량기반 기술 개발 필요

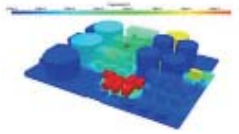
AI모델/클라우드/AI 반도체 디지털 생태계 문제

폭발적 Data 처리, Aging & Thermal 이슈, 에너지 소모 및 탄소 배출 문제

- ChatGPT로 촉발된 기하급수적인 AI 빅뱅과 Hyper-scale 클라우드 인프라 고도화에 따라, GPU 등의 반도체는 폭발적 Data 처리, Aging & Thermal 이슈, 에너지 소모 및 탄소 배출 문제 야기

Aging & Data 이슈 >> Chip 성능저하

- ✓ 기하급수적인 Data 처리를 위한 반도체에서 전력밀도와 발열문제 증가
- ✓ 실리콘 다이옥사이드 전기적 파괴로 파손, 수명단축 초래
- ✓ Data 폭발적 증가와 Interconnect 병목 현상 >> 시스템효율 저하



에너지 및 환경 문제 >> 에너지 탄소 배출

- ✓ AI 언어모델 ChatGPT, GPT-3 트레이닝 1,287MWh를 소비하고 모델 학습시 사용되는 에너지 외에도, 전력 소비와 열을 발생시키며 에너지 효율적인 칩 설계의 최적화는 필수
- ✓ ChatGPT는 GPT-3.5 버전의 트레이닝은 552톤 CO2e 배출

< Large deep learning 7 모델들의 에너지 소모 비교 >

Model	Energy consumption, MWh	CO2e emissions, tons
Evolved Transformer	2.5	3.3
T5	85.7	46.7
Meena	232	96.4
Gshard (600)	24.1	4.8
Switch Transformer	179	72.2
GPT-3	1,287	552.1
PALM	3,283	271

(Source: "The Carbon Footprint of Machine Learning Training Will Plateau)

생성형 AI 일상화 >> GPU 데이터 처리

- ✓ AI 추론모델은 성능 뿐 아니라 TCO(Total Cost of Ownership)의 중요성이 높아지기에, NVIDIA의 약점 - Applications 세분화되며, 특정 Workload 전용 솔루션 요구
- ✓ Event-driven / Asynchronous Analog/Digital/Mixed Spiking Signals구현

AI 반도체를 둘러싼 환경 변화에 따른 국가 경쟁 치열

국가 차원의 AI 반도체 혁신적 기술 개발 전략 필요

- 국외 메모리 반도체 및 파운드리 역량은 삼성전자와 TSMC가 주도권을 가지고 있으나, 국외 반도체 IP시장은 여전히 우위를 점하고 있음
- RISC-V의 특수성에 따라 미국·유럽·중국 등 지역별로 원천기술·활용성 확보를 위한 맞춤형 기술개발 추진 중
- 공정 의존성이 높은 AI 반도체의 특성으로 대형 기업 위주 재편화 중으로 반도체 기업이 아닌 대형 인터넷 기업 등이 반도체 설계 조직을 내부적으로 구성하고 자사를 위한 반도체 설계·제조를 본격적으로 시작
- (데이터센터) 범용성이 중요하여 GPU 위주 시장으로, 최근 초거대 인공지능모델 성능 등 대규모 기술 경쟁 예상
- (엣지) 최종 제품 요구사항에 최적화된 맞춤형 제품 및 서비스 제공을 통한 파이프라인 형성이 매우 중요한 요소로 부각 중

■ 글로벌 시스템반도체 기업들의 AI반도체 시장 진입

주요 기업	학습용(Traning), 클라우드용	추론용(Inference), EDGE 용
테슬라	Dojo 3rd Gen ('27)	FSD
아마존	Trainium	Inferentia
인텔	Gaudi 3	Goya, Greco
구글	TPU 6th Gen, Trillium	Edge TPU
퀄컴	-	Zeroth, Snapdragon X Neural
애플	-	M4 Neural Engine
삼성전자	-	Exynos NPU & MACH-1
ARM	-	Ethos-N78
그래프코어	IPU	IPU
하일로	-	Hallo-8, Hailo-15
Kneron	-	KL720, KNEO330
Cerebras	CS-2s, CS-3	WSE-2, WSE-3
사피온	X330(출시전)	X330(출시전)
퓨리오사	-	Warboy, RNGD
리벨리온	-	ATOM, Rebel
딥엑스	-	DX Series

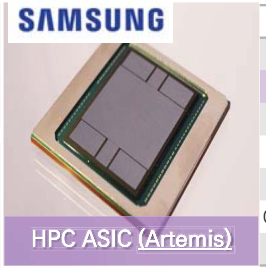
*출처: 각 사, 언론 종합, 다올투자증권

02 혁신과 기술 차별성 | 양산 적용 가능한 기술

ASIC Accelerator 솔루션 (핵심요소기술 주력제품)

- 우수한 반도체 설계기술로 ASIC/SoC 양산 칩 20여개 이상 개발 경험 풍부한 팀워크 및 검증된 Global Expert (삼성전자 수요/검증)

디지털• 아날로그 통합 설계	병렬 처리 전원 및 입출력 회로 설계
4nm FinFET 등 최첨단 파운드리 공정 및 Transistor level 비동기 회로 설계	Standard cell recharacterization 기술 기반 0.3V 이하 초저전력 게이트 구현



Fabless	제품	Foundry선단공정
APPLE	A15	3NM
Soteria	Artemis	4NM
NVIDIA	H100	4NM
Samsung	Exynos2400	4NM
APPLE	M1,A14	5NM
Qualcomm	5G CPU	5NM
TESLA	CPU	7NM

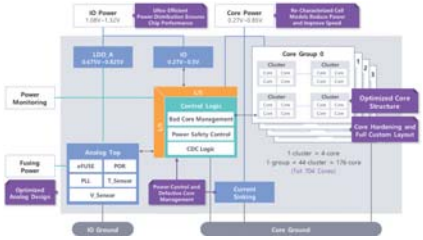
액침냉각에 최적화된 AI ASIC 클라우드 초저전력 주문형 HPC 가속기

- 고객사와의 협업으로 Eco-Friendly Data Center Immersion Cooling 유지 보수에 최적화된 Soteria Dependable Power Protection Circuit, Bad Core Management 기술 및 Firmware 요구
- Immersion Liquid Cooling (액침냉각) 최적화, 거대규모 고속 병렬연산을 실제 구현하는 양산에 필요한 기술을 개발 중 (기술혁신개발사업 '23~'26)

방열판 리브를 통해 유전체 유체가 전파 | 기존 냉각 수조(회로 기판 고정 장치)가 시간당 4,500L의 큰 유량

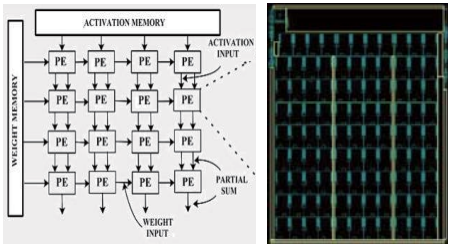
데이터센터 전력 40% 감축 | 물 사용 ZERO | 공간효율 증대

AI data centers shift to liquid cooling – MS, Meta, etc.



