(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0102651
(43) 공개일자 2025년07월07일

(51) 국제특허분류(Int. Cl.)
G06F 17/16 (2006.01) *G06F 7/523* (2006.01)
G06F 9/30 (2018.01) *G06F 9/38* (2024.01)
G06N 3/063 (2023.01)
(52) CPC특허분류
G06F 17/16 (2013.01)
G06F 7/523 (2013.01)
(21) 출원번호 10-2023-0194544
(22) 출원일자 2023년12월28일
심사청구일자 2023년12월28일

(71) 출원인
이화여자대학교 산학협력단
서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)
(72) 발명자
김지훈
서울특별시 강남구 언주로30길 57, E동 3105호 (도곡동, 타워팰리스)
김현지
서울특별시 서대문구 이화여대3길 14, 803호 (대현동)
(74) 대리인
리앤목특허법인

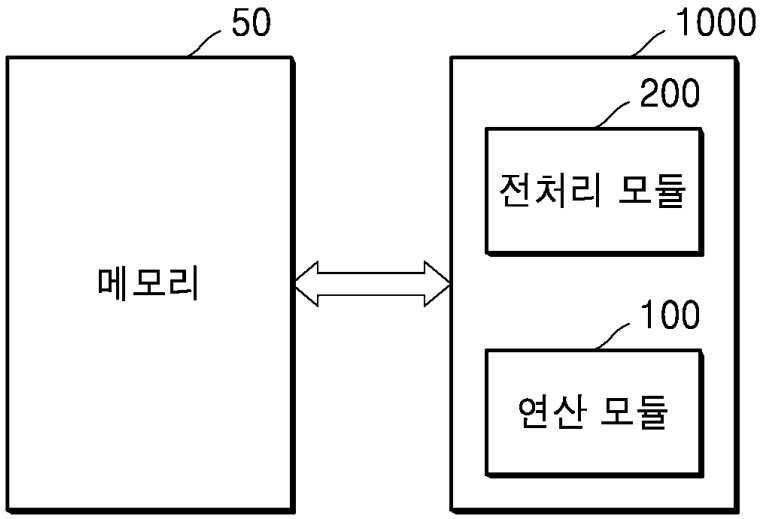
전체 청구항 수 : 총 11 항

(54) 발명의 명칭 **DRAM 대역폭에 따라 확장성을 갖는 희소행렬 연산 가속 방법, 연산 가속 장치 및 상기 방법을 실행시키기 위하여 기록매체에 저장된 컴퓨터 프로그램**

(57) 요 약

본 발명에 따른 연산 가속 방법은, 컴퓨팅 장치에 의해 수행되는 희소행렬의 연산을 가속하는 방법에 있어서, 메모리에 저장된 희소행렬의 요소들에 대하여 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 상기 메모리의 대역폭에 대응하는 숫자로 구비되는 복수의 프로세싱 엘리먼트 라인 단위로 정렬하는 전처리 단계, 및 상기 전처리 단계에서 전처리된 희소행렬의 요소들에 대하여 각 프로세싱 엘리먼트 라인에 구비되는 두개의 연산 유닛을 이용하여 각 프로세싱 엘리먼트 라인 단위로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 연산하는 연산 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06F 9/30123 (2013.01)
G06F 9/3885 (2013.01)
G06N 3/063 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711193299
과제번호 2020-0-01308-004
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 차세대지능형반도체기술개발(설계)
연구과제명 딥러닝 초소형 코어 어레이 기반 지능형 모바일 프로세서
기 여 율 1/1
과제수행기관명 한국전자통신연구원
연구기간 2023.01.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

컴퓨팅 장치에 의해 수행되는 희소행렬의 연산을 가속하는 방법에 있어서,

메모리에 저장된 희소행렬의 요소들에 대하여 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 상기 메모리의 대역폭에 대응하는 숫자로 구비되는 복수의 프로세싱 엘리먼트 라인 단위로 정렬하는 전처리 단계; 및

상기 전처리 단계에서 전처리된 희소행렬의 요소들에 대하여 각 프로세싱 엘리먼트 라인에 구비되는 두개의 연산 유닛을 이용하여 각 프로세싱 엘리먼트 라인 단위로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 연산하는 연산 단계;

를 포함하는, 연산 가속 방법.

청구항 2

제1항에 있어서,

상기 전처리 단계는,

상기 프로세싱 엘리먼트 라인이 구비되는 연산 모듈의 버퍼 사이즈를 기초로 상기 희소행렬을 블록 단위로 나누어 복수의 블록을 설정하는 단계; 및

상기 희소행렬의 각 블록 단위로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 각 프로세싱 엘리먼트 라인에 할당하여 정렬하는 단계;

를 포함하는, 연산 가속 방법.

청구항 3

제2항에 있어서,

상기 복수의 블록을 설정하는 단계는, 상기 프로세싱 엘리먼트 라인에 구비되는 벡터 버퍼의 사이즈에 대응하는 열의 개수를 갖는 블록 단위로 복수의 블록을 설정하는 단계를 포함하는, 연산 가속 방법.

청구항 4

제3항에 있어서,

상기 각 프로세싱 엘리먼트 라인에 할당하여 정렬하는 단계는, 상기 벡터 버퍼의 사이즈를 기초로 설정되는 벡터 요소에 매핑하여 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 정렬하는 단계를 포함하는, 연산 가속 방법.

청구항 5

제4항에 있어서,

상기 연산 단계는,

각 프로세싱 엘리먼트 라인의 벡터 버퍼에 구비되는 제1 뱅크 및 제2 뱅크에 대하여, 상기 제1 뱅크에 짝수인 상기 벡터 요소를 분배하고 상기 제2 뱅크에 홀수인 상기 벡터 요소를 분배하는 단계;

각 프로세싱 엘리먼트 라인에 구비되는 매트리스 버퍼로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 전송하는 단계; 및

각 프로세싱 엘리먼트 라인에 포함된 두개의 연산 유닛을 이용하여 각 프로세싱 엘리먼트 라인 단위로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 연산하는 단계를 포함하는, 연산 가속 방

법.

청구항 6

컴퓨터를 이용하여 제1항 내지 제5항 중 어느 한 항의 방법을 실행시키기 위하여 기록매체에 저장된 컴퓨터 프로그램.

청구항 7

메모리에 저장된 희소행렬의 연산을 가속하는 장치에 있어서,

희소행렬의 요소들에 대하여 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 상기 메모리의 대역폭에 대응하는 숫자로 구비되는 복수의 프로세싱 엘리먼트 라인 단위로 정렬하는 전처리 모듈; 및

복수의 프로세싱 엘리먼트 라인을 포함하고, 전처리된 희소행렬의 요소들에 대하여 각 프로세싱 엘리먼트 라인에 구비되는 두개의 연산 유닛을 이용하여 각 프로세싱 엘리먼트 라인 단위로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 연산하는 연산 모듈;

을 포함하는, 연산 가속 장치.

청구항 8

제7항에 있어서,

상기 전처리 모듈은, 상기 프로세싱 엘리먼트 라인이 구비되는 연산 모듈의 버퍼 사이즈를 기초로 상기 희소행렬을 블록 단위로 나누어 복수의 블록을 설정하고, 상기 희소행렬의 각 블록 단위로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 각 프로세싱 엘리먼트 라인에 할당하여 정렬하는, 연산 가속 장치.

청구항 9

제8항에 있어서,

상기 전처리 모듈은, 상기 프로세싱 엘리먼트 라인에 구비되는 벡터 버퍼의 사이즈에 대응하는 열의 개수를 갖는 블록 단위로 복수의 블록을 설정하는, 연산 가속 장치.

청구항 10

제9항에 있어서,

상기 전처리 모듈은, 상기 벡터 버퍼의 사이즈를 기초로 설정되는 벡터 요소에 매핑하여 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 정렬하는, 연산 가속 장치.

청구항 11

제10항에 있어서,

상기 연산 모듈은, 각 프로세싱 엘리먼트 라인의 벡터 버퍼에 구비되는 제1 뱅크 및 제2 뱅크에 대하여, 상기 제1 뱅크에 짝수인 상기 벡터 요소를 분배하고 상기 제2 뱅크에 홀수인 상기 벡터 요소를 분배하고,

각 프로세싱 엘리먼트 라인에 구비되는 매트리스 버퍼로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 전송하고,

각 프로세싱 엘리먼트 라인에 포함된 두개의 연산 유닛을 이용하여 각 프로세싱 엘리먼트 라인 단위로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 연산하는, 연산 가속 장치.

발명의 설명

기 술 분 야

[0001] 본 발명의 실시예들은, 희소 행렬-벡터 곱셈의 연산에 있어서 연산 가속 방법, 연산 가속 장치 및 상기 방법을 실행시키기 위하여 기록매체에 저장된 컴퓨터 프로그램에 관한 것이다.

배 경 기 술

[0002] 최근 많은 국가에서 과학, 산업, 국가 안보와 같은 다양한 분야를 강화하기 위해 고성능 컴퓨팅에 많은 관심을 가지고 있다. 고성능 컴퓨팅에는 기존의 과학 컴퓨팅뿐만 아니라 인공지능 학습 및 데이터 분석 방법의 응용 프로그램도 포함된다.

[0003] 희소 행렬(Sparse matrix)은 행렬의 값이 대부분 0인 행렬을 의미한다. 희소 행렬 연산은 기존의 조밀 행렬 연산과 다르게 메모리 접근량 대비 연산량이 적어 기존의 고성능 컴퓨팅 시스템에서 성능이 저하되고 있어 최근 고성능 컴퓨팅과 인공지능 어플리케이션에서 희소 행렬 연산의 중요성이 대두되고 있다. 희소 행렬 연산 중에서도 희소 행렬-벡터 곱셈(Sparse matrix-vector multiplication)은 기존의 고성능 과학 연산부터 최근 주목 받고있는 머신 러닝과 딥 러닝에서도 중요한 연산 중 하나이다.

