GPU 프로그램에 자료경합 치유기술 적용을 위한 warp 단위 처리기법

(Warp Unit Processing to Apply Healing Data Races in GPU Programs)

이 건 표[†], 최 으 뜸[†], 전 용 기[†‡*]
[†]국립 경상대학교 정보과학과, [‡]국립 경상대학교 항공우주 및 소프트웨어학과
(Keon-Pyo Lee, Eu-Teum Choi, Yong-Kee Jun)
([†]Department of Informatics, Gyeongsang National Univ.,
[‡]Department of Aerospace and Software Engineering, Gyeongsang National Univ.)

Abstract : Previous techniques for healing data race in concurrent program are developed for use in CPU programs. However, the central of parallel processing technology is now changing from CPU to GPU. That means data race healing technology for GPU program is needed. This paper presents the technique which can be used for developing data race healing for GPU program. And that makes it overcome existing technique's healing in unit of thread.

Keywords : data race healing, concurrent program, GPU

Ⅰ. 서론

병행프로그램의 수행 중에 비결정적 결과를 발생시킬 수 있는 자료경합을 배재하기 위해 자료경합 치유 기술이 활용되고 있다. 하지만 기존 기법들은 CPU 환경에서 동작을 전제로 개발되었기 때문에 GPU 프로그램의 자료경합을 치유할 수 없다.

본 논문은 기존 자료경합 치유 기술의 소개 후 GPU 프로그램에 적용될 수 없는 원인을 분석하고, GPU 프로그램의 자료경합 치유 기술을 개발하기 위한 warp 단위 처리기술을 제시한다.

Ⅱ. 배경

병행프로그램에서 동시성오류인 자료경합은 두 개 이상의 스레드에서 하나의 공유변수에 적절한 동기화 없이 최소 하나 이상의 쓰기 이벤트를 포함할 때 발생할 수 있다. 이러한 자료경합은 프로그래머가 의도하지 않은 결과를 발생시켜 프로그램의 신뢰성 보장을 어렵게 한다.

자료경합을 해결하는 기술은 탐지[1]와 치유[2]로 나뉜다. 탐지 기술은 과도한 자원적 비용을 발생시키므로 프로그램 수행 중에 자료경합을 용인하는 치유 기술이 보다 효과적이다.

자료경합 치유 기술은 CPU를 사용하는 프로그래밍을 중심으로 연구되고 있다. 하지만 높은 성능을 요구하는 소프트웨어의 증가로 인해 GPU 병렬처리의 활용이 증대되고 있으며 고 신뢰성을 만족해야하는 항공기 소프트웨어에서의 활용도 논의[3]되는 중이다. 따라서 자료경합과 같은 동시성 오류를 치유하는 기술이 GPU 프로그래밍 영역으로 확장되어야 한다.

GPU 프로그램은 CPU 대비 많은 스레드를 사용하며 스레드의 묶음인 warp를 최소 스케줄링 단위로 하는 SIMT 모델[4]로 실행된다. 이러한 특성 때문에 기존의 기법들은 감시할 스레드 수로 인해 오버헤드 증가가 발생하며 스레드 단위의 인터리빙 조정이 불가능하여 GPU 프로그램의 자료경합을 치유하기 어렵다.

Ⅲ. warp 단위 처리 기법

GPU 프로그램에서 발생하는 자료경합은 warp 단위로 스케줄링 되는 스레드들의 공유변수 접근으로 인해 발생한다. 본 논문은 warp 단위로 동작하는 자료경합 탐지 및 인터리빙 조정 기법을 제시한다.

* 교신저자(Corresponding Author)
이건표, 최으뜸 : 경상대학교
전용기: 경상대학교 정보과학과, 항공우주및소프트웨어학과, 항공기부품기술연구소, 공학연구원

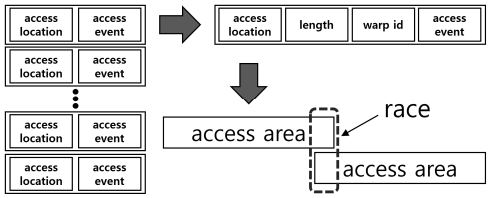

그림 1. warp 단위 자료경합 탐지 기법
Fig. 1. Data Race Detection in Unit of Warp

1. warp단위 자료경합 탐지 기법

GPU 프로그램의 스레드들을 서로 다른 공유변수에 접근하기 위해 스레드 id를 이용하여 접근 주소 값을 계산한다. 이러한 특성을 이용하여 본 연구에서 제안하는 감시 기법은 warp 내의 스레드들이 접근하는 공유 메모리 주소 값들을 영역 정보로 압축하여 표현하고 영역들의 교집합을 통하여 자료경합을 탐지한다.

