

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2021-0039600
(43) 공개일자 2021년04월12일

(51) 국제특허분류(Int. Cl.)

G06F 17/16 (2006.01)

(52) CPC특허분류

G06F 17/16 (2013.01)

(21) 출원번호 10-2019-0122023

(22) 출원일자 2019년10월02일

심사청구일자 2019년10월02일

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

정성우

서울특별시 강남구 도곡로28길 8, 104동 1802호 (도곡동, 도곡1차아이파크아파트)

두 콩 뚜안

서울특별시 동대문구 안암로 118-1, 402호 (제기동)

최정환

서울특별시 동대문구 무학로49길 35, 203호(용두동)

(74) 대리인

송인호, 윤형근, 최관락

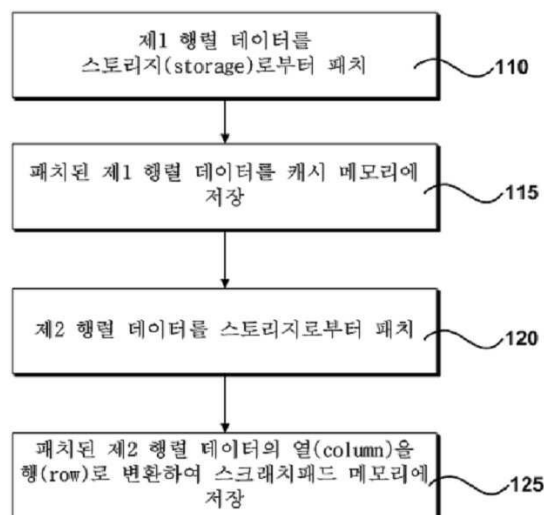
전체 청구항 수 : 총 5 항

(54) 발명의 명칭 행렬 연산시 메모리 데이터 변환 방법 및 컴퓨터

(57) 요약

행렬 연산시 메모리 데이터 변환 방법 및 컴퓨터가 개시된다. 행렬 연산시 메모리 데이터 변환 방법은, (a) 제1 행렬 데이터를 스토리지(storage)로부터 패치하여 캐시 메모리에 저장하는 단계; (b) 제2 행렬 데이터를 스토리지로부터 패치하는 단계; 및 (c) 상기 패치된 제2 행렬 데이터의 열(column)과 행(row)을 변환하여 스크래치패드 메모리에 저장하는 단계를 포함한다.

대표도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호	1711093591
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	(이공)중견연구자지원(3억초과~5억이하)
연구과제명	모노리틱 3D에 기반한 CPU/GPU 컴포넌트의 아키텍처적 설계
기 여 율	40/100
과제수행기관명	고려대학교 산학협력단
연구기간	2019.03.01 ~ 2020.02.29

이 발명을 지원한 국가연구개발사업

과제고유번호	1711093540
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	나노·소재기술개발(R&D)
연구과제명	초절전 IoT 소자 결합 3차원 플랫폼 집적 공정과 최적 아키텍처 개발, 검증
기 여 율	30/100
과제수행기관명	인하대학교 산학협력단
연구기간	2019.03.01 ~ 2019.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711092438
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	나노·소재기술개발(R&D)
연구과제명	모노리틱 3D 기반 3차원 고집적 뉴로모픽 인터커넥트 설계/분석 플랫폼
기 여 율	30/100
과제수행기관명	동국대학교 산학협력단
연구기간	2019.03.01 ~ 2019.12.31

명세서

청구범위

청구항 1

- (a) 제1 행렬 데이터를 스토리지(storage)로부터 패치하여 캐시 메모리에 저장하는 단계;
- (b) 제2 행렬 데이터를 스토리지로부터 패치하는 단계; 및
- (c) 상기 패치된 제2 행렬 데이터의 열(column)과 행(row)을 변환하여 스크래치패드 메모리에 저장하는 단계를 포함하는 행렬 연산시 메모리 데이터 변환 방법.

청구항 2

제1 항에 있어서,

상기 (c) 단계는,

상기 패치된 제2 행렬 데이터의 논리적인 열(column)에 위치한 데이터를 하나의 물리적 행(row)에 위치되도록 상기 스크래치패드 메모리에 저장하는 것을 특징으로 하는 행렬 연산시 메모리 데이터 변환 방법.

청구항 3

제1 항에 있어서,

상기 제2 행렬 데이터는 곱해지는 행렬 데이터인 것을 특징으로 하는 행렬 연산시 메모리 데이터 변환 방법.