[0004] 현재까지 DRAM 대역폭이 증가하고 있는 추세이며 앞으로도 기술의 발전으로 계속 증가할 것으로 예측된다. 하지만 기존에 제안된 희소행렬-벡터 곱셈 가속기는 주어진 하드웨어플랫폼의 DRAM 대역폭에만 최적화 되어있으며 사용하는 하드웨어 플랫폼의 DRAM 대역폭이 증가하였을때 이에 대응할 수 없는 구조로 성능 및 효율성이 떨어진다.

발명의 내용

해결하려는 과제

[0005] 본 발명이 해결하고자 하는 과제는 전술한 문제점을 극복하기 위한 것으로서, 희소 행렬-벡터 곱셈 연산에 있어서 증가하는 DRAM 대역폭에 따라 확장성을 갖는 연산 가속 방법, 컴퓨터 프로그램 및 장치를 제공하는 것이다. 그러나 이러한 과제는 예시적인 것으로, 이에 의해 본 발명의 범위가 한정되는 것은 아니다.

과제의 해결 수단

[0006] 상술한 기술적 과제들을 달성하기 위한 기술적 수단으로서, 본 발명의 일 측면에 따른 연산 가속 방법은, 컴퓨팅 장치에 의해 수행되는 희소행렬의 연산을 가속하는 방법에 있어서, 메모리에 저장된 희소행렬의 요소들에 대하여 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 상기 메모리의 대역폭에 대응하는 숫자로 구비되는 복수의 프로세싱 엘리먼트 라인 단위로 정렬하는 전처리 단계, 및 상기 전처리 단계에서 전처리된 희소행렬의 요소들에 대하여 각 프로세싱 엘리먼트 라인에 구비되는 두개의 연산 유닛을 이용하여 각 프로세싱 엘리먼트 라인 단위로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 연산하는 연산 단계를 포함한다.

[0007] 일 예에 따르면, 상기 전처리 단계는, 상기 프로세싱 엘리먼트 라인이 구비되는 연산 모듈의 버퍼 사이즈를 기초로 상기 희소행렬을 블록 단위로 나누어 복수의 블록을 설정하는 단계, 및 상기 희소행렬의 각 블록 단위로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 각 프로세싱 엘리먼트 라인에 할당하여 정렬하는 단계를 포함할 수 있다.

[0008] 다른 예에 따르면, 상기 복수의 블록을 설정하는 단계는, 상기 프로세싱 엘리먼트 라인에 구비되는 벡터 버퍼의 사이즈에 대응하는 열의 개수를 갖는 블록 단위로 복수의 블록을 설정하는 단계를 포함할 수 있다.

[0009] 또 다른 예에 따르면, 상기 각 프로세싱 엘리먼트 라인에 할당하여 정렬하는 단계는, 상기 벡터 버퍼의 사이즈를 기초로 설정되는 벡터 요소에 매핑하여 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 정렬하는 단계를 포함할 수 있다.

[0010] 또 다른 예에 따르면, 상기 연산 단계는, 각 프로세싱 엘리먼트 라인의 벡터 버퍼에 구비되는 제1 뱅크 및 제2 뱅크에 대하여, 상기 제1 뱅크에 짝수인 상기 벡터 요소를 분배하고 상기 제2 뱅크에 홀수인 상기 벡터 요소를 분배하는 단계, 각 프로세싱 엘리먼트 라인에 구비되는 매트리스 버퍼로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 전송하는 단계, 및 각 프로세싱 엘리먼트 라인에 포함된 두개의 연산 유닛을 이용하여 각 프로세싱 엘리먼트 라인 단위로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 연산하는 단계를 포함할 수 있다.

[0011] 상술한 기술적 과제들을 달성하기 위한 기술적 수단으로서, 컴퓨팅 장치를 이용하여 상술한 방법을 실행시키기 위하여 기록매체에 저장된 컴퓨터 프로그램이 제공된다.

[0012] 상술한 기술적 과제들을 달성하기 위한 기술적 수단으로서, 본 발명의 일 측면에 따른 연산 가속 장치는, 메모리에 저장된 희소행렬의 연산을 가속하는 장치에 있어서, 희소행렬의 요소들에 대하여 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 상기 메모리의 대역폭에 대응하는 숫자로 구비되는 복수의 프로세싱 엘리먼트 라인 단위로 정렬하는 전처리 모듈, 및 복수의 프로세싱 엘리먼트 라인을 포함하고, 전처리된 희소행렬의 요소들에 대하여 각 프로세싱 엘리먼트 라인에 구비되는 두개의 연산 유닛을 이용하여 각 프로세싱 엘리먼트 라인 단위로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 연산하는 연산 모듈을 포함한다.

[0013] 일 예에 따르면, 상기 전처리 모듈은, 상기 프로세싱 엘리먼트 라인이 구비되는 연산 모듈의 버퍼 사이즈를 기초로 상기 희소행렬을 블록 단위로 나누어 복수의 블록을 설정하고, 상기 희소행렬의 각 블록 단위로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 각 프로세싱 엘리먼트 라인에 할당하여 정렬할 수 있다.

[0014] 다른 예에 따르면, 상기 전처리 모듈은, 상기 프로세싱 엘리먼트 라인에 구비되는 벡터 버퍼의 사이즈에 대응하는 열의 개수를 갖는 블록 단위로 복수의 블록을 설정할 수 있다.

[0015] 또 다른 예에 따르면, 상기 전처리 모듈은, 상기 벡터 버퍼의 사이즈를 기초로 설정되는 벡터 요소에 매핑하여 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 정렬할 수 있다.

[0016] 또 다른 예에 따르면, 상기 연산 모듈은, 각 프로세싱 엘리먼트 라인의 벡터 버퍼에 구비되는 제1 뱅크 및 제2 뱅크에 대하여, 상기 제1 뱅크에 짝수인 상기 벡터 요소를 분배하고 상기 제2 뱅크에 홀수인 상기 벡터 요소를 분배하고, 각 프로세싱 엘리먼트 라인에 구비되는 매트리스 버퍼로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 전송하고, 각 프로세싱 엘리먼트 라인에 포함된 두개의 연산 유닛을 이용하여 각 프로세싱 엘리먼트 라인 단위로 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 연산할 수 있다.

[0017] 전술한 것 외의 다른 측면, 특징, 이점은 이하의 발명을 실시하기 위한 구체적인 내용, 청구범위 및 도면으로부터 명확해질 것이다.

발명의 효과

[0018] 상기한 바와 같이 이루어진 본 발명의 일 실시예에 따르면, 희소 행렬-벡터 곱셈 연산에 있어서 증가하는 DRAM 대역폭에 따라 확장성을 가지면서도 동일한 성능을 유지할 수 있는 연산 가속 장치, 연산 가속 방법 및 상기 방법을 실행시키기 위하여 기록매체에 저장된 컴퓨터 프로그램을 구현할 수 있다. 물론 이러한 효과에 의해 본 발명의 범위가 한정되는 것은 아니다.

도면의 간단한 설명

[0019] 도 1은 본 발명의 일 실시예에 따른 연산 가속 장치가 메모리에 액세스하는 구성을 개략적으로 도시하는 도면이다.

도 2 내지 도 4는 본 발명의 일 실시예에 따른 연산 가속 방법을 설명하기 위한 순서도이다.

도 5는 본 발명의 일 실시예에 따른 연산 가속 장치의 구성을 개략적으로 도시하는 도면이다.

도 6은 본 발명의 일 실시예에 따른 희소행렬을 전처리하는 방법을 설명하기 위한 도면이다.

도 7은 본 발명의 일 실시예에 따른 전처리된 희소행렬을 설명하기 위한 도면이다.

도 8은 본 발명의 일 실시예에 따른 전처리된 희소행렬을 연산하는 방법을 설명하기 위한 도면이다.

도 9 내지 도 11은 본 발명의 일 실시예에 따른 연산 가속 장치의 효과를 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0020] 아래에서는 첨부한 도면을 참조하여 본 개시가 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 다양한 실시예들을 상세히 설명한다. 그러나 본 개시의 기술적 사상은 다양한 형태로 변형되어 구현될 수 있으므로 본 명세서에서 설명하는 실시예들로 제한되지 않는다. 본 명세서에 개시된 실시예들을 설명함에 있어서 관련된 공지 기술을 구체적으로 설명하는 것이 본 개시의 기술적 사상의 요지를 흐릴 수 있다고 판단되

는 경우 그 공지 기술에 대한 구체적인 설명을 생략한다. 동일하거나 유사한 구성요소는 동일한 참조 번호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다.

[0021] 본 명세서에서 어떤 요소가 다른 요소와 "연결"되어 있다고 기술될 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라 그 중간에 다른 요소를 사이에 두고 "간접적으로 연결"되어 있는 경우도 포함한다. 어떤 요소가 다른 요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 요소 외에 또 다른 요소를 배제하는 것이 아니라 또 다른 요소를 더 포함할 수 있는 것을 의미한다.

[0022] 일부 실시예들은 기능적인 블록 구성들 및 다양한 처리 단계들로 설명될 수 있다. 이러한 기능 블록들의 일부 또는 전부는 특정 기능을 실행하는 다양한 개수의 하드웨어 및/또는 소프트웨어 구성들로 구현될 수 있다. 예를 들어, 본 개시의 기능 블록들은 하나 이상의 마이크로프로세서들에 의해 구현되거나, 소정의 기능을 위한 회로 구성들에 의해 구현될 수 있다. 본 개시의 기능 블록들은 다양한 프로그래밍 또는 스크립팅 언어로 구현될 수 있다. 본 개시의 기능 블록들은 하나 이상의 프로세서들에서 실행되는 알고리즘으로 구현될 수 있다. 본 개시의 기능 블록이 수행하는 기능은 복수의 기능 블록에 의해 수행되거나, 본 개시에서 복수의 기능 블록이 수행하는 기능들은 하나의 기능 블록에 의해 수행될 수도 있다. 또한, 본 개시는 전자적인 환경 설정, 신호 처리, 및/또는 데이터 처리 등을 위하여 종래 기술을 채용할 수 있다.