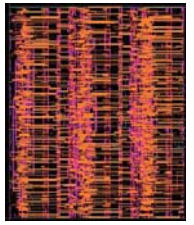
Seamless Logical/Physical Design

- 어플리케이션 워크로드별 타일링 및 하드닝/ 기본 PE(프로세싱 요소)의 그룹화



- 중요 연산자 (가산기, 승수 등)의 알고리즘 및 게이트 수준 최적화
  ex) 4나노 FinFET 공정 최적화
  Pulsed Latches 비동기연산

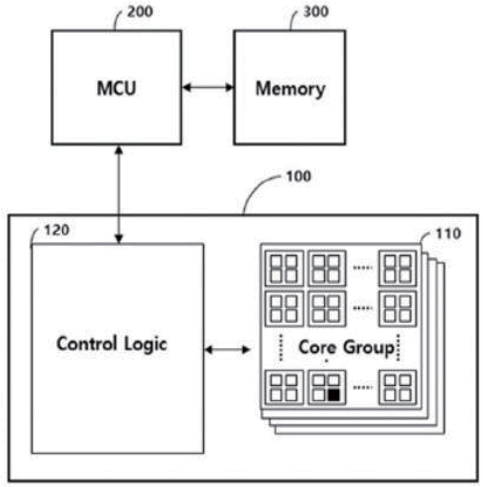
- PPAC (Power, Performance, Area, Cost) 최적화를 위한 PE 코어 강화 및 전체 칩 레이아웃



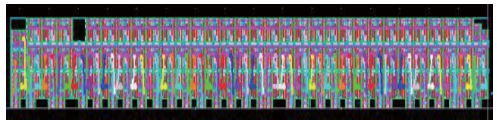
Optimized for Customer Operations

- Dependable Performance Harvesting Technology; Bad Core Management

원천 특허 기술



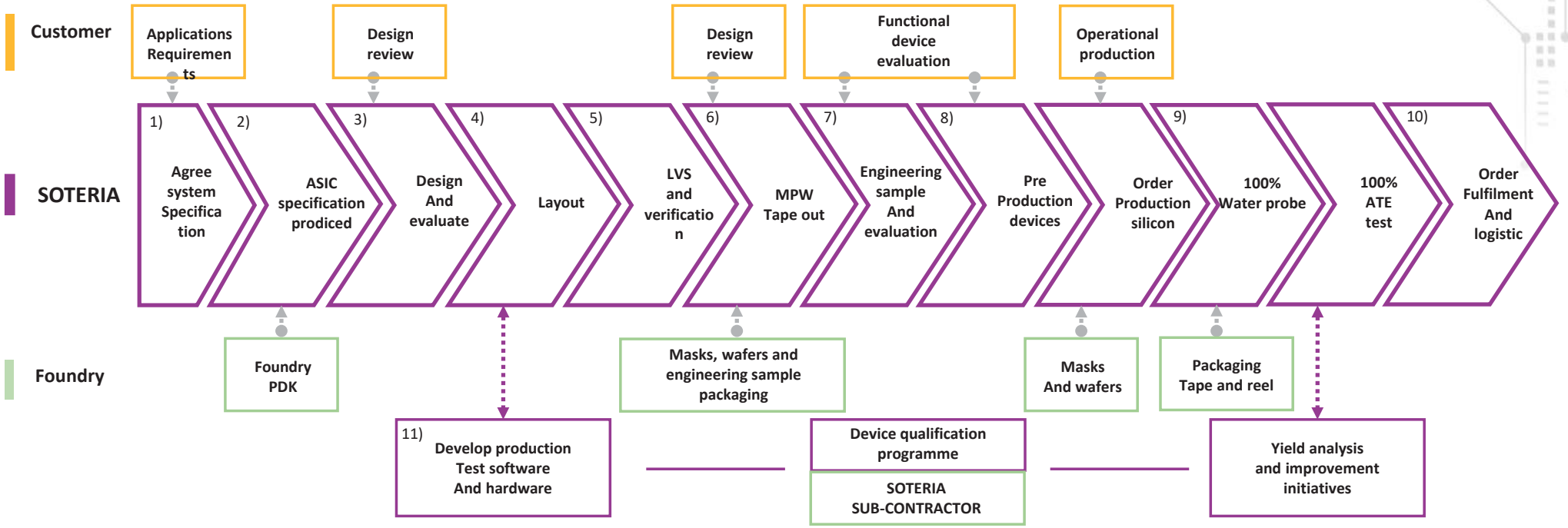
- Heat Dissipation Full Custom Layout



Systematic process

SOTERIA
Artificial Intelligence

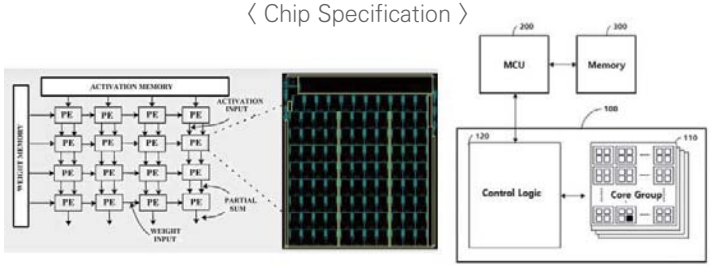
DTCO (Design Technology Co-Optimization) with global professional teams in ultra low-power, digital & analog expertise



02 강화된 고객 맞춤형 양산 개발 플랫폼

1단계 어플리케이션 워크로드별 타일링, 기본 프로세싱 요소(PE)의 그룹화, 중요 연산자의 알고리즘 최적화, dynamic latch 기반 비동기 연산 구현

〈 Chip Specification 〉



Process technology	Samsung 4nm Process
Number of PCB layers	12 layers
Number of ball count	64ea (Except Power and Ground)
Ball minimum pitch	0.1um
Operating voltage	4nm (Analog IP : 0.75V / Control logic : 0.3V / Core Logic : 0.3V / IO : 1.2V)
Operating frequency	4nm : 1GHz (@0.3V), (controller logic, @0.3V+-10%)
Total Power budget @PVT	4nm : ~6.3W (Nominal/105C condition, confirmation after PTPX result)

