

보도 시점 2025. 12. 10.(수) 10:00
(2025. 12. 10.(수) 석간) 배포 2025. 12. 9.(화) 16:00

수요-공급기업이 함께 마련한 국산 신경망 처리 장치(NPU) 성능지표, ‘K-Perf’ 발표

- 15개 수요-공급기관이 참여하는 ‘공동성능지표(K-Perf)’ 협의체 출범, 공동선언문 채택
- 국산 인공지능 반도체 혁신기술, 조기 상용화 성과에 대한 공유의 장 마련
- 1·2차 추경으로 국산 신경망 처리 장치(NPU) 관련 16개 기업, 26종 제품 개발·고도화
- 한국형-인공지능 반도체 미래 도약을 위한 ‘2025 인공지능 반도체 미래 기술 콘퍼런스’ 개최

【관련 국정과제】 22. 초격차 인공지능 선도기술·인재 확보

과학기술정보통신부(부총리 겸 과기정통부 장관 배경훈, 이하 ‘과기정통부’)는 12월 10일(수) 오전 9시 45분 서울 소공동 롯데호텔에서 국산 인공지능 반도체 조기 상용화 성과 및 기술 동향 등을 공유하는 ‘2025 인공지능 반도체 미래 기술 학술회의(콘퍼런스)’를 개최하였다.

이날 ‘인공지능 반도체 미래 기술 학술회의(콘퍼런스)’에서는 산업계 중심의 공동성능지표(K-Perf) 발굴·확산·고도화를 위해 주요 수요-공급기업 12개사와 3개 관계기관이 참여하는 ‘공동성능지표(K-Perf) 협의체’ 출범식도 함께 개최되었다. 과기정통부는 이날 출범식에서 그간 산업계 합의를 바탕으로 마련한 ‘공동성능지표(K-Perf)’를 발표하였다. ‘공동성능지표(K-Perf)’는 수요기업이 제시한 다양한 서비스 환경을 고려하여 측정 모델 및 조건·지표를 세분화한 것이 특징이며, 이를 기반으로 실제 사용 환경 중심의 신경망 처리 장치(NPU) 성능 데이터 공유·활용 체계 구축이 가능해질 것으로 기대된다.

< 공동성능지표(K-Perf) 협의체 참여기업·기관 >

구 분	참여 기관
[공급처] 설계 전문회사(팹리스)	■ 퓨리오사AI, 리벨레온, 하이퍼엑셀
[수요처] 클라우드 서비스 제공업체(CSP), 인공지능 개발사 등	■ (클라우드 서비스 제공<CSP>·시스템 통합<SI> 기업) 네이버클라우드, KT클라우드, 카카오엔터프라이즈, NHN클라우드, SKT, LG CNS, 삼성SDS ■ (인공지능·소프트웨어 기업) LG AI연구원, 모레
[유관기관] 시험인증기관, 연구개발 및 실증사업화지원기관	■ 한국정보통신기술협회(TTA), 정보통신기획평가원(IITP), 정보통신산업진흥원(NIPA)

<주요 측정 조건 및 측정 지표(안)>

측정 모델	측정 조건	측정 지표
글로벌 AI모델 + 국내 독자 AI모델 “협업체 기반 지속 최신화” ▪ (소형) 라마 3.1 8B ▪ (중형) 엑사원 3.5 32B, 라마 3.3 70B ▪ (대형) 라마 3.1 405B	“실제 사용환경 중심 조건 설정” ① 입출력 조건 세분화 ② AI 데이터 맞춤형 연산 활용 ③ 동시 처리 조건 다양화	성능 ▪ 출력 단계별 속도, 안전성, 확장성 효율성 ▪ PCIe 카드 전력 소모량, 시간당 처리량 등

출범식에서 ‘공동성능지표(K-Perf)’ 협의체는 ‘공동성능지표(K-Perf)’의 정착·고도화와 민간·공공 분야 현장 전반에 걸친 확산을 위해 공동선언문을 채택하였다. 선언문을 통해 수요·공급기업들은 국산 신경망 처리 장치(NPU) 성능 고도화 및 현장 중심 참고 사례(레퍼런스) 확보 등을 위해 공동성능지표(K-Perf) 지표를 적극 활용하기로 합의하였다. 이에 따라, ‘공동성능지표(K-Perf)’를 통해 그간 수요기업이 국산 신경망 처리 장치(NPU) 도입을 위해 필요로 하였던 구체적인 데이터를 수월하게 확보할 수 있게 되어 국산 신경망 처리 장치(NPU) 수요-공급기업 간 협력이 확대될 것으로 기대된다.