[0023] 도 1은 본 발명의 일 실시예에 따른 연산 가속 장치가 메모리에 액세스하는 구성을 개략적으로 도시하는 도면이다.

[0024] 도 1을 참조하면, 본 발명의 일 실시예에 따른 연산 가속 장치(1000)는 메모리(50)에 액세스할 수 있다. 예를 들어, 연산 가속 장치(1000)는 메모리(50)에 액세스하여 데이터의 읽기나 쓰기를 할 수 있다. 예컨대, 메모리(50)는 연산 가속 장치(1000)가 처리하는 데이터를 일시적 또는 영구적으로 저장할 수 있다. 예컨대, 메모리(50)는 RAM(Random Access Memory), ROM(Read Only Memory) 및 디스크 드라이브와 같은 비소멸성 대용량 기록 장치(Permanent Mass Storage Device)를 포함할 수 있으나, 본 발명이 이에 한정되는 것은 아니다.

[0025] 도 1에 도시된 바와 같이, 연산 가속 장치(1000)는 연산 모듈(100) 및 전처리 모듈(200)을 포함한다. 하지만 본 발명은 이에 한정되지 않으며, 연산 가속 장치(1000)는 다른 구성요소를 더 포함할 수도 있고 일부 구성요소가 생략될 수도 있다. 연산 가속 장치(1000)의 일부 구성요소는 복수의 장치로 분리될 수도 있고, 복수개의 구성요소들이 하나의 장치로 병합될 수도 있다.

[0026] 연산 가속 장치(1000)와 이에 포함된 모듈들은, 전적으로 하드웨어일 수 있고, 또는 부분적으로 하드웨어이고 부분적으로 소프트웨어인 측면을 가질 수 있다. 예컨대, 연산 가속 장치(1000)와 이에 포함된 각 모듈은 특정 형식 및 내용의 데이터를 처리하거나 전자통신 방식으로 주고받기 위한 하드웨어 및 이에 관련된 소프트웨어를 통칭할 수 있다. 본 명세서에서 "모듈" 또는 "장치" 등의 용어는 하드웨어 및 해당 하드웨어에 의해 구동되는 소프트웨어의 조합을 지칭하는 것으로 해석될 수 있다. 예컨대, 하드웨어는 CPU 또는 다른 프로세서(processor)를 포함하는 데이터 처리 기기일 수 있다. 또한, 하드웨어에 의해 구동되는 소프트웨어는 실행중인 프로세스, 객체(object), 실행파일(executable file), 실행 스레드(thread of execution), 프로그램(program) 등을 지칭할 수 있다.

[0027] 또한, 연산 가속 장치(1000)를 구성하는 각각의 요소는 반드시 서로 물리적으로 구분되는 별개의 장치를 지칭하는 것으로 한정되는 것은 아니다. 즉, 연산 가속 장치(1000)에 의해 수행되는 동작에 따라 기능적으로 구분한 것일 뿐, 반드시 서로 독립적으로 구분되어야 하는 것이 아니다. 물론, 실시예에 따라서는 이들은 서로 물리적으로 구분되는 별개의 장치로 구현되는 것도 가능하다. 이하에서는 상술한 것과 같은 구성을 취할 수 있는 연산 가속 장치(1000)의 각 모듈들의 동작 등에 대해, 연산 가속 방법과 함께 설명한다.

[0028] 도 2 내지 도 4는 본 발명의 일 실시예에 따른 연산 가속 방법을 설명하기 위한 순서도이다.

[0029] 도 2에 도시된 것과 같이, 본 실시예에 따른 연산 가속 방법은 전처리 단계(S210) 및 연산 단계(S220)를 거친다. 예컨대, 본 발명의 일 실시예에 따른 전처리 단계(S210)는 도 1의 전처리 모듈(200)에 의해 수행될 수 있다. 또한, 본 발명의 일 실시예에 따른 연산 단계(S220)는 도 1의 연산 모듈(100)에 의해 수행될 수 있다.

[0030] 본 발명의 일 실시예에 따른 전처리 단계(S210)에서 전처리 모듈(200)은 메모리에 저장된 희소행렬의 요소들에 대하여 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 메모리의 대역폭에 대응하는 숫자로 구비되는 복수의 프로세싱 엘리먼트 라인 단위로 정렬할 수 있다.

[0031] 도 3을 함께 참조하면, 본 발명의 일 실시예에 따른 전처리 단계(S210)는 프로세싱 엘리먼트 라인이 구비되는

연산 모듈의 버퍼 사이즈를 기초로 희소행렬을 블록 단위로 나누어 복수의 블록을 설정하는 단계(S310), 및 희소행렬의 각 블록 단위로 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 각 프로세싱 엘리먼트 라인에 할당하여 정렬하는 단계(S320)를 거칠 수 있다.

[0032] 본 발명의 일 실시예에 따른 전처리 모듈(200)은 프로세싱 엘리먼트 라인에 구비되는 벡터 버퍼의 사이즈에 대응하는 열의 개수를 갖는 블록 단위로 복수의 블록을 설정할 수 있다.

[0033] 본 발명의 일 실시예에 따른 전처리 모듈(200)은 벡터 버퍼의 사이즈를 기초로 설정되는 벡터 요소에 매핑하여 상기 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 정렬할 수 있다.

[0034] 본 발명의 일 실시예에 따른 연산 단계(S220)에서 연산 모듈(100)은 전처리 단계에서 전처리된 희소행렬의 요소들에 대하여 각 프로세싱 엘리먼트 라인에 구비되는 두개의 연산 유닛을 이용하여 각 프로세싱 엘리먼트 라인 단위로 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 연산할 수 있다.

[0035] 도 4를 함께 참조하면, 본 발명의 일 실시예에 따른 연산 단계(S220)는 각 프로세싱 엘리먼트 라인의 벡터 버퍼에 구비되는 제1 뱅크 및 제2 뱅크에 대하여, 제1 뱅크에 짝수인 벡터 요소를 분배하고 제2 뱅크에 홀수인 벡터 요소를 분배하는 단계(S410), 각 프로세싱 엘리먼트 라인에 구비되는 매트리스 버퍼로 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 전송하는 단계(S420), 및 각 프로세싱 엘리먼트 라인에 포함된 두개의 연산 유닛을 이용하여 각 프로세싱 엘리먼트 라인 단위로 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 연산하는 단계(S430)를 거칠 수 있다.

[0036] 도 5는 본 발명의 일 실시예에 따른 연산 가속 장치의 구성을 개략적으로 도시하는 도면이다.

[0037] 도 5를 참조하면, 본 발명의 일 실시예에 따른 연산 가속 장치는 메모리(50)에 저장된 희소행렬의 연산을 가속하는 장치일 수 있다. 본 발명의 일 실시예에 따른 연산 가속 장치의 연산 모듈(100)은 복수의 프로세싱 엘리먼트 라인을 포함하고, 전처리된 희소행렬의 요소들에 대하여 각 프로세싱 엘리먼트 라인에 구비되는 두개의 연산 유닛을 이용하여 각 프로세싱 엘리먼트 라인 단위로 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 연산할 수 있다.

[0038] 본 발명의 일 실시예에 따른 연산 모듈(100)은 복수의 프로세싱 엘리먼트 라인(110)을 포함할 수 있다. 예를 들어, 프로세싱 엘리먼트 라인(110)은 메모리(50)의 대역폭에 대응하는 숫자로 구비될 수 있다. 예컨대, 프로세싱 엘리먼트 라인(110)은 PE Line으로 지칭될 수 있다.

[0039] 프로세싱 엘리먼트 라인(110)은 벡터 버퍼(111) 및 매트릭스 버퍼(112)를 포함할 수 있다. 예를 들어, 도 5에 도시된 바와 같이, 벡터 버퍼(111)는 제1 뱅크(B0) 및 제2 뱅크(B1)를 포함할 수 있다. 또한, 프로세싱 엘리먼트 라인(110)은 연산 유닛(MAC)을 포함할 수 있다. 예컨대, 연산 유닛(MAC(Multiply-and-Accumulation))은 벡터 연산을 수행할 수 있다.

[0040] 본 발명의 일 실시예에 따른 각 프로세싱 엘리먼트 라인(110)은 두개의 뱅크(B0, B1)를 포함할 수 있다. 또한, 본 발명의 일 실시예에 따른 각 프로세싱 엘리먼트 라인(110)은 두개의 연산 유닛(MAC)을 포함할 수 있다.

[0041] 도 6은 본 발명의 일 실시예에 따른 희소행렬을 전처리하는 방법을 설명하기 위한 도면이다.

[0042] 도 6을 참조하면, 본 발명의 일 실시예에 따라 프로세싱 엘리먼트 라인이 2개인 경우의 전처리 과정이 도시되어 있다.

[0043] 전처리 모듈(200)은 메모리(50)에 저장된 희소행렬을 블록 단위로 나눌 수 있다. 전처리 모듈(200)은 연산 가속 장치 내 on-chip buffer (벡터 버퍼(Vector Buffer), PSUM buffer)의 사이즈에 맞추어 희소행렬을 블록 단위로 복수의 블록으로 나눌 수 있다. 예컨대, 전처리 모듈(200)은 벡터 버퍼의 사이즈를 기초로 단위 블록의 행 또는 열의 크기 또는 개수가 벡터 버퍼의 사이즈에 대응하는 크기 또는 개수를 갖도록 단위 블록을 나눌 수 있다. 예를 들어, 도 6에 도시된 바와 같이, 희소행렬은 6개의 블록(블록0 내지 블록5)으로 나뉠 수 있다.

