(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년06월23일
(11) 등록번호 10-2268179
(24) 등록일자 2021년06월16일

(51) 국제특허분류(Int. Cl.)
G06F 13/40 (2006.01) *G06F 13/16* (2006.01)
G11C 11/402 (2006.01)
(52) CPC특허분류
G06F 13/40 (2013.01)
G06F 13/1668 (2013.01)
(21) 출원번호 10-2017-0067968
(22) 출원일자 2017년05월31일
심사청구일자 2020년01월28일
(65) 공개번호 10-2018-0046346
(43) 공개일자 2018년05월08일
(30) 우선권주장
62/413,973 2016년10월27일 미국(US)
15/426,015 2017년02월06일 미국(US)
(56) 선행기술조사문헌
US05594698 A
US05847577 A

(73) 특허권자
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
(72) 발명자
리, 슈앙첸
미국 캘리포니아주 93117 골레타 이엘 콜레히오 로드 6510 아파트 1302
니우, 디민
미국 캘리포니아주 94087 서니베일 홀트하우스 테라스 527
(뒷면에 계속)
(74) 대리인
특허법인 고려

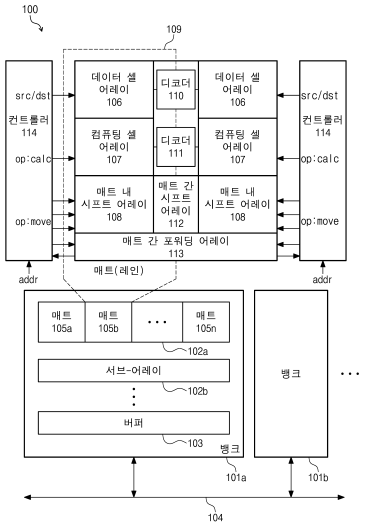
전체 청구항 수 : 총 20 항 심사관 : 김세영

(54) 발명의 명칭 **DPU 연산들을 위한 소프트웨어 스택 및 프로그래밍**

(57) 요 약

시스템은 라이브러리, 컴파일러, 드라이버 및 적어도 하나의 DRAM(dynamic random access memory) 프로세싱 유닛(DPU)을 포함한다. 라이브러리는 수신된 명령에 대응하는 적어도 하나의 DPU 연산을 판별할 수 있다. 컴파일러는 상기 수신된 명령에 대응하는 상기 판별된 적어도 하나의 DPU 연산에 대한 적어도 하나의 DPU 명령을 형성할 수 있다. 드라이버는 상기 적어도 하나의 DPU 명령을 적어도 하나의 DPU로 전송한다. DPU는 적어도 하나의 열을 포함하는 어레이로 배열된 복수의 DRAM-기반 컴퓨팅 셀들을 포함하는 적어도 하나의 컴퓨팅 셀 어레이를 포함하고, 상기 적어도 하나의 열은 논리 기능을 제공하기 위해 구성된 DRAM-기반 컴퓨팅 셀들의 적어도 3개의 행들을 포함하고, 상기 논리 기능은 상기 적어도 3개의 행들 중 제1 행 및 제2 행에서 동작하고, 상기 논리 기능의 결과를 상기 적어도 3개의 행들 중 제3 행에 저장한다.

대 표 도 - 도1



(52) CPC특허분류
G11C 11/4023 (2013.01)
(72) 발명자
말라디, 크리슈나
미국 캘리포니아주 95135 산호세 로트렉 드라이브 4196
정, 홍종
미국 캘리포니아주 95032 로스 가토스 칼튼 애비뉴 120 6호

명 세 서

청구범위

청구항 1

명령을 수신하는 인터페이스;

상기 수신된 명령에 대응하는 적어도 하나의 DRAM(dynamic random access memory) 프로세싱 유닛(DPU) 연산을 판별하는 라이브러리;

상기 수신된 명령에 대응하는 상기 판별된 적어도 하나의 DPU 연산에 대한 적어도 하나의 DPU 명령을 형성하는 컴파일러; 및

상기 적어도 하나의 DPU 명령을 적어도 하나의 DPU로 전송하는 드라이버를 포함하되,

상기 적어도 하나의 DPU는 적어도 하나의 열을 포함하는 어레이로 배열된 복수의 DRAM-기반 컴퓨팅 셀들을 포함하는 적어도 하나의 컴퓨팅 셀 어레이를 포함하고, 상기 적어도 하나의 열은 DRAM-기반 컴퓨팅 셀들의 적어도 3개의 행들을 포함하고, 상기 DRAM-기반 컴퓨팅 셀들은 상기 적어도 3개의 행들 중 제1 행 및 제2 행에서 동작하는 논리 기능을 제공하고, 상기 논리 기능의 결과를 상기 적어도 3개의 행들 중 제3 행에 저장하도록 구성된 시스템.

청구항 2

제 1 항에 있어서,

상기 적어도 하나의 열의 DRAM-기반 컴퓨팅 셀들 각각은 3개의 트랜지스터들, 1개의 커패시터(3T1C) DRAM 메모리 셀, 또는 1개의 트랜지스터 1개의 커패시터(1T1C) DRAM 메모리 셀을 포함하고, 그리고

상기 적어도 하나의 열의 상기 DRAM-기반 컴퓨팅 셀들은 NOR 논리 기능을 제공하는 시스템.

청구항 3

제 2 항에 있어서,

상기 적어도 하나의 DPU는 적어도 하나의 열을 포함하는 어레이로 배열된 복수의 DRAM-기반 스토캐스틱 컴퓨팅 셀들을 더 포함하고, 상기 적어도 하나의 열은 DRAM-기반 스토캐스틱 컴퓨팅 셀들의 적어도 3개의 행들을 포함하고, 상기 DRAM-기반 스토캐스틱 컴퓨팅 셀들은 상기 적어도 3개의 행들 중 제1 행 및 제2 행에서 동작하는 스토캐스틱 논리 기능을 제공하고, 상기 스토캐스틱 논리 기능의 결과를 상기 적어도 3개의 행들 중 제3 행에 저장하도록 구성된 시스템.

청구항 4

제 3 항에 있어서,

상기 적어도 하나의 열의 상기 DRAM-기반 스토캐스틱 컴퓨팅 셀들 각각은 3개의 트랜지스터들, 1개의 커패시터(3T1C) DRAM 메모리 셀, 또는 1개의 트랜지스터 1개의 커패시터(1T1C) DRAM 메모리 셀을 포함하는 시스템.

청구항 5

제 4 항에 있어서,

상기 적어도 하나의 DPU 연산은 스토캐스틱 컴퓨팅 연산을 포함하는 시스템.

청구항 6

제 1 항에 있어서,

상기 적어도 하나의 DPU는 적어도 하나의 열을 포함하는 어레이로 배열된 복수의 DRAM-기반 스토캐스틱 컴퓨팅 셀들을 더 포함하고, 상기 적어도 하나의 열은 DRAM-기반 스토캐스틱 컴퓨팅 셀들의 적어도 3개의 행들을 포함

하고, 상기 DRAM-기반 스토캐스틱 컴퓨팅 셀들은 상기 적어도 3개의 행들 중 제1 행 및 제2 행에서 동작하는 스토캐스틱 논리 기능을 제공하고, 상기 스토캐스틱 논리 기능의 결과를 상기 적어도 3개의 행들 중 제3 행에 저장하도록 구성되는 시스템.

청구항 7

제 1 항에 있어서,

상기 라이브러리는 상기 수신된 명령에 응답해서 전체 컨볼루션 계층의 매핑을 더 판별하는 시스템.

청구항 8

제 1 항에 있어서,

상기 라이브러리는 상기 수신된 명령에 대응하는 상기 적어도 하나의 DPU 연산을 수행하도록 다수의 DPU들을 매핑하는 상기 적어도 하나의 DPU 동작 내에서의 다중 병렬화를 더 판별하는 시스템.

청구항 9

제 1 항에 있어서,

상기 드라이버는 상기 수신된 명령에 근거해서 상기 적어도 하나의 DPU의 내외로의 데이터 이동을 더 제어하는 시스템.

청구항 10

제 1 항에 있어서,

상기 컴파일러는 상기 적어도 하나의 DPU의 상기 적어도 하나의 행에서 동작하는 상기 수신된 명령에 대응하는 단일 DPU 명령을 더 형성하는 시스템.

청구항 11

적어도 하나의 DRAM(dynamic random access memory) 프로세싱 유닛(DPU);

명령을 수신하는 인터페이스;

상기 수신된 명령에 대응하는 적어도 하나의 DPU 명령을 판별하는 라이브러리; 및

상기 수신된 명령에 대응하는 상기 적어도 하나의 DPU 명령을 상기 적어도 하나의 DPU로 전송하는 드라이버를 포함하되,

상기 적어도 하나의 DPU 각각은:

적어도 하나의 열을 포함하는 어레이로 배열된 복수의 DRAM-기반 컴퓨팅 셀들을 포함하는 적어도 하나의 컴퓨팅 셀 어레이; 및

적어도 하나의 열을 포함하는 어레이로 배열된 복수의 DRAM-기반 스토캐스틱 컴퓨팅 셀들을 포함하고,

상기 복수의 DRAM-기반 컴퓨팅 셀들의 상기 적어도 하나의 열은 DRAM-기반 컴퓨팅 셀들의 적어도 3개의 행들을 포함하고, 상기 DRAM-기반 컴퓨팅 셀들은 상기 적어도 3개의 행들 중 제1 행 및 제2 행에서 동작하는 논리 기능을 제공하고, 상기 논리 기능의 결과를 상기 적어도 3개의 행들 중 제3 행에 저장하도록 구성되고,

상기 복수의 DRAM-기반 스토캐스틱 컴퓨팅 셀들의 상기 적어도 하나의 열은 DRAM-기반 스토캐스틱 컴퓨팅 셀들의 적어도 3개의 행들을 포함하고, 상기 DRAM-기반 스토캐스틱 컴퓨팅 셀들은 상기 적어도 3개의 행들 중 제1 행에 의해 수신된 데이터의 제1 스트림 및 제2 행에 의해 수신된 데이터의 제2 스트림에서 동작하는 스토캐스틱 논리 기능을 제공하고, 상기 스토캐스틱 논리 기능으로부터 도출되는 데이터의 스트림을 상기 적어도 3개의 행들 중 제3 행에 저장하도록 구성된 시스템.

청구항 12

제 11 항에 있어서,

상기 적어도 하나의 열의 DRAM-기반 컴퓨팅 셀들 각각은 3개의 트랜지스터들, 1개의 커패시터(3T1C) DRAM 메모

리 셀, 또는 1개의 트랜지스터 1개의 커패시터(1T1C) DRAM 메모리 셀을 포함하고, 그리고

상기 적어도 하나의 열의 DRAM-기반 스토캐스틱 컴퓨팅 셀들 각각은 3개의 트랜지스터들, 1개의 커패시터(3T1C) DRAM 메모리 셀, 또는 1개의 트랜지스터 1개의 커패시터(1T1C) DRAM 메모리 셀을 포함하는 시스템.

청구항 13

제 12 항에 있어서,

상기 적어도 하나의 열의 상기 DRAM-기반 컴퓨팅 셀들은 NOR 논리 기능을 제공하는 시스템.

청구항 14

제 11 항에 있어서,

상기 적어도 하나의 DPU 명령은 스토캐스틱 컴퓨팅 연산을 포함하는 시스템.

청구항 15

제 11 항에 있어서,

상기 적어도 하나의 DPU의 적어도 하나의 행에서 동작하는 상기 수신된 명령에 대응하는 단일 DPU 명령을 형성하는 컴파일러를 더 포함하는 시스템.

청구항 16

적어도 하나의 열을 포함하는 어레이로 배열된 복수의 DRAM(dynamic random access memory)-기반 스토캐스틱 컴퓨팅 셀들을 포함하는 적어도 하나의 스토캐스틱 컴퓨팅 셀 어레이를 포함하는 적어도 하나의 DRAM 프로세싱 유닛(DPU);

제1 수신된 명령에 대응하는 적어도 하나의 스토캐스틱 DPU 명령을 판별하는 라이브러리; 및

상기 제1 수신된 명령에 대응하는 상기 적어도 하나의 스토캐스틱 DPU 명령을 상기 복수의 DRAM(dynamic random access memory)-기반 스토캐스틱 컴퓨팅 셀들을 포함하는 상기 적어도 하나의 DPU로 전송하는 드라이버를 포함하되,

상기 적어도 하나의 열은 DRAM-기반 스토캐스틱 컴퓨팅 셀들의 적어도 3개의 행들을 포함하고, 상기 DRAM-기반 스토캐스틱 컴퓨팅 셀들은 상기 적어도 3개의 행들 중 제1 행에 의해 수신된 데이터의 제1 스트림 및 제2 행에 의해 수신된 데이터의 제2 스트림에서 동작하는 스토캐스틱 논리 기능을 제공하고, 상기 스토캐스틱 논리 기능으로부터 도출되는 데이터의 스트림을 상기 적어도 3개의 행들 중 제3 행에 저장하도록 구성된 시스템.

청구항 17

제 16 항에 있어서,

상기 적어도 하나의 열의 상기 DRAM-기반 스토캐스틱 컴퓨팅 셀들 각각은 3개의 트랜지스터들, 1개의 커패시터(3T1C) DRAM 메모리 셀, 또는 1개의 트랜지스터 1개의 커패시터(1T1C) DRAM 메모리 셀을 포함하는 시스템.

청구항 18

제 16 항에 있어서,

상기 드라이버는 상기 제1 수신된 명령에 근거해서 상기 복수의 DRAM-기반 스토캐스틱 컴퓨팅 셀들을 포함하는 상기 적어도 하나의 DPU의 내외로의 데이터 이동을 더 제어하는 시스템.