청구항 4

스토리지;

캐시 메모리;

스크래치패드 메모리; 및

행렬 연산시, 제1 행렬 데이터를 상기 스토리지로부터 패치하여 상기 캐시 메모리에 저장하며, 제2 행렬 데이터를 상기 스토리지로부터 패치하여 상기 스크래치패드 메모리에 저장시, 하나의 물리적 행에 위치하도록 상기 제2 행렬 데이터를 저장하는 프로세서를 포함하는 컴퓨터.

청구항 5

제4 항에 있어서,

상기 프로세서는,

상기 패치된 제2 행렬 데이터의 열(column)과 행(row)을 변환하여 상기 스크래치패드 메모리에 저장하는 것을 특징으로 하는 컴퓨터.

발명의 설명

기술 분야

[0001] 본 발명은 행렬 연산시 메모리 데이터 변환 방법 및 컴퓨터에 관한 것이다.

배경 기술

[0003] GPU의 행렬 곱셈의 경우 L1D (L1 Data Cache) 및 SPM(Scratchpad Memory)과 같은 온칩 메모리는 두 개의 입력 행렬 (첫 번째 행렬의 행 데이터와 두 번째 행렬의 열 데이터)의 데이터를 제공하나, 온칩 메모리 아키텍처는 한 번에 하나의 행 액세스만 허용하기 때문에, 두 번째 행렬 열 데이터는 다중 액세스로 인해 행렬 곱셈 속도가 느려지는 한계점이 있다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) (01) 대한민국등록특허공보 제10-0909510(2009.07.20.)

발명의 내용

해결하려는 과제

[0006] 본 발명은 행렬 연산시, 연산 속도를 높일 수 있는 메모리 데이터 변환 방법 및 컴퓨터를 제공하기 위한 것이다.

[0007] 또한, 본 발명은 행렬 연산시, 곱해지는 행렬 데이터에 대해 하나의 액세스를 통해 판독하도록 할 수 있어 연산 성능을 향상시킬 수 있는 행렬 연산시 메모리 데이터 변환 방법 및 컴퓨터를 제공하기 위한 것이다.

과제의 해결 수단

[0009] 본 발명의 일 측면에 따르면, 행렬 연산시 메모리 데이터 변환 방법이 제공된다.

[0010] 본 발명의 일 실시예에 따르면, (a) 제1 행렬 데이터를 스토리지(storage)로부터 패치하여 캐시 메모리에 저장하는 단계; (b) 제2 행렬 데이터를 스토리지로부터 패치하는 단계; 및 (c) 상기 패치된 제2 행렬 데이터의 열(column)을 하나의 물리적 행(row)에 위치되도록 변환하여 스크래치패드 메모리에 저장하는 단계를 포함하는 행렬 연산시 메모리 데이터 변환 방법이 제공된다.

[0011] 상기 (c) 단계는, 상기 패치된 제2 행렬 데이터의 논리적인 열(column)에 위치한 데이터를 하나의 물리적 행(row)에 위치되도록 상기 스크래치패드 메모리에 저장할 수 있다.

[0012] 상기 제2 행렬 데이터는 곱해지는 행렬 데이터이다.

[0014] 본 발명의 다른 측면에 따르면, 행렬 연산 성능을 향상시킬 수 있는 컴퓨터가 제공된다.

[0015] 본 발명의 일 실시예에 따르면, 스토리지; 캐시 메모리; 스크래치패드 메모리; 및 행렬 연산시, 제1 행렬 데이터를 상기 스토리지로부터 패치하여 상기 캐시 메모리에 저장하며, 제2 행렬 데이터를 상기 스토리지로부터 패치하여 상기 스크래치패드 메모리에 저장시, 하나의 물리적 행에 위치하도록 상기 제2 행렬 데이터를 저장하는 프로세서를 포함하는 컴퓨터가 제공될 수 있다.

[0016] 상기 프로세서는, 상기 패치된 제2 행렬 데이터의 열(column)과 행(row)을 변환하여 상기 스크래치패드 메모리에 저장할 수 있다.

발명의 효과

[0018] 본 발명의 일 실시예에 따른 행렬 연산시 메모리 데이터 변환 방법 및 컴퓨터를 제공함으로써, 행렬 연산시 곱해지는 행렬 데이터를 하나의 워드 라인으로 판독 할 수 있는 이점이 있다.

[0019] 이를 통해, 본 발명은 행렬 연산시 연산 속도를 향상시킬 수 있는 이점이 있다.

도면의 간단한 설명

[0021] 도 1은 본 발명의 일 실시예에 따른 행렬 연산시 메모리 주소 변환 방법을 나타낸 순서도.