[0044] 전처리 모듈(200)은 단위 블록 단위로 전처리 과정을 진행할 수 있다. 예를 들어, 도 6에 도시된 바와 같이, 블록1에 대한 전처리 과정이 도시되어 있다.

[0045] 먼저, 전처리 모듈(200)은 하나의 단위 블록을 열 단위로 각 프로세싱 엘리먼트 라인(PE Line)에 균등하게 할당할 수 있다. 이어서, 전처리 모듈(200)은 각 PE Line의 연산 유닛(MAC unit)에 단위 블록의 행 하나씩 할당할 수 있다. 이어서, 전처리 모듈(200)은 각 PE Line의 제1 연산 유닛(MAC0)에 할당된 행의 남은 non-zero 요소들 중 가장 첫번째 요소를 선택(예컨대, 가장 첫번째는 가장 작은 열을 의미)할 수 있다. 예컨대, 도 6에 도시된

바와 같이, a 요소 및 b 요소가 선택될 수 있다.

[0046] 이어서, 전처리 모듈(200)은 제2 연산 유닛(MAC1)에서 연산할 희소행렬의 non-zero 요소를 선택할 수 있다. 여기서, 선택 순서는 희소행렬 행의 오름차순의 순서로 선택될 수 있다. 예컨대, 각 PE Line 단위로 희소 행렬의 열 인덱스가 분배될 수 있다. 예를 들어, 도 6에 도시된 바와 같이, 8개의 열에 대하여, PE Line#0에 0 내지 3의 열 인덱스가 분배되고, PE Line#1에 0 내지 3의 열 인덱스가 분배될 수 있다. 전처리 과정을 마친 희소행렬은 메모리(50)에 저장될 수 있다.

[0047] 도 7은 본 발명의 일 실시예에 따른 전처리된 희소행렬을 설명하기 위한 도면이다.

[0048] 도 7을 참조하면, 본 발명의 일 실시예에 따라 전처리 후 메모리(50)에 저장되어 있는 전처리된 희소행렬의 일 예가 도시되어 있다.

[0049] 전처리 모듈(200)은 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 정렬할 수 있다. 예를 들어, 전처리 모듈(200)은 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 각 프로세싱 엘리먼트 라인에 할당하여 정렬할 수 있다. 예를 들어, 전처리 모듈(200)은 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 각 프로세싱 엘리먼트 라인의 연산 유닛에 할당하여 정렬할 수 있다. 예컨대, 도 7에 도시된 바와 같이, 희소행렬의 요소 e, e 요소의 열 인덱스 3, e 요소의 행 인덱스 1의 쌍이 PE Line1의 MAC1에 할당되어 정렬될 수 있다.

[0050] 전처리 모듈(200)은 벡터 버퍼의 사이즈를 기초로 설정되는 벡터 요소에 매핑하여 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 정렬할 수 있다. 예컨대, 도 7에 도시된 바와 같이, 벡터 요소 x0 내지 x3은 PE Line0에 매핑되어, PE Line0에 할당되는 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍이 벡터 요소 x0 내지 x3에 매핑되어 정렬될 수 있다.

[0051] 도 8은 본 발명의 일 실시예에 따른 전처리된 희소행렬을 연산하는 방법을 설명하기 위한 도면이다.

[0052] 도 8을 참조하면, 본 발명의 일 실시예에 따라 전처리된 희소행렬이 두개의 프로세싱 엘리먼트를 구비하는 연산 모듈에 의해 연산되는 과정이 도시되어 있다.

[0053] 연산 모듈(100)은 전처리된 희소행렬의 요소들에 대하여 각 프로세싱 엘리먼트 라인에 구비되는 두개의 연산 유닛을 이용하여 각 프로세싱 엘리먼트 라인 단위로 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 연산할 수 있다.

[0054] 연산 모듈(100)은 각 프로세싱 엘리먼트 라인의 벡터 버퍼에 구비되는 제1 뱅크 및 제2 뱅크에 대하여, 제1 뱅크에 짝수인 벡터 요소를 분배하고 제2 뱅크에 홀수인 상기 벡터 요소를 분배할 수 있다. 예컨대, 도 8에 도시된 바와 같이, PE Line1의 제1 뱅크(Bank0)에 짝수인 x6, x4 벡터 요소가 분배될 수 있다. 또한, PE Line1의 제2 뱅크(Bank1)에 홀수인 x7, x5 벡터 요소가 분배될 수 있다.

[0055] 연산 모듈(100)은 각 프로세싱 엘리먼트 라인에 구비되는 매트리스 버퍼로 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 전송할 수 있다.

[0056] 연산 모듈(100)은 각 프로세싱 엘리먼트 라인에 포함된 두개의 연산 유닛을 이용하여 각 프로세싱 엘리먼트 라인 단위로 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 연산할 수 있다.

[0057] 도 9 내지 도 11은 본 발명의 일 실시예에 따른 연산 가속 장치의 효과를 설명하기 위한 도면이다.

[0058] 먼저, 도 9를 참조하면, 본 발명의 일 실시예에 따라 전처리 단계가 수행된 경우와 전처리 단계가 수행되지 않은 경우를 비교한 도면이 도시되어 있다.

[0059] 본 발명에 따르면 전처리 단계가 수행됨에 따라 효율적인 연산이 이루어질 수 있다. 예컨대, 도 9에 도시된 바와 같이, 연산 사이클이 6에서 4로 감소될 수 있다.

[0060] 각 뱅크의 포트는 두개이나 포트 하나는 메모리에서 읽어온 데이터를 쓰기 위해 사용되고 나머지 하나는 연산할 때 벡터 요소를 읽기 위해 사용되므로, 결국 뱅크 하나에서는 한 번에 한 개의 벡터 요소를 읽어 올 수 있다.

[0061] 전처리 단계가 수행되지 않은 경우, cycle1에서는 a를 위해 벡터요소 x0이 필요하고, b를 위해서 벡터요소 x2가 필요한 상황이나, x0와 x2는 모두 bank 0에 저장이 되어있으므로 한 사이클에 모두 읽어올 수 없다. 결국, 두 사이클에 걸쳐서 x0, x2를 읽어와야 하며, 이러한 상황을 bank conflict라 칭할 수 있다.

[0062] 전처리 단계가 수행되는 경우, 연산 과정에서 bank conflict가 발생하지 않으므로 효율적인 연산이 이루어질 수

있다.

[0063] 도 10을 참조하면, 본 발명에 따라 전처리된 희소행렬과 전처리되지 않은 희소행렬의 행렬 압축 결과를 비교한 결과가 도시되어 있다.

[0064] 도 10에 도시된 바와 같이, 본 발명의 기술을 적용하였을 때 전처리 과정은 행렬 압축 결과에 큰 영향을 끼치지 않았음을 확인할 수 있다.

[0065] 도 11을 참조하면, 본 발명에 따라 메모리 대역폭에 따라 프로세싱 엘리먼트 라인(PE Line)을 복수개로 조정하였을 때의 결과가 도시되어 있다.

[0066] 도 11에 도시된 바와 같이, PE Line을 여러개로 조정하였을 때 이론적 성능에 거의 도달하였으며 종래 기술에 비하여 최대 28.3% 성능 개선을 이루었음을 확인할 수 있다.

[0067] 본 발명에 따르면, 본 발명은 (1) DRAM 대역폭에 확장성을 가지는 희소행렬-벡터 가속기 하드웨어 구조를 갖으며, (2) 성능 저하를 막기위한 전처리 과정을 포함하고 있다.

[0068] DRAM 대역폭에 확장성을 가지는 희소행렬-벡터 가속기 하드웨어 구조는, PE line 단위로 구성될 수 있다. 또한, 각 PE Line은 두개의 MAC(Multiply and Accumulation) Unit을 포함하고 있어 동시에 두 개의 0이 아닌 행렬 요소 처리가 가능하다. 또한, DRAM 대역폭에 따라 PE Line의 개수가 조절될 수 있다. 또한, PE Line내에 MAC unit 개수가 2개이고 vector buffer가 2개의 bank로 구성되었기 때문에 On-chip SRAM port 개수와 PE의 개수가 무관해진다. 따라서 기존의 기술과 같이 Vector buffer에 data를 복제할 필요가 없어진다.

[0069] 성능 저하를 막기위한 전처리 과정은, 희소행렬의 0이 아닌 요소의 random 분포에 의한 bank-conflict 문제와, MAC 유닛에서 동시에 같은 accumulator에 접근하는 Read-After-Write (RAW) dependency를 해결할 수 있다. 또한, 전처리 과정은 동시에 bank에 접근하지 않도록 서로 다른 bank가 요구되는 0이 아닌 행렬 요소를 찾고 페어링할 수 있다. 또한, RAW dependency를 피하기 위해 전처리 과정 중 행렬의 각 행들은 서로 다른 MAC unit에 할당될 수 있다. 또한, 전처리 과정은 PE Line 단위로 이루어질 수 있다. 또한, 각 PE Line은 두개의 MAC을 포함하고 PE Line끼리는 독립적이기 때문에 PE 개수를 늘린다고 해서 기존 기술과 같이 경우의 수가 줄어들지 않아 성능 저하를 막을 수 있다.

[0070] 이상 설명된 다양한 실시예들은 예시적이며, 서로 구별되어 독립적으로 실시되어야 하는 것은 아니다. 본 명세서에서 설명된 실시예들은 서로 조합된 형태로 실시될 수 있다.