어플리케이션 워크로드 분석

- 특정 어플리케이션의 워크로드를 분석하여 가장 중요한 작업과 연산을 분석하여, PE에 Architecture, 어떤 연산자가 가장 많이 사용되는지 등 고객사와 Technical Meeting을 통해 결정

타일링 및 하드닝

- 각 작업에 대해 적절한 PE를 선택하거나 설계하고 이들을 타일 형태로 배열하고, Clustering 및 계층 그룹화 하여, Hybrid Hierarchical Structure 구현하여, 전체 칩 내에서 데이터 흐름과 처리 속도를 최적화

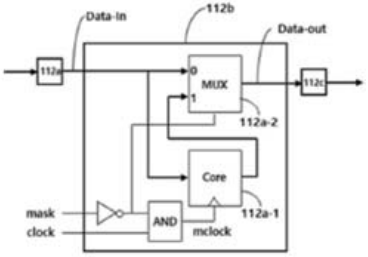
연산자 최적화

- 중요 연산자에 대해 알고리즘 및 게이트 수준에서의 최적화를 수행하여, MatMul, MSM, Adder의 경우, Tree-structure Parallel Prefix Computation 인 Lander-Fisher, Han-Carlson, Brent-Kung 방식들에 대한 분석과 최적화

Dynamic Latch 기반 비동기 연산 개발 구현

- 마지막으로 Dynamic Latch 기법을 사용하여 비동기식 회로 설계와 구현으로 전력 소모와 사이즈를 대폭 개선할수 있음. Intel의 Dynamic Logic, Self Shut-Off Pulsed Latch를 적용

▪ Bad Core and Defective Parts management (ASIC Fault)



02 혁신적인 세부 기술 구현

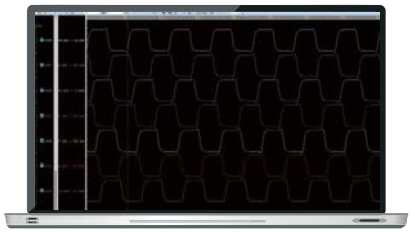
2단계

다중 전압 도메인, 0.3V 초저전압 라이브러리 재특성화, 저전압 IO 설계 및 언더슈트/오버슈트 보호 회로 구현

Analog-to-Digital Hybrid Integration

0.3~0.275V Near Threshold Voltage Library Re-Characterization

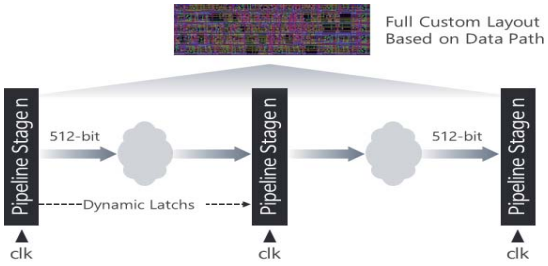
Simulation Results



- Soteria Innovative Design
- Ultra Low Power Operation – 0.3V Ring Oscillator

Unique Technological Prowess of Soteria

- Soteria Asynchronous Operation
- Advanced Self Shut-Off Dynamic Latch based Computation



다중 전압 도메인 지원

- 이는 칩 내에서 여러 개의 서로 다른 전압 도메인을 구현하여, 각 도메인에서 필요한만큼의 전력만 사용하도록 더욱 세밀하게 조정하여 전체 전력 소모를 저감

0.3V 라이브러리 재특성화

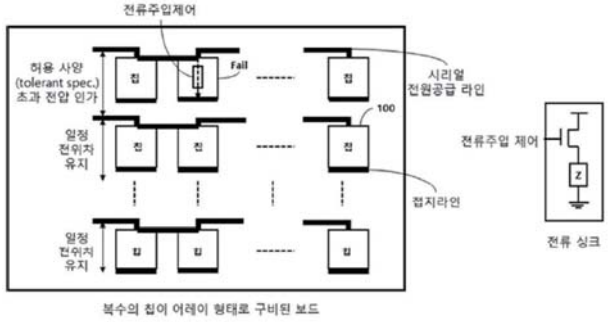
- 이 단계에서는 기존의 디지털 논리 게이트 라이브러리를 0.3V 운영 조건에 맞게 재특성화. 이 과정은 회로의 성능과 정확도를 유지하면서 작동 전압을 낮추기 위해 필요하며, 삼성 파운드리 공정기술팀과 협업

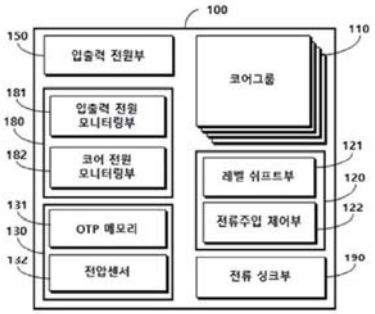
저전압 IO 설계

- 여기서 목표는 입출력(IO) 인터페이스를 저전압인 상태에서도 안정적으로 작동할 수 있도록 설계하며, 칩과 외부 세상 사이의 데이터 교환 시에도 낮은 에너지 소모

언더슈트/오버슈트 보호 회로 설계

- 입력 신호가 너무 낮거나(언더슈트) 너무 높아짐(오버슈트)을 방지하기 위해 보호 회로를 설계하며, 칩이 예상치 못한 조건 하에서도 안정적으로 작동하도록 보장하여 액침 냉각 (Immersion Liquid Cooling) 시스템 유지 보수에 최적화



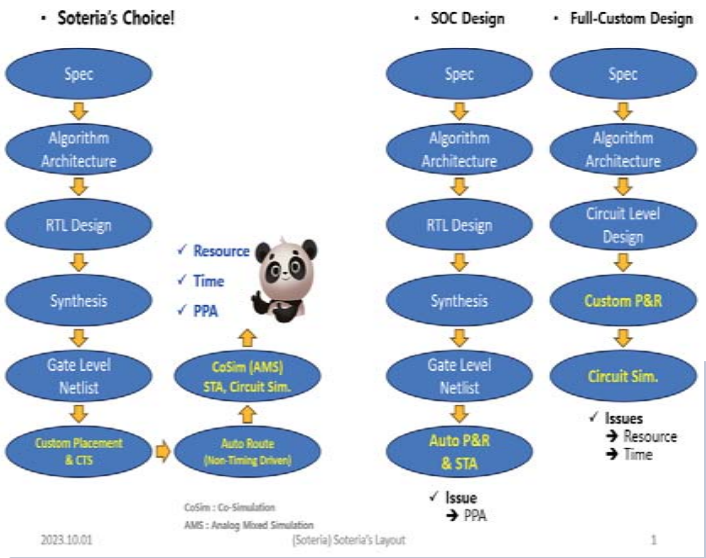


02 선도적인 첨단 개발 시스템

3단계

PE Core Reinforcement and Full Chip Layout for PPAC (Power, Performance, Area, Cost) Optimization & Heat Dissipation Mechanism

