이기종 가속기 기반 엣지 AI 시스템의 시간-공간 하이브리드 스케줄러 설계

김한결, *안재훈, 김영환

한국전자기술연구원

taylor4359@keti.re.kr, *corehun@keti.re.kr, yhkim93@keti.re.kr

Design of Time-Space Hybrid Scheduler for Edge AI Systems based on Heterogeneous Accelerators

Hangyeol Kim, Jaehoon An*, Younghwan Kim

Korea Electronics Technology Institute

요 약

엣지 환경에서는 제한된 하드웨어 자원 위에서 학습과 추론을 동시에 수행해야 하며, 이로 인한 자원 경쟁과 간섭이 추론 지연(SLA) 악화 및 전체 활용도 저하로 이어진다. 본 논문은 GPU와 NPU로 구성된 이기종 가속기 기반 엣지 시스템을 대상으로, **시간 분할(Time-Slicing)**과 **공간 분할(Spatial Partitioning)**을 결합한 하이브리드 스케줄러 아키텍처를 설계한다. 제안 스케줄러는 (1) 워크로드 관측(요청 도착률, 대기열 길이, p95 지연 등)과 모델 특성(지연 민감도, 메모리/연산 요구)을 실시간 수집하고, (2) 공간 분할 단계에서 GPU MIG 인스턴스 크기/개수와 NPU 오프로딩을 통해 격리를 보장하며, (3) 각 파티션 내부에서는 우선순위 선점형 시간 분할로 추론에는 짧은 타임슬라이스와 높은 우선순위를, 학습에는 긴 타임슬라이스와 백그라운드 실행을 부여한다. 또한 파티션 재구성의 오버헤드를 줄이기 위해 부하 추세 기반의 완만한 파티션 적응 정책을 적용하고, 모델 크기·정확도 요구에 따라 GPU/NPU 선택을 동적으로 병행한다.

Ⅰ. 서 론

엣지 컴퓨팅 환경에서 제한된 자원으로 AI 모델의 동시 학습(Training)과 추론(Inference)을 수행해야 하는 요구가 증가하고 있다. 그러나 하나의 장치에서 학습과 추론 작업을 병행하면 자원 경쟁으로 인해 추론 지연이 커지고 학습 속도가 느려지는 문제가 발생한다. 기존에는 학습 작업과 추론 서비스를 분리된 장치나 별도 단계로 운영하는 경우가 많았지만, 이러한 격리 운용은 GPU 유휴 시간 증가 등 자원 비효율과 새로운 데이터에 대한 모델 적응 지연을 초래한다 [1]. 예를 들어 추론 전용으로 GPU를 독점하면 지연은 낮아지지만, 전체 추론 처리량이 낮고, 반대로 학습 작업에 GPU를 전적으로 할당하면 추론 요청을 처리할 수 없어 서비스 공백이 생긴다. 따라서 엣지 단말의 이기종 가속기(예: GPU와 NPU)를 효율적으로 활용하여 동시에 학습과 추론을 수행하면서도 추론 지연은 최소화하고 자원 활용도를 높이는 스케줄링 기법이 필요하다.

본 논문에서는 시간분할(Time-Slicing)과 공간 분할(Spatial Partitioning) 기법을 병행한 하이브리드 스케줄러 아키텍처를 설계한다. 시간 분할은 하나의 가속기를 여러 작업이 시간상으로 공유하도록 하는 것으로 높은 활용도를 가능케 하나, 작업 간 간섭으로 지연 변동이 커질 수 있다. [2]. 공간 분할은 하드웨어 자원을 물리적으로 분할하여 작업별로 할당함으로써 격리와 예측 가능성을 제공하지만, 자원 단편화로 인한 비효율이 우려된다 [3]. NVIDIA Ampere GPU의 MIG(Multi-Instance GPU) 기술은 하나의 GPU를 하드웨어적으로 여러 인스턴스로 나누어 각 인스턴스에 전용 메모리와 연산 코어를 할당함으로써 워크로드 간 간섭을 줄이고 지연을 낮추는 대표적 공간 분할 기법이다 [4]. 한편 MIG로 분할된 각 인스턴스 내부에서도 다중 작업을 시분할로 공유할 수 있어 사용자 수용에 따라 유연하게 처리할 수 있다. [5]. 이러한 두 접근법을 결합한 스케줄링이 혼합된 작업 환경에서 성능과 효율의 균형을 달성할 것으로 기대된다.

본 연구에서 제안하는 스케줄러는 엣지 AI 디바이스의 GPU와 NPU 자원을 통합 관리하며, 작업 유형과 실시간 부하를 모니터링하여 공간적 자원 할당과 시간 슬라이스 조정을 동적으로 수행한다. 이를 통해 지연 민감한 추론 작업에는 전용 자원과 짧은 시간 조각을 우선 할당하고, 연산 집약적인 학습 작업은 잔여 자원을 활용하여 백그라운드로 진행한다. 이하에서 시스템 설계 내용을 자세히 설명하고 아키텍처 개요와 워크로드 유형별 자원 분할 정책에 대해서 설명한다.