청구항 19

제 16 항에 있어서,

적어도 하나의 열을 포함하는 어레이로 배열된 복수의 DRAM-기반 컴퓨팅 셀들을 포함하는 적어도 하나의 컴퓨팅 셀 어레이를 더 포함하되,

상기 복수의 DRAM-기반 컴퓨팅 셀들의 상기 적어도 하나의 열은 DRAM-기반 컴퓨팅 셀들의 적어도 3개의 행들을

포함하고, 상기 DRAM-기반 컴퓨팅 셀들은 상기 적어도 3개의 행들 중 제1 행 및 제2 행에서 동작하는 논리 기능을 제공하고, 상기 논리 기능의 결과를 상기 적어도 3개의 행들 중 제3 행에 저장하도록 구성되고,

상기 라이브러리는 제2 수신된 명령에 대응하는 상기 적어도 하나의 컴퓨팅 셀 어레이에 대한 적어도 하나의 DPU 명령을 더 판별하고,

상기 드라이버는 상기 제2 수신된 명령에 대응하는 상기 적어도 하나의 DPU 명령을 상기 복수의 DRAM-기반 컴퓨팅 셀들을 포함하는 상기 적어도 하나의 컴퓨팅 셀 어레이로 더 전송하는 시스템.

청구항 20

제 19 항에 있어서,

상기 적어도 하나의 열의 상기 DRAM-기반 스토캐스틱 컴퓨팅 셀들 각각은 3개의 트랜지스터들, 1개의 커패시터(3T1C) DRAM 메모리 셀, 또는 1개의 트랜지스터 1개의 커패시터(1T1C) DRAM 메모리 셀을 포함하는 시스템.

발명의 설명

기 술 분 야

[0001] 본 발명은 DRAM-기반 스토캐스틱 컴퓨팅 시스템에 관한 것이다.

배 경 기 술

[0002] GPU들(Graphics Processing Units) 및 TPU들(Tensor Processing Units)은 종래에 딥 러닝 프로세싱(deep learning processing)에사용되었다. 딥 러닝 프로세싱은 GPU들 또는 TPU들에 의해 효율적으로 수행될 수 없는 고도의 병렬화된 프로세싱을 포함한다.

발명의 내용

해결하려는 과제

[0003] 본 발명은 DRAM(dynamic random access memory) 프로세싱 유닛(DPU)을 포함할 수 있는 시스템을 제공한다.

과제의 해결 수단

[0004] 예시적인 실시예는 명령을 수신하는 인터페이스, 라이브러리, 컴파일러, 드라이버 및 DRAM(dynamic random access memory) 프로세싱 유닛(DPU)을 포함할 수 있는 시스템을 제공한다. 상기 라이브러리는 상기 인터페이스에 의해서 수신된 명령에 대응하는 적어도 하나의 DPU 연산을 판별한다. 상기 컴파일러는 상기 수신된 명령에 대응하는 상기 판별된 적어도 하나의 DPU 연산에 대한 적어도 하나의 DPU 명령을 형성할 수 있다. 상기 드라이버는 상기 적어도 하나의 DPU 명령을 적어도 하나의 DPU로 전송할 수 있다. 상기 DPU는 적어도 하나의 열을 포함하는 어레이 로 배열된 복수의 DRAM-기반 컴퓨팅 셀들을 포함하는 적어도 하나의 컴퓨팅 셀 어레이를 포함할 수 있고, 상기 적어도 하나의 열은 DRAM-기반 컴퓨팅 셀들의 적어도 3개의 행들을 포함하고, 상기 DRAM-기반 컴퓨팅 셀들은 상기 적어도 3개의 행들 중 제1 행 및 제2 행에서 동작하는 논리 기능을 제공하고, 상기 논리 기능의 결과를 상기 적어도 3개의 행들 중 제3 행에 저장하도록 구성된다. 일 실시예에서, 상기 적어도 하나의 열의 DRAM-기반 컴퓨팅 셀 각각은 3개의 트랜지스터, 1개의 커패시터(3T1C) DRAM 메모리 셀, 또는 1개의 트랜지스터, 1개의 커패시터(1T1C) DRAM 메모리 셀을 포함할 수 있고, 그리고 상기 적어도 하나의 열의 상기 DRAM-기반 컴퓨팅 셀들은 NOR 논리 기능을 제공한다. 일 실시예에서, 상기 DPU는 적어도 하나의 열을 포함하는 어레이로 배열된 복수의 DRAM-기반 스토캐스틱 컴퓨팅 셀들을 더 포함할 수 있고, 상기 적어도 하나의 열은 DRAM-기반 스토캐스틱 컴퓨팅 셀들의 적어도 3개의 행들을 포함할 수 있고, 상기 DRAM-기반 스토캐스틱 컴퓨팅 셀들은 상기 적어도 3개의 행들 중 제1 행 및 제2 행에서 동작하는 스토캐스틱 논리 기능을 제공하고, 상기 스토캐스틱 논리 기능의 결과를 상기 적어도 3개의 행들 중 제3 행에 저장하도록 구성된다. 일 실시예에서, 상기 적어도 하나의 열의 상기 DRAM-기반 스토캐스틱 컴퓨팅 셀들 각각은 3개의 트랜지스터들, 1개의 커패시터(3T1C) DRAM 메모리 셀 또는 1개의 트랜지스터, 1개의 커패시터(1T1C) DRAM 메모리 셀을 포함할 수 있다. 일 실시예에서, 적어도 하나의 DPU 연산은 스토캐스틱 컴퓨팅 연산을 포함할 수 있다.

[0005] 예시적인 실시예는, 적어도 하나의 DRAM(dynamic random access memory) 프로세싱 유닛(DPU), 상기 DPU 각각은, 적어도 하나의 열을 포함하는 어레이로 배열된 복수의 DRAM-기반 컴퓨팅 셀들을 포함하는 적어도 하나의 컴퓨팅 셀 어레이를 포함하고, 상기 적어도 하나의 열은 DRAM-기반 컴퓨팅 셀들의 적어도 3개의 행들을 포함하고, 상기 DRAM-기반 컴퓨팅 셀들은 상기 적어도 3개의 행들 중 제1 행 및 제2 행에서 동작하는 논리 기능을 제공하고, 상기 논리 기능의 결과를 상기 적어도 3개의 행들 중 제3 행에 저장하도록 구성되며; 그리고 적어도 하나의 열을 포함하는 어레이로 배열된 복수의 DRAM-기반 스토캐스틱 컴퓨팅 셀들을 포함하고, 상기 적어도 하나의 열은 DRAM-기반 스토캐스틱 컴퓨팅 셀들의 적어도 3개의 행들을 포함하고, 상기 DRAM-기반 스토캐스틱 컴퓨팅 셀들은 상기 적어도 3개의 행들 중 제1 행 및 제2 행에서 동작하는 스토캐스틱 논리 기능을 제공하고, 상기 스토캐스틱 논리 기능의 결과를 상기 적어도 3개의 행들 중 제3 행에 저장하도록 구성되며; 명령을 수신하는 인터페이스; 상기 수신된 명령에 대응하는 적어도 하나의 DPU 연산을 판별하는 라이브러리; 및 상기 수신된 명령에 대응하는 적어도 하나의 DPU 명령을 상기 적어도 하나의 DPU로 전송하는 드라이버를 포함한다. 일 실시예에서, 상기 적어도 하나의 열의 DRAM-기반 컴퓨팅 셀 각각은 3개의 트랜지스터, 1개의 커패시터(3T1C) DRAM 메모리 셀, 또는 1개의 트랜지스터, 1개의 커패시터(1T1C) DRAM 메모리 셀을 포함하고, 그리고 상기 적어도 하나의 열의 상기 DRAM-기반 스토캐스틱 컴퓨팅 셀들 각각은 3개의 트랜지스터들, 1개의 커패시터(3T1C) DRAM 메모리 셀 또는 1개의 트랜지스터, 1개의 커패시터(1T1C) DRAM 메모리 셀을 포함한다. 일 실시예에서, 상기 적어도 하나의 열의 상기 DRAM-기반 컴퓨팅 셀들은 NOR 논리 기능을 제공한다. 일 실시예에서, 상기 적어도 하나의 DPU 연산은 스토캐스틱 컴퓨팅 연산을 포함할 수 있는 시스템을 제공한다.

[0006] 예시적 실시예는, 적어도 하나의 열을 포함하는 어레이로 배열된 복수의 DRAM(dynamic random access memory)-기반 스토캐스틱 컴퓨팅 셀들을 포함하는 적어도 하나의 스토캐스틱 컴퓨팅 셀 어레이를 포함하는 DRAM 프로세싱 유닛(DPU), 상기 적어도 하나의 열은 DRAM-기반 스토캐스틱 컴퓨팅 셀들의 적어도 3개의 행들을 포함하고, 상기 DRAM-기반 스토캐스틱 컴퓨팅 셀들은 상기 적어도 3개의 행들 중 제1 행 및 제2 행에서 동작하는 스토캐스틱 논리 기능을 제공하고, 상기 스토캐스틱 논리 기능의 결과를 상기 적어도 3개의 행들 중 제3 행에 저장하도록 구성되며; 수신된 명령에 대응하는 적어도 하나의 스토캐스틱 DPU 연산을 판별하는 라이브러리; 및 상기 수신된 명령에 대응하는 상기 적어도 하나의 스토캐스틱 DPU 명령을 상기 적어도 하나의 DPU로 전송하는 드라이버를 포함할 수 있는 시스템을 제공한다. 일 실시예에서, 상기 적어도 하나의 열의 상기 DRAM-기반 스토캐스틱 컴퓨팅 셀들 각각은 3개의 트랜지스터들, 1개의 커패시터들(3T1C) DRAM 메모리 셀 또는 1개의 트랜지스터, 1개의 커패시터(1T1C) DRAM 메모리 셀을 포함할 수 있다.

도면의 간단한 설명

[0007] 이하의 설명에서, 본 명세서에 개시된 주제의 양상들은 도면들에 도시된 예시적인 실시예들을 참조하여 설명될 것이다.

도 1은 본 명세서에 개시된 주제에 따른 DRAM(dynamic random access memory) 기반 프로세싱 유닛(DPU)의 예시적인 실시예의 블록도를 도시한다.

도 2a는 컴퓨팅 셀 어레이에서컴퓨팅 셀로 사용될 수 있는 3-트랜지스터, 1-커패시터 DRAM 컴퓨팅-셀 토포그래피의 예시적인 실시 예를 도시한다.

도 2b는 컴퓨팅 셀 어레이에서 컴퓨팅 셀로 사용될 수 있는 1-트랜지스터, 1-커패시터 DRAM 컴퓨팅-셀 토포그래피의 다른 예시적인 실시예를 도시한다.

도 3은 본 명세서에 개시된 주제에 따른 매트 간 시프트 어레이의 예시적인 실시예를 도시한다.

도 4a는 본 명세서에 개시된 주제에 따른 매트 간 시프트 어레이의 실시예를 도시한다.

도 4b는 본 명세서에 개시된 주제에 따른 좌측 매트 간 시프트에 대한 인접한 컴퓨팅-셀 열들 내의 2개의 동일하게 위치한 컴퓨팅 셀 사이의 매트 간 시프트 상호 접속 구성을 개념적으로 도시한다.

도 4C는 본 명세서에 개시된 주제에 따른 좌측 매트 간 시프트에 대한 인접한 컴퓨팅-셀 열들 내의 2개의 다르게 위치한 컴퓨팅 셀 사이의 매트 간 시프트 상호 접속 구성을 개념적으로 도시한다.

도 5는 본 명세서에 개시된 주제에 따른 매트 간 포워딩 어레이의 실시예를 도시한다.

도 6a 내지 도 6g는 본 명세서에 개시된 주제에 따라 DPU에 의해 제공될 수 있는 NOR-논리-기반 연산들을 도시한다.

도 7은 본 명세서에 개시된 주제에 따른 스토캐스틱(stochastic) 데이터 영역을 포함하는 DPU의 예시적인 실시예의 블록도를 도시한다.

도 8a 및 도 8b는 멀티플렉싱 연산으로 변환될 수 있는 덧셈 연산 및 AND 논리 연산으로 변환될 수 있는 곱셈 연산에 대한 스토캐스틱 컴퓨팅 연산들을 각각 도시한다.

도 9는 본 명세서에 개시된 주제에 따라 DPU들을 포함하는 시스템 구조를 도시한다.

발명을 실시하기 위한 구체적인 내용

[0008] 이하의 상세한 설명에서, 본 발명의 완전한 이해를 제공하기 위해 다수의 특정 세부 사항들이 설명된다. 그러나, 당업자는 개시된 양상들이 이러한 특정 세부 사항들 없이 실시될 수 있음을 이해할 것이다. 다른 예들에서, 공지된 방법들, 절차들, 구성 요소들 및 회로들은 여기에 개시된 주제를 모호하게 하지 않기 위해 상세하게 설명되지 않았다.

[0009] 본 명세서에서 "일 실시예" 또는 "실시예"는 본 실시예와 관련하여 기술된 특정 특징, 구조 또는 특성이 본 명세서에 개시된 적어도 하나의 실시예에 포함될 수 있다는 것을 의미한다.