Hybrid P&R + CoSim Verification



PE 코어 강화와 Full custom layout을 통한 PPAC 최적화 및 효과적인 방열 메커니즘 개발

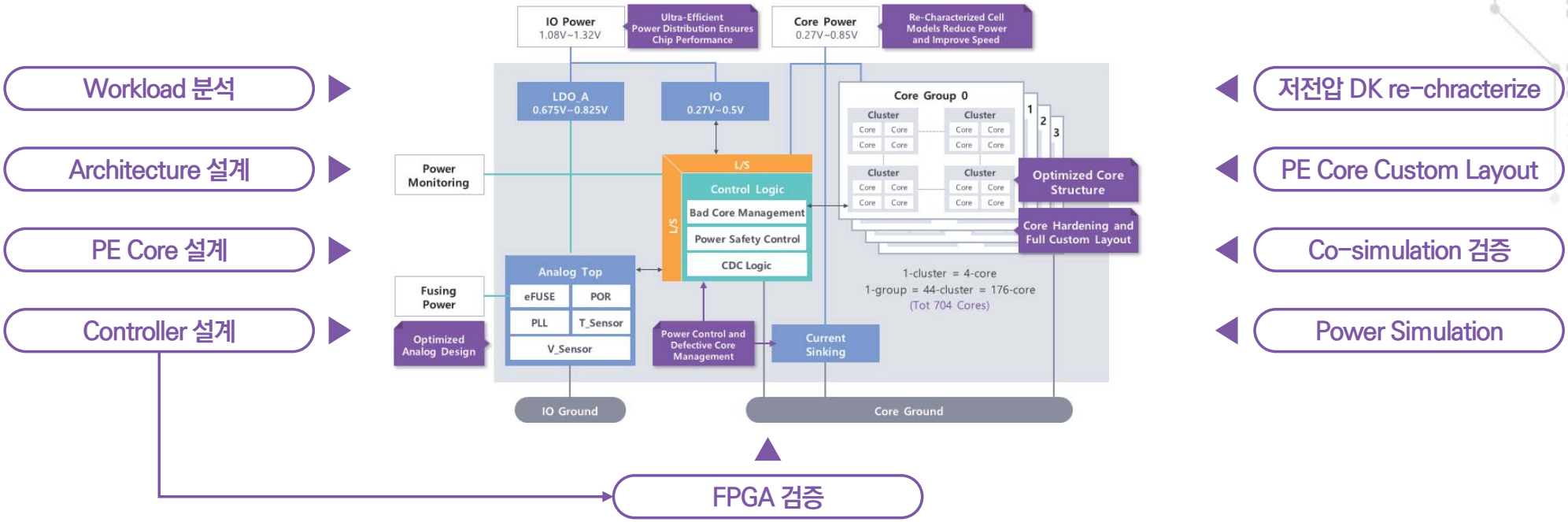
- Clustered PE Core : Speculative Parallel Processing 을 효과적으로 구현 할수 있고, Explicit expressiveness task 기반의 SW/HW 통합 architecture 구조화, FinFET 효율성을 극대화하는 super pipeline 개발
- Hybrid custom layout: PE Core layout P&R에서는, DRAM 설계 접근법을 도입하여, 스택핑, Virtuoso 최신 버전과 고밀도 3D manual layout을 통해, size 최소화하고 Interconnect delay 최적화
- 지능형 PPAC 최적화: 핵심 코어 프로세싱에서는 Full custom layout을 통해 트랜지스터 수준에서 회로를 제어하고, 전력 소비, 성능, 발열 및 면적 등 중요한 파라미터에 대해 최대한의 제어력을 가지고, 최적 성능 극대화가 가능하고, 개별 트랜지스터와 인터커넥트를 정확히 배치할 수 있으므로, 디자인 과정에서 큰 유연성을 얻을 수 있으나, 시간 및 인력 소요 등을 감안하여, AI 머신러닝을 활용
- 첨단 열 방출 메커니즘: 실시간 온도 모니터링과 동시에 열 분산 경로를 자동으로 조정하는 혁신적인 열 관리 시스템을 개발

액침냉각에 최적화된
AI ASIC 클라우드 초저전력
주문형 HPC 가속기

02 주요 결과물

연구개발 주요결과물

Immersion Liquid Cooling(액침냉각)에 최적화된 AI ASIC 클라우드 초저전력 주문형 HPC 가속기



OUTPUT

디자인 Document
단계별 설계 문서

시뮬레이션 결과
PE 코어 강화, 칩 레이아웃 최적화 및 열 관리 메커니즘의 효과를 보여주는 시뮬레이션 결과

칩 설계 파일
실제 제조 과정에서 사용되는 반도체 칩 설계파일, GDSII 형식으로 제공되며 마스크 제작에 필요

FPGA 프로토타입
초기 설계 검증을 위해 제작된 프로토타입

HPC 가속기 칩
실리콘 패키징된 상용버전 칩

제품 특성

Product Code:	MIK-100
Key Features:	High integration density of MPPA → *Excellent peak performance*
Process:	4nm low-power customized process
Package Type:	Flip-chip exposed die packaging

시장 기술 차별화의 구체성

High Performance Dependable Ultra Low Power Architecture & Library

- 내부 기준 전압 및 비교기를 사용하여 코어 압력전압을 0.3V NTV (Near Threshold Voltage) 설정
- Full custom layout & Virtuoso AI 기반 Hybrid 디자인
- 동작 전압 NTV 맞춤형 Library 재특성화, Schematics 최적화
- MPPA (Massively Parallel Processing Array) scheduling & distribution hardening
- Performance harvesting & fault-tolerant 기술

- 유럽 지역 데이터 센터들과 Immersion Cooling 프로모션
- 2억불 공급 계약 체결 및 수주
- Target TAM 1.2GW 친환경 데이터 센터 Project
- 삼성 파운드리, DSP와 전략적 파트너쉽 구축
- 2024년 Q4 부터 4NM Chip 양산, 3년간 약 5억불 매출 기대

