다중 테넌트 환경에서 GPU 커널 스케줄링 기반 적응형 자원 할당 시스템 연구

권용진, 안재훈*, 김영환
한국전자기술연구원 지능형 IDC 사업단
(rrnjs0721, corehun, yhkim)@keti.re.kr

A Study on Adaptive Resource Allocation Systems Based on GPU Kernel Scheduling in Multi-Tenant Environments

YongJin Kwon, JaeHoon An*, YoungHwan Kim
Intelligent IDC Project Office, Korea Electronics Technology Institute

요 약

현대 클라우드에서는 여러 사용자가 GPU를 공유하는 다중 테넌트 환경이 일반화되면서, 효율적인 GPU 자원 할당과 성능 보장이 중요한 과제로 떠오르고 있다. 하지만 기존 GPU 스케줄링 기법들은 주로 단일 고정 정책만을 사용하여 다양한 워크로드 특성과 실시간 사용자 요구 변화에 효과적으로 대응하지 못했다. 본 연구에서는 사용자 정의 정책에 따라 실행 중에도 동적으로 GPU 커널 스케줄링을 조정함으로써 이러한 한계를 극복하는 적응형 다중 테넌트 GPU 스케줄링 시스템을 제안한다. 제안 시스템은 JCT최소화, SLO보장, 하이브리드 최적화의 세 가지 핵심 정책 모드를 단일 프레임워크에서 지원하며, Bubble-less 커널 스케줄링과 시공간 하이브리드 자원 분할 기법을 결합하여 GPU 유휴 시간을 제거하고 자원 활용률을 극대화한다. 또한 실시간 모니터링 및 우선순위 조정을 통해 워크로드 실행 도중에도 정책 목표 달성을 위해 정책 전환과 자원 재할당을 수행한다. 이러한 동적 스케줄링 접근법은 기존 정적 방식의 한계를 극복하여, 단일 GPU 내에서 여러 테넌트의 다양한 요구사항을 유연하고 효율적으로 만족시킬 수 있음을 보인다.

Ⅰ. 서 론

GPU 컴퓨팅의 급속한 발전과 함께 클라우드 환경에서의 다중 테넌트 GPU 활용이 필수적인 요구사항으로 자리잡았다. 그러나 현재 생산 환경에서 GPU 활용률은 평균 52%에 머물러 있으며, 이는 비효율적인 자원 할당과 워크로드 간 간섭으로 인한 성능 저하가 주요 원인이다. [1]

기존 GPU 스케줄링 연구들은 주로 단일 최적화 목표에 집중되어 왔다. NVIDIA MPS, MIG[2]와 같은 하드웨어 기반 솔루션은 정적 자원 분할로 인한 유연성 부족 문제를 겪고 있으며, Bless[3], Gpulet[4]와 같은 소프트웨어 기반 접근법들도 고정된 정책 적용으로 인해 다양한 워크로드 특성 변화에 적응하지 못하는 한계를 보인다.

현재 시스템들이 직면한 핵심 문제점은 세 가지로 요약된다. 첫째, 시스템 초기화 시점에 정해진 단일 정책만 사용하여 다양한 워크로드 특성 변화에 유연하게 대응하지 못한다. 둘째, JCT 최소화, SLO 보장, 공정성 등 서로 다른 목표를 가진 사용자들의 요구사항을 동시에 만족시키기 어렵다. 셋째, 워크로드 실행 중 성능 목표나 우선순위가 변경되어도 이를 반영할 수 있는 동적 조정 메커니즘이 부족하다.

본 연구는 이러한 문제점들을 해결하기 위해 사용자 정의 정책에 따라 동적으로 커널 스케줄링을 수행할 수 있는 적응형 다중 테넌트 GPU 스케줄링 시스템을 설계한다. JCT 최소화, SLO 보장, 하이브리드 최적화라는 세 가지 핵심 정책을 단일 시스템에서 동적으로 전환할 수 있는 다중 정책 지원 아키텍처를 설계한다. Bubble-less 커널 스케줄링을 결합하여 워크로드 특성 변화와 사용자 요구사항 변경에 즉시 대응할 수 있는 실시간 정책 적응 메커니즘을 개발한다.

Ⅱ. 본론

본 논문에서 제안하는 적응형 다중 테넌트 GPU 스케줄링 시스템은 정책 관리 계층, 스케줄링 엔진 계층, 자원 할당 계층으로 구성된 계층적 구조로 설계된다.

정책 관리 계층은 사용자가 정의한 성능 목표와 제약 조건을 해석하고 적절한 스케줄링 전략을 선택한다. 세 가지 핵심 정책은 독립적인 모듈로 구현되어 런타임에 동적으로 전환될 수 있다.