[0010] 따라서, 본 명세서 전체의 다양한 곳에서 "일 실시예에서" 또는 "실시예에서" 또는 "일 실시예에 따라"(또는 유사한 의미를 갖는 다른 표현들)의 표현은 반드시 모두 동일한 실시예를 지칭하지는 않는다. 또한, 특정한 특징들, 구조들 또는 특성들은 하나 이상의 실시 예들에서 임의의 적절한 방식으로 연결될 수 있다. 이와 관련하여, 본 명세서에 사용된 바와 같이, "예시적인"이라는 단어는 "예, 예시 또는 실례로서 제공되는"을 의미한다. "예시적으로" 본 명세서에 기재된 임의의 실시 형태는 다른 실시예들보다 반드시 바람직하거나 유리한 것으로 해석되어서는 안된다. 또한, 본 명세서의 논의 내용에 따라, 단수는 상응하는 복수의 형태를 포함할 수 있고, 복수의 용어는 상응하는 단수 형태를 포함할 수 있다. 또한, 여기에 도시되고 논의된 다양한 도면들(구성 요소 포함)은 단지 예시적인 목적을 위한 것이며, 실제 스케일(scale)로 그려진 것은 아니라는 점에 유의해야 한다. 마찬가지로, 다양한 파형들 및 타이밍 도들이 단지 예시적인 목적을 위해 도시된다. 예를 들어, 일부 요소의 치수는 명확성을 위해 다른 요소에 비해 과장될 수 있다. 또한, 적절한 것으로 고려되는 경우, 참조 부호들은 대응하는 및/또는 유사한 요소들을 나타내기 위해 도면들 사이에서 반복되었다.

[0011] 본 명세서에서 사용된 용어는 특정 예시적인 실시예들 만을 설명하기 위한 것이며, 청구된 주제를 제한하려는 것이 아니다. 본 명세서에서 사용된 단수 형태들 및 "상기"는 문맥 상 다르게 지시하지 않는 한 복수 형태를 포함하고자 한다. 본 명세서에서 사용되는 "포함하는" 및/또는 "구비하는"이라는 용어는 명시된 특징들, 정수들, 단계들, 동작들, 구성 요소들 및/또는 컴포넌트들의 존재를 나타내지만, 하나 이상의 다른 특징들, 정수들, 단계들, 동작들, 구성 요소들, 컴포넌트들 및/또는 그룹들의 존재 또는 부가를 배제하지 않는다는 것이 더 이해될 것이다. 여기에 사용 된 "제1", "제2" 등의 용어는 명사들의 앞에서 명사들에 대한 레이블들로 사용되며 명시적으로 정의되지 않은 한 모든 유형의 순서(예를 들어, 공간적, 시간적, 논리적 등)를 암시하지 않는다. 또한, 동일하거나 유사한 기능을 갖는 부품들, 구성 요소들, 블록들, 회로들, 유닛들 또는 모듈들을 지칭하기 위해 2개 이상의 도면들에 걸쳐 동일한 참조 번호들이 사용될 수 있다. 그러나, 이러한 사용법은 설명의 단순화 및 논의의 용이함을 위해서만 사용된다. 그러한 구성 요소들 또는 유닛들의 구성 또는 구조적 세부 사항들이 모든 실시예들에서 동일하다는 것을 의미하지 않고 또는 공통으로 참조된 부품들/모듈들이 본 명세서에 개시된 특정 실시예들의 교시들을 구현하는 유일한 방법이라는 것을 의미하지는 않는다.

[0012] 다르게 정의되지 않는 한, 본 명세서에서 사용된 모든 용어(기술 및 과학 용어 포함)는 이 주제가 속하는 기술 분야의 당업자에 의해 일반적으로 이해되는 것과 동일한 의미를 갖는다. 일반적으로 사용되는 사전에서 정의된 용어와 같은 용어는 관련 기술의 맥락에서 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며 이상적이거나 과도하게 공식적인 의미로 해석되지 않는다는 점을 이해해야 한다.

[0013] 여기에 개시된 주제는 덧셈, 곱셈, 시프팅, 최대/최소 및 비교와 같은(그러나 이에 한정되지 않는) 각각 다른 연산들에 대해 프로그램 가능하고 재구성 가능한 DRAM(dynamic random access memory) 기반 프로세싱 유닛(DPU)을 제공한다. 일 실시예에서, DPU는 3개의 트랜지스터, 1개의 커패시터(3T1C) DRAM 프로세스 및 구조에 기초한다. 다른 실시예에서, DPU는 사소한 수정을 갖는 트랜지스터 1개, 커패시터 1개(1T1C) DRAM 프로세스 및 구조에 기초한다. 따라서, DPU는 특정 컴퓨팅 논리 회로(가산기(adder)와 같은)를 포함하지 않지만, 고도의 병렬 연산들을 사용하는 메모리 셀들을 사용하여 계산들을 제공한다. 일 실시예에서, DPU는 덧셈(addition)이 멀티플렉싱 연산(multiplexing operation)으로 변환될 수 있고, 그리고 곱셈(multiplication)이 AND 논리 연산으로

변환될 수 있는 스토캐스틱 컴퓨팅 어레이(stochastic computing array)를 포함할 수 있다.

[0014] 또한 본 명세서에 개시된 주제는 프레임워크(framework) 확장, 라이브러리, 드라이버, 컴파일러 및 DPU들을 프로그램하고 재구성하기 위한 명령 세트 아키텍처(instruction set architecture, ISA)를 갖는 환경(생태계)을 포함하는 시스템 아키텍처를 제공한다.

[0015] 또한, 여기에 개시된 주제는 데이터 센터 및/또는 모바일 애플리케이션들에 적합하고, 이진(binary) 및 고정 소수점 계산들 모두를 위한 머신-러닝 애플리케이션들(machine-learning applications)을 위한 PIM(Processor-in-Memory) 솔루션을 제공하는 시스템 아키텍처이며, GPU/ASIC (TPU)/FPGA머신-러닝 애플리케이션들의 대안이다. 일 실시예에서, 여기에 개시된 주제는, 예를 들어 고성능, 에너지 효율, 및 이진 가중 신경 네트워크(Binary Weight Neural Network)에 대한 가속 딥 러닝(accelerated deep learning)을 제공하는 저비용 시스템을 제공한다.

[0016] 여기에 개시된 주제는 DRAM(dynamic random access memory) 기술을 사용하여 형성될 수 있고, 재구성 가능하고 그리고 프로그램 가능한 DRAM-기반 프로세싱 유닛(DPU)에 관한 것이다. 일 실시예에서, DPU는 DRAM-기반 메모리 셀 어레이 및 DRAM-기반 컴퓨팅 셀 어레이를 포함할 수 있으며, 이는 덧셈, 곱셈, 정렬(sort) 등과 같은 상이한 연산들을 수행하도록 구성될 수 있다.

[0017] DPU의 내부 아키텍처에는 서브-어레이들(sub-arrays)의 다중 뱅크들에 연결된 시스템 버스가 포함될 수 있다. 일 실시예에서, 시스템 버스는 서브-어레이들의 H-트리-연결된(H-tree-connected) 뱅크들을 제공하도록 구성될 수 있다. 각각의 서브-어레이는 로컬 컨트롤러를 포함할 수 있고, 각각의 개별 서브-어레이는 개별적으로 또는 동시에 활성화될 수 있다. 일 실시예에서, DRAM-기반 셀들은 2개의 어레이들 즉, 데이터 셀 어레이 및 컴퓨팅 셀 어레이로 분할될 수 있다. 일 실시예에서, 컴퓨팅 셀 어레이는 DRAM-기반 메모리 셀들에 의해 구현될 수 있다. 다른 실시예에서, 컴퓨팅 셀 어레이는 논리 회로를 갖는 DRAM-기반 메모리 셀들에 의해 구현될 수 있다. 또한 DPU 내부 아키텍처는 데이터-시프팅 및 데이터-이동 회로들을 포함할 수 있다. 일부 실시예들에서, 스토캐스틱(stochastic) 데이터 계산들을 위해 구성될 수 있는 제3 DRAM-기반 셀 어레이가 있을 수 있다.

[0018] 도 1은 본 명세서에 개시된 주제에 따른 DPU(100)의 예시적인 실시예의 블록도를 도시한다. DPU(100)는 하나 이상의 뱅크들(101a-101m)을 포함할 수 있으며, 그 중 뱅크들(101a, 101b)만이 도 1에 도시되어 있다. 각 뱅크(101)는 하나 이상의 서브-어레이들(102a-102n)을 포함할 수 있으며, 서브 어레이들(102a, 102b)만이 도 1에 도시되어 있다. 또한 각 뱅크(101)는 버퍼(103)를 포함할 수 있다. 버퍼(103)는 개별적인 서브-어레이들(102) 및 시스템 버스(104)에 연결될 수 있다. 버퍼(103)는 뱅크(102) 내 전체 행을 독출하고, 같은 뱅크나 다른 뱅크에 그 독출된 행을 다시 기입한다. 또한, 버퍼 (103)는 서브-어레이(102) 내의 다수의 매트들(mats, 105a-105n)에 행 데이터의 복사본을 브로드캐스트(broadcast) 할 수 있다. 일 실시예에서, 뱅크(101) 및 시스템 버스(104)는 H-트리-연결된 뱅크들을 제공하도록 구성될 수 있다.

[0019] 각 서브-어레이(102)는 하나 이상의 매트들(또는 레인들(lanes)) (105)을 포함할 수 있고, 매트(102a)의 매트들(105a-105n)이 도 1에 도시된다. 각 매트(105)는 데이터 셀 어레이(106), 컴퓨팅 셀 어레이(107) 및 매트 내 시프트 어레이(intra-mat shift array)(108)를 포함할 수 있는 DPU (100)의 영역이다. 예시적인 매트(105)는 도 1에서 점선(109)으로 둘러싸인 것으로 표시되어 있다. 각 매트(105)는 데이터 셀 어레이 디코더(110), 컴퓨팅 셀 어레이 디코더(111), 매트 간 시프트 어레이(inter-mat shift array)(112) 및 매트 간 포워딩 어레이(inter-mat forwarding array)(113)를 인접 매트와 공유할 수 있다. 일 실시예에서, 데이터 셀 어레이 디코더(110), 컴퓨팅 셀 어레이 디코더(111) 및 매트 간 시프트 어레이(112)는 인접 매트들(105) 사이의 서브-어레이 컨트롤러(114)와 물리적으로 교대로 배치될 수 있다. 일 실시예에서, 디코더들(110, 111)은 종래의 DRAM-타입 메모리 디코더들로서 동작할 수 있다.

[0020] 일 실시예에서, 각 매트(105)는 서브-어레이 컨트롤러(114)에 통신 가능하게 연결된다. 각 서브-어레이 컨트롤러(114)는 다른 서브-어레이 컨트롤러들(114)로부터 독립되도록 구성될 수 있다. 서브-어레이 제어기(114)는 DRAM 어드레스 버스로부터 명령들을 어드레스들(addr)로서 수신할 수 있다. 어드레스들(즉, 어드레스 신호들)에 응답하여, 서브-어레이 컨트롤러(114)는 데이터 셀 어레이(106) 및 컴퓨팅 셀 어레이(107) 중 하나 또는 모두에 디코딩된 어드레스를 출력으로서 제공할 수 있다. 즉, 서브-어레이 컨트롤러(114)는 관련된 데이터 셀 어레이(106)에 대해 디코더 (110)에 의해 디코딩된 소스/목적지(source/destination, src/dst) 어드레스들을 출력할 수 있고, 컴퓨팅 셀 어레이(107)의 경우에 디코더 (110)에 의해 디코딩된 연산/계산(operation/calculation, op/calc) 어드레스들을 출력할 수 있다. 또한 서브-어레이 컨트롤러(114)는 2개 이상의 서브-어레이 컨트롤러들(114)이 조직화된 방식으로 동작하게 하는 명령들을DRAM 버스로부터 어드레스로서 수신할 수 있다. 또한 서브-

어레이 컨트롤러(114)는 매트 내 시프트 어레이(108), 매트 간 시프트 어레이(112) 및 매트 간 포워딩 어레이(113)를 제어하는 것과 같은 데이터 이동 회로들을 제어할 수 있다.

[0021] 각각의 데이터 셀 어레이(106)는 적어도 하나의 열 및 적어도 하나의 행으로 배열된 하나 이상의 DRAM(dynamic random access memory) 셀들을 포함할 수 있다. 일 실시예에서, 데이터 셀 어레이(106)는 종래의 DRAM 셀 어레이로서 구성될 수 있다. 일 실시예에서, 데이터 셀 어레이(106)는 2K개의 열들 및 16개의 행들을 포함할 수 있다. 다른 실시예에서, 데이터 셀 어레이(106)는 2K보다 적거나 많은 열들 및/또는 16개보다 적거나 많은 행들을 포함할 수 있다.

[0022] 각각의 컴퓨팅 셀 어레이(107)는 적어도 하나의 열 및 적어도 하나의 행으로 배열된 하나 이상의 컴퓨팅 셀들을 포함할 수 있다. 컴퓨팅 셀 어레이(107) 내 열들의 수는 데이터 셀 어레이(106) 내 열들의 수와 동일하다. 일 실시예에서, 컴퓨팅 셀 어레이(107)는 2K 열들 및 16 행들을 포함할 수 있다. 또 다른 실시예에서, 컴퓨팅 셀 어레이(107)는 2K보다 적거나 많은 열들 및/또는 16개보다 적거나 많은 행들을 포함할 수 있다.

[0023] 도 2a는 컴퓨팅 셀 어레이(107) 에서 컴퓨팅 셀로 사용될 수 있는 3-트랜지스터, 1-커패시터(3T1C) DRAM 컴퓨팅-셀 토포그래피(topography)(201)의 예시적인 실시예를 도시한다. 도 2a에 도시 된 바와 같이, 행(Row X) 내 3T1C 컴퓨팅 셀은 기입 비트라인(Write BL)에 전기적으로 연결된 소스 단자, 커패시터(C1)의 제1 단자와 제2 트랜지스터(T_2)의 게이트 단자에 전기적으로 연결된 드레인 단자 및 기입 인에이블(WEN) 라인에 전기적으로 연결된 게이트 단자를 포함하는 제1 트랜지스터(T_1)를 포함한다. 제1 커패시터(C_1)의 제2 단자는 접지 라인과 전기적으로 연결된다. 제2 트랜지스터(T_2)는 그라운드 라인에 전기적으로 연결된 소스 단자 및 제3 트랜지스터(T_3)의 소스 단자에 전기적으로 연결된 드레인 단자를 포함한다. 제3 트랜지스터(T_3)는 워드라인(WL)에 전기적으로 연결된 게이트 단자 및 독출 비트라인(Read BL)에 전기적으로 연결된 드레인 단자를 포함한다. 3T1C 컴퓨팅-셀 토포그래피(201)는 독출 비트라인(Read BL) 전기적으로 연결된 입력 및 기입 비트라인(Write BL)에 전기적으로 연결된 출력을 갖는 감지 증폭기(SA)를 포함한다.

