인공지능 반도체
최고위 전략대화

국산 AI반도체를 활용한 『K-클라우드』 추진방안(안)(요약본)

2022. 12. 12.

과학기술정보통신부

왜 AI반도체인가?

딥러닝의 창시자, 제프리 힌튼

"반드시 **AI혁명은 한 번 더 일어날 것**이다.
다가올 혁명은 **하드웨어의 혁신**을 통해
이뤄질 가능성이 높다"

대통령과 AI 석학과의 대화
'22.9월, 캐나다 토론토大

AI 확산 가속화로 **AI반도체**는
'21년 $347억→**'26년 $861억**(現 **메모리 반도체 시장의 50%**)로
연 16% 급성장이 예상되는 분야 (가트너, '22)

「K-클라우드」 프로젝트란?

◇ **세계 최고 수준**의 **초고속·저전력 국산 AI반도체** 개발과 **데이터센터 적용**을 통해 **국내 클라우드 경쟁력**을 **강화**하고 국민들에게 **향상된 AI서비스를 제공**하는 프로젝트

1. 추진 배경

○ **국내 클라우드 경쟁력 제고**를 위해서는 최근 AI·5G 등으로 인해 수요가 대폭 증가하는 **데이터센터**에서 **기회**를 **모색**할 필요

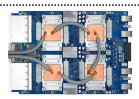
○ **데이터센터**는 **반도체 성능 향상**을 통해 **클라우드** 서비스의 **경쟁력**을 **획기적**으로 **개선**할 수 있는 **핵심적인 요소***

* (사례) 아마존(AWS)은 데이터센터에 **자사 AI반도체**를 적용하여 **기존 GPU보다** ① **최대 70% 저렴한 비용**에 **컴퓨팅 서비스**를 제공하고 있으며, ② **Alexa 음성인식** 서비스 비용은 **30% 절감**, **영상인식** 서비스(Rekognition)의 **속도는 8배 향상**시켰다고 밝힘

- 특히, **많은 전력**이 **소요***되는 **데이터센터**에 **저전력 AI반도체**를 적용, 데이터센터의 **탄소중립 달성**과 **친환경적 운영**도 기대 가능

* 데이터센터 내 전력 소비: **IT장비(52%, 이 중 44%가 서버)** > 냉각(38%) > 전력시스템(10%)

○ 이에 **글로벌 클라우드社**들은 **자사 전용 AI반도체 개발 경쟁 중**

amazon web services*	Azure (마이크로소프트)	Google Cloud Platform
 **Inferentia**	 **Brainwave**	 **TPU**

* '22.11.30, 아마존(AWS)은 기존 칩 대비 계산능력이 2배 향상된 '그래비톤3E' 및 'Inferentia2'도 공개, 엔비디아·AMD 등 반도체 회사와 경쟁 본격화

○ 한편, 우리나라에서도 **AI반도체(NPU) 출시**가 **시작**되었으며, 시장 성장에 대한 기대를 바탕으로 이들 **기업들에 대한 투자가 본격화***

* 퓨리오사AI社('21.6월, 네이버 등 800억원 투자), 사피온社('22.1월 SK하이닉스·SKT 등 800억원 투자), 리벨리온社('22.6월, KT·카카오벤처스 등 620억원 투자)

< (참고) 국내 AI반도체 현황 >

기업명 제품명	연산성능	전력 소모	기업명 제품명	연산성능	전력 소모
SAPEON X220 	87 TOPS*	65W (서버용)	**리벨리온** ION 	4 TFLOPS* (≒16 TOPS)	2.1W (서버용)
퓨리오사AI Warboy 	64 TOPS	60W (서버용)	**딥엑스** DX-L2 	6.4 TOPS	개발중 (엣지용)

* **(TOPS)** Tera Operations Per Second, **1초에 1조번 정수 연산**
(TFLOPS) Tera Floating-Point Operations Per Second, **1초에 1조번 실수 연산 (1 TFLOPS ≒ 2 TOPS)**
※ 이외에 텔레칩스, 아이닉스 등 7개 기업·기관이 정부 R&D로 엣지(단말)형 AI반도체 개발 중

☞ **국내 클라우드 경쟁력 제고**를 위해 메모리 반도체 기술력을 바탕으로 **글로벌 1위 수준**의 **초고속·저전력 AI반도체 개발**에 **승부**를 **걸 필요**

