최신 인공지능 반도체 및 컴파일러 지원 기술 동향

Trends in Supporting Technologies for Advanced AI Semiconductors and Compilers

김용주 (Y. Kim, y.kim@etri.re.kr) 지능디바이스·시뮬레이션연구실 책임연구원
하영목 (Y. Ha, ymha@etri.re.kr) 지능디바이스·시뮬레이션연구실 선임연구원
정영준 (J.Y. Joon, jjing@etri.re.kr) 지능디바이스·시뮬레이션연구실 책임연구원/실장

ABSTRACT

Recent advancements in artificial intelligence (AI) across diverse sectors have been widely supported by developments in AI semiconductors, with NVIDIA graphics processing units leading the market. However, concerns over market diversity and high energy consumption of AI workloads have prompted the development of next-generation AI semiconductors toward improving performance and energy efficiency. We discuss the latest trends in AI semiconductor and compiler technologies, both domestically and internationally. Key local companies, such as SAPEON, Rebellions, and Furiosa AI, and overseas giants, such as Google, Meta, and Tesla, are innovating in this field. Moreover, compiler technologies, such as MLIR, TVM, and XLA, are crucial for optimizing the performance of AI solutions across hardware platforms. Such developments are essential for enhancing AI applications and demand active research. This study offers insights into the current and future landscape of AI semiconductors and compilers, and it provides a foundation for future technological strategies in the AI industry.

KEYWORDS AI, 가속기, 반도체, 컴파일러

Ⅰ. 서론

최근 몇 년간, 인공지능(AI: Artificial Intelligence) 기술은 각 산업 분야에서 획기적인 발전을 이루며 새로운 혁신의 물결을 이끌고 있다. 자동차, 금융, 제조, 헬스케어 등의 산업 분야에서부터 미술, 음악, 문학 등의 예술 분야에 이르기까지, AI는 비즈니스 모델을 재정의하고 운영 효율성을 높이며, 사용자 경험을 개선하는 중추적인 역할을 담당하고 있다. 이러한 변화는 AI 기술의 성능 향상과 더불어 다양한 산업에서 그 적용 가능성을 극대화하는 촉매제 역할을 하고 있다.

AI 기술의 발전과 함께 이를 지원하는 핵심 요소인 AI 반도체 시장도 주목받고 있다. 특히, NVIDIA

* DOI: https://doi.org/10.22648/ETRI.2024.J.390501
* 본 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. RS-2023-00228255, 거대인공신경망 처리 PIM-NPU 지원 시스템 SW 기술개발), (No. RS-2023-00277060, 개방형 엣지 AI 반도체 설계 및 SW 플랫폼 기술개발)과 한국전자통신연구원(24ZS1230, 메모리-연산 융합 뉴로모픽 컴퓨팅)의 지원을 받아 수행됨.

의 그래픽 처리 장치(GPU: Graphics Processing Unit)는 AI 모델의 학습과 추론을 위한 처리 능력을 제공함에 있어 타 업체보다 앞선 시장 선점과 소프트웨어 지원 능력을 앞세워 시장을 장악하고 있는 실정이다. 최근 시장 조사에 따르면 2024년 현재 NVIDIA는 AI를 위한 하드웨어 시장 점유율이 92%에 달하고 있다[1]. 이러한 NVIDIA의 인공지능 반도체 시장 독점은 시장 내 다양성 부족과 기술 집중도의 측면에서 잠재적인 문제점으로 지적되고 있다. 게다가 AI는 모델 학습 및 추론 작업에 많은 전력을 소모하고 있어 에너지 부족 문제 및 탄소 배출 증가라는 중요한 환경 관련 문제를 동반한다[2]. NVIDIA의 GPU는 고성능을 제공하지만, 그에 따른 에너지 효율성은 상대적으로 낮은 편이다. 이는 지속 가능한 기술 발전을 추구하는 현대 사회에서 큰 도전 과제로 부상하고 있으며, 더욱 효율적인 솔루션에 대한 요구가 증가하고 있는 실정이다.

AI의 최적 활용과 에너지 소비 절감을 위해 다양한 인공지능 반도체 업체들이 차세대 AI 반도체 개발에 노력을 기울이고 있다. 이 기업들은 전력 소비를 줄이면서도 성능을 극대화할 수 있는 새로운 반도체 아키텍처를 개발하여 NVIDIA의 아성에 도전하고 있다. 또한, 이러한 반도체의 효율적 사용을 가능하게 하는 컴파일러 기술이 개발되고 있어, 하드웨어와 소프트웨어의 조화를 통한 최적의 성능 달성을 지원하고 있다.

본 동향 분석 자료에서는 국내외 AI 반도체 시장의 최신 개발 동향을 분석하고, 인공지능 반도체의 최적 수행을 지원하기 위해 개발되고 있는 컴파일러 기술의 현재 개발 동향을 분석한다. 이를 통해 기술 개발자, 연구자, 그리고 산업계 이해관계자들에게 유용한 인사이트를 제공하고 향후 기술 전략 수립에 기여할 수 있는 정보를 제공하고자 한다.