Ⅱ. 본론

1. 관련 연구

엣지 AI 시스템의 이기종 가속기 스케줄링에 관한 최근 연구들이 활발히 보고되고 있다. Corun 프레임워크는 하나의 GPU에서 다수의 추론 요청과 지속적 모델 재학습을 동시 처리하는 기법으로, 사전 오프라인 프로파일링을 통해 메모리 초과나 지연 급증 없이 최적의 동시 실행 개수를 찾아내는 스케줄링 방법을 제안하였다. 그 결과 Corun은 단일 GPU로 여러 작업을 병행하여 추론 처리량을 크게 높이면서도 지연 증가는 최소화하였으며, 별도 GPU를 각각 사용할 때 대비 비용 효율적임을 보였다.

LeMix 시스템은 대규모 언어 모델(LLM)의 서빙과 훈련 작업을 멀티 GPU 환경에서 동일 노드에 혼재시키는 방안을 제시하였다. LeMix는 오프라인 프로파일러와 실행 예측 기법, 계층적 자원 할당 및 런타임 메모리-인지 스케줄러를 통합하여 작업들의 상호 간섭을 완화하면서 동적으로 자원을 할당한다. 이를 통해 사용률이 낮은 시기에 노드 수를 줄이고, 부하가 높을 때도 추론 응답성을 크게 저하시키지 않으면서 GPU 자원 활용

도를 22% 이상 향상 시켰다고 보고 하였다.

2. 시스템 아키텍처 설계

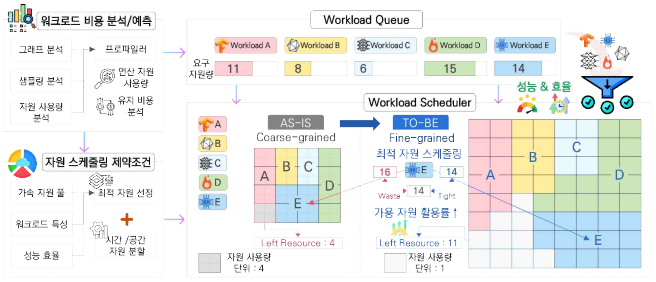

그림 1. 이종 AI 가속기 환경 최적 자원 배치 스케줄러 설계

본 논문에서 제안하는 하이브리드 시간 - 공간 자원 분할 스케줄러의 구조. 입력 대기열에서 추론 요청과 학습 작업이 도착하면, 모니터링 모듈이 작업 유형과 성능 지표(요청 빈도, 마감 시간 등)를 추적한다. 스케줄링 정책 모듈은 이를 바탕으로 자원 할당 방식을 결정하는데, 한편으로는 공간 분할 관점에서 GPU를 여러 파티션으로 나누거나(NVIDIA MIG 등의 기술 활용) 작업을 NPU로 오프로딩할지를 결정하고, 다른 한편으로는 시간 분할 관점에서 각 작업의 실행 우선순위와 타임슬라이스 길이를 조정한다. 최종적으로 하드웨어 분할 제어기가 이러한 결정 사항을 하드웨어 레벨에서 적용하여, GPU의 MIG 인스턴스를 구성하거나 작업을 해당 자원으로 디스패치(dispatch)한다.