스케줄링 엔진 계층은 선택된 정책에 따라 실제 커널 스케줄링을 수행한다. Bubble-less 커널 스쿼드 생성, 우선순위 기반 작업 선택, 동적 정책 적응의 세 가지 주요 기능을 제공한다.

자원 할당 계층은 물리적 GPU 자원을 논리적으로 분할하고 할당한다. SM 할당, 메모리 분할, 대역폭 관리를 통해 워크로드 간 격리를 보장하면서도 최대한의 자원 활용률을 달성한다.

1. 다중 정책 지원 메커니즘

JCT 최소화 정책은 수정된 SJF 방식을 사용하되, 예상 완료 시간과 자원 공유 가능성을 종합적으로 고려한다. 작업 간 간섭 모델링을 통해 동시 실행 가능한 작업 조합을 식별하고, 전체 시스템 처리량을 최대화한다. 실행 중 성능 모니터링을 통해 예상과 실제 진행 상황의 차이를 추적하여 스케줄링 순서를 동적으로 재조정한다.

SLO 보장 정책은 지연 시간, 처리량, 가용성이라는 세 가지 핵심 SLO 메트릭을 지원한다. 실시간 성능 모니터링을 통해 SLO 위반 위험을 감지하고, 해당 작업의 우선순위를 동적으로 조정한다. 중요도가 높은 작업에 대해서는 자원 예약 메커니즘을 적용하여 시스템 부하 증가 시에도 최소한의 서비스 수준을 유지한다.

하이브리드 최적화 정책은 Sia 스케줄러의 설계 철학을 차용하여 처리량과 공정성을 동시에 고려한 전체 Goodput 최대화를 목표로 한다.[5] 이를 위해 수식1의 다중 모델 간섭을 고려한 실효 처리량 예측 모델을 도입한다. 여기서 T_m^{eff}는 모델 m이 다른 모델들과 동시 실행될 때의 실제 처리량이며, p_{mk}는 모델 간 간섭 계수를 나타낸다. 이는 기존의 단순한 SM

분할 방식과 달리 메모리 대역폭 경합 등의 실제 간섭 효과를 정량화하여 보다 정확한 성능 예측을 가능하게 한다.

$$T_m^{eff}(s_m, C) = T_m^{solo}(s_m) \times \Pi_{k \in C, k \neq m}(1 - \rho_{mk} \times \phi(s_m, s_k))$$

[수식1] 다중 모델 간섭을 고려한 처리량 예측

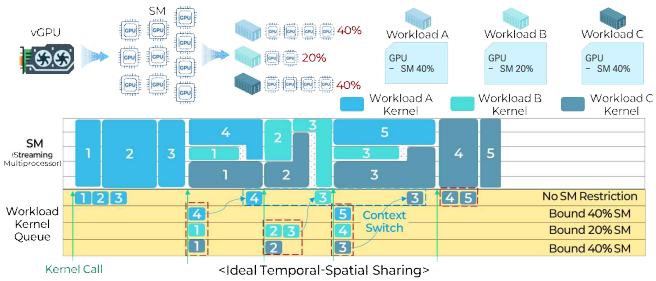
실효 처리량 모델을 기반으로 SLO 달성도를 고려한 다중 모델 최적화 목적 함수를 구성한다. 수식2의 목적 함수는 각 모델의 가중 Goodput의 합을 최대화하되, SLO 달성도 기반 품질 보정 함수 Q_m을 통해 단순 처리량이 아닌 서비스 품질을 반영한 실질적 성능 지표를 사용한다. 가중치 기반 목적함수를 통해 JCT 최소화와 SLO 보장 간의 트레이드오프를 관리한다. 이를 통해 이질적인 워크로드 환경에서도 안정적인 성능을 제공한다.

$$\max_{s_1,\ldots,s_M} \sum_{m=1}^{M} w_{m(t)} \times T_m^{eff}(s_m, C) \times Q_{m(SLO_{m(t)})}$$

[수식2] SLO 인식 다중 모델 Goodput 최적화

2. Bubble-less 커널 스쿼드 메커니즘

Bubble-less 커널 스쿼드는 GPU 자원의 유휴 시간을 최소화하기 위해 서로 다른 작업의 커널들을 효율적으로 조합한다. 스쿼드 생성 알고리즘은 커널 실행 시간 예측, 자원 요구량 분석, 간섭 영향 평가를 통해 호환성 매트릭스를 생성하고, 현재 적용된 정책의 목표에 따라 최적의 커널 조합을 선택한다.