II. AI 반도체 개발 동향

1. 국내 AI 반도체 개발 동향

가. 사피온

사피온은 2016년 SK텔레콤에서 사내 AI칩 R&D 팀에서 시작하여 2021년 분사하여 설립한 SK 그룹의 자회사이다. 2017년에 처음으로 프로토타입인 X110을 발표하였고, 2020년에는 데이터 센터용으로 특화한 X220을 출시하였다. 가장 최근 발표한 AI 반도체는 2023년 11월에 공개한 X330이다.

X330[3]은 데이터 센터 및 서버 환경에서 AI 추론 작업을 위한 고효율과 강력한 성능을 제공하는 제품이다. 이 칩은 7nm 공정으로 제작되었으며 이전 제품들에 비해 최대 4배 향상된 컴퓨팅 성능을 제공한다. X330은 낮은 전력 소모를 유지하면서도 피크 전력 사용량 증가는 최소화하도록 설계되었다. X330은 정수 및 부동소수점 연산을 지원하고, 다양한 8비트와 16비트 포맷을 포함하여 FP8 포맷을 지원하여 AI 추론을 보다 효율적으로 수행할 수 있다. 또한, 4K 60fps의 멀티 스트림 비디오 코덱을 갖추고 있어 멀티모달 AI를 효율적으로 지원한다. 제품은 PCIe 카드 형태로 제공되며, X330 Compact와 X330 Prime 두 가지 버전이 있다. Compact 버전은 PCIe FHHL(Full Height Half Length) 표준을, Prime 버전은 PCIe FHFL(Full Height Full Length) 표준을 사용한다. 두 카드 모두 PCIe Gen 5 인터페이스를 통해 호스트 CPU(Central Processing Unit)에 연결된다. X330의 아키텍처는 64K MAC 매트릭스 블록과 16개의 신경 벡터 프로세서를 통합하고, 16개의 RISC-V(Reduced Instruction Set Computer-V) 기반 CPU와 고성능 비디오 코덱 클러스터를 포함하여 데이터 센터에서 AI 추론 연산을 지원한다.

사피온은 차세대 AI 반도체로 X430을 설계 중으로 해당 제품은 HBM(High Bandwidth Memory)과 칩

렛, CXL(Compute Express Link)을 적용할 계획이다. X430에는 HBM3 이상의 최신 버전이 탑재될 예정이며, 칩렛 구조를 도입하여 여러 개의 다이를 연결해 하나의 반도체로 만드는 방식을 채택하여 생산 비용 절감과 수율 상승을 목표로 한다. 생산 공정은 TSMC 5nm 이상을 사용할 계획이다. X430은 2025년 말에서 2026년 초에 출시될 예정이다.

표 1 국내외 주요 AI반도체 비교

	사피온 [3]	리벨리온 [4]	퓨리오사 AI [5]	SK Hynix [6,7]	Google [8]	Meta [9]	Graph core [10]	Tesla [11,12]	AWS [13]
제조국가	대한민국	대한민국	대한민국	대한민국	미국	미국	영국	미국	미국
제품명	X330	ATOM	Warboy	AiM	TPU v5p	MTIA v2	Colossus GC200	D1	Inferentia2
발표시기	2023.11	2023.2	2021.9	2022.3	2023.12	2024.4	2021.8	2024	2023.4
사용공정	TSMC 7nm	삼성 5nm	삼성 14nm	SK Hynix 1y	–	TSMC 5nm	TSMC 7nm	TSMC 7nm	–
Precision	INT8, FP8, FP16	INT4, INT8, FP16	INT8	BF16	–	INT8, FP16, BF16, FP32	FP16, FP32	FP8, FP16, FP32	INT8, INT16, INT32, FP8, FP16, BF16, TF32, FP32
성능 8-bit TOPS	734	128	64	–	918	–	–	–	380
성능 16-bit TOPS	–	–	–	–	–	354	–	–	–
성능 32-bit TFLOPS	–	–	–	–	–	–	–	22.6	190
성능 16-bit TFLOPS	368	32	–	1	459	177	250	362	190
성능 8-bit TFLOPS	734	–	–	–	–	–	62	362	47.5
메모리 종류	GDDR6	GDDR6	LPDDR4X	GDDR6	HBM	LPDDR5	SRAM	SRAM	HBM
메모리 사이즈	32GB	16GB	16GB	1GB	95GB	128GB	900MB	440MB	32GB
메모리 Bandwidth	512 GB/s	256 GB/s	66 GB/s	512 GB/s	2,765 GB/s	204.8 GB/s	47.5 TB/s	512 GB/s	820 GB/s
Host interface	PCIe Gen5 x16	PCIe Gen5 x16	PCIe Gen4 x8	PCIe Gen3 x8x8 (bifurcated)	–	PCIe Gen5 x8	PCIe Gen4 x16	–	–
TDP	250W	130W	60W	–	–	90W	300W	400W	–
주요 적용분야	Data center (inference)	Computer vision, Language processing	Computer vision	Generative AI	LLM, Diffusion model, Generative AI	Generative AI	Computer vision, Generative AI	AI model training	Generative AI (inference)

* –: 미공개 정보

나. 리벨리온

리벨리온은 2020년에 창업한 한국의 AI 반도체 기업으로 금융 특화 AI 반도체 전문 회사로 시작했으나 최근에는 자율주행, IoT, 스마트 시티 등 다양한 분야에 사용되는 AI 반도체를 설계하고 제조한다. 리벨리온은 'ION'과 'ATOM' 두 가지 AI 반도체를 개발했다. ION은 금융기술(핀테크)에 특화됐고, ATOM은 클라우드용으로 개발되어 최신 AI 모델을 가속하기 위해 설계된 고성능 AI칩이다.