2. AI반도체 기술개발 현황

○ **현황** AI반도체 기술개발을 위한 **대규모 연구개발** 추진 중

- (NPU) 차세대지능형반도체 기술개발사업('20~'29, 총 1조 96억원) 등을 통해 서버·모바일·엣지 분야에 활용 가능한 **고성능·저전력 NPU 개발 중**
- (PIM) PIM인공지능반도체 핵심기술개발사업('22~'28, 총 4,027억원) 등을 통해 **프로세서와 메모리를 융합**한 **PIM 반도체 개발 착수**

※ 'PIM 설계연구센터(HUB)' 개소('22.6월), PIM인공지능반도체 사업단 출범('22.7월)

○ **향후 방향** **국산 AI반도체**가 출시되는 단계에서, 활용 극대화를 위한 **SW 개발** 및 **레퍼런스 확보**를 위한 **실증사업 병행 추진** 필요

- **극저전력 데이터센터 구현**을 위해 **AI반도체**와 **스토리지**를 모두 **비휘발성 메모리(Non-Volatile Memory) 기반**으로 개발 추진 필요
- 아울러, **AI반도체-클라우드-AI서비스**로 구성된 **산업 생태계 조성**을 위해 **산·학·연 간 협업 체계**를 마련할 필요

< AI 반도체 사업 추진 현황 및 'K-클라우드 적용'을 위한 향후 방향 >

구 분	추진 현황	향후 방향
①AI반도체 고도화	· NPU 지속 고도화 및 PIM 개발 (사업) 차세대지능형반도체기술개발('20~'29), PIM 인공지능반도체 핵심기술개발('22~'28)	· 엣지용으로 한정된 NVM 기반 AI 반도체 개발을 서버용까지 확대
②AI반도체용 SW개발	· 서버용 NPU/PIM 컴파일러·SDK 개발 (사업) 거대인공신경망 AI SW기술개발('23~'27)	· 이기종·다중 AI 반도체(가속기) 지원용 병렬처리 및 분산처리 SW · Smart SSD용 분산스토리지 SW
	-	· NPU 통합 서비스 활용 플랫폼 · 고성능 VM 및 컨테이너 제공 기술 · NPU/PIM 가상서버 클러스터 기술
③데이터센터 실증	· AI반도체 응용실증지원('21~계속)	· NPU Farm 구축사업('23년 신규) · PIM Farm 구축 사업 · NVM PIM Farm 구축 사업
④산학연협력 강화	· PIM-HUB 운영 (대기업-산학연 협력 및 인력양성 등 중심)	· AI반도체-SW·클라우드-AI서비스 기업 간 협력체계 마련 · AI반도체 대학원 신설('23년~, 3개교)

☞ **'K-클라우드' 추진**을 위한 **초고속·저전력 국산 AI반도체** 기반 **데이터센터 고도화**를 목표로 **기존 사업 재편 및 필요사업 신설** 등 **사업 체계화 추진**

3. 비전 및 목표

비 전

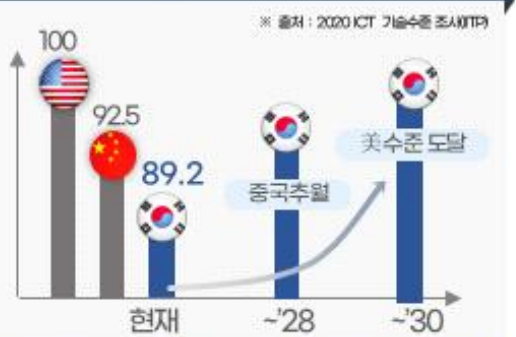
메모리 반도체 기술력을 바탕으로 **글로벌 최고 수준**의 **초고속·저전력 AI반도체 개발**에 승부를 걸어 **국내 클라우드 경쟁력**을 **혁신적**으로 **개선**

목 표

국내 시장 점유율 확대(~80%)

국내 데이터센터 시장에
국산 AI반도체 적용 확대

국가 기술수준 향상(89.2 → 100)