도 2는 본 발명의 일 실시예에 따른 컴퓨터의 내부 구성을 개략적으로 도시한 블록도.

도 3은 본 발명의 일 실시예에 따른 행렬 곱셈 연산 성능 개선을 도시한 그래프.

도 4는 종래 대비 본 발명의 일 실시예에 따른 행렬 크기에 따른 스크래치패드 메모리 및 전체 시스템의 에너지 소비 결과를 나타낸 그래프.

발명을 실시하기 위한 구체적인 내용

[0022] 본 명세서에서 사용되는 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "구성된다" 또는 "포함한다" 등의 용어는 명세서상에 기재된 여러 구성 요소들, 또는 여러 단계들을 반드시 모두 포함하는 것으로 해석되지 않아야 하며, 그 중 일부 구성 요소들 또는 일부 단계들은 포함되지 않을 수도 있고, 또는 추가적인 구성 요소 또는 단계들을 더 포함할 수 있는 것으로 해석되어야 한다. 또한, 명세서에 기재된 "...부", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어 또는 소프트웨어로 구현되거나 하드웨어와 소프트웨어의 결합으로 구현될 수 있다.

[0023] 이하, 첨부된 도면들을 참조하여 본 발명의 실시예를 상세히 설명한다.

[0025] 도 1은 본 발명의 일 실시예에 따른 행렬 연산시 메모리 주소 변환 방법을 나타낸 순서도이다. 이하에서는 행렬 곱셈 연산을 수행하는 방법에 대해 설명하기로 한다.

[0026] 단계 110에서 컴퓨터(100)는 제1 행렬 데이터를 스토리지(storage)로부터 패치한다.

[0027] 단계 115에서 컴퓨터(100)는 패치된 제1 행렬 데이터를 캐시 메모리에 저장한다. 여기서, 캐시 메모리는 L1 데이터 캐시일 수 있다.

[0028] 제1 행렬 데이터를 스토리지로부터 패치하여 L1 데이터 캐시에 저장하는 방법은 종래와 동일할 수 있다.

[0029] 단계 120에서 컴퓨터(100)는 제2 행렬 데이터를 스토리지로부터 패치한다.

[0030] 단계 125에서 컴퓨터(100)는 패치된 제2 행렬 데이터의 열(column)을 행(row)로 변환하여 스크래치패드 메모리에 저장한다.

[0031] 즉, 본 발명의 일 실시예에 따르면, 컴퓨터(100)는 패치된 제2 행렬 데이터의 논리적 열(column)을 하나의 물리적 행(row)에 위치되도록 스크래치패드 메모리에 저장할 수 있다.

[0032] 이를 통해, 본 발명의 일 실시예에 따르면, 행렬 연산시 연산 성능을 향상시킬 수 있다. 즉, 일반적으로 곱해지는 제2 행렬 데이터가 종래와 같이 저장되는 경우 제2 행렬 데이터의 열(column) 데이터는 다중 액세스를 통해 판독되게 된다. 이는 물리적 메모리의 특성상 열(column) 데이터를 판독하기 위해서는 물리적 메모리의 행(row)를 모두 오픈(open)해야만 하는 특성에 기인한다.

[0033] 따라서, 본 발명의 일 실시예에서는 곱해지는 제2 행렬 데이터의 열 데이터에 대한 다중 액세스를 최소화하기 위해 패치된 제2 행렬 데이터의 논리적 열 데이터를 물리적 메모리의 하나의 행(row)에 위치하도록 저장함으로써 한번의 액세스로 제2 행렬 데이터의 논리적 열 데이터에 대한 판독이 가능하도록 할 수 있다.

[0034] 이로 인해, 본 발명의 일 실시예에서는 행렬 연산시, 곱해지는 행렬 데이터의 열(column)을 행(row)로 변환하여 GPU의 스크래치패드 메모리에 저장함으로써 연산 속도를 높일 수 있는 이점이 있다.