[0071] 이상 설명된 다양한 실시예들은 컴퓨터 상에서 다양한 구성요소를 통하여 실행될 수 있는 컴퓨터 프로그램의 형태로 구현될 수 있으며, 이와 같은 컴퓨터 프로그램은 컴퓨터로 판독 가능한 매체에 기록될 수 있다. 이때, 매체는 컴퓨터로 실행 가능한 프로그램을 계속 저장하거나, 실행 또는 다운로드를 위해 임시 저장하는 것일 수도 있다. 또한, 매체는 단일 또는 수개 하드웨어가 결합된 형태의 다양한 기록수단 또는 저장수단일 수 있는데, 어떤 컴퓨터 시스템에 직접 접속되는 매체에 한정되지 않고, 네트워크 상에 분산 존재하는 것일 수도 있다. 매체의 예시로는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM 및 DVD와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical medium), 및 ROM, RAM, 플래시 메모리 등을 포함하여 프로그램 명령어가 저장되도록 구성된 것이 있을 수 있다. 또한, 다른 매체의 예시로, 애플리케이션을 유통하는 앱 스토어나 기타 다양한 소프트웨어를 공급 내지 유통하는 사이트, 서버 등에서 관리하는 기록매체 내지 저장매체도 들 수 있다.

[0072] 본 명세서에서, "부", "모듈" 등은 프로세서 또는 회로와 같은 하드웨어 구성(hardware component), 및/또는 프로세서와 같은 하드웨어 구성에 의해 실행되는 소프트웨어 구성(software component)일 수 있다. 예를 들면, "부", "모듈" 등은 소프트웨어 구성 요소들, 객체 지향 소프트웨어 구성 요소들, 클래스 구성 요소들 및 태스크 구성 요소들과 같은 구성 요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 및 변수들에 의해 구현될 수 있다.

