

2024년도 대한전자공학회
하계종합학술대회 특별세션 강연 발표정보

■ 발표자 정보

성명	최규현	사진	
소속(학교)	한국전자기술연구원		
부서(학과)	SoC플랫폼연구센터		
직위	책임연구원		
발표분야	KETI세션		
약력	* 학력, 경력, 연구실적, 주요연구 및 관심분야를 기술해주세요. 학력 • 2012.02 : 고려대학교 전기전자전파공학부 학사 • 2018.02 : 고려대학교 전기전자전파공학과 박사 경력 • 2018.01 ~ 2021.10 : SK하이닉스 메모리시스템연구 책임연구원 • 2021.10 ~ 현재 : 한국전자기술연구원 SoC플랫폼연구센터 책임연구원 주요연구 및 관심분야 • AI반도체, PIM (Processing in memory), 거대언어모델 등		

■ 강연 정보

제 목	온디바이스 LLM 구동을 위한 PIM 기술 동향 및 연구 방향
Abstract	최근 언어모델의 경량화 기술에 대해 지속적으로 연구되고 있으며, 특히 온디바이스 환경에서 언어모델을 구동하기 위한 모델 구조 및 가속기 연구가 지속적으로 이루어지고 있다. 전력 소모, 컴퓨팅 파워, 메모리 용량이 제한된 온디바이스 환경에서는 기존 서버 환경 대비 제한적인 리소스에서 언어 모델을 추론하기 위한 다양한 아키텍처가 연구되고 있으나, 그 중 저지연으로 언어모델 추론을 수행하기 위해 메모리 내에서 연산을 수행하는 PIM(Processing in memory) 기술의 중요성이 증가하고 있다. 현재 주요 언어모델 개발 동향 및 PIM 기술의 최신 동향을 알아보고, 온디바이스 환경에서 언어 모델을 구동하기 위한 PIM 플랫폼 구성 방법을 소개한다.