리벨리온의 ATOM[4] 칩은 최신 5nm 공정으로 제조된 다목적 추론칩이다. 이 칩은 엣지와 클라우드 컴퓨팅 환경에서 최고의 추론 성능을 제공하도록 설계되었다. ATOM은 머신러닝 전문 데이터플로우 아키텍처와 다중 코어 SoC 아키텍처를 결합하여 사용자 수준 병렬화를 통한 최적의 추론 성능을 구현한다. ATOM은 삼성전자의 최첨단 EUV(Extreme UltraViolet) 공정을 통해 제작되었으며, PCIe Gen5와 GDDR6 고속 I/O 기술을 탑재하여 엣지 컴퓨팅부터 데이터센터까지 다양한 시장에 대응할 수 있다. 주요 사양으로는 FP16에서 32TFLOPS(Tera Floating Point Operations Per Second), INT8에서 128TOPS(Tera Operations Per Second)의 성능을 제공하며, 16GB의 GDDR6 메모리와 256GB/s의 대역폭을 갖추고 있다. TDP(Thermal Design Power)는 30W에서 130W까지 조절이 가능하다. ATOM의 ION 코어는 최대 16개의 독립 작업을 동시에 처리할 수 있는 멀티 인스턴스 NPU(Neural Processing Unit)를 지원한다. 단일 슬롯 카드로 최대 128TOPS, 듀얼 슬롯 카드로 최대 256TOPS의 성능을 제공하며, 각각 256GB/s와 512GB/s의 외부 대역폭을 지원한다.

리벨리온의 차세대 AI 반도체 칩인 리벨(Rebel)은 삼성전자와 협력하여 개발되고 있다. 리벨은 LLM(Large Language Model)을 가속하기 위한 제품으로, 삼성전자의 4나노 공정을 이용해 생산되며, HBM3E 고대역폭 메모리가 탑재될 예정이다. 이 칩은 2024년 출시를 목표로 하고 있으며, NVIDIA의 H100이 장악하고 있는 250W급 AI 반도체 시장을 겨냥하고 있다. 리벨 칩은 고성능 AI 반도체로서 H100 이상의 메모리 대역폭을 확보하고, 빠른 데이터 처리를 위한 내부 대역폭을 극대화할 예정이다.

최근에는 사피온과 리벨리온이 글로벌 AI 반도체 시장에서 경쟁력을 강화하기 위하여 합병을 발표했다. 이번 합병은 두 회사의 기술적 시너지를 통해 한국의 AI 반도체 사업에 큰 영향을 줄 수 있을 것으로 보인다.

다. 퓨리오사AI

퓨리오사AI는 2017년에 설립된 한국의 AI 반도체 스타트업으로, 고성능 AI 추론 가속기를 개발하고 있다. 퓨리오사AI에서 발표한 가장 최근 제품은 Warboy[5]이다. Warboy는 고성능 AI 가속기로, 데이터 센터 및 엣지 서버에서 컴퓨터 비전 처리를 위해 설계되었다. 이 제품은 삼성의 14nm 공정을 사용하여 제조되었다. 메모리는 16GB의 LPDDR4X와 32MB의 온칩 SRAM을 갖추고 있으며, PCIe Gen4 x8 인터페이스를 통해 10GB/s의 데이터 전송률을 제공한다. 또한, INT8 성능은 64TOPS에 달하며, TDP는 40~60W로 설정 가능하다. 추가적으로 가상화 소프트웨어 및 ECC 메모리 지원도 포함되어 있다.

퓨리오사AI의 차세대 제품인 RNGD(Renegade)는 2024년 출시 예정으로 고성능 LLM 및 멀티모달 배포를 위해 설계된 데이터 센터용 AI 가속기이다. 이 제품은 TSMC의 5nm 공정을 통해 제조되며, FP8에서 512TFLOPS, BF16에서 256TFLOPS, INT8에서 512TOPS, INT4에서 1024TOPS의 성능을 제공한다. 또한 48GB의 HBM3 메모리를 탑재하고 있

으며, 메모리 대역폭은 1.5TB/s에 달한다. RNGD는 PCIe Gen5 x16 인터페이스를 통해 64GB/s의 데이터 전송률을 지원하고, TDP는 150W로 계획되어 있다.

라. SK하이닉스

SK하이닉스의 AiM(Accelerator-in-Memory)[6,7] 아키텍처는 메모리 내에서 연산을 수행하여 머신러닝 모델의 효율적인 실행을 목표로 하는 기술이다. AiM은 DRAM(GDDR6)에 최소한의 연산 장치와 버퍼를 배치해 전체 프로세서 코어의 면적과 전력 오버헤드를 줄이면서도 높은 성능과 낮은 에너지 소비를 구현한다. 이 아키텍처는 DRAM 명령 인터페이스를 통해 호스트 CPU가 연산 명령을 제어할 수 있게 하여 일정한 지연 시간과 효율적인 데이터 처리를 보장한다. AiM은 특히 LLM과 같은 메모리 중심의 머신러닝 모델에 적합하며, 시뮬레이션 결과 평균 10배에서 최대 54배의 속도 향상을 보여준다. 또한, AiM은 GDDR 외에도 DDR, LPDDR, HBM 등 다양한 DRAM 제품군에도 적용 가능하다.

