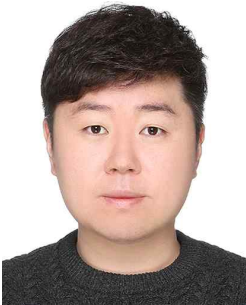


[Computer and AI Innovations]

<p>초청연사 1</p>	
	<p>유인재 교수 (부산대 전기전자공학부)</p>
<p>제 목</p>	<p>Huffman 코딩을 이용한 인공지능 신경망 데이터 압축 방법과 초고속 해제 하드웨어 아키텍처</p>
<p>요약문</p>	<p>생성형 인공지능 모델의 발달로 인해 최신 신경망 매개변수의 크기가 매년 10배씩 기하급수적으로 증가하고 있습니다. 또한, 미세 공정 기술의 발전으로 시스템 반도체의 집적도와 연산 능력이 비약적으로 향상되면서, 시스템 반도체와 메모리 반도체 간 데이터 이동이 전체 시스템 성능의 주요 병목으로 작용하고 있습니다. 이에 따라, 최신 인공지능 신경망 매개변수를 압축하여 DRAM에 저장함으로써 메모리와 인공지능 신경망 가속기 간 데이터 이동을 줄이는 방법에 대한 연구가 활발히 진행되고 있습니다.</p> <p>하지만 이러한 접근 방식은 인공지능 신경망 가속기 내에 압축된 데이터를 해제할 수 있는 회로를 필요로 하여 다음과 같은 어려움이 있습니다. 첫째, 추가적인 회로의 면적과 전력 소모를 상쇄하기 위해서, 다양한 신경망 매개변수뿐만 아니라 인공지능 활성화 데이터까지도 효과적으로 압축할 수 있는 유연한 압축 알고리즘 및 하드웨어 아키텍처가 필요합니다. 둘째, DRAM에서 읽어오는 데이터를 실시간으로 해제하기 위해서는 DRAM 대역폭과 대등한 속도를 갖는 초고속 압축 해제 하드웨어 아키텍처가 필요합니다. 본 발표에서는 이러한 문제들을 해결하기 위해 개발된 Huffman 코딩 기반의 인공지능 신경망 매개변수 및 활성화 데이터의 무손실 압축 기법과 초고속 압축 해제 하드웨어 아키텍처를 소개합니다.</p>
<p>초청연사 2</p>	
	<p>윤명국 교수 (이화여대 인공지능대학)</p>
<p>제 목</p>	<p>Energy-Efficient Register File Architecture on GPUs</p>
<p>요약문</p>	<p>Graphics processing units (GPUs) are among the most popular computing devices for AI applications, including inferencing and training. In high-end GPUs, the number of streaming multiprocessors tends to increase to</p>

	<p>improve performance through higher thread-level parallelism (TLP). As TLP increases, more registers are required to store the context of concurrently scheduled threads. Consequently, register files have become one of the most power-hungry components in GPUs, consuming over 20% of the entire GPU chip's power. This presentation introduces the register file architecture in recent GPUs and explores several schemes to address the high power consumption of register files.</p>
<p>초청연사 3</p>	<div style="text-align: center;">  <p>이어진 교수 (인하대학교 컴퓨터공학과)</p> </div>
<p>제 목</p>	<p>Architecting Compatible PIM Protocol for CPU-PIM Collaboration</p>
<p>요약문</p>	<p>프로세싱-인-메모리(PIM) 기술은 메모리 병목 워크로드를 가속할 수 있는 효율적인 아키텍처로, 실제 프로토타입 제품이 출시될만큼 주목받고 있습니다. 그러나 기존PIM 디바이스의 프로토콜은 PIM 작업을 하는 동안 정상적인 메모리 요청을 장시간 지연시켜 CPU 성능을 과도하게 저해할 수 있습니다. 또한, 각 디바이스에서 사용하는 프로토콜이 서로 달라CPU와 같은 호스트 프로세서는 특정 디바이스를 위한 전용 프로토콜을 지원하는 것이 부담될 수 있습니다. 본 연구에서는 명령 공간이 제한된DRAM 표준에 초점을 맞춰, 다양한PIM 디바이스와 호스트 프로세서 간 호환성을 위해 설계된 새로운PIM 명령 및 프로토콜을 제안합니다. 여러 PIM 디바이스 동작의 공통적인 특징을 추출하여, 여러 디바이스에서 공통적으로 사용하여PIM을 동작시킬 수 있는 최소한의 명령어를 제안합니다. 또한 제안된 프로토콜 하에서 일반 메모리 요청과PIM 메모리 요청 처리량 간의 균형을 맞추는 메모리 스케줄링 정책도 함께 제시합니다.</p>