[0024] 행(Row Y)의 컴퓨팅 셀 및 행(Row R)의 컴퓨팅 셀은 모두 행(Row X) 내 컴퓨팅 -셀의 배열과 유사한 3T1C DRAM 구성으로 배열된 3개의 트랜지스터들(T_1-T_3) 및 커패시터(C)를 포함할 수 있다. 예를 들어, 도 2a에 도시된 3개의 컴퓨팅 셀들 및 감지 증폭기(SA)는NOR 논리 연산, 즉 결과가 행R(Row R)에 저장되는 X NOR Y 논리 연산을 제공하도록 구성된다. 비록3T1C DRAM 컴퓨팅 셀들의 단지 하나의 열만이 도 2a에 명시적으로 도시되어 있지만, 다른 실시예에서, 3T1C 컴퓨팅 셀들은 다수의 열들(즉, 2K 개 열들)로 구성될 수 있다. 다른 실시예에서, 3개 이상의 행들이 제공될 수 있음을 이해해야 한다. 또한, 도 2a에 도시 된 3T1C DRAM 컴퓨팅-셀 구성은 NOR 논리 연산을 제공하지만, 3T1C DRAM 컴퓨팅-셀 토포그래피(201)의 NOR 논리 연산은 XNOR(exclusive NOR), ADD(addition), SET(select), MAX, SIGN, MUX(multiplex), CSA(conditional-sum addition logic), 곱셈(multiply), 팝카운트(popcount) 및 COMPARE등과 같은 기능적 연산들을 제공하기 위해 활용될 수 있으나, 이들에 한정되지 않는다. 또한 시프트 어레이들(108, 112)은 시프팅 기능을 제공한다.

[0025] 도 2b는 도 1의 컴퓨팅 셀 어레이(107)에서컴퓨팅 셀로 사용될 수 있는 1-트랜지스터, 1-커패시터(1T1C) DRAM 컴퓨팅-셀 토포 그래피(202)의 대안적인 실시예를 도시한다. 도 2 b에 도시된 바와 같이, 1T1C DRAM 컴퓨팅 셀은 커패시터(C_2)의 제1 단자에 전기적으로 연결된 소스 단자, 비트라인(BL)에 전기적으로 연결된 드레인 단자 및 워드라인(WL)에 전기적으로 연결된 게이트 단자를 포함하는 트랜지스터(T_4)를 포함한다. 비트라인(BL)에 전기적으로 접속 된 드레인 단자 및 은 워드라인(WL)에 전기적으로 접속된다. 커패시터(C_2)의 제2 단자는 접지 라인에 전기적으로 연결(couple)된다. 비트라인(BL)은 감지 증폭기(SA)의 입력에 전기적으로 연결된다. 감지 증폭기(SA)의 출력은 멀티플렉서(MUX)의 제 1 입력, 트랜지스터(T_5)의 드레인 단자 및 산술 논리 유닛(arithmetic logic unit) (ALU)의 입력에 전기적으로 연결 된다. 멀티플렉서(MUX)의 출력은 래치(LATCH)의 입력에 전기적으로 연결된다. 트랜지스터(T_5)의 소스 단자는 래치(LATCH)의 출력에 전기적으로 연결된다. ALU의 출력은 멀티플렉서(MUX)의 제2 입력에 전기적으로 연결된다. 도 2b의 트랜지스터(T_5), 멀티플렉서(MUX), 래치(LATCH) 및 산술 논리 유닛(ALU)은 컨트롤러(114)로부터 제어 신호들(CNTL1-CNTL4)을 각각 수신한다. 일 실시예에서, 산술 논리 유닛(ALU)은 NOR 기능을 제공하도록 구성될 수 있다. 도 2b의 비트라인(BL)에 전기적으로 연결된 논리 회로는 NOR 논리 연산을 제공할 수 있으나, 비트라인 (BL)에 전기적으로 연결된 논리 회로 즉, 산술 논리 유닛(ALU)은 XNOR(exclusive NOR), ADD(addition), SET(select), MAX, SIGN, MUX(multiplex), CSA(conditional-sum addition logic), 곱셈, 팝카운트(popcount) 및 COMPARE등과 같은 다른 기능적 연산들을 제공할 수 있으나, 이

들에 한정되지 않는다. 또한 시프트 어레이들(108, 112)은 시프팅 기능을 제공한다. 단지 하나의 1T1C 컴퓨팅 셀이 도 2b에 도시되어 있으나, 1T1C 컴퓨팅 셀들의 다수의 열들 및 행들이 제공될 수 있다는 것을 이해해야 한다.

[0026] 도 2a 및 도 2b에 도시된 바와 같이, DPU의 컴퓨팅 셀들은 특정의 복잡한 컴퓨팅 로직들(computing logics)을 포함하지 않지만, 대신 다수의 다양한 타입들의 계산들을 수행하는 능력을 제공하는 재-프로그래밍 가능한 특성(re-programmable nature)을 갖는 비교적 간단한 토포그래피를 포함한다. 또한 DPU의 토포그래피는 메모리 구조에 내재된 대량 병행성들(massive parallelisms)을 활용하여 더 많은 계산들을 더 빠르고 효율적으로 수행하도록 배열될 수 있다.

[0027] 도 3은 본 명세서에 개시된 주제에 따른 매트 내 시프트 어레이(108)의 예시적인 실시예를 도시한다. 매트 내 시프트 어레이(108)의 설명을 단순화하기 위해, 도 3에 도시된 것과 같은, 컴퓨팅 셀들(107)의 4 개의 열들 너비인 매트(105)를 고려한다. 매트 내 시프트 어레이(108)는 어레이 내에 배열된 복수의 트랜지스터들(T_6)(복수의 트랜지스터들 중 단지 하나의 트랜지스터(T_6)만 도 3에 도시되어 있음), 2n개의 시프트 라인들(SLs) (여기서 n은 매트(105) 내 컴퓨팅 셀들의 열들), n+2개의 시프트 레프트 컨트롤 라인들(shift left control lines, SLcLs), 2개의 시프트 라이트 컨트롤 라인들(shift right control lines, SRcLs) 및 n개의 시프트 마스크 라인들(shift mask lines, SMLs)을 포함한다. 매트 내 시프트 어레이(108)의 일부 트랜지스터들(T_6)은 기입 비트라인들(BLs) 및 2n개의 시프트 라인들(SLn) 사이에 전기적으로 연결되고, 매트 내 시프트 어레이(108)의 다른 트랜지스터들(T_6)은 독출 비트라인들(BLs) 및 2n개의 시프트 라인들(SLn) 사이에 연결된다. 이들 트랜지스터들(T_6)의 게이트들은 n+2개의 시프트 레프트 제어 라인들(SLcLs) 및 2개의 시프트 라이트 제어 라인들(SRcLs)에 전기적으로 연결된다. 매트 내 시프트 어레이 내 다른 트랜지스터들(T_6)은 n개의 시프트 마스크 라인들(SMLs) 및 2n개의 시프트 라인들(SLs) 사이에 전기적으로 연결된다. 매트 내 시프트 어레이(108) 내 제어 라인들은 매트(105)와 관련된 서브-어레이 컨트롤러(114)와 전기적으로 연결된다.

[0028] 매트 내 시프트 어레이(108)는 제어 라인들(SLcLs, SRcLs) 상의 적절한 신호들에 의해 매트(105) 내에서 데이터를 좌측(left) 또는 우측(right)으로 시프트할 수 있다. 왼쪽 시프트의 경우, 데이터는 부호(sign) 비트로 채워질 수 있으며, 데이터는 연산마다 1 비트 또는 (n-1) 비트들 시프트되고, n은 매트(105)당 열들의 수이다. 오른쪽 시프트의 경우, 명령들에 의한 제어대로 데이터는 0 또는 1으로 채워지며, 매트(MAT) 당 열들의 수 2^0, 2^1, ..., 2^{k-1}, 2^k까지 시프트되고, 2^k 는 열들의 수이다.

[0029] 도 4a는 본 명세서에 개시된 과제에 따른 매트 간 시프트 어레이(112)의 실시예를 도시한다. 매트 간 시프트 어레이(112)의 간단한 설명을 위하여, 도 4a 내지 도 4c에 도시된 바와 같이, 컴퓨팅 셀들(107)의 2개의 열들의 너비인 매트(15)의 구성을 고려한다. 즉, 각 매트(105)는 제1 컴퓨팅 셀들 열(107a) 및 제2 컴퓨팅 셀들 열(107b)을 포함한다. 매트 간 시프트 어레이(112)는 트랜지스터들(T_{112a} 및 T_{112b}), 트랜지스터들(T_{112c} 및 T_{112d}), 데이터 시프트 라인들(112e, 112f) 및 매트 내 시프트 제어 라인들(ISLcLs)을 포함한다. 매트 내에서, 트랜지스터(T_{112a})는 컴퓨팅 셀들(107a)의 제1 열의 독출 비트라인(Read BL)에 전기적으로 연결되는 소스 단자, 데이터 시프트 라인(112e)에 전기적으로 연결된 드레인 단자를 포함한다. 트랜지스터(T_{111b})는 컴퓨팅 셀(107b)의 제2 열의 독출 비트라인(Read BL)에 전기적으로 연결되는 소스 단자, 데이터 시프트 라인(112f)에 전기적으로 연결되는 드레인 단자를 포함한다. 데이터 시프트 라인들(112e, 112f)은 버퍼(103)(도 4a에 미 도시됨)에 전기적으로 연결된다. 서로 다른 매트들 사이에서, 트랜지스터(T_{112c})는 인접한 매트들 내의 데이터 시프트 라인들(112e)에 각각 전기적으로 결합되는 소스 및 드레인 단자들을 포함한다. 트랜지스터(T112d)는 인접한 매트들 내의 데이터 시프트 라인들(112f)에 각각 전기적으로 연결되는 소스 및 드레인 단자들을 포함한다. 트랜지스터들(T_{112c}, T_{112d})의 게이트들은 각각 다른 매트 간 시프트 제어 라인들(inter-mat shift control lines)(ISLcLs)에 전기적으로 연결된다. 매트 간 시프트 어레이(112)는 제어 라인들(ISLcLs) 상의 적절한 신호들에 의해 상이한 매트들 사이에서 데이터를 좌측 또는 우측으로 시프트할 수 있다. 매트 간 시프트 어레이(112)의 제어 라인들은 매트(105)와 관련된 서브-어레이 컨트롤러(114)에 전기적으로 연결된다.

[0030] 도 4b는 본 명세서에 개시된 주제에 따른 좌측 매트 간 시프트를 위해 인접한 컴퓨팅-셀 열들(105a, 105b) 내의 동일하게 위치된 2 개의 컴퓨팅 셀들 사이의 매트 간 시프트 상호 접속 구성을 개념적으로 도시한다. 도 4b의 상호 접속 구성은 강조되는 상호 접속 노드들의 예시에 의해서 개념적으로 도시될 수 있다. 예를 들어, 트랜지

스터들(T_{112c}, T_{112d})은 각 트랜지스터 사이에 도전 경로(conductive path)가 존재하도록 활성화되어, 데이터 시프트 라인들(112e, 112f)을 (좌측의) 컴퓨팅-셀 열들(105a) 및 (우측의) 컴퓨팅-셀 열들(105b) 사이에 연결한다. 트랜지스터들(T_{112c}, T_{112d})의 게이트 단자들은 액티브 매트 간 시프트 컨트롤 라인(ISLcL)에 전기적으로 연결된다. 매트(105b) 내 트랜지스터들(T_{112a}, T_{112b})이 활성화되어서 매트(105b) 내 컴퓨팅 셀(107a)의 독출 비트라인(Read BL)이 매트(105b)의 좌측 매트(105a)의 컴퓨팅 셀(107a)의 기입 비트라인(Write BL)에 전기적으로 연결되고, 매트(105b) 내의 컴퓨팅 셀(107b)의 독출 비트라인(Read BL)은 매트 (105b)의 좌측의 매트(105a) 내 컴퓨팅 셀(107a)의 기입 비트라인(Write BL)에 전기적으로 연결된다.

[0031] 도 4c는 본 명세서에 개시된 주제에 따른 좌측 매트 간 시프트에 대한 인접한 컴퓨팅-셀 열들(105a, 105b) 내의 2개의 동일하지 않게 배열된 컴퓨팅 셀들 간의 매트 간 시프트 상호 접속 구성을 개념적으로 도시한다. 도 4c의 상호 접속 구성은 강조되는 상호 접속 노드들의 예시에 의해서 개념적으로 도시될 수 있다. 예를 들어, 트랜지스터들(T_{112c}, T_{112d})은 각 트랜지스터 사이에 도전 경로가 존재하도록 활성화되어, 데이터 시프트 라인들(112e, 112f)을 (우측의) 컴퓨팅-셀 열들(105a) 및 (좌측의) 컴퓨팅-셀 열들(105b)에 연결한다. 트랜지스터들(T_{112c}, T_{112d})의 게이트 단자들은 액티브 매트 간 시프트 제어 라인(ISLcL)에 전기적으로 연결되어 있다. 매트(105a) 내의 트랜지스터들(T_{112a}, T_{112b})은 활성화되어서 매트(105a) 내 컴퓨팅 셀(107a)의 독출 비트라인(Read BL)이 매트(105a)의 좌측의 매트(105b) 내의 컴퓨팅 셀(107a)의 기입 비트라인(Write BL)에 전기적으로 연결되고, 매트(105a) 내 컴퓨팅 셀(107b)의 독출 비트라인(Read BL)이 매트(105a)의 좌측 매트(105b) 내 컴퓨팅 셀(107a)의 기입 비트라인(Write BL)에 전기적으로 연결된다.