이와 함께, 인공지능 반도체 연구 개발 및 실증·사업화 성과를 공유하기 위한 오찬 간담회와 성과전시관(성과전시 부스)도 열렸다. 과기정통부는 올해 연구 개발에 1,425억 원을 투자하여 국산 인공지능 반도체 기술 고도화를 집중적으로 지원하였으며, 국산 신경망 처리 장치(NPU)가 황금시간(적기) 내 상용화될 수 있도록 1·2차 추경을 통해 국산 인공지능 반도체 실증·사업화 관련 예산을 대폭 증액(794억 원) 하여 1,103억 원을 투자하는 등 총 2,500억 원 이상을 지원하였다.

특히, 과기정통부는 올해 추경을 통해 확대한 실증·사업화 사업을 통해 16개 기업의 조기 상용화를 지원하였으며, 국산 인공지능 반도체 관련 제품 26종을 개발·고도화하는 성과를 창출하였다. 기업들은 오찬 간담회에서 이번 추경 사업을 통해 적시에 국산 신경망 처리장치(NPU) 사업화에 실질적으로 필요한 지원을 받을 수 있었으며 상용화 기간 단축, 해외 진출 등에 큰 도움이 되었음을 강조하였다.

성과전시관(성과전시 부스)에서는 특히 2PF 급 차세대 국산 신경망 처리 장치(NPU) 개발 성과, 독자 인공지능 기초 모형(AI 파운데이션 모델)과 국산 신경망 처리장치(NPU) 간 상용화 수준의 최적화 사례 등 우리 인공지능 반도체 기술력의 우수성을 입증하는 다양한 사례들도 선보였다.

한편, 이날 개막식에서는 올 한해 인공지능 반도체 기술 및 산업 고도화에 우수한 성과를 창출한 인공지능 반도체 산업 발전 유공자에 대한 부총리 표창 등에 대한 수여가 함께 진행되었다.

이번 표창은 대형 거대 언어 모델(LLM) 지원에 최적화된 국산 신경망 처리 장치(NPU) 개발에 기여한 오진욱 리벨리온 최고기술책임자(CTO)와 국산 인공지능 반도체의 국제협력 확산에 기여한 김한준 퓨리오사AI 최고기술책임자(CTO), 서웅 딥엑스 이사 등 총 6명에게 수여되었으며, 부총리 상인 인공지능 반도체 경진대회 대상은 대회에서 가장 우수한 성적을 거둔 자충수팀(서울 과기대)의 대표로 조승우 학생이 수상하였다.

개막식 이후에는 인공지능 반도체 국내·외 기술 동향 및 산업계 동향 공유를 위한 발표 분과(세션)도 마련되었으며, ①수요기업의 인공지능 반도체 응용/활용 사례, ②초청 강연, ③차세대 인공지능 반도체 설계 고도화 사례, ④인공지능 반도체 생태계 확산 등 4개 분과(세션)에서 최고 전문가의 발표가 진행되었다. 특히, 초청공연에는 엔비디아와 SK하이닉스가 인공지능 반도체와 메모리 관점에서 인공지능 시대 반도체 시장의 미래를 전망하는 발표를 통해 시선을 끌었다.

이날 행사에 참석한 배경훈 부총리 겸 과기정통부 장관은 “국가 인공지능 대전환이 본격화되는 상황에서 인공지능 반도체는 이를 뒷받침하는 핵심 기반 시설(인프라)”이라며, “전 국민이 비용 부담 없이 인공지능 서비스를 활용할 수 있도록 저전력 국산 인공지능 반도체의 고도화를 집중적으로 지원하고, 인공지능 반도체 설계회사(팹리스)가 크게 도약할 수 있도록 적극 지원할 것” 이라고 밝혔다.

담당 부서	정보통신정책실 정보통신 방송 기술정책과	책임자	과장	장두원 (044-202-6230)
		담당자	사무관	김휘태 (044-202-6232)