AiMX는 AiM을 기반으로 한 AI 가속기로, LLM의 추론을 효율적으로 처리하는 것을 목표로 한다. AiMX는 높은 대역폭과 낮은 에너지 소비로 메모리 중심의 연산을 가능하게 하여, 특히 메모리 집약적인 작업에서 큰 성능 향상을 제공한다. AiMX는 고정된 메모리 집약적 함수(예: 행렬-벡터 곱셈, GEMV)를 효율적으로 처리하며, 다양한 작은 작업은 AiM 컨트롤 허브를 통해 유연하게 처리할 수 있다. 이는 고성능을 요구하는 LLM 추론 작업에서 비용 효율성과 에너지 효율성을 동시에 제공하는 데 중점을 두고 있다. 또한, AiMX는 PCIe Gen3 인터페이스를 통해 호스트 시스템과 통합되며, 여러 AiM 패키지와 FPGA(Field Programmable Gate Array)를 사용하여 확장 가능한 구조를 갖추고 있다.

2. 해외 AI 반도체 개발 동향

가. Google

Google의 텐서 프로세싱 유닛(TPU: Tensor Processing Unit)[8]은 머신러닝 작업을 가속화하기 위해 설계된 맞춤형 ASIC(Application-Specific Integrated Circuit)이다. TPU는 특히 구글의 딥러닝 프레임워크인 텐서플로우(TensorFlow)와 최적화되어 있으며, 인공지능 모델의 학습과 추론 속도를 크게 향상시킨다. TPU는 높은 성능과 에너지 효율성을 제공하며, 구글 클라우드 플랫폼을 통해 외부 개발자들도 사용할 수 있도록 제공되고 있다. TPU의 주요 목표는 대규모 데이터 처리와 복잡한 연산을 빠르고 효율적으로 처리하는 것이다.

Google의 TPU v5p는 고성능 AI 추론과 학습 작업을 위한 최신 텐서 처리 유닛이다. TPU v5p는 단일 팟에서 최대 8,960개의 칩을 결합할 수 있어 대규모 연산 작업에서 탁월한 확장성을 제공한다. BF16 정밀도에서 459TFLOPS의 성능을 제공하여 고속 AI 학습과 추론이 가능하다. 8비트 정수 연산에서는 918TOPS의 성능을 제공한다. 95GB의 HBM를 탑재하여 2,765GB/s의 대역폭을 지원하며 이를 통해 대규모 데이터 처리와 모델 학습을 지원한다. 칩당 인터칩 인터커넥트 대역폭이 4,800GB/s로, 칩 간의 빠른 통신을 지원하여 전체 시스템의 효율성을 높인다.

구글은 최근 여섯 번째 세대 TPU인 트릴리움의 개발 계획을 공개했다. 트릴리움 TPU는 이전 세대인 TPU v5e에 비해 칩당 최고 연산 성능이 4.7배 향상될 것이라고 한다. 또한, 큰 임베딩을 처리하기 위해 설계된 구글의 3세대 SparseCore 기술을 통합하였고, 메모리 용량과 대역폭도 두 배로 증가하여 각 칩이 32GB의 HBM를 1.6TB/s 속도로 운영할 수 있으며, 칩 간 인터커넥트 대역폭도 두 배로 증가하여 더 큰 모델의 학습 및 추론 작업을 위해 여러 TPU를

효율적으로 연결할 수 있다. 구글은 2024년 하반기에 트릴리움 TPU를 공개할 예정이며, 이 TPU를 AI 하이퍼컴퓨터 플랫폼의 일부로 제공할 예정이다.

나. Meta

Meta의 최신 AI 칩인 MTIA(Meta Training and Inference Accelerator)는 AI 워크로드를 최적화하기 위해 설계된 커스텀 칩 시리즈이다. MTIA v1은 2023년에 처음 도입되었고, 개선된 버전인 MTIA v2는 2024년 4월에 공개되었다.

MTIA v2[9]는 TSMC의 5nm 공정을 사용하여 제조되었으며, 이전 세대에 비해 성능이 세 배 향상되었다. MTIA v2는 특히 Meta의 랭킹 및 추천 시스템에서 높은 성능을 제공하도록 설계되었으며, 메모리와 대역폭을 두 배 이상 증가시켜 효율적인 데이터 처리를 가능하게 한다. MTIA v2는 1.35GHz의 클럭 속도와 90W TDP를 갖춘 고성능 AI 추론 가속기로, 8x PCIe Gen5 인터페이스를 통해 32GB/s의 호스트 연결 속도를 제공한다. 메모리 용량은 로컬 메모리 384KB, 온칩 메모리 256MB, 오프칩 LPDDR5 128GB를 갖추고 있으며, 메모리 대역폭은 로컬 메모리 1TB/s, 온칩 메모리 2.7TB/s, 오프칩 LPDDR5 204.8GB/s에 달한다. INT8에서 708TFLOPS, FP16/BF16에서 354TFLOPS의 GEMM 성능을 제공한다.