[0032] 도 5는 본 명세서에 개시된 주제에 따른 매트 간 포워딩 어레이(113)의 일 실시예를 도시한다. 매트 간 포워딩 어레이(113)의 설명을 단순화하기 위해, 도 5에 도시된 바와 같이, 매트들(105)은 컴퓨팅 셀들(107)의 2개의 열들의 너비인 매트(15)의 구성을 고려한다. 즉, 각 매트(105)는 제1 컴퓨팅 셀들 열(107a) 및 제2 컴퓨팅 셀들 열(107b)을 포함한다. 매트(105)와 함께, 매트 간 포워딩 어레이(113)는 트랜지스터들(T_{113a}, T_{113b}), 트랜지스터들(T_{113c}, T_{113d}) 트랜지스터들(T_{113e}, T_{113f}), 2^n개 데이터 포워딩 라인들(FDL)(여기서, n은 매트 내 컴퓨팅-셀 열들의 수), 포워딩 컨트롤 라인들(FCL) 및 2^m개 포워딩 섹션 라인들(FSL) (여기서, m은 섹션들의 수)를 포함한다. 트랜지스터들(T_{113a}, T_{113b})의 소스 단자들은 컴퓨팅 셀들(107a)의 제1 열의 기입 비트라인(Write BL) 및 독출 비트라인(Read BL)에 전기적으로 각각 연결된다. 트랜지스터들(T_{113a}, T_{113b})의 드레인 단자들은 제1 데이터 포워딩 라인(FDL)(113g)에 전기적으로 연결된다. 트랜지스터들(T_{113c}, T_{113d})의 소스 단자들은 컴퓨팅 셀들(107b)의 제2 열의 기입 비트라인(Write BL) 및 독출 비트라인(Read BL)에 전기적으로 각각 연결된다. 트랜지스터들(T_{113a}, T_{113b})의 드레인 단자는 제 2 데이터 포워딩 라인(FDL)(113h)에 전기적으로 연결된다. 트랜지스터들(T_{113e}, T_{113f})의 소스 단자들은 트랜지스터들(T_{113a}, T_{113b})의 게이트 단자들에 각각 전기적으로 연결된다. 트랜지스터들(T_{113e}, T_{113f})의 드레인 단자들은 모두 동일한 포워딩 제2 라인들(FSLs)에 연결된다. 트랜지스터들(T_{113e}, T_{113f})의 게이트 단자들은 각각 다른 포워딩 제어 라인들(FCLs)에 연결된다. 매트 간 포워딩 어레이(113)는 포워딩 제어 라인들(FCLs) 상의 적절한 신호들에 의해 매트들 간에 데이터를 포워드(forward) 할 수 있다. 매트 간 포워딩 어레이(113)의 제어 라인들은 데이터가 포워드되는 매트들(105)과 관련된 서브-어레이 컨트롤러(114)에 전기적으로 연결된다.

[0033] 도 6a 내지 도 6g는 본 명세서에 개시된 주제에 따라 DPU에 의해 제공될 수 있는 NOR-기반 연산들을 나타낸다. 도 6a내지 도 6g에서, 제 1 피연산자가 행 X(Row X)에 저장될 수 있고 제 2 피연산자가 행 Y(Row Y) 또는 행 W(Row W)에 저장될 수 있다. 도 6a 내지 도 6g내의 화살표 들은 컴퓨팅 셀들의 전체 행에 대한 NOR 논리 연산의 입력 및 출력 흐름들을 나타낸다. 예를 들어, 도 6a내의 행 X(Row X)는 행X (Row X)의 컴퓨팅 셀들에 저장된 피연산자들 전체 행을 나타낼 수 있다. 행X(Row X)에 저장된 피연산자들 및 행Y(Row Y)에 저장된 피연산자들에 대한 NOR 논리 연산의 결과는 결과 행(Row R)에 저장된다. 일 실시예에서, 행X(Row X) 및 행Y(Row Y) 내 피연산자들은 예컨대, 100개의 열들(즉, x_1, x_2, ... , x_{100} 및 y_1, y_2, ..., y_{100})을 포함하고, 결과는 행R(Row R)(즉, r_1, r_2, ... , r_{100})에 저장될 수 있다. 즉, x_i nor y_i = r_i 이고, 여기서 i는 열 인덱스이다. 다른 실시예에서, 행X(Row X)는 한 행의 컴퓨팅 셀들의 선택된 그룹만을 나타낼 수 있다.

[0034] 도 6b는 프리픽스 Kogge-Stone 덧셈기에 기초한 N-비트 수에 대한 예시적인 전가산기(full adder) 연산을 나타

낸다. 도 6b에서, 제1 N-비트 피연산자는 행X(Row X)에 저장되고 제2 N-비트 피연산자는 행Y(Row Y)에 저장된다. 도 6B에 도시된 가산 연산의 예에서, 중간 항들 G_0, P_0, G_1, P_1, G_2, P_2, ... , G_{logN+1} 및 P_{logN+1} 이 계산된다. 도 6b의 최상단 블록은, 행X(Row X) 및 행Y(Row Y)으로부터의 입력 피연산자들을 사용하여 G_0 및 P_0 을 판별하는 5개의 개별 연산들을 나타낸다. 제1 연산에서 최상단 블록은 행(Row X) 의 역수(즉, ~X)를 판별하고, 행1(Row 1)에 저장된다. 제2 연산은 행(Row Y)의 역수(즉, ~Y)를 판별하고, 행2(Row 2)에 저장된다. 제3 연산은 Row X NOR Row Y를 판별하고, 행3(Row 3)에 저장된다. 제4 연산은 연산 G_0 = Row 1 NOR Row 2를 판별하고, 행4(Row 4)에 저장된다. 제5 연산은 P_0 = Row 3 NOR Row 4를 판별하고, 행5(Row 5)에 저장된다.

[0035] 도 6b의 중간 블록에서, 최상위 블록으로부터의 중간 결과 G_0 및 P_0 는 i가 열 인덱스인 중간 결과들 G_{i+1} 및 P_{i+1} 을 판별하는데 사용된다. 즉, 도 6a의 최상단 블록에서 판별된 중간 결과들 G_0 및 P_0는 중간 결과들 G_1 및 P_1을 판별하기 위해 사용된다. 중간 결과 G_1 및 P_1은 중간 결과 G_2 및 P_2 등을 판별하기 위해 사용되며, 중간 결과들 G_{logN+1} 및 P_{logN+1}을 판별한다. 도 6b의 최하단 블록에서, 결과 행들(R1, R2)은 캐리 결과 및 전 가산기 연산에 대한 합산 결과를 각각 저장한다.

[0036] 도 6c는 3T1C DRAM 컴퓨팅-셀 토포그래피(201)에 의해 제공될 수 있는 예시적인 선택기 동작을 도시한다. 행 1은 행X(Row X)의 역수(즉, ~X)의 중간 결과를 저장한다. 행 2는 행Y(Row Y)의 역수(즉, ~Y)의 중간 결과를 저장한다. 행 3은 행S(Row S)의 역수(즉, ~S)의 중간 결과를 저장한다. 행 4는 행1 NOR 행 3의 중간 결과를 저장한다. 행 5는 행 2 NOR 행S(Row S)의 중간 결과를 저장한다. 행 6은 행 4 NOR 행 5의 중간 결과를 저장한다. 행 R(Row R)은 행 6의 역수를 저장한다. 즉, S?X:Y 이다.

[0037] 도 6d는 3T1C DRAM 컴퓨팅-셀 토포그래피(201)에 의해 제공될 수 있는 다른 예시적인 선택기 동작을 도시한다. 행 1은 행X(Row X)의 역수(즉, ~X)의 중간 결과를 저장한다. 행 2는 행 S(Row S)의 역수(즉, ~S)의 중간 결과를 저장한다. 행 3은 행 1 NOR 행 3의 중간 결과를 저장한다. 행 4는 행 X의 역수(즉, ~X)의 중간 결과를 저장한다. 행 (Row R)은 행 3 NOR 행 4의 결과, 즉 S?X:~X를 저장한다.

[0038] 도 6e는 3T1C DRAM 컴퓨팅-셀 토포그래피(201)에 의해 제공될 수 있는 예시적인 MAX/MIN 동작을 도시한다. 행 1은 행Y(Row Y)의 역수(즉, ~ Y) 의 중간 결과를 저장한다. 행 2는 행 X+(~Y+1)의 중간 결과를 저장한다. 행 3은 C_{out} >> n의 중간 결과를 저장한다. 행 4는 C_{out}X:Y의 중간 결과를 저장한다. 행 R(Row R)은 MAX (X:Y)의 결과를 저장한다.

[0039] 도 6f는 3T1C DRAM 컴퓨팅-셀 토포그래피(201)에 의해 제공될 수 있는 예시적인 1-비트 곱셈 연산을 도시한다. 행 1은 행 X NOR 행W(Row W)의 중간 결과를 저장한다. 행 2는 행X(Row X) NOR 행 1의 중간 결과를 저장한다. 행 3은 행W(Row W) NOR 행 1의 중간 결과를 저장한다. 결과 행R(Row R)은 행 2 NOR 행 3의 결과, 즉 행(Row X) XNOR 행(Row W)의 결과를 저장한다.

[0040] 도 6g는 3T1C DRAM 컴퓨팅-셀 토포그래피 201)에 의해 제공될 수 있는 예시적인 다중-비트 곱셈 연산을 도시한다. 도 6g의 상측 블록에서, 행 1은 행W(Row W)의 역수(즉, ~W)의 중간 결과를 저장한다. 행 2는 2^i 번 왼쪽으로 시프트된 행X(Row X)의 역수의 중간 결과(즉, ~X<<2^i)를 저장한다. 여기서 i는 인덱스이다. 행 3은 행 1 NOR 행 2의 중간 결과, 즉 PP_i = ~W NOR ~X<<2^i를 저장한다. 도 6g의 하측 블록에서, 행 1은 행 PP_0 (Row PP_0) SUM 행 PP_i (Row PP_i)의 중간 결과, 즉, PP_i를 저장한다. 행 2는 행 2 NOR 행 W_{sign} (Row W_{sign})의 중간 결과를 저장한다. 행R(Row R)은 X*W의 결과를 저장한다.

[0041] 도 7은 본 명세서에 개시된 주제에 따른 스토캐스틱(stochastic) 데이터 영역(715)을 포함하는 DPU(700)의 예시적인 실시예의 블록도를 도시한다. 도 1에 도시된 DPU(100)의 구성 요소들과 유사한 DPU(700)의 구성 요소들은 DPU(100)의 구성 요소들과 동일한 참조 부호들을 가지며, 도 1의 구성 요소와 유사한 구성 요소들에 대한 설명을 생략하였다. DPU (700)의 서브-어레이(102)는 (실제) 데이터 셀 어레이(106), 컴퓨팅 셀 어레이(107) 및 매트 간 시프트 어레이(108)와 함께 스토캐스틱 데이터 어레이(715) 및 컨버터-스토캐스틱 어레이(716)를 포함한다.

[0042] 각각의 스토캐스틱 데이터 어레이(715)는 적어도 하나의 열 및 적어도 하나의 행으로 배열되는 하나 이상의 스토캐스틱 컴퓨팅 셀들을 포함할 수 있다. 일 실시예에서, 스토캐스틱 데이터 어레이(715)는 2K 개의 열들 및 16

개의 행들을 포함할 수 있다. 스토캐스틱 데이터 어레이(715)의 열들 수는 데이터 셀 어레이(106) 및 컴퓨팅 셀 어레이(107)의 열들의 수와 동일하다. 다른 실시예에서, 스토캐스틱 데이터 어레이(715)는 2K개 보다 적거나 또는 많은 열들 그리고/또는 16개보다 적거나 또는 많은 행들을 포함할 수 있다. 스토캐스틱 데이터 어레이(715)에서, "1"의 존재의 확률이 사용되고, n-비트 값을 나타내기 위해 2^n-비트가 사용된다. 컨버터-스토캐스틱 어레이(converter-to-stochastic array)(716) 내의 난수 발생기는 실수를 스토캐스틱 수로 변환하는데 사용될 수 있다. 팝카운트(popcount) 연산은 스토캐스틱 수를 실수로 다시 변환하기 위해 사용될 수 있다.

[0043] 스토캐스틱 컴퓨팅 접근법을 사용함으로써, 덧셈(addition)은 멀티플렉싱(multiplexing) 연산으로 변환될 수 있고, 곱셈(multiplication)은 AND 논리 연산으로 변환될 수 있다. 예를 들어, 도 8a는 멀티플렉싱 연산으로서 스토캐스틱 덧셈 연산을 제공하는 회로를 도시하고, 도 8b는 AND 논리 연산으로서 스토캐스틱 곱셈 연산을 제공하는 회로를 도시한다. 스토캐스틱 컴퓨팅을 위한 종래의 기술은 상당한 메모리 용량을 필요로 한다. 그러나, 본 명세서에 개시된 주제는 DRAM-기반 DPU들이 큰 병렬 AND 및 MUX 연산들을 수행할 수 있기 때문에 매우 효율적인 스토캐스틱 컴퓨팅을 제공하는데 사용될 수 있다. 또한 본 명세서에 개시된 DPU를 사용하는 스토캐스틱 컴퓨팅은 복잡한 연산들(딥 러닝이 대표적인 애플리케이션)이 가속화되는 것을 가능하게 하였다.