[0073] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

[0074] 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

[0075] 50: 메모리

1000: 연산 가속 장치

100: 연산 모듈

110: 프로세싱 엘리먼트 라인

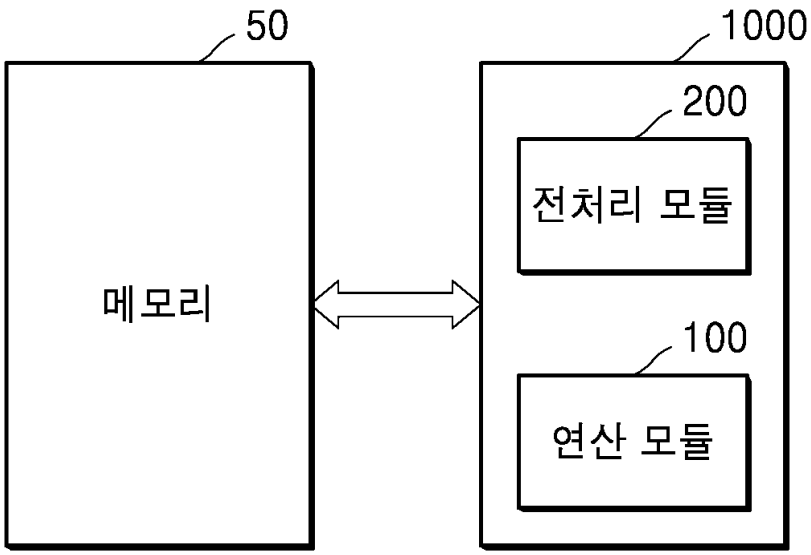
111: 벡터 버퍼

112: 매트릭스 버퍼

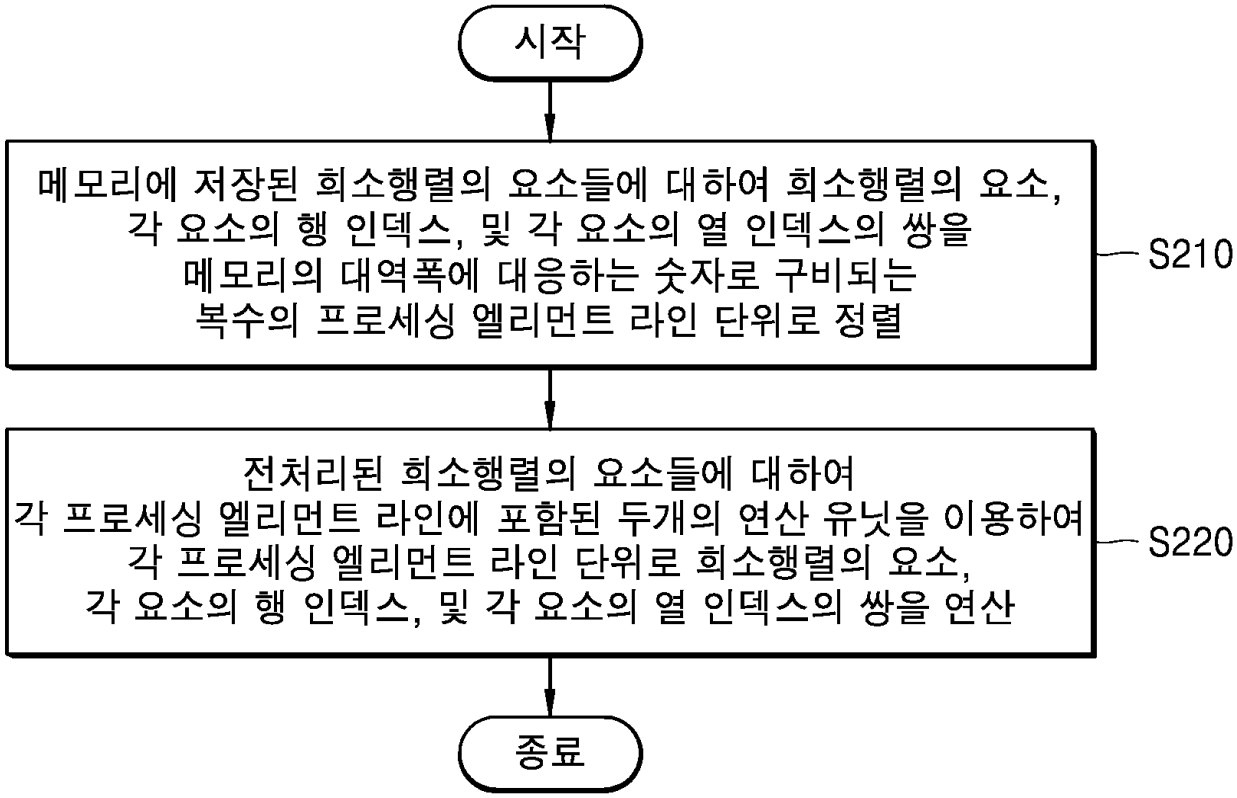
200: 전처리 모듈

도면

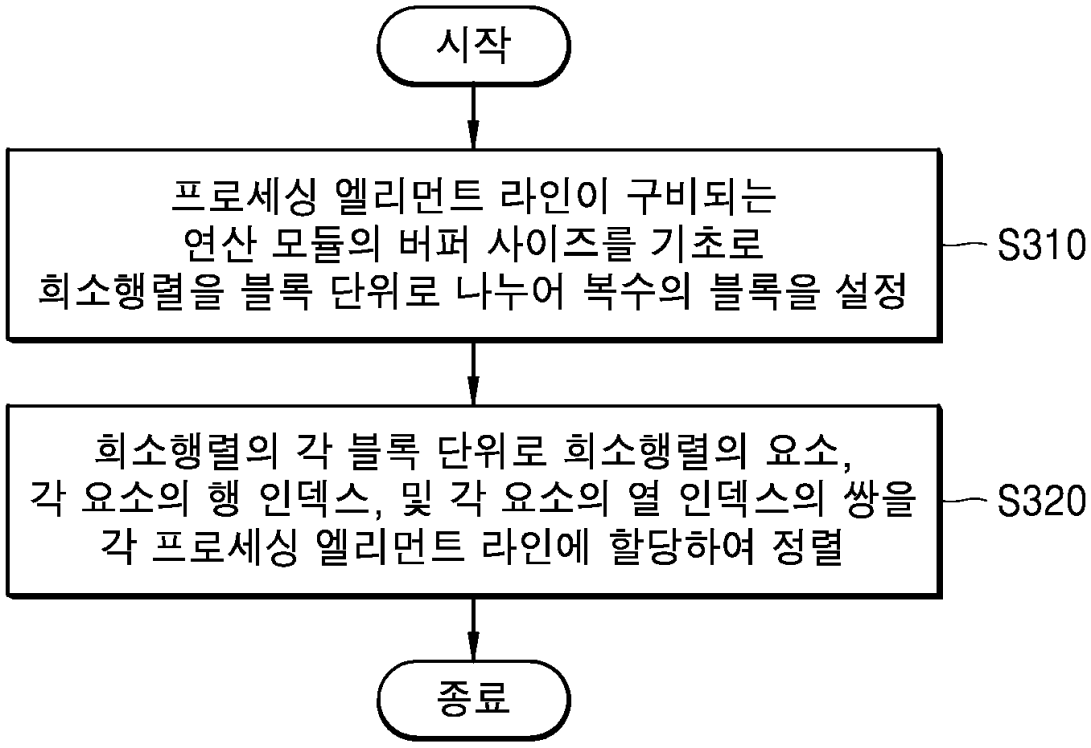
도면1



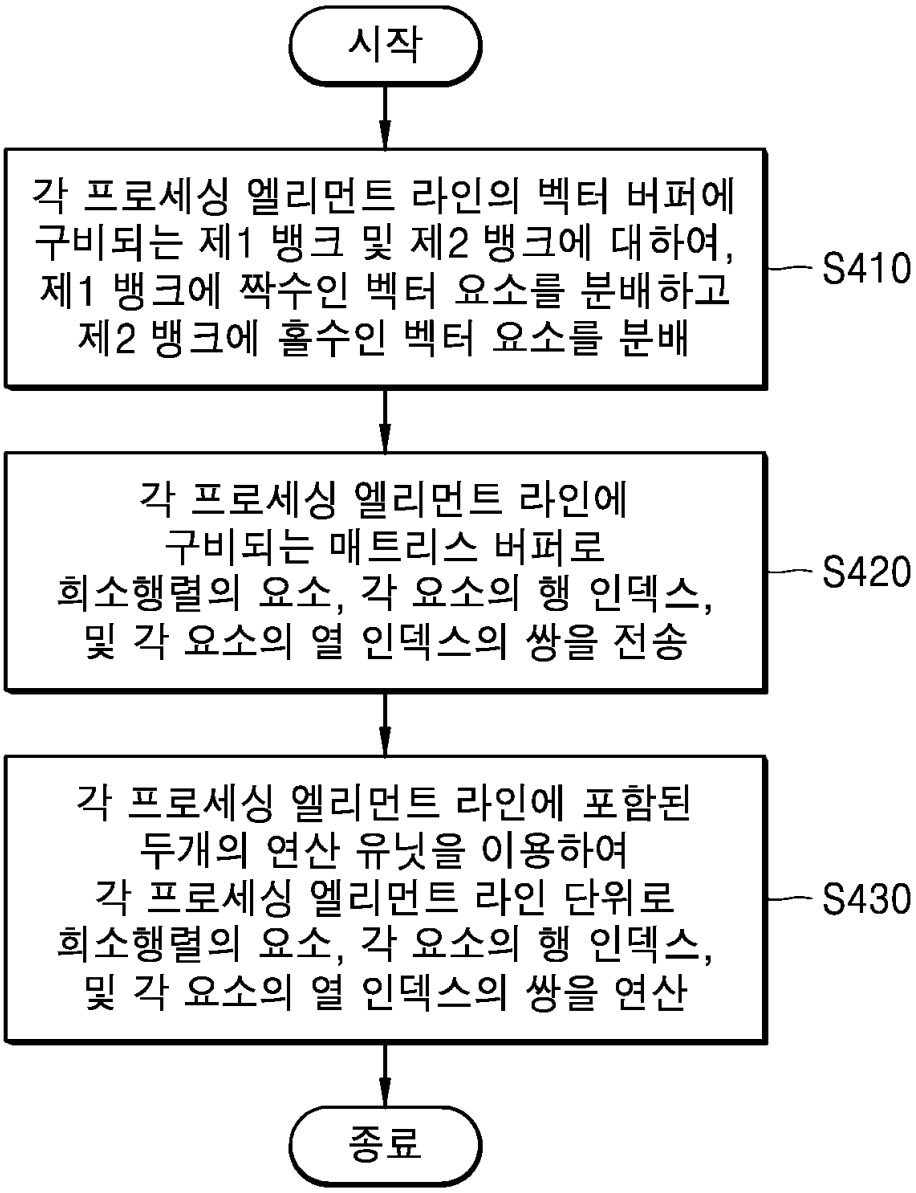
도면2


시작
메모리에 저장된 희소행렬의 요소들에 대하여 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 메모리의 대역폭에 대응하는 숫자로 구비되는 복수의 프로세싱 엘리먼트 라인 단위로 정렬
S210
전처리된 희소행렬의 요소들에 대하여 각 프로세싱 엘리먼트 라인에 포함된 두개의 연산 유닛을 이용하여 각 프로세싱 엘리먼트 라인 단위로 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 연산
S220
종료

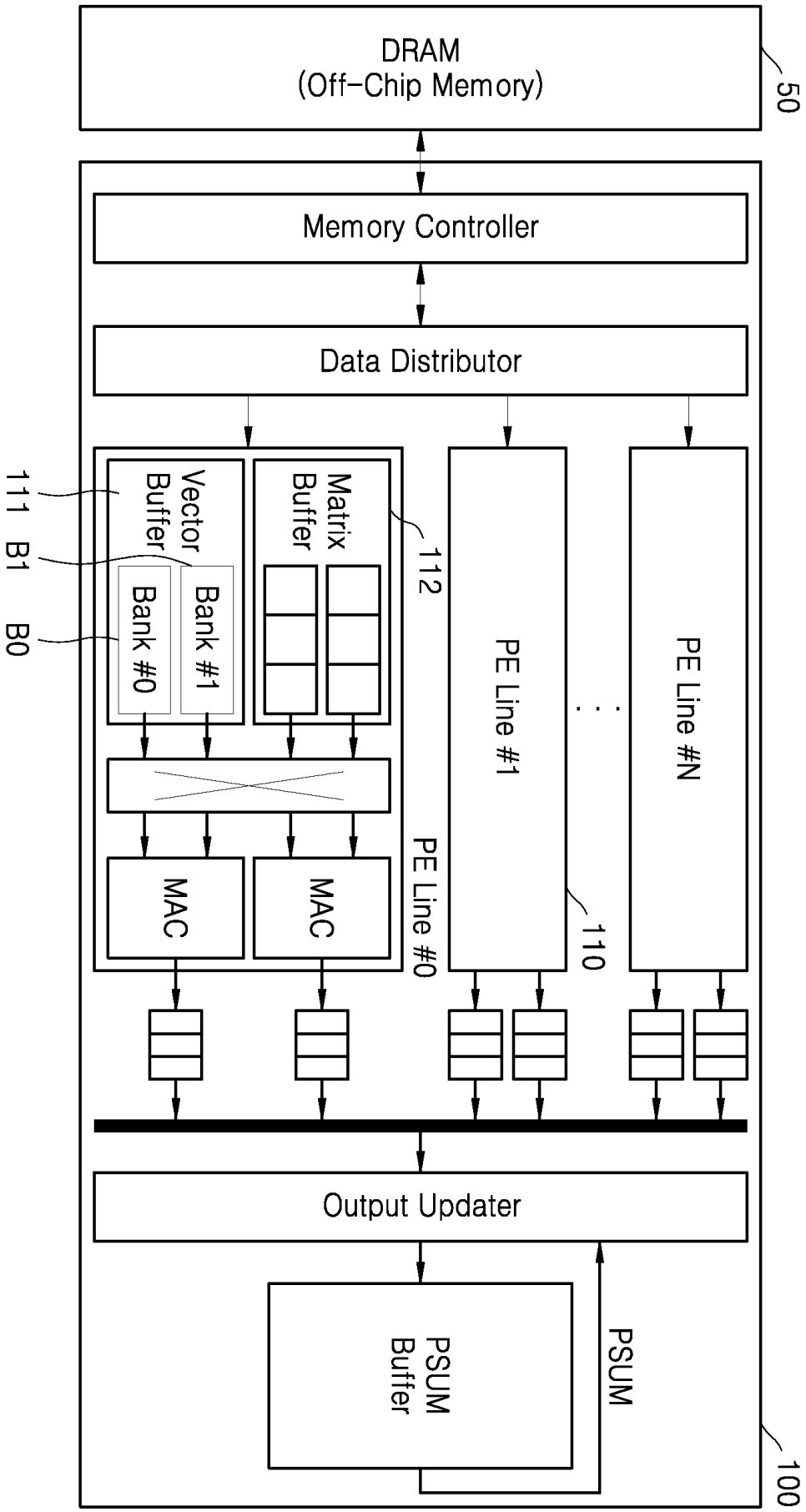
도면3


시작
프로세싱 엘리먼트 라인이 구비되는 연산 모듈의 버퍼 사이즈를 기초로 희소행렬을 블록 단위로 나누어 복수의 블록을 설정
S310
희소행렬의 각 블록 단위로 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 각 프로세싱 엘리먼트 라인에 할당하여 정렬
S320
종료

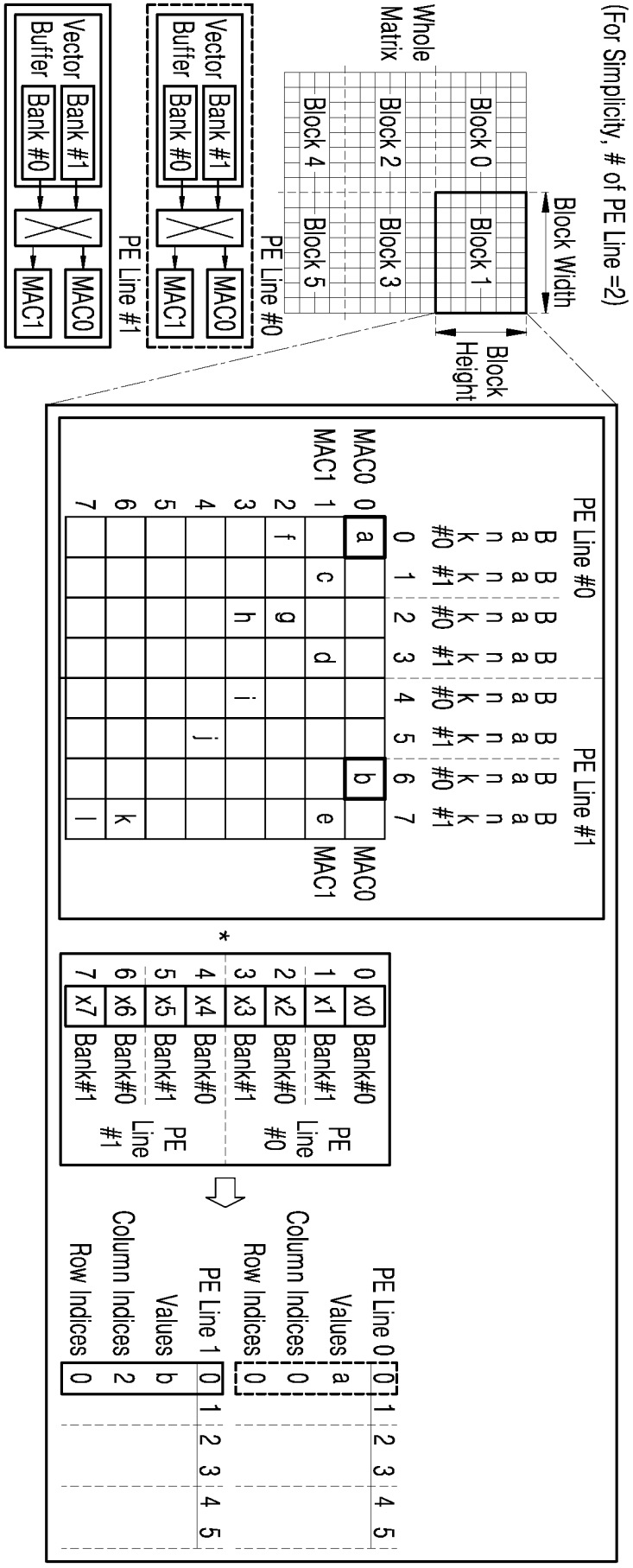
도면4


시작
각 프로세싱 엘리먼트 라인의 벡터 버퍼에 구비되는 제1 뱅크 및 제2 뱅크에 대하여, 제1 뱅크에 짝수인 벡터 요소를 분배하고 제2 뱅크에 홀수인 벡터 요소를 분배
S410
각 프로세싱 엘리먼트 라인에 구비되는 매트리스 버퍼로 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 전송
S420
각 프로세싱 엘리먼트 라인에 포함된 두개의 연산 유닛을 이용하여 각 프로세싱 엘리먼트 라인 단위로 희소행렬의 요소, 각 요소의 행 인덱스, 및 각 요소의 열 인덱스의 쌍을 연산
S430
종료