다. Graphcore

Graphcore는 AI와 머신러닝을 위한 가속기를 개발하는 영국의 반도체 회사이다. 이 회사는 머신러닝 모델 전체를 프로세서 내부에 보유할 수 있는 대규모 병렬 IPU(Intelligence Processing Unit)를 개발했다.

Graphcore는 2017년 7월 첫 번째 칩인 Colossus GC2를 발표하였고, 2020년 7월에는 TSMC의 7nm 공정으로 제작된 2세대 프로세서인 GC200[10]을 발표했다. GC200은 590억 개의 트랜지스터와 823mm^{2}의 집적 회로, 1,472개의 연산 코어 및 900MB의 로컬 메모리를 갖추고 있으며 FP16 기준 250TFLOPS의 성능을 제공한다. 2022년, Graphcore와 TSMC는 GC200 다이를 전력 공급 다이와 페이스 투 페이스로 본딩한 3D 패키지인 Bow IPU를 발표했으며, 이를 통해 낮은 코어 전압에서 더 높은 클럭 속도를 제공한다. Graphcore는 인간 뇌의 시냅스보다 더 많은 파라미터를 가진 AI 모델을 가능하게 하는 Good 머신의 제작을 목표로 하고 있다.

라. Tesla

Tesla의 D1[11,12] 칩은 Tesla에서 추진 중인 Dojo 슈퍼컴퓨터 프로젝트의 핵심 구성 요소이다. 이 칩은 7nm 공정으로 제조되었으며, 500억 개 이상의 트랜지스터를 포함하고, 다이 크기는 645mm^{2}이다. D1 칩은 FP16/CFP8 정밀도에서 최대 362TFLOPS, FP32 작업에서는 약 22.6TFLOPS의 연산 성능을 제공한다. 또한, 칩 내 대역폭은 10TBps, 칩 외 대역폭은 4TBps로 높은 데이터 처리 능력을 제공한다. 아키텍처 측면에서 D1 칩은 64비트 CPU를 포함한 여러 개의 기능 유닛(FU)으로 구성되어 있으며, 각 유닛은 1.25MB의 SRAM을 갖추고 있다. 이 칩은 25개의 D1 칩이 모여 하나의 '트레이닝 타일'을 형성하며, 여러 타일을 연결하여 최대 1.1EFLOPS에 도달할 수 있는 확장 가능한 시스템인 엑사팟을 구성한다. 주로 AI 학습에 사용되며, 자율주행 시스템을 위한 대규모 비디오 데이터 처리를 효율적으로 수행할 수 있도록 설계되었다.

마. AWS

Amazon Web Services(AWS)에서 AI와 머신러닝 모델의 추론 작업을 가속화하기 위하여 AWS Inferentia라는 전용 하드웨어 가속기를 개발하였다. 2019

년에 첫 번째 버전인 Inferentia1을 출시하였으며 2022년에는 Inferentia2를 출시하였다.

Inferentia2[13]는 LLM과 생성 AI 모델을 효율적으로 실행할 수 있도록 최적화되어 있다. 32GB의 HBM을 탑재하고 있으며, 메모리 대역폭은 820GB/s에 이른다. 이 칩은 FP8, FP16, BF16, TF32, FP32, INT8, INT16, INT32 등의 다양한 데이터 타입을 지원하며, 높은 처리량과 낮은 지연 시간을 제공한다. Inferentia2 기반 인스턴스는 최대 12개의 Inferentia2 칩을 장착할 수 있으며, 이는 대규모 모델의 분산 추론을 가능하게 한다.

Ⅲ. AI 반도체 지원 컴파일러 기술 동향

1. MLIR

MLIR(Multi-Level Intermediate Representation)[14]은 소프트웨어와 하드웨어의 다양한 수준에서 컴파일러 개발을 최적화하기 위해 설계된 유연하고 확장 가능한 컴파일러 인프라다. MLIR은 LLVM 프로젝트의 일부로, 다양한 수준의 추상화를 지원하는 IR(Intermediate Representation) 시스템을 제공하며, 특히 이기종 컴퓨팅 환경에서 필요한 다양한 고수준 최적화와 IR 간 변환이 가능하다. 주요 특징으로는 다중 수준 추상화, 모듈러 설계, 확장성, 통합 백엔드 지원 등이 있다. MLIR은 다양한 하드웨어 가속기를 위한 고수준 연산을 지원하고, 사용자가 정의한 Dialect를 통해 특정 도메인이나 언어에 특화된 중간 표현을 설계할 수 있다. 이러한 유연성 덕분에 MLIR은 머신러닝, 고성능 컴퓨팅, GPU 및 TPU와 같은 다양한 하드웨어 아키텍처뿐만 아니라 양자 컴퓨팅 및 고수준 합성 같은 다양한 영역에서도 활용되고 있다. MLIR은 또한 TensorFlow와 같은 머신러닝 프레임워크의 컴파일러 백엔드로 사용되어 고성능 ML 모델을 실행하기 위한 중요한 역할을 하며, 다양한 컴퓨팅 자원을 효율적으로 활용할 수 있는 경로를 제공한다. MLIR의 구조는 고수준의 추상화에서 저수준의 세부 레벨로의 변환을 지원하며, 각 레벨은 자체적인 최적화 및 분석 패스를 가지고 있어 복잡한 최적화와 변환 작업을 효율적으로 수행할 수 있다.