내일을 만드는 과학기술
내일을 채우는 디지털·AI

대한민국
정책브리핑



시간	발표 주제	발표자(안)
9:30~09:45 ('15)	참가자 등록	
개 회 식		
09:45~10:40 ('55)	국민 의례 및 내빈 소개(5')	
	개 회 사(5')	부총리
	K-Perf 협의체 출범식(25')	부총리 및 협의체 구성원
	시 상 식(20') (AI반도체 산업발전유공(부총리 표창) 등)	-
10:40~11:10 ('30)	휴식 / 성과 전시 관람	
Session 1 : AI반도체 응용/활용 사례		
11:10~11:40 ('30)	AX 성공의 열쇠:EXAONE AI Stack 전략 사례	LG AI연구원 (전기정 부문장)
11:40~12:10 ('30)	피지컬 AI 시대의 핵심, 온디바이스 AI 반도체 시장을 선점하라	마음AI (최홍섭 대표)
12:10~12:40 ('30)	AI반도체 스케일아웃 및 클라우드 최적화 개발	SK텔레콤 (김영진 리더)
12:40~14:00 ('80)	중식 / 성과 전시 관람	
Session 2 : 초청강연		
14:00~14:40 ('40)	글로벌 AI반도체 제조 및 응용	엔비디아코리아 (정소영 대표)
14:40~15:20 ('40)	AI시대 메모리와 컴퓨팅	SK하이닉스 (김천성 부사장)
15:20~15:40 ('20)	휴식 / 성과 전시 관람	
Session 3 : 차세대 AI반도체 설계 고도화		
15:40~16:00 ('20)	효율적으로 구동되는 프런티어 LLM - AI 인프라의 새로운 시대	리벨리온 (오진욱 CTO)
16:00~16:20 ('20)	지속가능한 AI 인프라와 LLM 서비스를 위한 LPU 반도체	하이퍼엑셀 (이진원 CTO)
16:20~16:40 ('20)	온디바이스 반도체 시대로의 도약	딥엑스 (김정욱 부사장)
16:40~17:00 ('20)	From Chip to Market : 레니게이드로 실현하는 AI Inference 효율화	퓨리오사AI (김한준 CTO)
17:00~17:10 ('10)	휴식 / 성과 전시 관람	
Session 4 : AI반도체 생태계 확산		
17:10~17:30 ('20)	'26년도 AI반도체 R&D사업 추진 계획	IITP (강호석 팀장)
17:30~17:50 ('20)	AI반도체 실증 등 사업화 지원 성과 및 계획	NIPA (조재홍 팀장)
17:50~	폐 회	

□ 공동 성능평가지표(K-Perf)(안)

○ 측정 대상 모델 및 크기(안)

- (소형) Llama 3.1 8B, (중형) Exaone 4.0 32B 및 Llama 3.3 70B, (대형) Llama 3.1 405B

※ 독자 AI모델 등 향후 개발 동향에 따라 신규 모델 추가·변경 추진

○ 입력 변수 범위(안)

- (입·출력 크기) 대표 측정 시나리오별 입·출력 토큰 수 차이 및 글로벌 빅테크 요청 사항 등을 종합하여 입·출력 크기 설정*

* Input Sequence Length : 1,024(1K), 2,048(2K), 4,096(4K), 32,768(32K)

Output Sequence Length : 1,024(1K), 4,096(4K)

※ 테스트 대상 모델이 Sequence Length 미지원 시 Output Sequence Length에서 감산 (테스트 모델의 지원 범위 내에서 ISL, OSL을 조정, 예: 1k+31k, 16k+16k)

< 대표 측정 시나리오(예시) >

구분	Input	Output
① 생성형(Reasoning)	적음(1K~4K)	많음(4K~32K)
② 번역형(Translation)	입·출력 동일(1~4K)	입·출력 동일(1~4K)
③ 요약형(Summarization)	많음(4~32K)	적음(1~4K)

- (동시 사용 환경) 다양한 요청에 대한 일괄 처리 크기인 배치(Batch) 크기를 다양화*하여 다수 사용자의 동시 요청 처리 성능 확인

* Number of User(Batch) : 1, 2, 4, 8, 16, 32, 64, 128

- (정밀도) AI모델의 일반적인 데이터 특성 등을 고려한 최적 연산 정밀도 BF16을 기준으로 측정*하되, FP8 등 추가 정밀도 제시 가능

* weight/activation/key/value bit수 제시

○ 측정 지표(안)

- Prefill(첫 생성) 단계와 Decode(지속 생성) 단계별 성능 확인을 위한 대표 지표와 구동 시 소모전력·시간당 처리량 등 공통 지표 측정

측정단계	측정치표	활용도
Prefill	TTFT(Time to First Token)	사용자 응답 지연 최소화 여부 확인
Decode	TPOT(Time per Output Token)	긴 답변을 생성시 속도 안정성 확인
공통	TPS/1 user(Throughput per Second/1 user)	단일 사용자 시간당 처리량 확인
	PCIe 카드 TDP max·평균 전력소모	NPU 카드 전력효율성 확인

산업계 중심의 국산 NPU 공동성능지표 발굴 및
민간·공공분야 적용·확산을 위한
‘K-Perf’ 협력 공동 선언문

「K-Perf 협의체」에 참여하는 국내 NPU 공급사, 클라우드·인프라 기업 및 AI·SW 기업, 시험인증기관 등 유관기관은 산업계 합의를 통해 마련한 국산 NPU 성능지표 ‘K-Perf’의 정착·확산·고도화와 우리나라 AI반도체 산업 경쟁력 확보를 위하여 상호 신뢰와 협력을 바탕으로 다음과 같이 공동으로 선언한다.