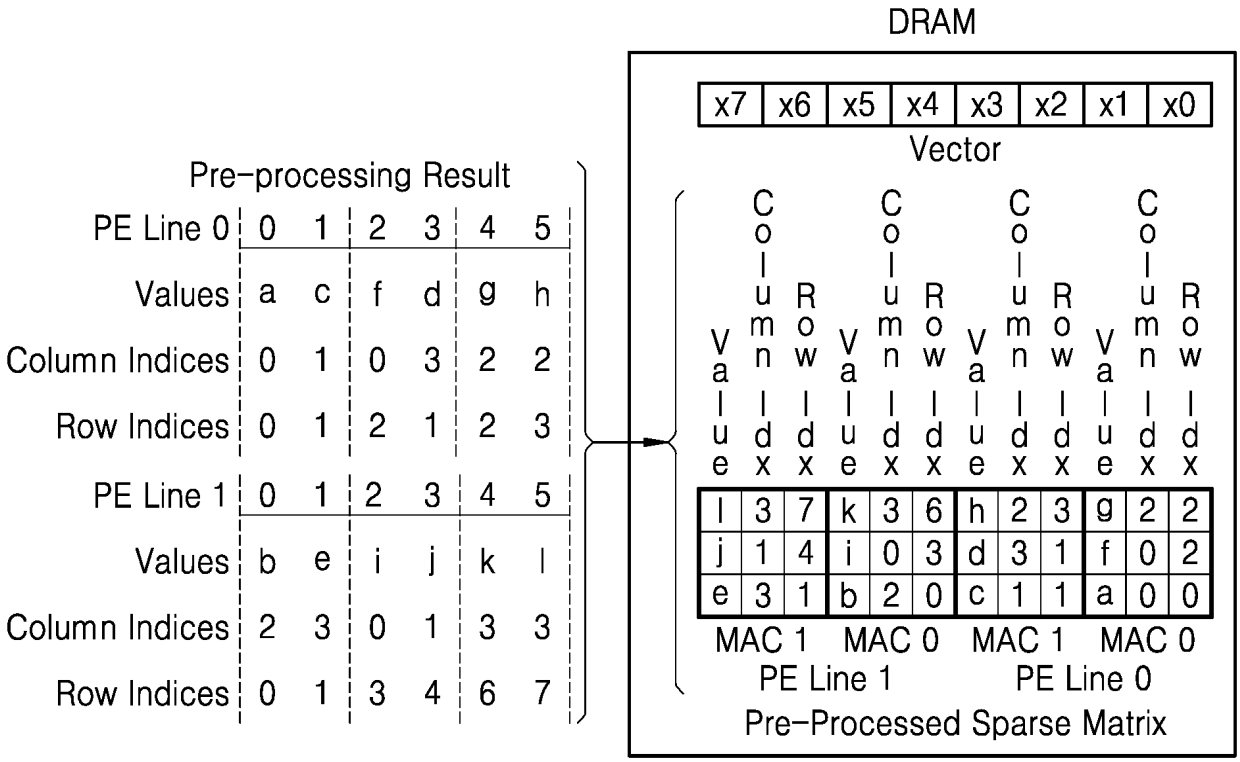
도면5


DRAM
(Off-Chip Memory)
50
Memory Controller
Data Distributor
111
B1
B0
Vector Buffer
Bank #0
Bank #1
Matrix Buffer
112
MAC
MAC
PE Line #0
PE Line #1
PE Line #N
110
Output Updater
PSUM Buffer
PSUM
100

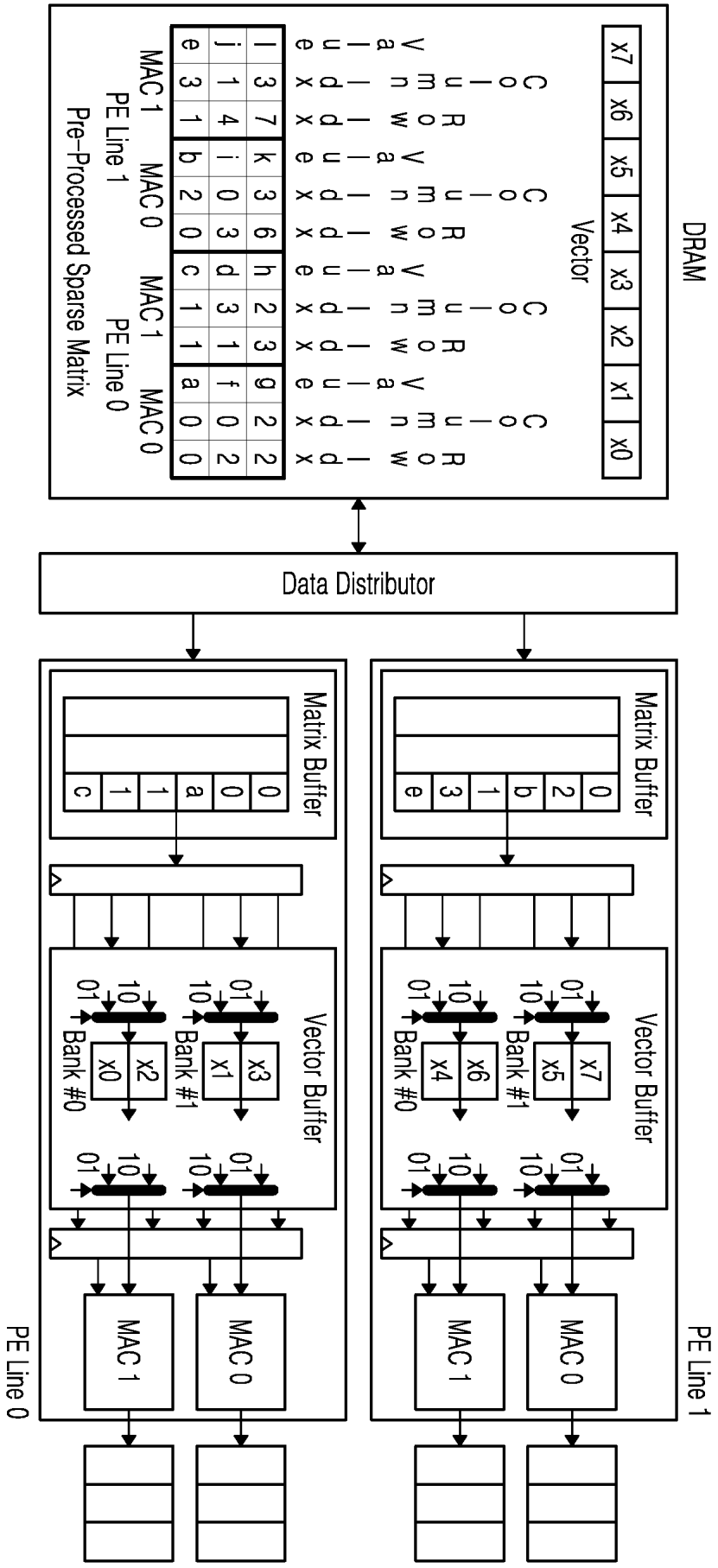
도면6


(For Simplicity, # of PE Line =2)
Whole Matrix
Block 0
Block 1
Block 2
Block 3
Block 4
Block 5
Block Width
Block Height
PE Line #0
PE Line #1
Vector Buffer
Bank #0
Bank #1
MAC0
MAC1
PE Line #0
PE Line #1
Values
Column Indices
Row Indices
PE Line 0
PE Line 1
Bank#0
Bank#1