03 Domain Specific Architecture

CXL기반 Pooled memory 및 NPU를 사용한 NDP(Near Data Processing)의 필요성

- BERT와 같은 대규모 AI 모델의 경우 대용량의 크기로 이해 시스템 메모리 계층내에 모두 적재 될 수 없으므로, 모델 추론 시 디스크로의 Swap으로 인해 처리량 및 응답 시간이 하락하는 문제가 발생
- 특히, 언어모델에서 짧은 응답 시간 및 긴 시퀀스 처리에는 심각한 저하를 발생시킴
- 이를 해결하기 위해 CXL 기반의 Memory Pool 서버 같은 확장성 있는 메모리 시스템 구축이 필수적임.
- 이러한 메모리 시스템 서버의 구축은 추론 연산을 위해 필요한 데이터 이동을 줄이며, NPU와 같은 계산 자원을 활용한 Near Data Processing을 가능하게 함으로써 Host의 CPU,GPU 및 메모리 자원의 절약, 전력감소 및 짧은 응답시간 및 대량의 데이터 처리에 이점을 갖음

인공지능 신경망 연산을 위한 새로운 컴퓨터 아키텍쳐 도래 및 Ecclesia 프로젝트

- 인공지능 신경망 연산에 특화된 GPU를 판매하는 Nvidia의 시장 독점과 공급부족 등과 온디바이스 AI의 탄생 및 시장 확대의 이유로 전세계 수많은 대기업 및 스타트업들이 Nivida를 대체하는 NPU를 개발하거나 온디바이스 AI에 특화된 edge-AI processor를 개발 중
- 하지만 Nivida가 오랜 시간동안 만들어 놓은 소프트웨어 생태계와 고객과의 관계는 새로운 회사가 진입하기는 극히 어려운 상황이며 온디바이스 AI 시장은 24년 최초로 삼성전자 갤럭시S24가 온디바이스AI를 표방하며 출시 되었지만 대기업들이 자체 솔루션을 사용하고 있기 때문에 새로운 팹리스 업체 입장에서는 큰 시장은 접근하기 힘든 것이 현실
- 따라서 소테리아는 기존 Nivida나 글로벌 대기업들의 자체 솔루션과 경쟁하지 않고, 오히려 생태계 속에서 협력하는 사업모델로 Ecclesia 프로젝트를 기획

선단 공정인 4나노를 통해 MPW test chip 구현

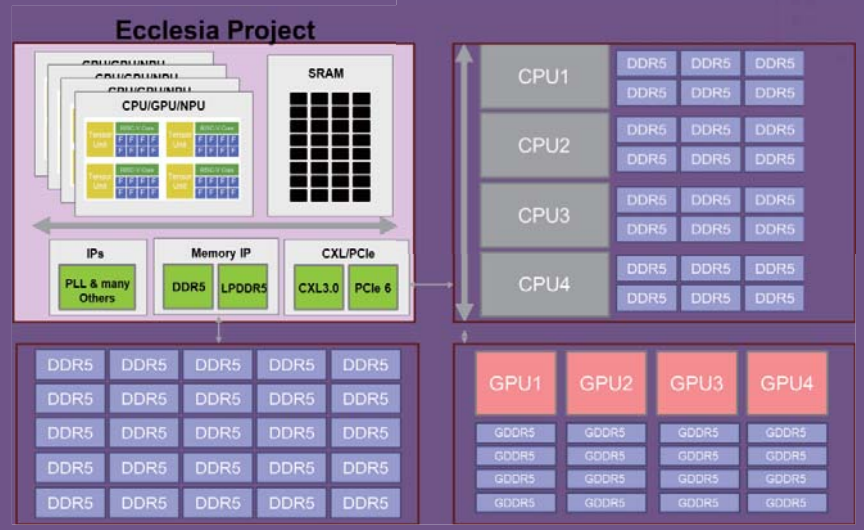
BERT-Large를 latency 4.35ms 이하 성능 달성을 목표

현재 약 284억 규모 개발비

Semidynamics, SKYECHIP, 조지아공대 공동개발로 글로벌 R&D

[Ecclesia 프로젝트 구성 목표]

Ecclesia는 자체 CPU/GPU/NPU를 포함하여 충분한 SRAM 캐시메모리를 탑재하여 하나의 보드에 함께 구성되어 있는 메모리를 사용하여 인공지능 신경망 연산 처리



기존 CPU와 GPU 연산처리 대폭 삭감

잦은 데이터 인터페이스 필요 감소

효과

시스템의 저전력화

전체 시스템의 TCO (Total Cost Operation) 감소

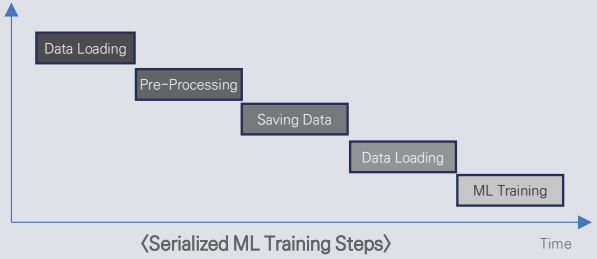
03 소테리아 NDP "Ecclesia Project"

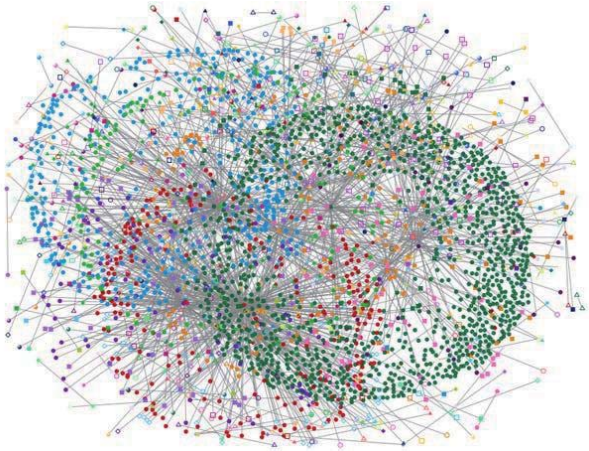
- Memory-based deep-learning Solutions: Higher energy efficiency

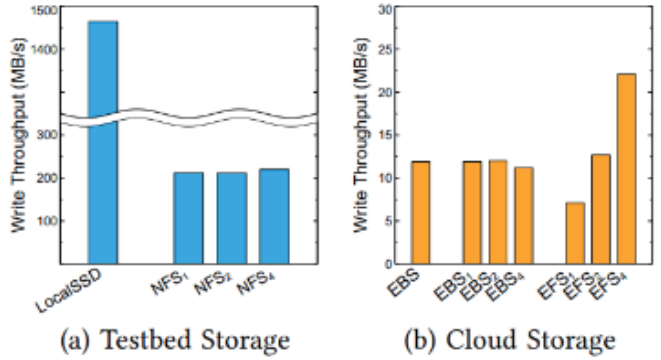
➢ 대규모 LLM, GNN 모델 및 학습

➢ Slow Storage
  - AWS 클라우드의 EFS, EBD 등 스토리지의 I/O 대역폭 매우 낮음
  - Write bandwidth 는 30MB/s 보다도 낮음