[0044] 도 9는 본 명세서에 개시된 주제에 따른 DPU들을 포함하는 시스템 아키텍처(900)를 도시한다. 시스템 아키텍처(900)는 하드웨어 계층(910), 라이브러리 및 드라이버 계층(920), 프레임워크 계층(930) 및 애플리케이션 계층(940)을 포함할 수 있다.

[0045] 하드웨어 계층(910)은 여기에 설명된 DPU들과 같은, 내장된 DPU들을 갖는 하드웨어 장치들 및/또는 컴포넌트들을 포함할 수 있다. 디바이스 및/또는 컴포넌트의 일 실시예는 하나 이상의 내장된 DPU들을 포함할 수 있는 PCIe(Peripheral Component Interconnect Express) 디바이스(911)일 수 있다. 디바이스 및/또는 컴포넌트의 다른 실시예는 하나 이상의 내장된 DPU들을 포함할 수 있는 DIMM(Dual In-line Memory Module)(912)일 수 있다. 시스템 아키텍처(900)의 하드웨어 계층(910)은 PCIe 장치 및/또는 DIMMs에 제한되지 않지만, SOC(System on a Chip) 장치들 또는 DPU들을 포함할 수 있는 다른 메모리-타입 장치들을 포함할 수 있음을 이해해야 한다. 하드웨어 계층(910)에서 디바이스들 및/또는 컴포넌트들에 내장될 수 있는 DPU들은 도 1의 DPU(100)와 유사하게 구성될 수 있고, 그리고/또는 도 7의 DPU(700)와 유사하게 구성될 수 있다. 임의의 실시예에서, DPU의 특정 컴퓨팅 셀 어레이들은 3T1C 컴퓨팅-셀 토포그래피(201)(도 2a) 또는 1T1C 컴퓨팅-셀 토포그래피(202)(도 2b)를 포함하도록 구성될 수 있다.

[0046] 시스템 아키텍처(900)의 라이브러리 및 드라이버 계층(920)은 DPU 라이브러리(921), DPU 드라이버(922) 및 DPU 컴파일러(923)를 포함할 수 있다. DPU 라이브러리(921)는, 애플리케이션 계층(940)에서 동작할 수 있는 상이한 애플리케이션들에 대해 하드웨어 계층(910) 내 DPU에서 각 서브-어레이에 대해 최적의 매핑 기능, 리소스 할당 기능 및 스케줄링 기능을 제공하도록 구성될 수 있다.

[0047] 일 실시예에서, DPU 라이브러리(921)는 이동(move), 덧셈(add), 곱셈(multiply) 등과 같은 연산들을 포함할 수 있는 프레임 워크 계층(930)에 대한 하이-레벨 API(application programming interface)를 제공할 수 있다. 예를 들어, DPU 라이브러리(921)는 순방향 및 역방향 컨볼루션(convolution), 풀링(pooling), 정규화(normalization) 및 가속화된 딥 러닝 프로세스에 적용 할 수 있는 활성화 계층들과 같은 (그러나 이에 한정되지 않는) 표준-타입 루틴들에 대한 구현들도 포함할 수 있다. 일 실시예에서, DPU 라이브러리(921)는 CNN(convolution neural network)의 전체 컨볼루션 계층에 대한 계산을 매핑하는 API-유사 기능을 포함할 수 있다. 또한, DPU 라이브러리(921)는 DPU 상에 컨볼루션 계층 계산의 맵핑을 최적화하기 위한 API-유사 기능을 포함할 수 있다.

[0048] DPU 라이브러리(921)는, 칩, 뱅크, 서브-어레이 및/또는 매트 레벨에서, 태스크(배치(batch), 출력 채널, 픽셀들, 입력 채널들, 컨볼루션 커널들(convolution kernels)) 내의 임의의 개별(individual) 또는 다중 병행화들(multiple parallelisms)을 대응하는 DPU 병행화들로 매핑함으로써 자원 할당을 최적화하기 위한 API-유사 기능들을 포함할 수 있다. 또한, DPU 라이브러리(921)는 초기화시 및/또는 실행시(runtime) 성능(즉, 데이터 이동 흐름) 및 전력 소비의 균형을 유지하는 최적의 DPU 구성을 제공하는 API-유사 기능을 포함할 수 있다. DPU 라이브러리(921)에 의해 제공되는 API-유사 다른 기능들은 뱅크 당 액티브 서브 어레이들의 수, 액티브 서브 어레이들 당 입력 특징 맵들의 수, 특징 맵의 파티셔닝, 그리고/또는 컨볼루션 커널의 재사용 스킴 등을 설정하는 것과 같은 디자인-노브-타입(design-knob-type) 기능들을 포함할 수 있다. 또 API-유사 다른 기능들은, 각 서브어레이에 대해 컨볼루션 컴퓨팅, 채널 썸 업(channel sum up) 및/또는 데이터 디스패치(dispatch)와 같은 특정 태스크를 할당함으로써 부가적인 자원 할당 최적화를 제공할 수 있다. 피연산자들이 정수와 스토캐스틱 수 사이에

서 변환되는 경우, DPU 라이브러리(921)는 정밀도 제약들(precision constraints)을 만족시키면서 오버 헤드를 최소화하는 API-유사 기능들을 포함한다. 정밀도가 예상보다 낮을 경우, DPU 라이브러리(921)는 스토캐스틱 표현을 위한 부가적 비트들을 사용하여 값을 다시 계산하거나 또는 CPU와 같은 다른 하드웨어로 태스크를 넘기는(offload) API-유사 기능들을 포함할 수 있다.

[0049] 또한 DPU 라이브러리(921)는 DPU에서 활성화된 서브-어레이들을 동시에 스케줄링하고, 컴퓨팅 연산들에 의해 숨겨지도록 데이터 이동을 스케줄링하는 API-유사 기능들을 포함할 수 있다.

[0050] DPU 라이브러리(921)의 다른 양태는 추가의DPU 개발을 위한 확장 인터페이스를 포함한다. 일 실시예에서, DPU 라이브러리(921)는 NOR 및 시프트 논리(shift logic)를 사용하여 기능을 직접 프로그램하기 위한 인터페이스를 제공하여 표준-타입 연산들(즉, 덧셈, 곱셈, MAX/MIN 등) 이외의 연산들이 제공될 수 있다. 또한, DPU 라이브러리(921)에 의해 특별히 지원되지 않는 연산이 라이브러리 및 드라이버 계층(920)에서 SoC 컨트롤러(도시되지 않음), CPU/GPU(central processing unit/graphics processing unit) 컴포넌트 및/또는 CPU/TPU(CPU/Tensor Processing Unit) 컴포넌트로 오프로드(offload)될 수 있도록 확장 인터페이스는 인터페이스를 제공할 수 있다. DPU 라이브러리(921)의 또 다른 양상은 DPU 메모리가 컴퓨팅을 위해 사용되지 않을 때 메모리 확장으로서 DPU의 메모리를 사용하는 API-유사 기능을 제공한다.

[0051] DPU 드라이버(922)는 하드웨어 계층(910)에서 DPU, DPU 라이브러리(921), 및 상위 계층의 오퍼레이팅 시스템(OS) 사이의 인터페이스 연결을 제공하여 DPU 하드웨어 계층을 시스템에 통합할 수 있도록 구성될 수 있다. 즉, DPU 드라이버(922)는 DPU를 시스템 OS 및 DPU 라이브러리(921)에 노출시킨다. 일 실시예에서, DPU 드라이버(922)는 초기화시 DPU 제어를 제공할 수 있다. 일 실시예에서, DPU 드라이버(922)는 DRAM-타입 어드레스들의 형태 또는 DRAM-타입 어드레스들의 시퀀스들로 명령들을 DPU에 전송할 수 있고, DPU 내외로의 데이터 이동을 제어할 수 있다. DPU 드라이버(922)는 DPU-CPU 및/또는 DPU-GPU 통신들을 핸들링하는 것과 함께 다중-DPU 통신을 제공할 수 있다.

[0052] DPU 컴파일러(923)는, DPU를 제어하기 위해 DPU 드라이버(922)에 의해 사용되는 메모리 어드레스들의 형태로 DPU 라이브러리(921)로부터의 DPU 코드를 DPU 명령들로 컴파일 할 수 있다. DPU 컴파일러(923)에 의해 생성된 DPU 명령들은 DPU의 하나 및/또는 두 개의 행들 상에서 동작하는 단일 명령들일 수 있다. 단일 명령어들은 벡터 명령들 및/또는 수집된 벡터, 리드-온-연산(read-on-operation) 명령들을 포함할 수 있다.

[0053] 프레임워크 계층(930)은 라이브러리 및 드라이버 계층(920) 그리고 하드웨어 계층(910)에 사용자-친화적인 인터페이스를 제공하도록 구성될 수 있다. 일 실시예에서, 프레임워크 계층(930)은 애플리케이션 계층(940)에서의 광범위한 애플리케이션들과 호환되는 사용자-친화적인 인터페이스를 제공하고 DPU 하드웨어 계층(910)을 사용자에게 투명하게 만든다. 다른 실시예에서, 프레임워크 계층(930)은 토치-7(Torch-7)-타입 애플리케이션들 및 텐서플로(TensorFlow)-타입 애플리케이션들과 같은 기존의 종래의 방법들에 정량화 기능들을 추가하는 프레임워크 확장을 포함할 수 있으나, 이에 한정되지 않는다. 일 실시예에서, 프레임워크 계층(903)은 트레이닝 알고리즘에 정량화 기능들을 부가하는 것을 포함할 수 있다. 다른 실시예에서, 프레임워크 계층(930)은 나눗셈, 곱셈 및 제곱근의 시프트 근사화 된 방법들인 기존의 배치-정규화(batch-normalization) 방법들에 대한 오버라이드(override)를 제공 할 수 있다. 또 다른 실시예에서, 프레임워크 계층(930)은 사용자가 계산에 사용된 비트들의 수를 설정할 수 있게 하는 확장을 제공할 수 있다. 또 다른 실시예에서, 프레임워크 계층(930)은 DPU 라이브러리 및 드라이버 계층(920)으로부터 프레임 워크 계층(930)으로 다중-DPU API를 랩핑 (wrapping)하는 능력을 제공함으로써, 사용자가 하드웨어 계층에서 다중 GPU들을 사용하는 것과 유사하게 다중-DPU들을 사용할 수 있다. 프레임워크 계층(930)의 또 다른 특징은 사용자가 기능을 하드웨어 계층 (910)에서 DPU 또는 GPU에 할당하는 것을 허용한다.

[0054] 애플리케이션 계층(940)은 이미지 태그 프로세싱, 셀프-운전/조종 차량들, 알파고(AlphaGo)-타입 딥-마인드(deep-mind) 애플리케이션들 및/또는 음성 연구와 같은 광범위한 애플리케이션을 포함할 수 있으나, 이에 한정되지 않는다.

[0055] 당업자가 인식할 수 있는 바와 같이, 여기서 설명된 혁신적인 개념들은 광범위한 애플리케이션들에 걸쳐 변형 및 변경될 수 있다. 따라서, 청구된 주제의 범위는 전술한 임의의 특정 예시적인 교시에 제한되어서는 안되며, 다음의 청구 범위에 의해 정의된다.