- [0036] 도 2는 본 발명의 일 실시예에 따른 컴퓨터의 내부 구성을 개략적으로 도시한 블록도이다.
- [0037] 도 2를 참조하면, 본 발명의 일 실시예에 따른 컴퓨터(100)는 스토리지(210), 캐시 메모리(220), 스크래치패드 메모리(230) 및 프로세서(240)를 포함한다.
- [0038] 스토리지(210)는 제1 행렬 데이터 및 제2 행렬 데이터를 저장한다. 물론, 스토리지(210)에는 제1 행렬 데이터 및 제2 행렬 데이터 이외에도 다양한 데이터들이 저장됨은 당연하다.
- [0039] 캐시 메모리(220)는 L1(level-1) 캐시일 수 있다. L1 캐시는 당업자에게는 자명한 사항이므로 이에 대한 상세한 설명은 생략하기로 한다.
- [0040] 스크래치패드 메모리(230)는 주기억장치와는 별도로 실행중의 명령에 필요한 오퍼랜드, 중간데이터의 유지, 인터럽트 처리 등에 사용하는 고속처리용 메모리이다.
- [0041] 프로세서(240)는 본 발명의 일 실시예에 따른 컴퓨터(100)의 내부 구성 요소들(예를 들어, 스토리지(210), 캐시 메모리(220), 스크래치패드 메모리(230) 등)을 제어한다.
- [0042] 또한, 프로세서(240)는 행렬 연산을 위해, 스토리지(210)로부터 제1 행렬 데이터 및 제2 행렬 데이터를 패치(fetch)한다. 이어, 프로세서(240)는 제1 행렬 데이터는 L1 캐시에 저장할 수 있다. 또한, 프로세서(240)는 패치된 제2 행렬 데이터의 논리적인 열(column)을 하나의 물리적 행에 위치되도록 스크래치패드 메모리(230)에 저장할 수 있다. 즉, 프로세서(240)는 패치된 제2 행렬 데이터의 열(column)을 행(row)로 변환하여 스크래치패드 메모리(230)에 저장할 수 있다.
- [0043] 이로 인해, 프로세서(240)는 행렬 연산시, 스크래치패드 메모리(230)에 한번만 액세스함으로써 제2 행렬 데이터를 판독할 수 있어, 연산 성능을 향상시킬 수 있는 이점이 있다.
- [0044] 도 3은 본 발명의 일 실시예에 따른 행렬 곱셈 연산 성능 개선을 도시한 그래프이다. 본 발명의 일 실시예에 따르면, 8×8 , 16×16 , 24×24 및 32×32 의 행렬 크기에 대해 종래 대비 행렬 곱셈 성능이 7.6%, 17.8%, 31.6% 및 34.9% 향상된 것을 알 수 있다. 이는 주로 행렬 재적에 대해 제2 행렬의 열 단위 데이터를 판독하는데 소요되는 시간이 짧아진 것에 기인한다. 즉, 제2 행렬의 열 단위 데이터를 스크래치패드 메모리의 열에 정렬하는 대신, 하나의 물리적 행에 위치되도록 저장함으로써 한번의 액세스로 행렬 연산이 가능하도록 할 수 있다. 따라서, 행렬의 크기가 커질 경우 종래의 경우 열에 순차적으로 액세스하는데 많은 시간이 소요되기 때문에 본 발명의 일 실시예에 따른 메모리 데이터 변환 방법이 많은 이점이 있는 것을 알 수 있다.
- [0045] 도 4는 종래 대비 본 발명의 일 실시예에 따른 행렬 크기에 따른 스크래치패드 메모리 및 전체 시스템의 에너지 소비 결과를 나타낸 그래프이다.
- [0046] 본 발명의 일 실시예에 따른 방법이 스크래치패드 메모리 및 전체 시스템의 에너지 소비를 각각 92.7 ~ 93.9% 및 6.8 ~ 19.4% 절감하는 것을 알 수 있다. 이는 스크래치패드 메모리에서 단 한번의 워드 라인만을 제2 행렬의 열 단위 요소에 대해 판독하기 때문이다.
- [0048] 본 발명의 실시 예에 따른 장치 및 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 컴퓨터 판독 가능 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 분야 통상의 기술자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media) 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.
- [0049] 상술한 하드웨어 장치는 본 발명의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.
- [0050] 이제까지 본 발명에 대하여 그 실시 예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을