국산 AI 반도체 기술수준을
세계 최고 수준까지 향상

추진 전략

국산 AI반도체 고도화

국산 AI반도체를
3단계에 걸쳐 고도화

① ('23~'25) NPU →
② ('26~'28) 저전력 PIM →
③ ('29~'30) 극저전력 PIM

AI반도체용 SW 개발

국산 AI반도체를
데이터센터에 적용하기 위한
SW 추가 개발

국산 AI반도체 고도화 단계별
데이터센터 적용 및
4대 분야 AI서비스 선도 적용

데이터센터 · AI서비스 실증

'K-클라우드 얼라이언스' 구성
'PIM-HUB' 역할 강화
R&D 및 지원조직 강화
AI반도체 대학원 신설

산 · 학 · 연 협력 강화

4. 추진 전략

1 국산 AI반도체 고도화* : NPU → 저전력 PIM → 극저전력 PIM

(현재, '22년) GPU+DRAM+SSD ⇒ **(목표, '30년)** [NPU + PIM] + **Smart SSD**
(NVM 기반)

* '23~'30년 차세대지능형반도체 및 PIM인공지능반도체 기술개발 사업으로 8,262억원 투자 예정(현재 예타 기준)

○ **1단계 ('23~'25) : NPU (국내점유율 목표 23%)** 상용화 초기의 **국산 NPU를 지속 고도화**(추론→학습), **데이터센터에 적용***하여 **성공 레퍼런스 확보** 및 **초기 시장 창출**

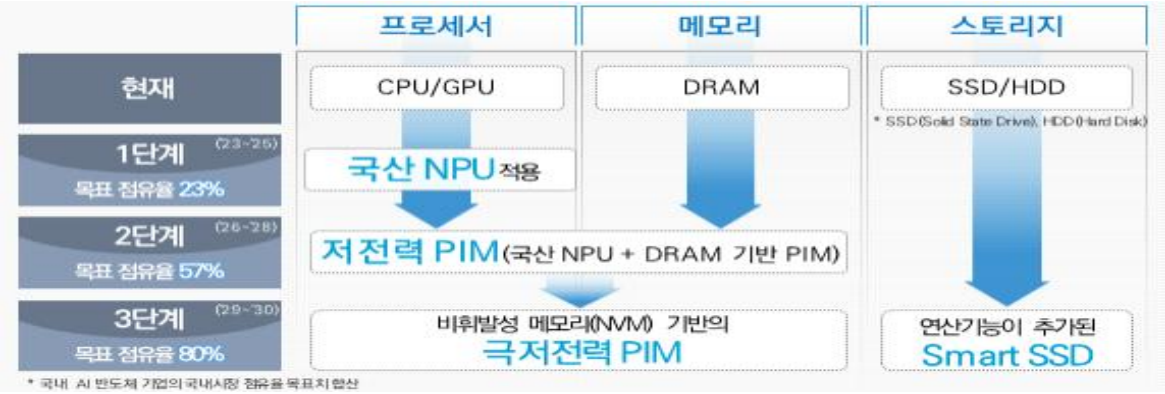
* AI반도체·클라우드·AI서비스 기업이 연합체를 구성하여 실증사업 추진('23~)

○ **2단계 ('26~'28) : 저전력 PIM (57%)** **DRAM 기반 상용 PIM**과 **국산 NPU**를 **접합(패키징)***하여 **외산 GPU급 성능**을 **저전력**으로 구현(**DRAM 기반 PIM**)

* (예시) ①삼성 HBM-PIM을 패키징으로 단일칩化 ②SK하이닉스 GDDR6-AiM을 컴포넌트 방식으로 연결

○ **3단계 ('29~'30) : 극저전력 PIM (80%)** **비휘발성 메모리(NVM)**를 활용한 **아날로그 MAC 연산 기반**의 **극저전력 NPU·PIM** 개발, **극저전력화 달성(NVM 기반 PIM)**

- 스토리지 내에서 데이터 연산처리가 가능한 Near-Memory PIM 형태의 **Smart SSD 스토리지도 적용**하여 **저전력 극대화**