부호의 설명

[0056] 100: DPU

101: 뱅크

102: 서브-어레이

103: 버퍼

104: 시스템 버스

105: 매트

106: 데이터 셀 어레이

107: 컴퓨팅 셀 어레이

108: 매트 내 시프트 어레이

110: 데이터 셀 어레이 디코더

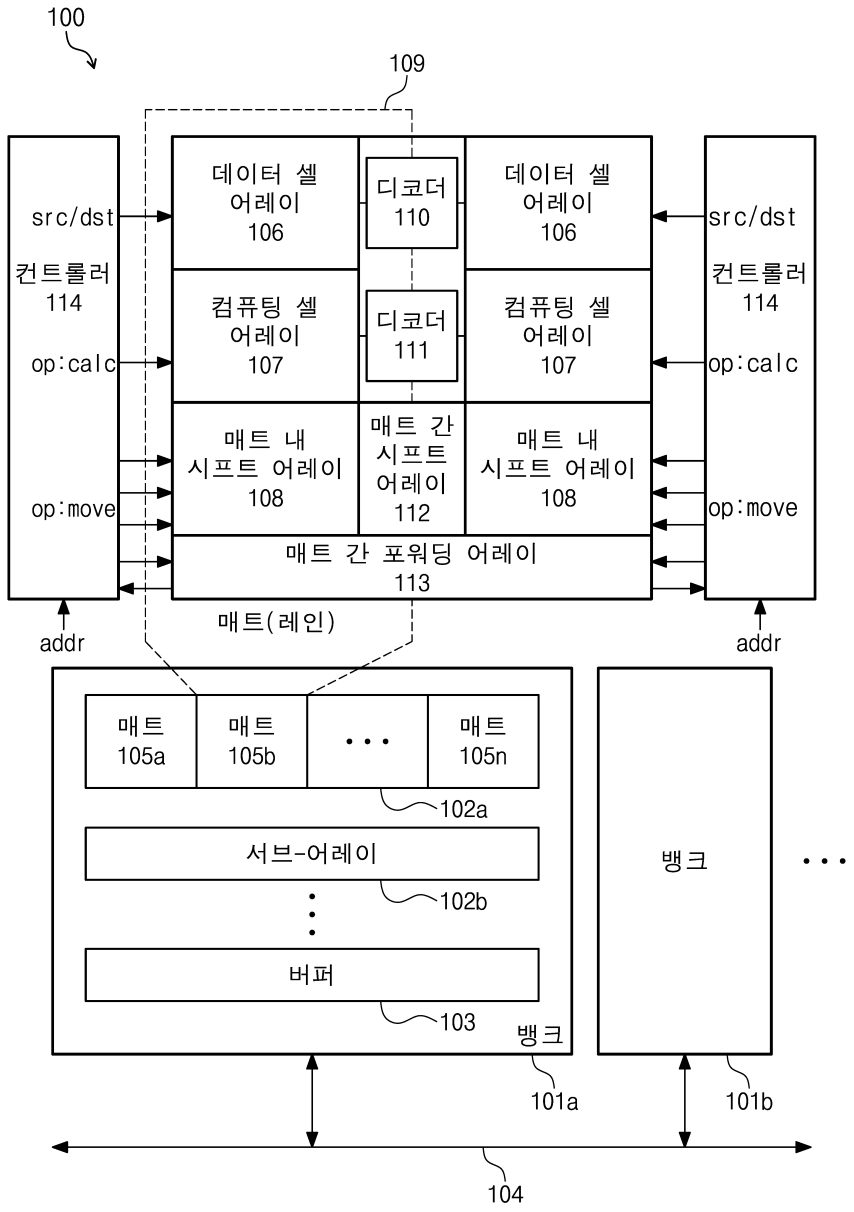
111: 컴퓨팅 셀 어레이 디코더

112: 매트 간 시프트 어레이

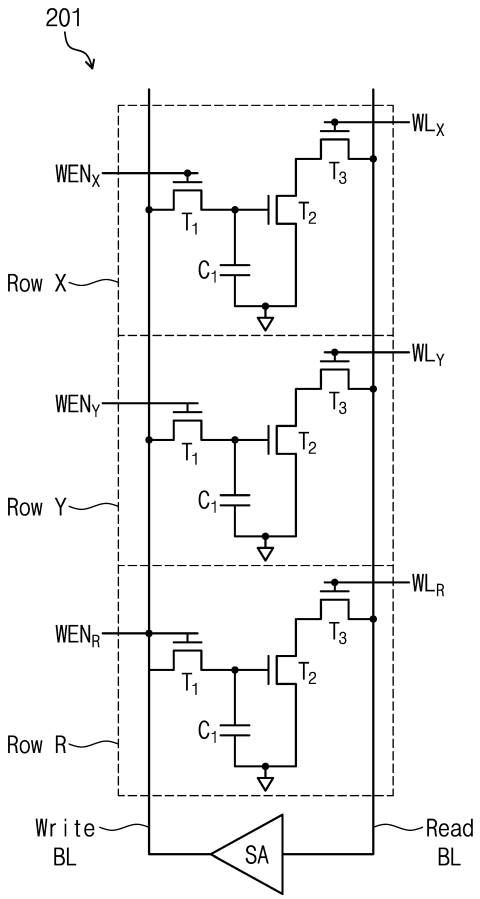
113: 매트 간 포워딩 어레이

도면

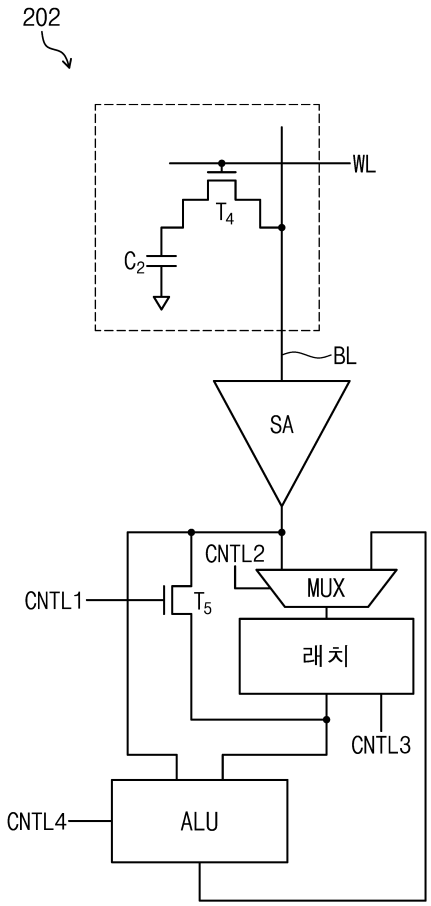
도면1


100
109
src/dst
컨트롤러
114
op:calc
op:move
addr
데이터 셀
어레이
106
디코더
110
데이터 셀
어레이
106
컴퓨팅 셀
어레이
107
디코더
111
컴퓨팅 셀
어레이
107
매트 내
시프트 어레이
108
매트 간
시프트
어레이
112
매트 내
시프트 어레이
108
매트 간 포워딩 어레이
113
매트(레인)
src/dst
컨트롤러
114
op:calc
op:move
addr
매트
105a
매트
105b
• • •
매트
105n
102a
서브-어레이
102b
버퍼
103
뱅크
101a
뱅크
• • •
101b
104

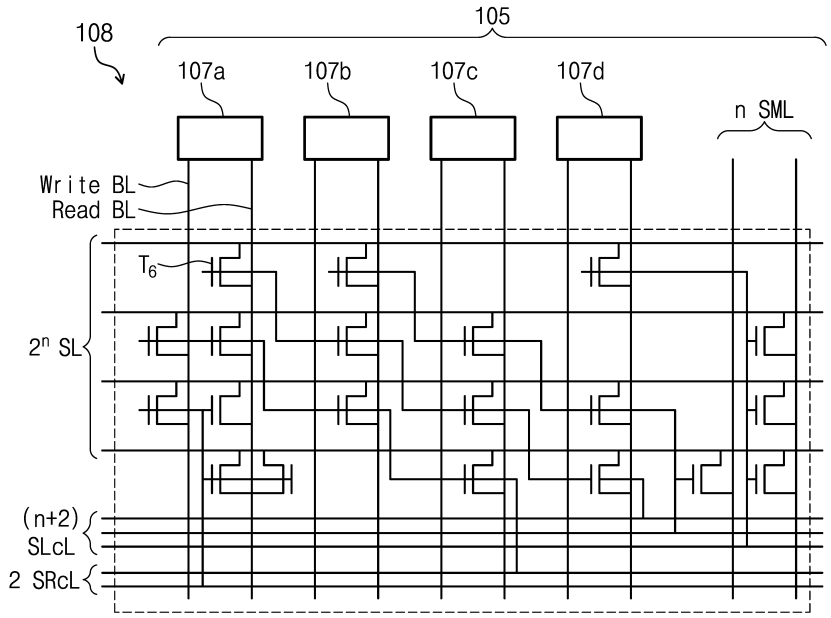
도면2a


201
WLX
WENX
T3
T1
T2
C1
Row X
WLY
WENY
T3
T1
T2
C1
Row Y
WLR
WENR
T3
T1
T2
C1
Row R
Write BL
SA
Read BL

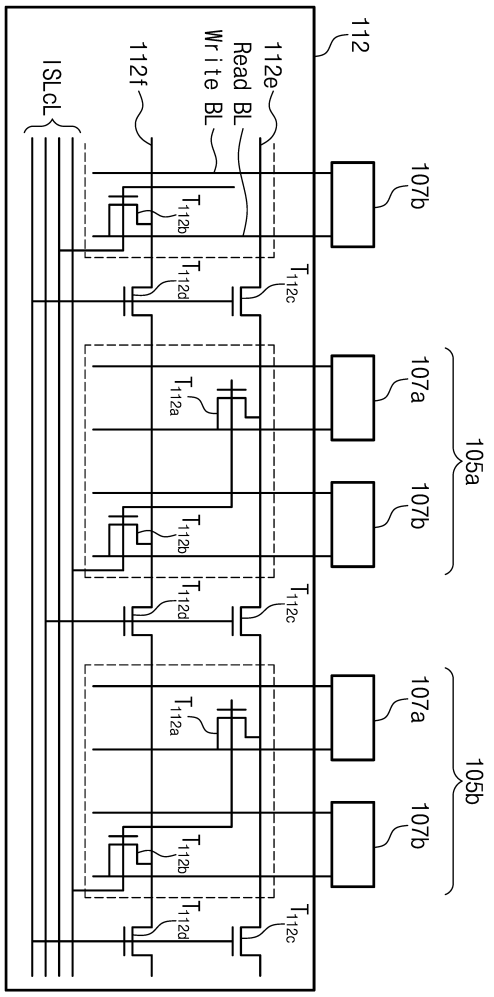
도면2b


202
WL
T4
C2
BL
SA
CNTL2
MUX
CNTL1
T5
래치
CNTL3
CNTL4
ALU

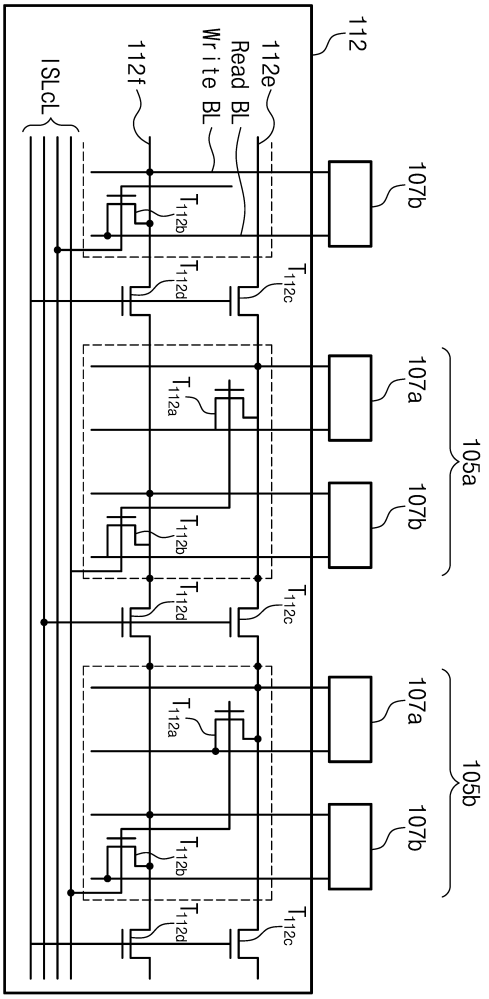
도면3


108
105
107a
107b
107c
107d
n SML
Write BL
Read BL
T6
2n SL
(n+2)
SLcL
2 SRcL

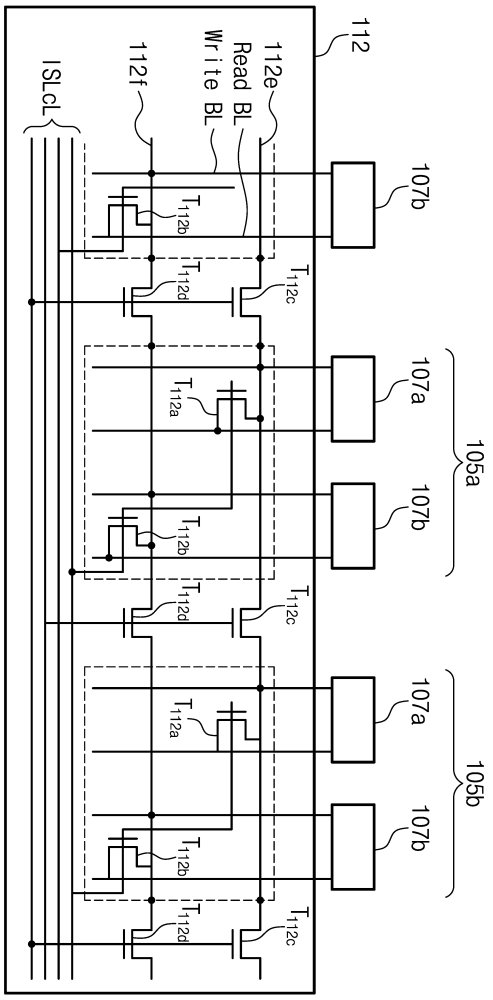
도면4a


112
112e
Read BL
Write BL
112f
ISLcL
107b
107a
105a
105b
T_{112b}
T_{112d}
T_{112c}
T_{112a}

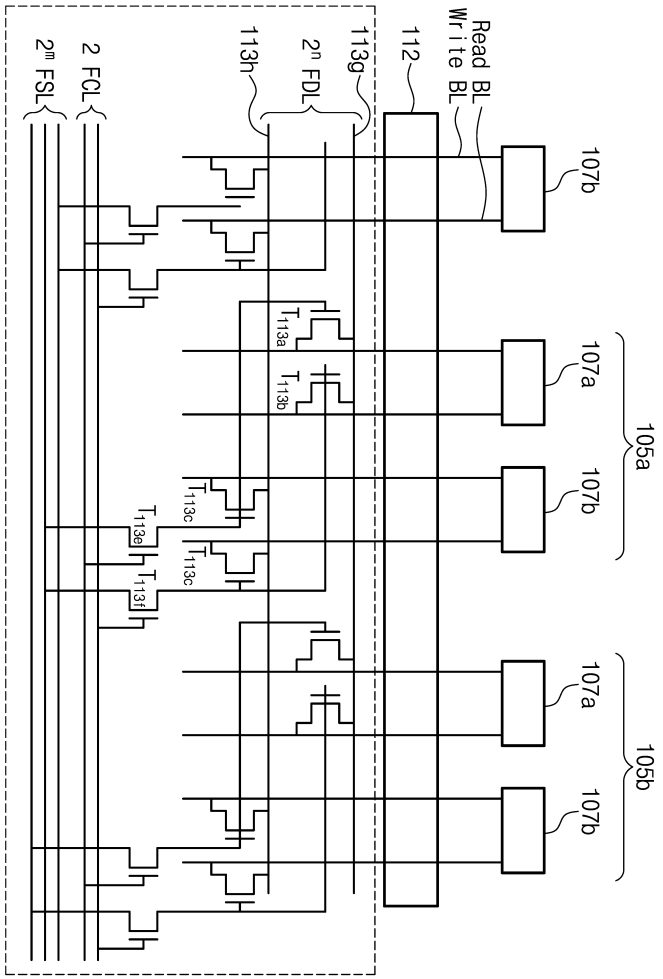
도면4b


112
112e
Read BL
Write BL
112f
ISLcL
107b
107a
105a
105b
T_{112b}
T_{112d}
T_{112c}
T_{112a}