2. 주소 기반 인터리빙 조정 기법

warp 단위 스케줄링으로 인하여 특정 스레드의 스케줄링만을 조정할 수 없다. 이러한 문제점을 해결하기 위해 제안하는 인터리빙 조정 기법은 warp 분기를 이용하여 자료경합이 발생되는 주소 값에 접근하는 스레드를 지연시킨다. 이러한 방식은 압축된 접근 영역에 warp가 접근하지 않는 주소 값이 포함되어 자료경합의 오 탐지가 발생하더라도 스레드가 실제로 자료경합이 발생되는 주소 값에 접근할 경우에만 지연된다.

IV. 실험

실험은 자료경합을 발생시키는 합성 GPU 프로그램에 제안된 기법을 적용하여, 자료경합 탐지기법의 오버헤드, warp 내 스레드 인터리빙 조정 여부, 오 탐지 발생 여부를 확인하였다. 사용된 GPU 프로그램은 128개의 warp와 4096개 공유변수 접근 스레드를 사용하였고, 실험 환경은 Intel i7-3770K CPU, Nvidia GeForce GTX 650, 8GB RAM 환경에서 진행되었다.

1. 실험 결과

표 1의 탐지 열은 기존 자료경합 치유기술과 본 기법의 자료경합 탐지 측면의 비교 분석 결과이다. 제안된 기법의 시간, 공간적 오버헤드는 기존 기법 대비 각 54.9%, 94.8% 감소되었다. 또한 인터리빙 조정기법은 warp 내의 스레드 인터리빙을 조정할 수 있으며, False Positive가 발생하지 않았다.

표 1 전체 실험 결과
Table 1. Overall Results

	자료경합 탐지		인터리빙 조정	
	공간 오버헤드	수행 시간	조정 여부	FP
기존기법	49392 byte	46ms (4.5x)	X	-
제안기법	2528 byte	20ms (2x)	O	X

2. 결과 분석

warp 단위 자료경합 탐지 기법의 소요시간은 원 프로그램의 약 2배로 스레드 단위 기법 대비 약 2.2배의 성능향상을 보였으며, 공간 오버헤드는 약 5%로 적었다. 자료경합의 오 탐지로 인한 불필요한 스레드 지연이 발생하지 않아 수행시간에는 영향을 미치지 못하였다.

IV. 결 론

GPU를 이용한 대규모 병렬처리 기술은 범용 프로그램을 넘어 항공기 소프트웨어와 같은 고 신뢰성이 요구되는 분야에서도 필요성이 대두되고 있다. GPU 프로그램에의 자료경합 치유 기술 적용은 이러한 추세에 부합한다. 본 논문은 GPU 자료경합 치유를 위해 사용할 수 있는 새로운 자료경합 탐지 및 인터리빙 조정 기법을 제시하였다. 또한 자료경합 치유 기술의 적용 범위를 GPU 프로그램으로 확장할 수 있는 가능성을 보인다.

향후 과제로서 GPU 프로그램의 자료경합을 실제 치유하기 위한 연구가 필요하며, 현재 진행 중이다.

참 고 문 헌

[1] O. Ha, and Y. Jun, "An Efficient Algorithm for On-the-Fly Data Race Detection Using an Epoch-Based Technique". Scientific Programming, Vol. 2015, No. 13, 2015 .

[2] M. Zhang, Y. Wu, S. Lu, S. Qi, J. Ren, W. Zheng, "A Lightweight System for Detecting and Tolerating Concurrency Bugs", IEEE Transactions on Software Engineering, Vol. 42, pp. 899-917, Oct 2016

[3] M. Dutton, "The challenges of graphics processing in the avionics industry", Digital Avionics Systems Conference, January 2010.

[4] C. Nvidia, "Programming guide." 2010.