2. IREE

IREE(Intermediate Representation Execution Environment)[15,16]는 Google에서 개발한 오픈소스 프로젝트로, 머신러닝 모델의 최적화와 실행을 위한 컴파일러 및 런타임 환경이다. IREE의 주요 목표는 다양한 하드웨어 플랫폼에서 고성능으로 머신러닝 모델을 실행할 수 있도록 하는 것이다. 이를 위해 IREE는 MLIR 기반의 컴파일러 인프라를 사용하여 모델을 다양한 수준에서 최적화하고 변환한다. IREE는 JIT(Just-In-Time) 컴파일과 AOT(Ahead-Of-Time) 컴파일을 모두 지원하여 유연성과 높은 실행 성능을 제공한다.

IREE는 CPU, GPU, DSP(Digital Signal Processor), TPU 등 다양한 하드웨어 가속기를 지원하며, 특정 하드웨어에 최적화된 코드를 생성할 수 있다. 또한, IREE는 크로스 플랫폼 지원을 통해 데스크톱, 모바일, 임베디드 디바이스 등 다양한 플랫폼에서 일관된 성능 최적화가 가능하다.

IREE는 JAX, ONNX, PyTorch, TensorFlow 및 TensorFlow Lite와 같은 인기 있는 머신러닝 프레임워크를 지원하며, 이러한 프레임워크에서 모델을 내보낸 후 IREE 컴파일러를 사용하여 IR로 변환할 수 있다. 개발자들은 IREE의 다양한 디버깅 및 프로파일링 도구를 사용하여 CPU 및 GPU 성능을 최적화할 수 있으며, 이를 통해 효과적인 ML 모델 구축 및 배포를 할 수 있다.

3. TVM

TVM(Tensor Virtual Machine)[17]은 다양한 하드웨어 플랫폼에서 딥러닝 모델을 최적화하고 배포하기 위한 오픈소스 머신러닝 컴파일러 스택이다. TVM은 Apache Software Foundation에서 개발된 프로젝트로, CPU, GPU, FPGA, ASIC 등 다양한 디바이스에서 딥러닝 모델을 효율적으로 실행할 수 있도록 설계되었다. TVM의 주요 목표는 딥러닝 모델의 성능을 극대화하면서도, 모델을 다양한 디바이스에 쉽게 배포할 수 있는 기능을 제공하는 것이다.

TVM은 Relay라는 고수준의 IR을 사용하여 다양한 딥러닝 프레임워크의 모델을 수용하고 변환할 수 있다. 이를 통해 많은 모델을 표현할 수 있고 다양한 최적화가 가능하다. AutoTVM이라는 자동 최적화 도구를 포함하여, 하드웨어 특성에 맞춘 최적의 커널을 자동으로 검색하고 튜닝이 가능하다. TVM은 크로스 플랫폼을 지원하여 개발자가 특정 하드웨어에 종속되지 않고 다양한 환경에서 모델을 사용할 수 있도록 지원한다.

TVM의 또 다른 중요한 특징은 VTA(Versatile Tensor Accelerator)로, 이는 FPGA와 같은 프로그래머블 하드웨어에서 직접 하드웨어 가속기를 설계하고 프로그래밍할 수 있는 프레임워크를 제공한다. TVM은 TensorFlow, PyTorch, MXNet, Keras 등 여러 머신러닝 프레임워크의 모델을 직접 가져와서 최적화할 수 있으며, 이를 통해 클라우드 서버, 모바일 기기, 임베디드 시스템 등 다양한 환경에 모델을 효율적으로 배포할 수 있다.

4. XLA

XLA(Accelerated Linear Algebra)[18,19]는 Google에서 개발한 오픈소스 컴파일러로, 주로 TensorFlow와 같은 딥러닝 프레임워크에서 기계 학습 모델의 성능을 최적화하고 실행 속도를 높이기 위해 설계되었다. XLA는 기계 학습 연산을 최적화하고 다양한 하드웨어 가속기에서 효율적으로 실행할 수 있는 저수준 코드를 생성하는 데 중점을 둔다. 이를 위해 계산 그래프를 분석하고 최적화하여 중복된 연산을 제거하고, 연산 순서를 재배열하며, 연산을 병합함으로써 실행 효율성을 극대화한다.

XLA는 NVIDIA GPU, TPU, CPU 등 다양한 하드웨어를 지원하며, 각 하드웨어의 특성에 맞춘 최적화를 수행하여 동일한 모델이 다양한 환경에서 최상의 성능을 발휘할 수 있게 한다. 또한 XLA는 JIT 컴파일과 AOT 컴파일을 지원하여, 실행 중에 모델의 연산을 최적화하거나 사전에 컴파일하여 배포 시 최적화된 코드를 실행할 수 있다. TensorFlow와 긴밀하게 통합된 XLA는 사용자가 별도의 설정 없이 TensorFlow 코드 내에서 쉽게 사용할 수 있다. TensorFlow에서 XLA를 활성화하면 모델의 연산 그래프가 자동으로 XLA를 통해 최적화되고 실행된다. 또한 PyTorch 및 JAX 같은 다른 프레임워크를 지원하기 위해 OpenXLA 프로젝트로 확장되었다.