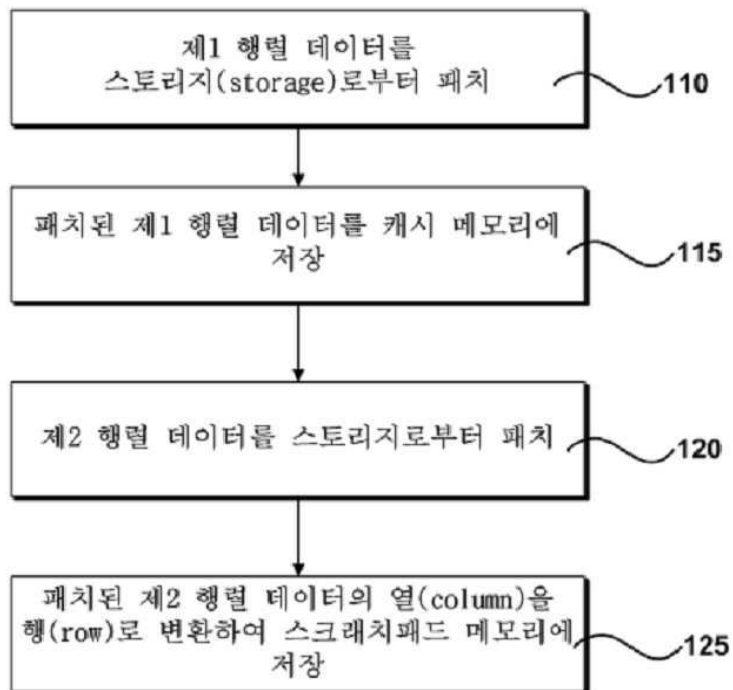
이해할 수 있을 것이다. 그러므로 개시된 실시 예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

부호의 설명

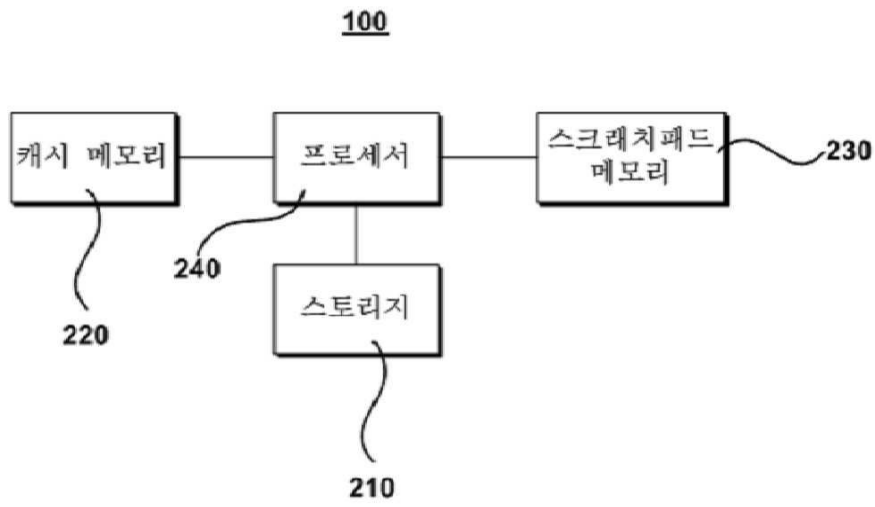
- [0052]
- 100: 컴퓨터
 - 110: 스토리지
 - 120: 캐시 메모리
 - 130: 스크래치패드 메모리

도면

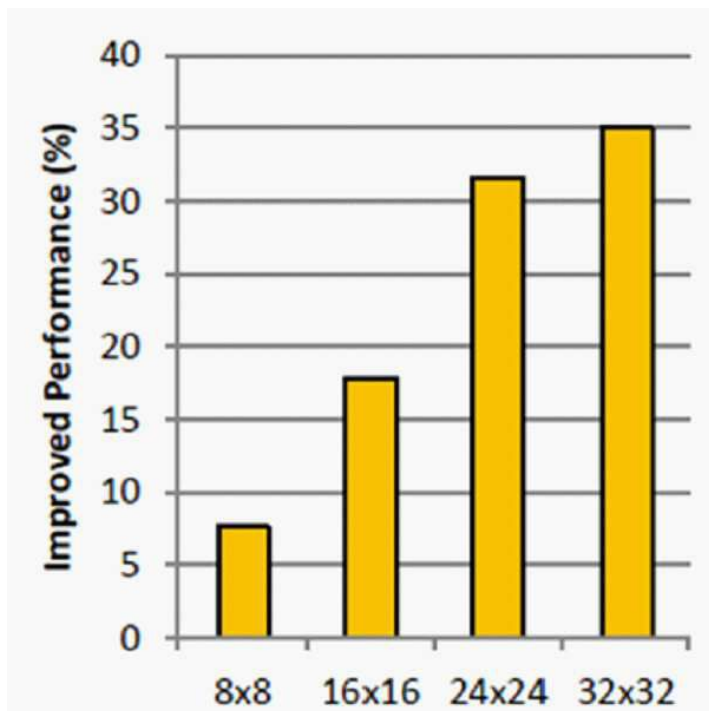
도면1



도면2



도면3



도면4