➢ 대규모 모델 기계 학습 시 Data 이동으로 인한 학습 병목







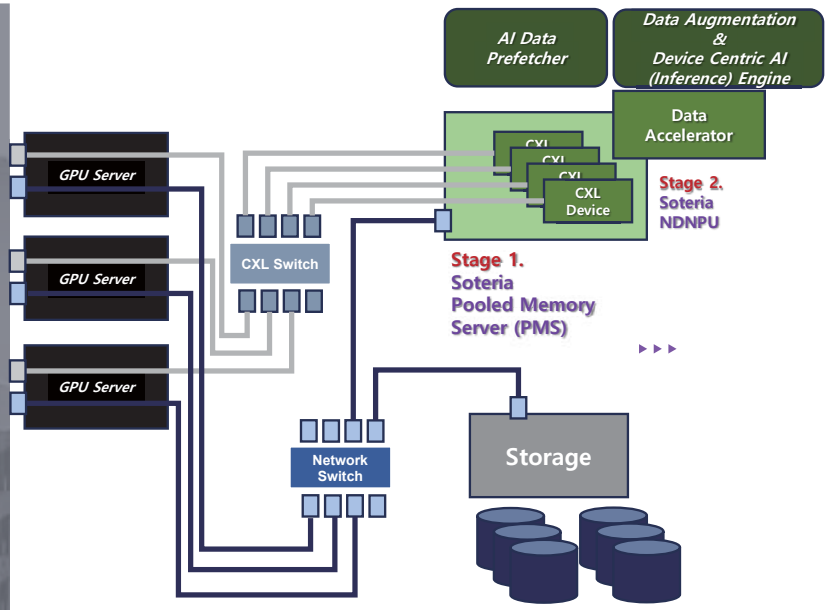
(a) Testbed Storage (b) Cloud Storage

Figure 5: Write throughput measurements of in-house testbed and AWS storage. We used 16 writer threads.

TABLE I
VARIOUS TYPES OF INSTANCES PROVIDED BY AWS. THIS INFORMATION WAS AVAILABLE ON OCTOBER 1, 2023 IN N.VIRGINIA REGION.

Name	Type	On-Demand	Spot VM	vCPU (#)	Mem (GiB)	Network (Gbps)	Storage (Gbps)
g3s.lxarge	GPU	0.75	0.225	4	30.5	- 10	N/A
g4dn.xlarge	GPU	0.526	0.1941	4	16	- 25	- 3.5
g5.xlarge	GPU	1.006	0.3018	4	16	- 10	- 3.5
p2.xlarge	GPU	0.2704	0.225	4	61	High	0.75
p2.8xlarge	GPU	7.2	2.1165	32	488	10	- 5
p2.16xlarge	GPU	14.4	4.2262	64	732	25	- 10
c3.xlarge	CPU	0.21	0.1281	4	7.5	Moderate	N/A
c4.xlarge	CPU	0.199	0.1257	4	7.5	High	0.75
t3.xlarge	CPU	0.1664	0.1033	4	16	- 5	N/A
c5.xlarge	CPU	0.17	0.0999	4	8	- 10	- 4.75
m6i.xlarge	CPU	0.192	0.12	4	16	- 12.5	- 10
d3.xlarge	CPU	0.499	0.1584	4	32	- 15	0.85
c5n.xlarge	CPU	0.216	0.127	4	10.5	- 25	- 4.75
c6in.xlarge	CPU	0.2268	0.1272	4	8	- 30	- 20

03 소테리아 NDP “Ecclesia Project”



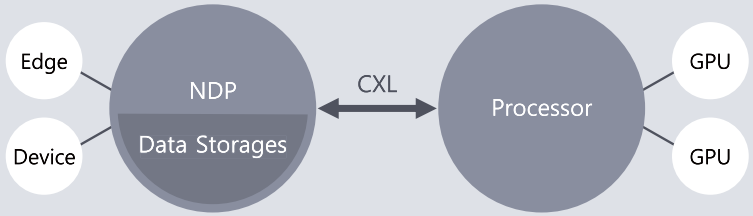
Hardware Solution

- The memory server acts as an intermediate server for ML pipelining.

Software Solution

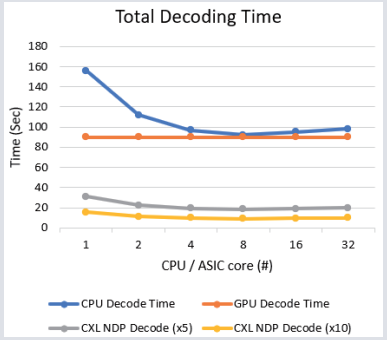
- Responsible for data prefetching and preprocessing.
- Responsible for communication and scheduling for DDL framework and data movement

소테리아 Near Memory LLM Accelerator (NM-LLMA)



- LLM 모델 크기 확장으로 GPU HBM 메모리 부족과 고비용 HBM으로 하부 메모리 시스템의 적극적인 활용이 중요함
- CXL 기술로 노드 메모리 확장 가능, 하지만, 높은 Latency
- Transformer 기반 LLM 모델을 Prefill 단계와 반복적인 Decode 단계로 실행됨. LLM 추론 시 CXL 메모리에 있는 데이터의 병렬 연산은 CPU-GPU 간의 데이터 이동으로 GPU 효율성이 낮아짐
- Soteria Near Memory LLM Accelerator(NM-LLMA)은 Nvidia GPU 사용율을 극대화 하며 고속의 Token 생성을 가능하게 함

〈Near-Memory 기반 병렬 연산 가속 하드웨어 프로토타입〉



▷ 서버 구성
- NVIDIA 2080 Super GPUx2
- AMD Ryzen CPU, 64GB
- NUMA 서버 기반 CXL NDP 하드웨어 에뮬레이션
- CPU를 사용한 병렬 연산 가속기 하드웨어 에뮬레이션

▷ 실험 환경
- FlexGen on Pytorch / CUDA 12.1
- Facebook OPT1.3B, 50 Layers
- 모든 병렬 연산 중간 데이터는 하부 메모리 시스템이 오프로드 되어 있음

03 연구개발 성과 활용 방안

대규모 언어 모델(Large Language Models, LLMs) 및 Data intensive 인공지능(AI) 모델의 개발 고도화로 교육 · 의료 · 서비스 등 수많은 영역으로 확장 가능