[그림1] 커널 스쿼드 기반 GPU 커널 스케줄링

자원 할당은 SM 개수, 메모리 용량, 대역폭이라는 세 가지 주요 자원을 관리한다. SM 할당은 공간적 분할과 시간적 공유를 동적으로 조합하고, 메모리 할당은 각 작업의 사용 패턴을 분석하여 최적화한다. SLO 보장이 필요한 작업에게는 별도의 메모리 영역을 예약하여 성능 예측 가능성을 높인다.

스쿼드 실행 중 작업 완료로 인해 자원이 해제되면, 남은 작업들에게 즉시 재분배하여 bubbles를 최소화한다. 이는 그림에서 보이는 바와 같이 기존 방식에서 발생하는 자원 낭비를 효과적으로 제거한다. 이러한 메커니즘을 통해 전체 GPU 활용률을 극대화하면서도 개별 작업의 성능 요구사항을 만족시킬 수 있다.

3. 실시간 정책 적응 시스템

실시간 정책 적응을 위해 하드웨어 성능 카운터, 소프트웨어 메트릭, 사용자 피드백을 통합한 다층 모니터링 구조를 채택한다. 모니터링 데이터는 시계열 데이터베이스에 저장되어 패턴 분석과 예측에 활용되며, 기계학습 기반 이상 감지 알고리즘을 통해 성능 저하나 SLO 위반 위험을 사전에 감지한다. 수집된 모니터링 데이터를 바탕으로 수식3의 실시간 가중치 업데이트 메커니즘을 적용한다.

$$w_{m(t+1)} = w_{m(t)} \times \exp(\alpha \times (SLO_{target}^{m} - SLO_{actual}^{m(t)}))$$

[수식3] 실시간 정책 적응을 위한 가중치 업데이트

$w_{m(t+1)}$는 다음 스쿼드에서 적용할 모델 m의 우선순위 가중치이며, $SLO_{actual}^{m(t)}$는 지연 시간과 처리량을 종합한 현재 SLO 달성률을 나타낸다. 가중치 조정을 통해 SLO 위반 위험이 감지된 모델의 우선순위를 즉시 상승시켜 다음 커널 스쿼드에서 더 많은 자원을 할당받도록 한다.

Ⅲ. 결론

본 논문에서는 사용자 정의 정책에 따라 동적으로 커널 스케줄링을 수행할 수 있는 적응형 다중 테넌트 GPU 스케줄링 시스템을 설계하였다. 제안 시스템의 핵심 기여는 기존 정적 스케줄링 방식의 한계를 극복하고 다양한 사용자 요구사항을 동시에 만족시킬 수 있는 유연한 프레임워크를 제공한다는 점이다. JCT 최소화, SLO 보장, 하이브리드 최적화라는 세 가지 핵심 정책을 단일 시스템에서 지원함으로써 사용자는 자신의 요구사항에 가장 적합한 최적화 전략을 선택할 수 있다.

Bubble-less 커널 스쿼드와 정책 기반 우선순위 조정을 결합한 통합 최적화 접근법은 기존 시스템 대비 향상된 성능과 예측 가능성을 제공할 것으로 기대된다. 특히 다중 테넌트 환경에서 워크로드 간 격리를 보장하면서도 최대한의 자원 활용률을 달성할 수 있는 구조를 제시했다.

향후 연구에서는 실제 구현과 성능 평가를 통해 이론적 설계의 유효성을 검증할 예정이다. 기계학습 기반 예측 모델의 정확도 향상, 다중 GPU 클러스터 환경으로의 확장, 보안과 프라이버시 측면에서의 추가 연구가 필요하다.

ACKNOWLEDGMENT

이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구 결과임 (No.RS-2025-02220502, AI 반도체 컴퓨팅 자원분해 및 자원풀링 기술 개발)

참 고 문 헌

[1] Jeon M. et al. "Analysis of Large-Scale Multi-Tenant GPU Clusters for DNN Training Workloads," 2019 USENIX Annual Technical Conference (USENIX ATC 19), pp. 947-960, July 2019.

[2] NVIDIA Corporation, "Multi-Instance GPU User Guide," NVIDIA Documentation, 2025, (https://docs.nvidia.com/datacenter/tesla/mig-user-guide/).

[3] Zhang S. et al. "Improving GPU Sharing Performance through Adaptive Bubbleless Spatial-Temporal Sharing," Proceedings of the Twentieth European Conference on Computer Systems (EuroSys 25), pp. 573-588, March 2025.

[4] Choi S. et al. "Serving Heterogeneous Machine Learning Models on Multi-GPU Servers with Spatio-Temporal Sharing," 2022 USENIX Annual Technical Conference (USENIX ATC 22), pp. 199-216, July 2022.

[5] Subramanya S. J. et al. "Sia: Heterogeneity-aware, goodput-optimized ML-cluster scheduling," Proceedings of the 29th Symposium on Operating Systems Principles (SOSP 23), pp. 642-657, October 2023.