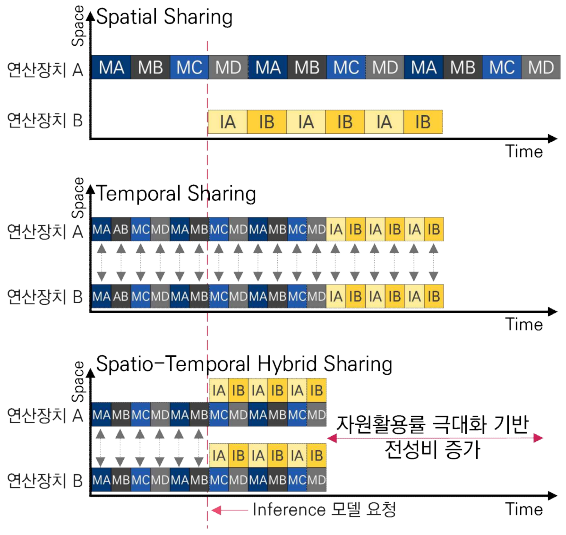

그림 2. 시간/공간 분할 스케줄러 설계

스케줄러의 공간 분할 구성은 현재 워크로드 구성에 따라 동적으로 변경된다. 예를 들어 지연에 민감한 추론 요청이 지속적으로 발생하면 GPU 자원의 일정 부분을 소형 MIG 인스턴스로 확보하거나 NPU 전용으로 할당하여 해당 인스턴스에서는 오직 추론 작업만 실행하도록 격리한다. 반대로 추론 부하가 적을 때는 더 큰 비중의 GPU 자원을 학습 작업에 할당한다. 이러한 적응형 GPU 파티셔닝을 통해 불필요한 리소스 재할당(overhead)을 줄이면서도 다양한 작업에 대응할 수 있다. 단, MIG 파티션 크기를 조정하는 작업은 잦은 변경 시 오버헤드가 발생할 수 있으므로, 일정 기간 모니터링한 부하 추세에 기반하여 완만하게 조정된다. NPU 역시 전용 메모리 및 연산 엔진을 가지고 있으므로, 지원하는 모델의 추론 작업은 가능하면 NPU로 보내 GPU 부하를 줄인다. 다만 NPU는 온보드 메모리 용량이 작아 대형 모델을 수용하기 어렵고 복잡한 DNN 실행 시 정확도 저하를 초래할 수 있다는 한계가 있다. 그러므로 모델 크기나 정확도 요구사항에 따라 GPU와 NPU를 선택적 병행 활용하여 성능과 정확도의 균형을 잡는다.

한편 시간 분할 스케줄링 측면에서는, 각 자원 파티션 내에서 다중 작업이 존재할 경우 선점형(priority preemptive) 스케줄링을 적용한다. 특히 실시간 추론 작업에는 높은 우선순위를 부여하여 실행 중인 학습 커널보다 우선 배치하고, 필요한 경우 학습 작업을 일시 중단하여라도 즉각 추론을 처리한다. 또한 추론 작업에는 짧은 타임슬라이스를 적용해 응답 지연을 낮추고, 학습 작업에는 비교적 긴 타임슬라이스를 부여하여 문맥 전환 overhead를 줄이면서 연산 효율을 높인다. 이러한 QoS-인식 시간조율을 통해 추론 요청의 응답시간 SLA를 보장하면서도 남는 시간에는 학습을 진행하여 자원을 최대한 활용한다. 스케줄러는 모니터링을 통해 추론 대기열 길이나 마감 시간 준수 여부를 계속 점검하며, 상황에 따라 학습 작업의 실행 슬라이스를 더 잘게 쪼개거나 일시 정지하는 등 동적 조정을 수행한다.

Ⅲ. 결론

본 논문에서는 GPU와 NPU로 구성된 엣지 AI 장치에서 동시 학습 및 추론을 효율적으로 지원하기 위한 하이브리드 자원 스케줄러 설계를 제안하였다. 제안된 스케줄러는 작업 특성 모니터링을 통해 공간적 자원 분할(MIG 기반 GPU 파티셔닝, NPU 할당)과 시간적 분할(우선순위 및 슬라이스 조정)을 동적으로 병행함으로써, 제한된 자원 환경에서도 추론 서비스의 지연을 보장하면서 남은 성능으로 학습 작업을 진행할 수 있다. 향후에는 실제 프로토타입 구현을 통해 제안 기법의 유효성 검증을 진행할 예정이다. 또한 FPGA, DSP 등 더 다양한 종류의 가속기에 대한 지원이나, 워크로드 예측을 통한 선제적 스케줄링 최적화 기법, 그리고 분산 엣지 환경에서의 자원 조율 등으로 연구를 확장할 계획이다.

ACKNOWLEDGMENT

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.RS-2025-25441574, Development of an On-Device Integrated Edge AI Server System)

참 고 문 헌

[1] NVIDIA, “NVIDIA Multi-Instance GPU User Guide (RN-08625-v2.0),” 2024.

[2] T. Wang, S. Li, B. Li, Y. Dai, A. Li, G. Yuan, Y. Ding, and Y. Zhang, “Improving GPU Multi-Tenancy Through Dynamic Multi-Instance GPU Reconfiguration,” arXiv:2407.13126, 2024..

[3] Y. Li, Z. Li, Y. Zhu, and C. Liu, “LeMix: Unified Scheduling for LLM Training and Inference on Multi-GPU Systems,” RTSS 2025.

[4] Y.-M. Tang, W.-F. Sun, H.-T. Ting, M.-H. Chen, et al., “PCIe Bandwidth-Aware Scheduling for Multi-Instance GPUs,” in Proc. HPCASIA ’25, ACM, 2025.

[5] S. Zhang, A. Xu, Q. Chen, H. Zhao, W. Cui, Z. Wang, Y. Li, L. Xiao, and M. Guo, “Efficient Performance-Aware GPU Sharing with Compatibility and Isolation through Kernel Space Interception,” in USENIX ATC 2025, 2025.