Dis-aggregated Pooled Memory Server(with CXL)용 NDNPU(Near Data Neural Processing Unit)는LLMs 및 Data intensive 인공지능(AI) 모델의 개발 및 실행에 있어 성공적인 개발, 통합될 때 효율성 및 성능향상, AI 모델 고도화, 연구개발 촉진 등 다양한 활용이 가능

효율성 및 성능 향상

NDNPU의 처리 속도 증가는 클록 속도, 병렬 처리 능력, 메모리 효율성을 통해 달성되며, AI 작업을 더 빠르고 효율적으로 수행

- NDNPU는 데이터 근처에서 직접 계산을 수행해 데이터 이동 시간을 줄여 처리 속도를 높이고, 병렬처리를 통해 대규모 AI 모델 학습 및 추론 속도를 크게 향상시는 장점이 있으나, 전력 소모에 대한 단점이 있는데 이를 해결

고도화된 AI 모델 제공

고도화된 AI 모델 개발로 의료 · 금융 · 자동차 · 엔터테인먼트 등 다양한 분야에서 새로운 응용 프로그램과 서비스를 창출

- NDNPU는 데이터 근처에서 직접 계산을 수행해 데이터 이동 시간을 줄여 처리 속도를 높이고, 병렬처리를 통해 대규모 AI 모델 학습 및 추론 속도를 크게 향상시는 장점이 있으나, 전력 소모에 대한 단점이 있는데 이를 해결

실시간 개인화된 의료 진단 및 치료 계획 | 실시간 금융 거래 및 위험 관리 | 지능형 도시 및 스마트 인프라 관리

AI 연구개발 가속화

더 빠른 실험 반복과 더 큰 모델 실험이 가능함으로 연구진이 새로운 아이디어를 더 빨리 시험하고 발전시킬 수 있음

- (학습 데이터셋의 확장) 더 큰 메모리 풀과 더 빠른 처리 능력은 더 큰 학습 데이터셋을 활용할 수 있어 Data layered processing 가능

04 연구개발 효과 기존 강자가 없는 Domain Specific Architecture

과학 · 기술적 측면

NDNPU(Near Data Neural Processing Unit) 기술은 컴퓨팅 아키텍처 근본 변화, 인공지능 연산 극대화, 엣지 컴퓨팅 등 Deep Learning Domain Specific Acceleration 으로 새로운 컴퓨팅 패러다임 제시

컴퓨팅 아키텍처 근본 변화

- NDNPU는 데이터 중심의 컴퓨팅 아키텍처로의 전환을 가속화하여 Von Neumann Architecture의 구조적 한계를 극복하고, 데이터가 저장된 위치에서 직접 연산을 수행함으로써 3세대 AI Neuromorphic Computing으로 에코시스템을 구축하는데 효과적

인공지능 연산 극대화

- NDNPU는 특히 AI 연산에 특화되어 복잡한 신경망 모델 연산을 메모리 단에서 직접 처리할 수 있으며, 대규모 AI 모델 학습과 추론 시간을 현저히 단축시키는 등 기존 하드웨어 한계를 뛰어넘는 방향 제시

엣지 컴퓨팅과 IoT의 혁신

- 데이터 처리를 위한 에너지 소비와 지연 시간을 최소화하여 소형 IoT 장치에서도 고도의 AI 기능을 구현할 수 있도록 응용될 수 있으며, 다양한 분야에서의 혁신적인 응용 가능성으로 확장

경제 · 사회적 측면

NDNPU 기술개발의 경제. 사회적 측면 기대효과는 산업경쟁력 강화, 에너지 소비 감소, 수입 대체 및 수출, 전문인력 양성, 국가 브랜드 제고 등

산업 경쟁력 강화

- 데이터 처리와 인공지능 연산의 효율성 향상을 통해 다양한 산업 분야에서 기업들의 경쟁력 강화

에너지 소비 감소

- 에너지 효율적인 데이터 처리를 통한 전력 소비, 탄소배출 감소에 기여

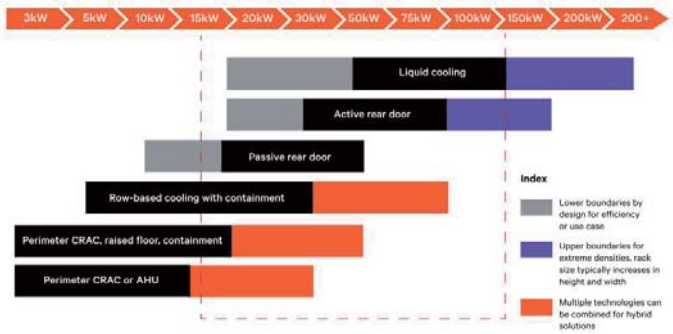
기술 접근성 향상

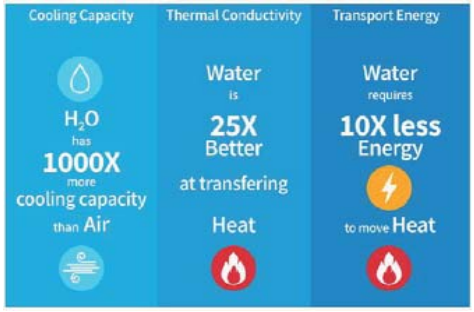
- 데이터 처리와 인공지능 연산의 효율성을 획기적으로 높이며, Data layered 컴퓨팅 인프라의 혁신을 가져옴
- 3세대 AI, SNN (Spiking Neural Network) 고도화에 응용될 수 있음
- 스마트 시티 구축 가속화: NDNPU를 활용하여 실시간 데이터 처리와 분석이 가능해짐에 따라, 교통 관리, 에너지 관리, 공공 안전 등 스마트 시티의 핵심 요소들이 크게 개선될 수 있음

인프라 측면

- 데이터 센터에 본사 HPC 칩과 액침 냉각 기술을 함께 공급
- 데이터 센터 환경에 최적화된 Immersion Cooling 시스템과 AI ASIC 가속기를 함께 고려하여 통합된 시스템을 제공함으로써 효율성과 성능을 극대화
- 또한, 관리 및 모니터링 기능을 갖춘 소프트웨어와 연동하여 데이터 센터의 효율적인 운영을 지원
- 데이터센터 에너지 수요는 폭발적으로 증가하여 2021년 전 세계 에너지 수요의 최대 2%를 소비하고 있어 당사의 액침 냉각에 최적화된 주문형 HPC 가속기는 이러한 문제를 해결할 수 있는 미래선도 기술임

- **조용한 작동**
- **CAPEX 및 OPEX 절감(kW당)**
- **더 나은 TCO(kW당)**
- **열 방출 용량이 약 10배 증가**
- **필요한 공간 감소**
- **에너지 효율성 및 지속 가능성 향상**
- **칩, 보드등 하드웨어의 신뢰성 향상**
- **환경 오염 물질 저감**
- **하드웨어의 부식 문제 감소**
- **전기화학적 문제 감소**