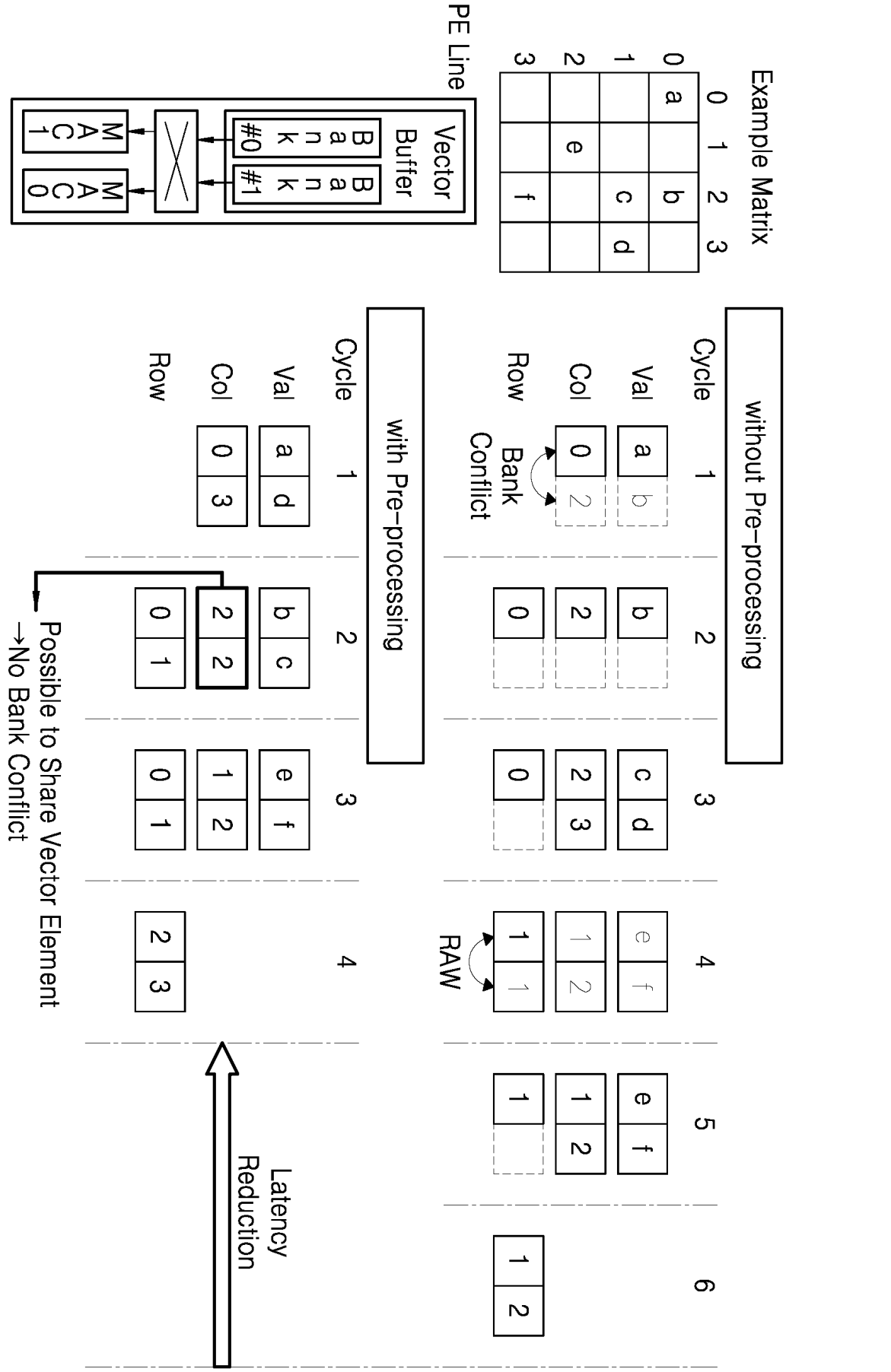
도면7


DRAM
x7 x6 x5 x4 x3 x2 x1 x0
Vector
Pre-processing Result
PE Line 0
Values
Column Indices
Row Indices
PE Line 1
Values
Column Indices
Row Indices
Value
Column Idx
Row Idx
MAC 1
MAC 0
MAC 1
MAC 0
PE Line 1
PE Line 0
Pre-Processed Sparse Matrix

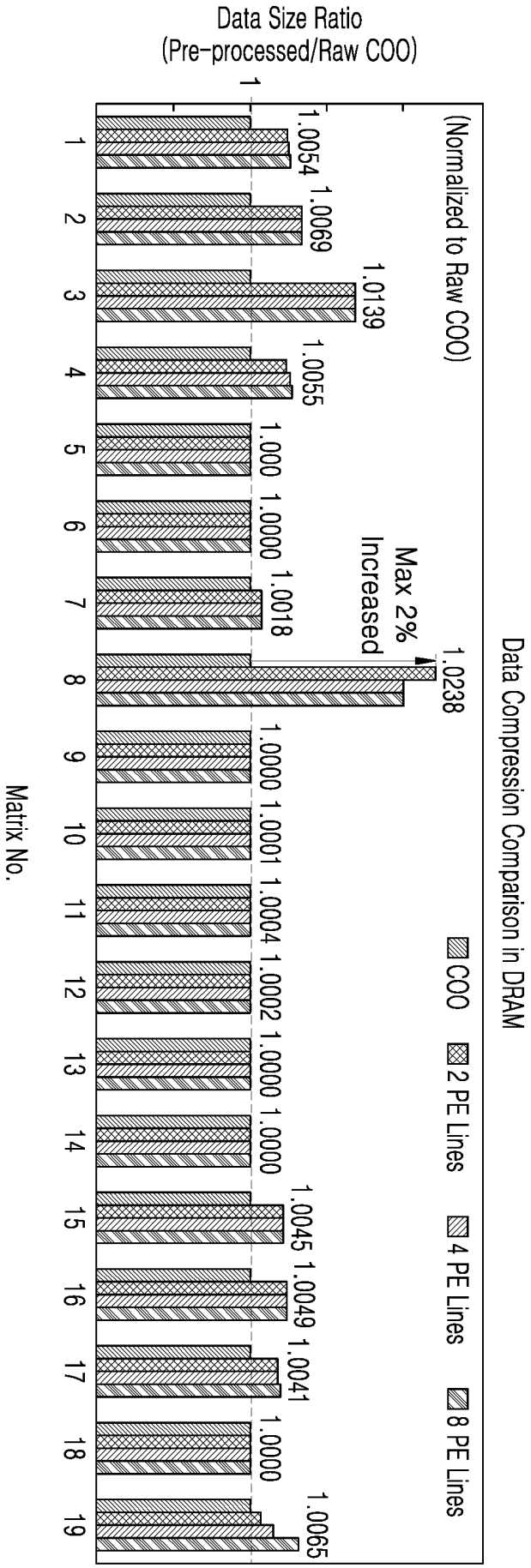
도면8


DRAM
Vector
x7
x6
x5
x4
x3
x2
x1
x0
Pre-Processed Sparse Matrix
PE Line 1
PE Line 0
MAC 1
MAC 0
Data Distributor
Matrix Buffer
Vector Buffer
Bank #0
Bank #1
01
10

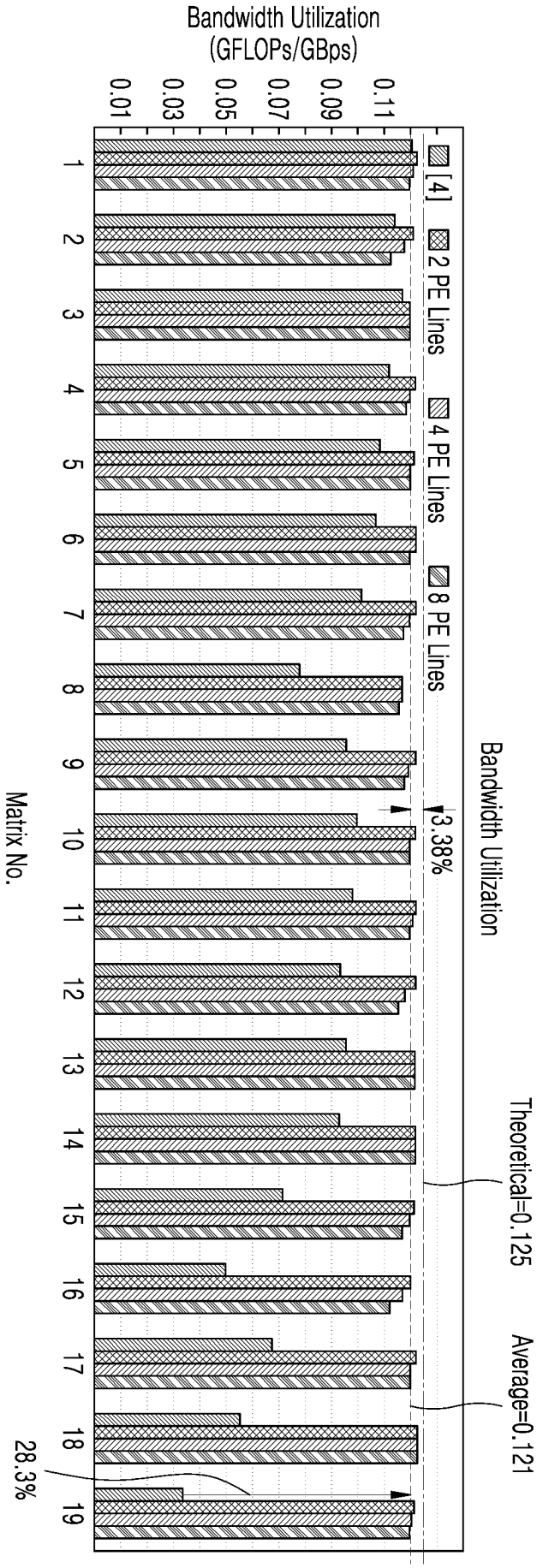
도면9


Example Matrix
PE Line
Vector Buffer
Bank #0
Bank #1
MAC 1
MAC 0
without Pre-processing
Cycle
Val
Col
Row
Bank Conflict
RAW
with Pre-processing
Possible to Share Vector Element
→No Bank Conflict
Latency Reduction

도면10


Data Size Ratio
(Pre-processed/Raw COO)
1
(Normalized to Raw COO)
Data Compression Comparison in DRAM
COO
2 PE Lines
4 PE Lines
8 PE Lines
Max 2% Increased
1.0054
1.0069
1.0139
1.0055
1.000
1.0000
1.0018
1.0238
1.0000
1.0000
1.0001
1.0004
1.0002
1.0000
1.0000
1.0045
1.0049
1.0041
1.0000
1.0065
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Matrix No.

도면11


Bandwidth Utilization
(GFLOPs/GBps)
0.01
0.03
0.05
0.07
0.09
0.11
[4]
2 PE Lines
4 PE Lines
8 PE Lines
Bandwidth Utilization
3.38%
Theoretical=0.125
Average=0.121
28.3%
Matrix No.
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19