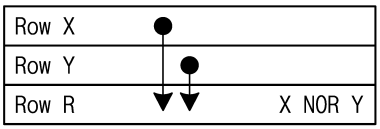
도면4c


112
112e
Read BL
Write BL
112f
ISLcL
107b
107a
105a
105b
T112b
T112d
T112c
T112a

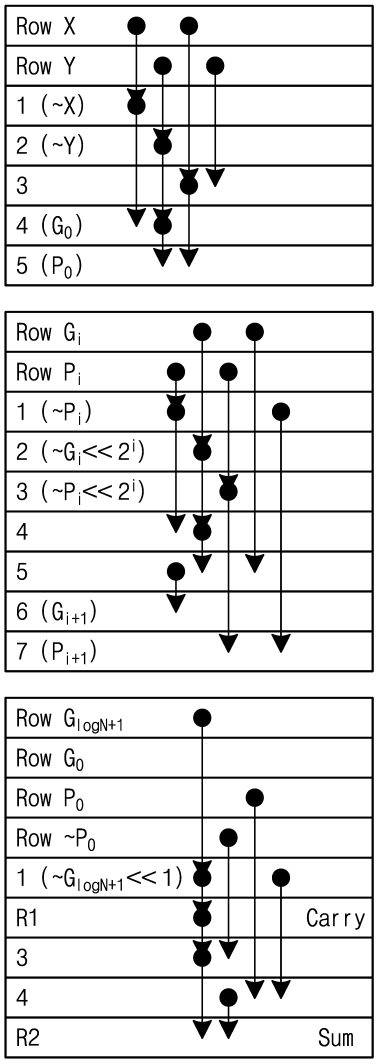
도면5


2m FSL
2 FCL
113h
2n FDL
113g
112
Read BL
Write BL
107b
107a
105a
107b
107a
105b
107b
T113a
T113b
T113c
T113e
T113c
T113f

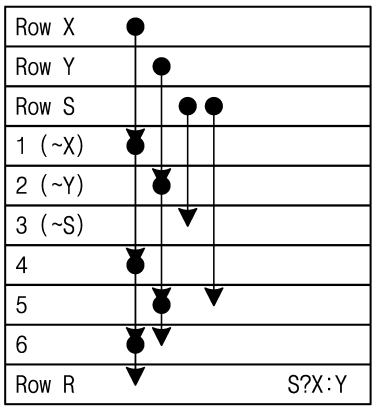
도면6a


Row X
Row Y
Row R
X NOR Y

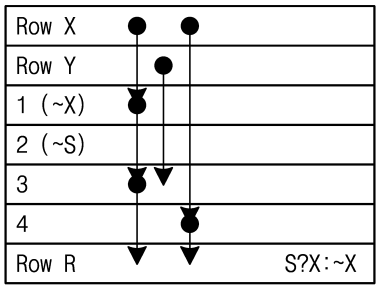
도면6b


Row X
Row Y
1 (~X)
2 (~Y)
3
4 (G0)
5 (P0)
Row Gi
Row Pi
1 (~Pi)
2 (~Gi<< 2i)
3 (~Pi<< 2i)
4
5
6 (Gi+1)
7 (Pi+1)
Row GlogN+1
Row G0
Row P0
Row ~P0
1 (~GlogN+1<< 1)
R1
Carry
3
4
R2
Sum

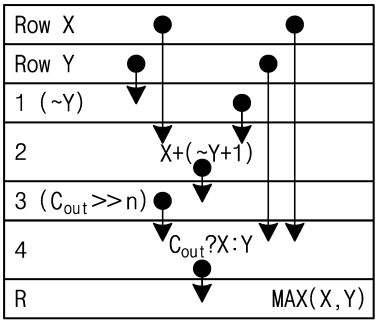
도면6c


Row X
Row Y
Row S
1 (~X)
2 (~Y)
3 (~S)
4
5
6
Row R
S?X:Y

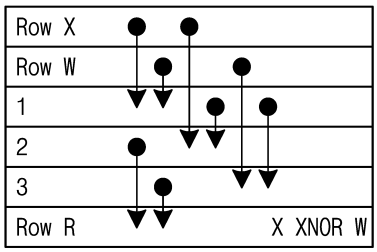
도면6d



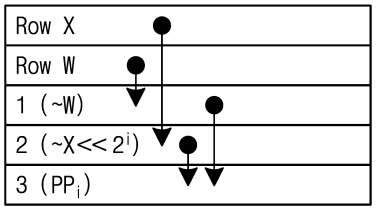
도면6e

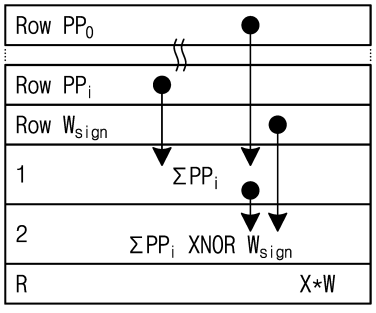


도면6f

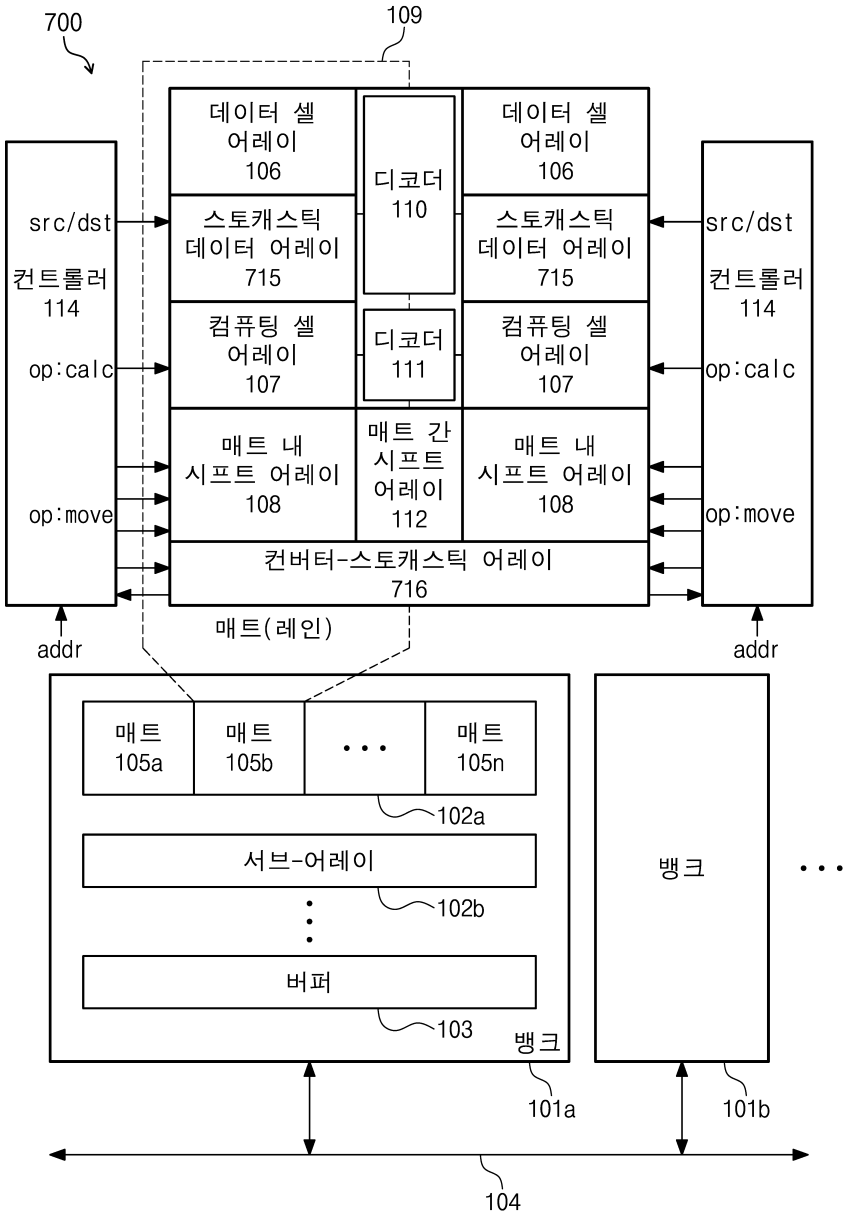


도면6g

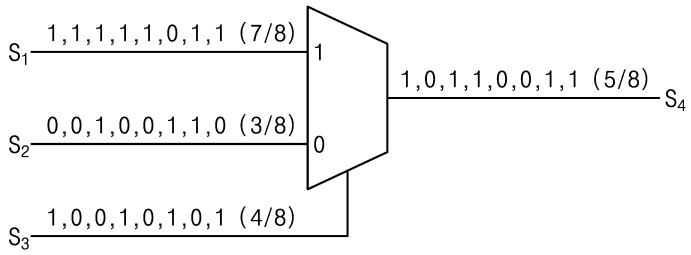




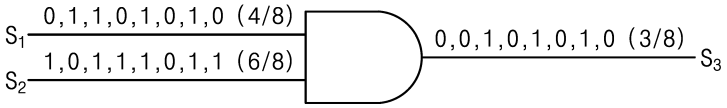
도면7


700
109
데이터 셀 어레이 106
디코더 110
데이터 셀 어레이 106
src/dst 컨트롤러 114
스토캐스틱 데이터 어레이 715
스토캐스틱 데이터 어레이 715
src/dst 컨트롤러 114
컴퓨팅 셀 어레이 107
디코더 111
컴퓨팅 셀 어레이 107
op:calc
op:calc
매트 내 시프트 어레이 108
매트 간 시프트 어레이 112
매트 내 시프트 어레이 108
op:move
op:move
컨버터-스토캐스틱 어레이 716
addr
매트(레인)
addr
매트 105a
매트 105b
매트 105n
102a
서브-어레이
102b
버퍼
103
뱅크
뱅크
101a
101b
104

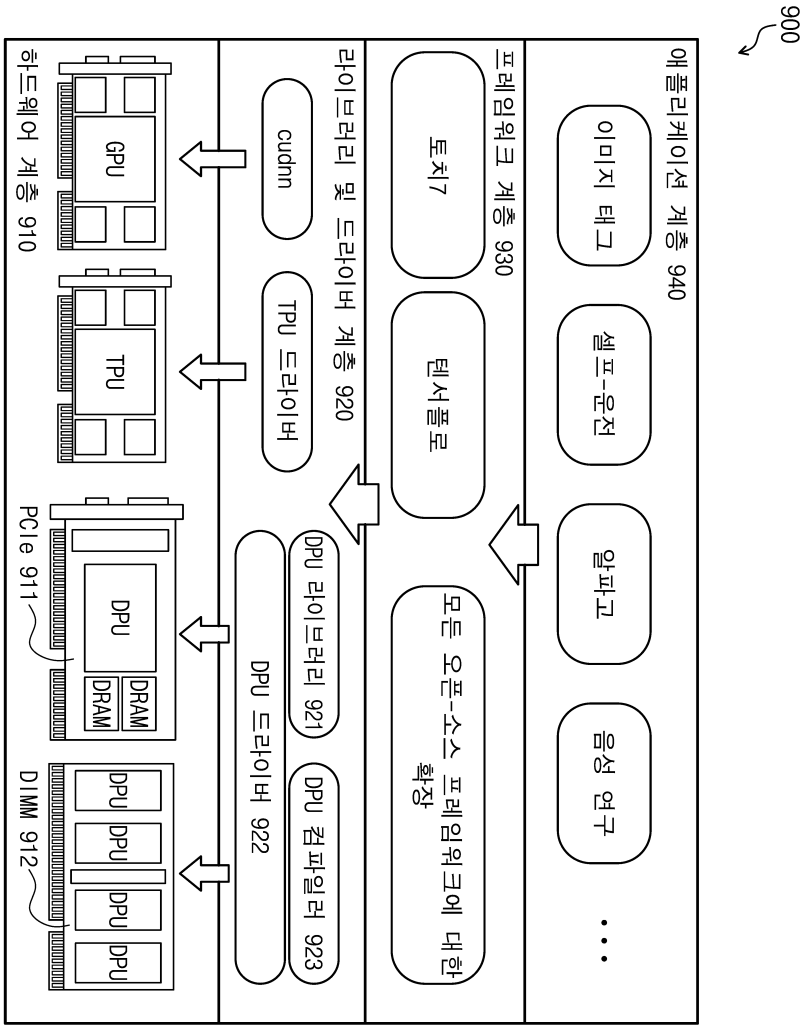
도면8a


S1
1,1,1,1,1,0,1,1 (7/8)
1
S2
0,0,1,0,0,1,1,0 (3/8)
0
S3
1,0,0,1,0,1,0,1 (4/8)
1,0,1,1,0,0,1,1 (5/8)
S4

도면8b


S1
0,1,1,0,1,0,1,0 (4/8)
S2
1,0,1,1,1,0,1,1 (6/8)
0,0,1,0,1,0,1,0 (3/8)
S3

도면9


900
애플리케이션 계층 940
이미지 태그
셀프-운전
알파고
음성 연구
...
프레임워크 계층 930
토치7
텐서플로
모든 오픈-소스 프레임워크에 대한 확장
라이브러리 및 드라이버 계층 920
cudnn
TPU 드라이버
DPU 라이브러리 921
DPU 컴파일러 923
DPU 드라이버 922
하드웨어 계층 910
GPU
TPU
DPU
DRAM
DRAM
PCIe 911
DPU
DPU
DPU
DPU
DIMM 912