1. 우리는 K-Perf를 활용한 국산 NPU 실증 및 검증을 활성화하고, 데이터에 기반한 피드백체계 구축 등 수요-공급 기업 간 상호 협력을 강화하여 실효성 있는 NPU 고도화를 달성한다.
1. 우리는 국산 NPU가 현장 중심의 레퍼런스를 확보할 수 있도록 K-Perf를 기업 활동에 적극 활용하여 K-Perf의 민간 및 공공 분야 확산에 기여한다.
1. 우리는 상호 협력을 바탕으로 K-Perf를 지속적으로 고도화하여 우리나라 산업계에 정착될 수 있도록 노력한다.
1. 우리는 수요-공급 기업 간 상호 신뢰와 호혜 관계를 바탕으로 K-Perf를 통해 확보한 데이터의 공유·활용 체계를 마련하고, 이를 적극 활용한다.

2025년 12월 10일

「K-Perf 협의체」 참여자 일동

붙임4

시반도체 산업발전유공 및 시반도체 경진대회 대상 수여자

□ 시반도체 산업 발전 유공(부총리 표창) 수여자(6인)

소속	이름	추천사유
리벨리온	 오진욱 CTO	○ 대형 LLM 지원에 최적화된 국산 NPU '리벨-쿼드' 개발에 핵심 역할 수행
퓨리오사AI	 김한준 CTO	○ OpenAI, LG AI연구원 등 글로벌 파트너십을 바탕으로 '레니게이드'의 성공적인 LLM 실증 등 상용 데모 성과 창출
송실대	 이찬호 교수	○ 정부 R&D 예타 사업 전문위원으로 활동하며 AI반도체 과제 컨설팅을 통해 R&D 성과 창출에 기여
ETRI	 구본태 본부장	○ AI반도체 대형 국가 R&D(K-클라우드 예타 사업) 기획 및 예타 조사 대응에 주도적 역할 수행
딥엑스	 서웅 이사	○ 해외 현지 수요처(중국 바이두 등)와 공동 개발 등을 통해 국산 NPU의 글로벌 시장 진출에 기여
NIPA	 조재홍 팀장	○ 주요 산업계 대표들과 지속적인 소통 등을 통해 정책 수요 발굴 및 실증·사업화 사업 기획에 기여

□ 시반도체 경진대회 대상(부총리 상) 수상팀

팀 명	대표 수상자	프로젝트 내용
자충수 (서울과학기술대학교)	 조승우	· (목표) 새로운 데이터에도 높은 성능을 내는 LeNet-5 모델과 이를 고속 추론하는 가속기 설계·구현 · (성과) 이전 레이어의 출력을 기다리지 않고 실시간 추론 가속기 구현 및 모델 parameter 전송 및 갱신 가능

순번	구분	전시 내용 (과제명)	운영기관
1	R&D	◆ 초저전력 SRAM 기반 PIM 연구성과	서울대학교
2		◆ 차세대 eMRAM 기반 PIM 기술개발 성과	KAIST
3		◆ RRAM 기반 저전력 뉴로모픽 프로세서 개발 성과	한양대학교
4		◆ 2,000 TFLOPS급 차세대 NPU 개발 성과	리벨리온
5		◆ 피지컬 AI 시대를 위한 온디바이스 AI 반도체	딥엑스
6		◆ 상황예측형 모바일 인공지능 프로세서 개발	텔레칩스
7	실증, 사업화	◆ Edge SoC 성능 극대화를 위한 AI모델 통합 최적화 설계 성과	아이닉스
8		◆ 국산 NPU의 복합 자율주행 AI 모델 최적화 설계	보스반도체
9		◆ 국산 NPU의 EXAONE 최적화 성과	퓨리오사AI
10		◆ LPDDR 기반 LLM 특화 AI반도체 및 서버 시스템 개발	하이퍼엑셀
11		◆ 온디바이스AI향 저전력 고성능 AI반도체 설계검증 시제품	모빌린트
12		◆ 국산 AI반도체 기반 AI 양돈 관리 솔루션 'EdgeFarm' 실증성과	인트플로우
13		◆ 국산 AI반도체의 상용화 검증 및 성능 평가 지원	AICA, TTA
14		◆ AI 반도체 기반 보행환경 플랫폼의 글로벌 시장 진출	엘비에스테크
15		◆ 국산 AI반도체를 활용한 해양 청소로봇(드론) 장비 개발	코아이
16		◆ 국산 AI반도체를 활용한 AI기반 치안 스마트 헬멧 개발·실증	아날로그 플러스
17	공공조달	◆ NPU 특화 품목 신설에 따른 나라장터 제품 등록 성과 전시	디퍼아이
18	경진대회	◆ AI반도체 경진대회 수상작	대상·최우수상

< 성과전시 부스 배치도 >