04 사업화 계획 | 해외시장 진출 계획

✓ 시장조사기관 (HPC Consulting)등과 연계하여 미국 중심 Data Center 고객의 기술 수요 및 제품 spec 파악
✓ Cloud computing, legal, finance등 중소규모 Data center와 선 개발계약을 체결 목표
✓ NRE 선금을 입수하고 특정 고객의 Data center 워크로드와 시스템 needs를 바탕으로 Customized 서버 및 Ecclesia 반도체 개발을 통해 ASIC 및 ASSP의 하이브리드 제품 전략 수립

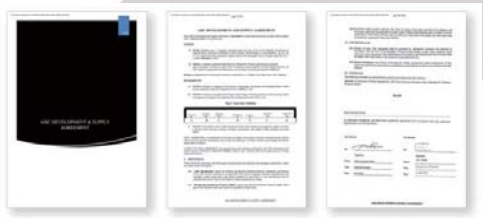
@ USA

- 실리콘밸리 연구소에서 혁신적인 반도체 설계 노하우를 가진 엔지니어를 바탕으로 연구 개발과 시장 거점을 확보
- 북미지역 유통채널을 통해 공급 계약 체결 및 구매 의향서
- Top-notch Research Institutes와 산학연구 진행으로 기술사업화 파트너쉽 추진

〈 북미지역 Channel Partnership 구매의향서 〉

@ EU

- Neurotron™ 알고리즘 탑재하여 데이터 센터(ASDC)로부터 Customer Specific Standard Product(CSSP) 수주
- 머신 러닝을 위한 MPPA 기반 고속 연산 가속 ASIC Chip 및 하부 메모리 시스템에서의 LLM Accelerator
- 연산 속도를 50~200배로 높여 이미지 분류, 객체 감지 및 추론 엔진 등 다양한 분야에 적용



유통 채널명	판매 내용	진출 시기	판매 금액
ACME 데이터센터	Neurotron AI Engine	'24.12	2억불 수주

@ China

- 10nm/8nm AI ASIC 개발 Fabless 업체인 Cynosure와 개발 협력관계를 구축하여 ASIC 뿐 아니라 HW solution 구성 까지 진행해서 중국 시장 개척 예정
- 홍콩 기반 전자 유통 회사 채널을 통해 Major End-user 확보

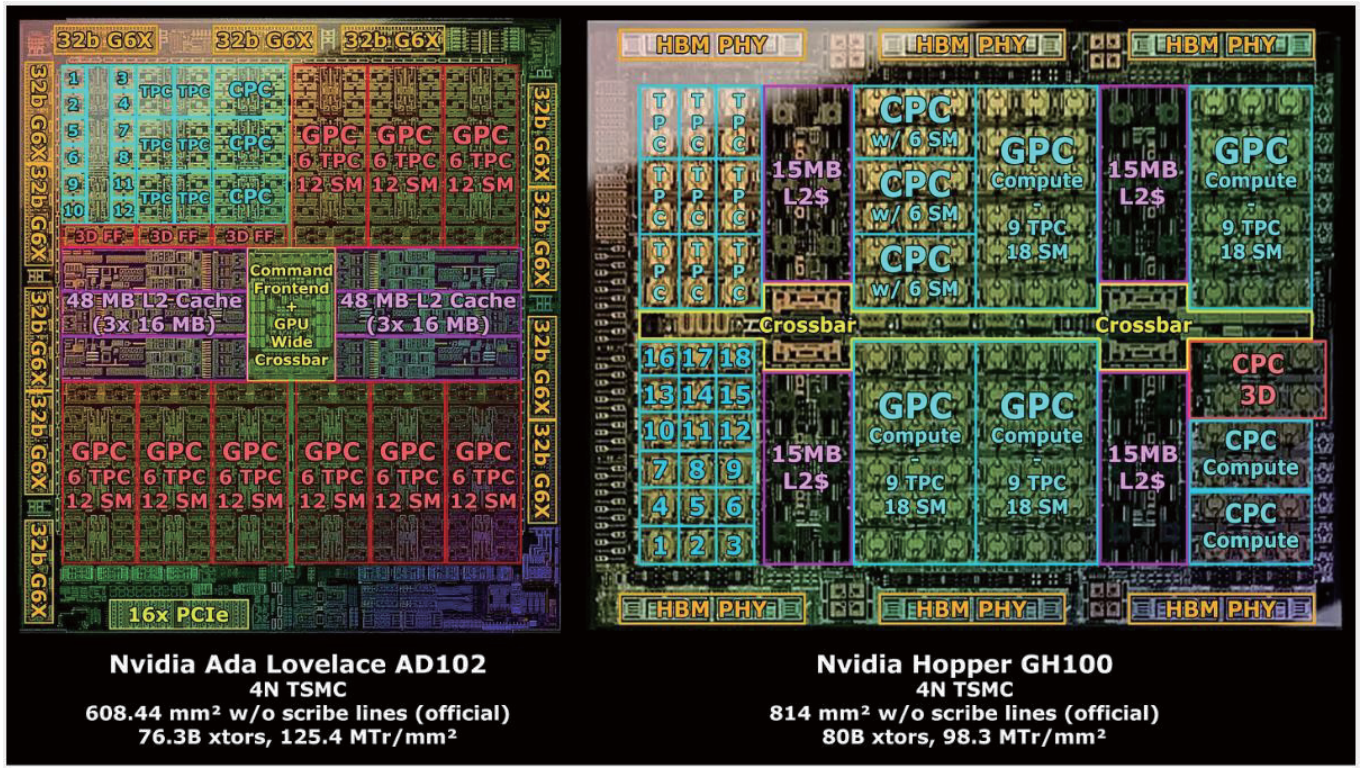
Major Partners / Customers

ZTE中兴 · HUAWEI · vivo · oppo · MEIZU · X&F兴飞科技 · TCL · 东莞通信 · UMEOX · innos · malata · 英華達 · HEDY.七喜 · AEON · HTT · CK Telecom · CHANGHONG长虹 · GOSO · neoway · REACH · GAOXINQI · Beatone


THANK YOU
Q & A

AD102 and GH100 Renderings from Nvidia

Appendix



Nvidia Ada Lovelace AD102
4N TSMC
608.44 mm² w/o scribe lines (official)
76.3B xtors, 125.4 MTr/mm²

Nvidia Hopper GH100
4N TSMC
814 mm² w/o scribe lines (official)
80B xtors, 98.3 MTr/mm²