5. Glow

Glow(Graph Lowering)[20]는 Facebook AI Research에서 개발한 오픈소스 딥러닝 컴파일러로, 다양한 하드웨어 백엔드에서 딥러닝 모델의 최적화와 실행을 위해 설계되었다. Glow는 성능 최적화와 메모리 효율성을 강조하며, 그래프 로어링 기술을 통해 고수준의 딥러닝 모델 그래프를 더 낮은 수준의 중간 표현으로 변환하여 하드웨어가 이해할 수 있는 형태로 만든다. 이 과정에서 연산 병합, 메모리 최적화, 명령 스케줄링 등의 다양한 최적화 기법을 적용하여 모델의 실행 속도를 높이고 자원 사용을 최적

화한다.

Glow는 높은 수준 IR과 낮은 수준 IR이라는 두 단계의 IR을 사용하여, 모델의 전반적인 구조적 최적화와 세부적인 연산 최적화를 모두 수행한다. 다양한 하드웨어 백엔드를 지원하며, CPU, GPU, 특수 목적의 AI 가속기 등을 포함하여, 동일한 모델이 여러 종류의 디바이스에서 효율적으로 실행될 수 있게 한다.

또한, Glow는 메모리 관리에 최적화된 기능을 제공하여 모델 실행 중 메모리 사용을 최소화하고, 메모리 할당과 해제를 효율적으로 관리한다. 모듈형 설계로 새로운 최적화 기법이나 하드웨어 지원을 쉽게 추가할 수 있어 확장성과 유지보수성이 뛰어나며, 오픈소스로 제공된다.

6. Thunder

Lightning AI는 PyTorch 프레임워크를 사용하는 AI 모델의 학습을 최적화하기 위해 Thunder[21]라는 소스 대 소스 컴파일러를 개발했다. 이 컴파일러는 여러 GPU를 활용하여 트레이닝 속도를 최대 40%까지 향상시킬 수 있도록 설계되었으며, nvFuser와 cuDNN 등의 기술을 통합하여 하드웨어 간의 계산을 효율적으로 관리한다. Thunder는 단일 GPU와 다중 GPU 구성을 모두 지원하며, 다양한 AI 개발 규모에 적합한 유연성을 제공한다. Thunder는 Apache 2.0 라이선스로 오픈소스로 제공되어 개발자와 연구자들이 프로젝트에 통합하고 수정할 수 있다.

또한, Lightning Studios와 통합되어 프로파일링 및 최적화 도구를 제공함으로써 GPU 성능 병목 현상을 식별하고 해결할 수 있도록 돕는다. 이를 통해 모델 성능을 이해하고 최적화하려는 딥러닝 사용자에게 유용하다.

7. PlaidML

PlaidML[22]은 인텔에서 개발한 오픈소스 딥러닝 컴파일러로, 다양한 하드웨어 플랫폼에서 딥러닝 모델을 효율적으로 실행할 수 있도록 설계되었다. 주된 목표는 하드웨어 독립적인 딥러닝 컴파일러를 제공하여 여러 종류의 디바이스에서 성능을 최적화하는 것이다. PlaidML은 특히 GPU와 같은 가속기를 활용하여 성능을 극대화하며, NVIDIA의 CUDA와 AMD의 ROCm, OpenCL 등을 지원하여 다양한 GPU와 함께 작동할 수 있다. 또한, CPU와의 호환성을 제공하여 GPU가 없는 환경에서도 모델을 실행할 수 있다. PlaidML은 Keras와 같은 고수준의 딥러닝 프레임워크와 통합하여 사용하기 쉽게 설계되었으며, Tile이라는 자체 언어를 사용하여 모델을 최적화한다. Tile은 수학적 연산을 효율적으로 표현하고 이를 다양한 하드웨어 백엔드로 컴파일할 수 있게 하여, 각 하드웨어의 특성에 맞는 최적화된 코드를 생성하여 성능을 극대화한다. PlaidML은 또한 자동 튜닝 기능을 통해 최적의 실행 설정을 자동으로 찾아내어 성능을 극대화한다.

8. NEST-C

NEST-C[23]는 한국전자통신연구원(ETRI)에서 개발한 딥러닝 컴파일러로, 다양한 신경망 처리 유닛(NPU)을 위한 코드 생성 및 최적화를 목적으로 한다. NEST-C는 CPU와 GPU를 포함한 여러 종류의 AI 칩과 애플리케이션을 지원할 수 있다. NEST-C의 주요 기능으로는 각 백엔드 매개변수의 자동 조정, 다양한 NPU를 지원하기 위한 동적 양자화, GCC/LLVM 컴파일러를 사용한 임베디드 하드웨어용 C/C++ 코드 생성 등이 있다. 이러한 다양한 하드웨어 지원과 범용성은 자율주행, 사물 인터넷

(IoT), 센서 기술과 같은 애플리케이션 개발에 적용되었다. NEST-C는 한국 AI 기술을 둘러싼 생태계를 구축하고 AI 애플리케이션의 신속한 개발 및 상용화를 촉진하기 위하여 GitHub에서 오픈소스로 공개되었다.

최근에는 다양한 레벨의 모듈형 NPU를 지원하며 HW 및 SW 동시 최적화가 가능한 개방형 AI 플랫폼을 개발 중으로 개발 완료 후 결과물을 공개하여 HW, SW 개발 생태계 구성에 기여할 계획이다.