2 AI반도체용 SW 개발 : 신규 예타 사업 추진

○ 기존 AI반도체 사업들은 프로세서(HW) 개발*을 목표로 진행되고 있어, 이를 데이터센터에 적용하기 위해서는 **SW 추가 개발 필요**

* '22년도 AI반도체 R&D 사업 (1,037억원) 중 SW사업 예산은 약 9.6% 규모

< (가칭) K-클라우드용 AI반도체 SW 기술개발 >

◇ **(목표)** 3단계에 걸쳐 고도화되는 **국산 AI반도체**에서 딥러닝 등 AI 알고리즘을 초고속·극저전력으로 실행하는 **컴파일러, 라이브러리, AI모델 자동 병렬화 기술** 등과 이를 상용 **클라우드**에 적용하기 위한 **가상머신(VM)** 및 **컨테이너, 가상 서버 클러스터 기술 등** 개발

◇ **(추진방안) 국산 AI반도체 고도화 단계별**로 **필요한 SW 기술개발** 중심, 신규 예타 사업 추진('23년 중)

3 데이터센터 실증 및 AI서비스 제공

○ **국산 AI반도체**를 **단계별로 데이터센터에 적용**하여 클라우드 기반 AI서비스를 실증함으로써 국산 AI반도체의 **레퍼런스 확보** 지원

○ **1단계로 국산 NPU 데이터센터 구축사업**과 **기존의 AI·클라우드 서비스 개발 사업**을 연계하여 **패키지**로 지원('23년 428억원(안), '25년까지 약 1천억원(잠정))

※ ('23.1월) 통합 공고 → ('23.3월) 국산 NPU 데이터센터 구축사업 선정 → ('23.4월) 그 외 사업 선정

< 국산 NPU 데이터센터 구축사업 개요(안) >

AI 반도체 시험 검증 환경조성 (광주)	AI 반도체 Farm 구축 및 실증
국산 AI 반도체 ➡ 데이터센터 적용	국산 AI 반도체 ➡ 데이터센터 적용
광주 AI 집적단지에 구축 공공분야 중심 AI서비스 실증	민간 데이터센터에 구축 민간분야 중심 AI서비스 실증
AI 반도체 시험·검증 시설	
'23~'24년, 200억원 (국비140억원, 지방비60억원) - '23년 예산 100억원(국비70억원, 지방비30억원) -	'23~'25년, 200억원 (국비 100%) - '23년 예산 70억원 -

○ **파급력·수요**가 높은 **4대 분야**에 서비스 **선도 적용, 주요 분야**로 확대

< K-클라우드 연합체 구성 및 AI-SaaS 모델 선도 적용 분야 (예시) >

분 야	안 전	보 건	교 육	국 방
AI반도체社 NPU	이미지·영상 처리 (객체 탐지·분석)	이미지·영상 처리 (이미지 세분화·분석)	자연어·영상 처리 (요약, 분류, 의도분석)	이미지·영상 처리 (객체 탐지·분석, 행동 추정)
클라우드社 AI-IaaS	A 클라우드社	B 클라우드社	C 클라우드社	D 클라우드社
서비스社 AI-SaaS	• 경비·보안서비스 • 지하철 관제 • 미아찾기 • 재난 예방 등	• 전염병 확산예측 • 암 지도 작성 • 부정맥 위치 추적 • 조산 예측 등	• AI 교육 (맞춤형 학습코스 등) • AI 보이스봇 (통화비서 등)	• 지능·능동형 감시 • 위기상황별 대응전략 수립 • 비상상황 전파 등

○ **안정적**인 클라우드 기반의 AI서비스 제공을 위해 **보안성 강화 추진**

4 산·학·연 협력 강화를 위한 추진체계 마련

○ **(K-클라우드 얼라이언스)** AI반도체·클라우드·AI서비스 기업과 정부·연구기관 등 **민·관 협업 창구 마련** 및 **주요 분야별 과제 발굴**

○ **(PIM-HUB 역할 강화)** 'K-클라우드' 2·3단계의 **PIM 고도화·실증**을 위한 **메모리 반도체 대기업과 산·학·연 간 기술 연계** 등 지원

○ **(R&D 및 지원조직 강화) 관련 분야별**(AI반도체·클라우드·AI 등) **협업을 강화**할 수 있도록 **ETRI와 IITP의 전담 조직 강화**

○ **(AI반도체 대학원) 대학**과 **AI반도체 기업**이 **협력**하여 교육과정 개발, **현장**이 요구하는 **설계역량**을 갖춘 최고급 인재양성('23년~, 3개교 신설)

별첨 **단계별 추진 과제(안)**

※ : 신규로 구축할 사업