Ⅳ. 결론

AI 반도체와 이를 지원하는 컴파일러 기술은 AI 모델의 성능 최적화와 효율적인 실행을 위해 필수적인 요소로, 다양한 하드웨어 플랫폼에서의 성능 향상을 가능하게 한다. 국내에서는 사피온, 리벨리온, 퓨리오사AI와 같은 기업들이 AI 반도체 개발을 선도하고 있으며, 해외에서는 Google, Meta, Graphcore, Tesla, AWS 등의 기업들이 고성능 AI 반도체를 개발하고 있다. 이와 함께 MLIR, IREE, TVM, XLA, Glow, Thunder, PlaidML, NEST-C 등의 컴파일러 기술이 AI 모델의 최적화와 실행을 지원하고 있다. 이러한 기술들은 AI 반도체의 성능을 극대화하고, 다양한 하드웨어 환경에서 효율적인 AI 모델 실행을 가능하게 한다. 향후 AI 반도체와 컴파일러 기술의 발전은 AI 응용 분야의 확장과 성능 향상에 중요한 역할을 할 것이다. 따라서, AI 반도체와 컴파일러 기술의 지속적인 연구와 개발이 필요하며, 이를 통해 AI 기술의 혁신과 발전을 촉진할 수 있을 것이다.

본고는 이러한 기술 동향을 종합적으로 분석함으로써 AI 반도체와 컴파일러 기술의 현재와 미래를 조망하고, 향후 연구 방향을 제시하는 데 기여하고자 한다.

약어 정리

AI	Artificial Intelligence
AOT	Ahead-Of-Time
ASIC	Application Specific Integrated Circuit
CPU	Central Processing Unit
CXL	Compute eXpress Link
DSP	Digital Signal Processor
EUV	Extreme UltraViolet
FHFL	Full Height Full Length
FHHL	Full Height Half Length
FLOPS	FLoating point Operations Per Second
FPGA	Field Programmable Gate Array
GPU	Graphics Processing Unit
HBM	High Bandwidth Memory
IR	Intermediate Representation
JIT	Just-In-Time
LLM	Large Language Model
MLIR	Machine Learning Intermediate Representation
MTIA	Meta Training and Inference Accelerator
NPU	Neural Processing Unit
RISC	Reduced Instruction Set Computer
TDP	Thermal Design Power
TOPS	Tera Operations Per Second

참고문헌

[1] N. Asia, “Led by Nvidia, U.S. dominates in generative AI tech,” 2024. 3. 28.

[2] K. Crawford, “Generative AI’s environmental costs are soaring – and mostly secret,” WORLD VIEW, 2024. 2. 20., https://doi.org/10.1038/d41586-024-00478-x

[3] https://www.sapeon.com/products/sapeon-x330

[4] https://rebellions.ai/rebellions-product/atom-2/

[5] https://furiosa.ai/warboy/specs

[6] M. He et al., “Newton: A DRAM-maker’s Accelerator-in-Memory (AiM) Architecture for Machine Learning,” in Annual IEEE/ACM Int. Symp. Microarchitecture (Athens, Greece), Oct. 2020, https://doi.org/10.1109/MICRO50266.2020.00040

[7] Y. Kwon, “Cost effective LLM inference solution using SK hynix’s AiM (Accelerator-in-Memory),” in SC

(DENVER, CO, USA), Nov. 2023.

[8] C. Robinson, "Google TPU v5p AI Chip Launches Alongside Gemini," STH, 2023, https://www.servethehome.com/google-tpu-v5p-ai-chip-launches-alongside-gemini/

[9] E. Tal et al., "Our next-generation Meta Training and Inference Accelerator," Meta, 2024, https://ai.meta.com/blog/next-generation-meta-training-inference-accelerator-AI-MTIA/

[10] S. Knowles, "Graphcore Colossus Mk2 IPU," in IEEE Hot Chips 33 Symp. (Palo Alto, CA, USA), 2021, https://doi.org/10.1109/HCS52781.2021.9567075

[11] Tesla Dojo Technology, "A Guide to Tesla's Configurable Floating Point Formats & Arithmetic," https://cdn.motor1.com/pdf-files/535242876-tesla-dojo-technology.pdf

[12] A. McFarland, "Tesla's Dojo Supercomputer Explained," HashDork, 2024, https://hashdork.com/tesla-dojo-supercomputer/

[13] AWS, "inferentia2 Architecture," AWS Neuron Documentation, 2024, https://awsdocs-neuron.readthedocs-hosted.com/en/latest/general/arch/neuron-hardware/inferentia2.html

[14] MLIR Homepage, https://mlir.llvm.org/

[15] IREE Homepage, https://iree.dev/

[16] The IREE Authors, "IREE [Computer software]," GitHub, 2019, https://github.com/iree-org/iree

[17] TVM Homepage, https://tvm.apache.org/

[18] OpenXLA Homepage, https://openxla.org/

[19] Google, "OpenXLA," GitHub, https://github.com/openxla/xla

[20] Facebook AI Research, "Glow," GitHub, https://github.com/pytorch/glow

[21] Lightning-AI, "Lightning-Thunder," GitHub, https://github.com/Lightning-AI/lightning-thunder

[22] Intel, "PlaidML," GitHub, https://github.com/plaidml/plaidml

[23] ETRI, "NEST-C," GitHub, https://github.com/etri/nest-compiler