(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0066953
(43) 공개일자 2020년06월11일

(51) 국제특허분류(Int. Cl.)
G06F 13/16 (2006.01) *G06F 15/78* (2006.01)
(52) CPC특허분류
G06F 13/16 (2013.01)
G06F 15/7821 (2013.01)
(21) 출원번호 10-2018-0153725
(22) 출원일자 2018년12월03일
심사청구일자 없음

(71) 출원인
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
(72) 발명자
강신행
경기도 수원시 영통구 신원로294번길 40-15, 403호 (원천동)
오성일
경기도 수원시 영통구 영통로514번길 53, 104동 904호(영통동, 황골마을주공2단지아파트)
(74) 대리인
리앤목특허법인

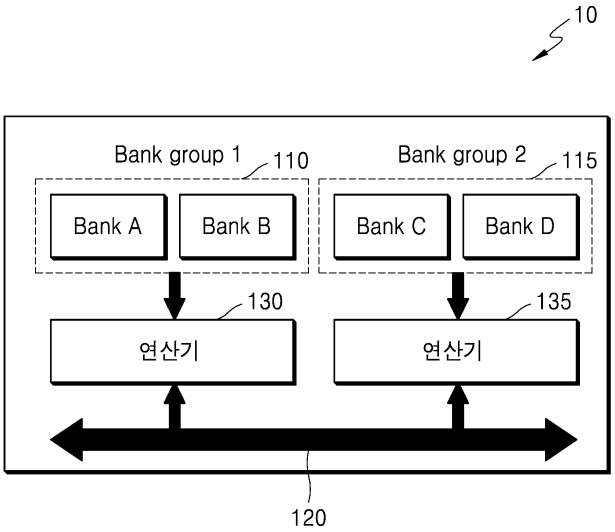
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **PIM을 채용하는 반도체 메모리 장치 및 그 동작 방법**

(57) 요 약

제1 주기마다 메모리 뱅크 그룹으로부터 메모리 뱅크 그룹에 저장된 내부 데이터를 수신하는 단계, 제1 주기보다 짧은 제2 주기마다 내부 메모리 버스를 통해 메모리 뱅크 그룹의 외부에 저장되거나 메모리 뱅크 그룹의 외부에서 처리되는 외부 데이터를 수신하는 단계 및 제2 주기마다 내부 데이터 및 외부 데이터에 대한 PIM 연산을 수행하는 단계를 포함하는, 반도체 메모리 장치에 포함되는 연산기의 동작 방법이 개시된다.

대 표 도 - 도1



명 세 서

청구범위

청구항 1

반도체 메모리 장치에 있어서,

각각이 병렬적으로 접근 가능한 복수의 메모리 뱅크 그룹들;

상기 복수의 메모리 뱅크 그룹들의 외부로부터 외부 데이터를 수신하는 내부 메모리 버스; 및

제1 주기마다 상기 복수의 메모리 뱅크 그룹들 중 제1 메모리 뱅크 그룹으로부터 내부 데이터를 수신하고, 상기 제1 주기보다 짧은 제2 주기마다 상기 내부 메모리 버스를 통해 상기 외부 데이터를 수신하며, 상기 제2 주기마다 상기 내부 데이터 및 상기 외부 데이터에 대한 연산을 수행하는 제1 연산기를 포함하는, 반도체 메모리 장치.

청구항 2

제 1항에 있어서,

상기 제1 주기는 동일한 메모리 뱅크 그룹을 연속으로 접근하는 경우의 지연시간(tCCD_L)에 대응되고,

상기 제2 주기는 서로 다른 메모리 뱅크 그룹들을 번갈아 접근하는 경우의 지연시간(tCCD_S)에 대응되는, 반도체 메모리 장치.

청구항 3

제 1항에 있어서,

상기 제1 주기는 상기 제2 주기의 n배(n은 2 이상의 자연수)인, 반도체 메모리 장치.

청구항 4

제 3항에 있어서,

상기 내부 데이터는 상기 내부 데이터 및 상기 외부 데이터에 대한 연산을 위해 n번 재사용되는, 반도체 메모리 장치.

청구항 5

제 1항에 있어서,

상기 외부 데이터는 상기 반도체 메모리 장치에 포함되는 메모리 채널들 중 상기 제1 메모리 뱅크 그룹을 포함하는 메모리 채널 외의 다른 메모리 채널에 저장된 데이터, 상기 반도체 메모리 장치의 외부의 다른 메모리 장치에 저장된 데이터 및 상기 반도체 메모리 장치 외부의 호스트 프로세서에서 처리되는 데이터 중 적어도 하나에 대응되는, 반도체 메모리 장치.

청구항 6

제 1항에 있어서,

상기 제1 연산기는 상기 내부 데이터 및 상기 외부 데이터에 대한 연산의 결과를 저장하는 복수의 버퍼들을 포함하고,

상기 복수의 버퍼들의 개수는 상기 제1 주기를 상기 제2 주기로 나눈 몫인 n에 대응되는, 반도체 메모리 장치.

청구항 7

제 6항에 있어서,

상기 n개의 버퍼들 각각은 상기 내부 데이터 및 상기 외부 데이터에 대한 연산의 결과를 한번씩 번갈아 가며 저

장하는, 반도체 메모리 장치.

청구항 8

제 6항에 있어서,

상기 제1 연산기는,

상기 n개의 버퍼들 중 상기 내부 데이터 및 상기 외부 데이터에 대한 연산의 결과를 저장할 버퍼를 선택하기 위한 제어 신호를 출력하는 셀렉터(selector); 및

상기 셀렉터에 의해 출력되는 제어 신호에 기초하여 상기 n개의 버퍼들 중 상기 제1 연산기와 연결될 버퍼를 결정하는 멀티플렉서(multiplexer) 및 디멀티플렉서(demultiplexer) 중 적어도 하나를 더 포함하는, 반도체 메모리 장치.

청구항 9

제 8항에 있어서,

상기 셀렉터는 상기 제2 주기마다 계수하는 카운터를 포함하고,

상기 셀렉터에 의해 출력되는 제어 신호는 상기 카운터의 출력 신호에 대응되는, 반도체 메모리 장치.

청구항 10

제 1항에 있어서,

상기 제1 연산기는 상기 복수의 메모리 뱅크 그룹들 중 제2 메모리 뱅크 그룹과 연결되는 제2 연산기와 병렬적으로 동작하는, 반도체 메모리 장치.

청구항 11

반도체 메모리 장치에 포함되는 연산기의 동작 방법에 있어서,

제1 주기마다 메모리 뱅크 그룹으로부터 상기 메모리 뱅크 그룹에 저장된 내부 데이터를 수신하는 단계;

상기 제1 주기보다 짧은 제2 주기마다 내부 메모리 버스를 통해 상기 메모리 뱅크 그룹의 외부에 저장되거나 상기 메모리 뱅크 그룹의 외부에서 처리되는 외부 데이터를 수신하는 단계; 및

상기 제2 주기마다 상기 내부 데이터 및 상기 외부 데이터에 대한 PIM(Processing in memory) 연산을 수행하는 단계를 포함하는, 방법.

청구항 12

제 11항에 있어서,

상기 제1 주기는 동일한 메모리 뱅크 그룹을 연속으로 접근하는 경우의 지연시간(tCCD_L)에 대응되고,

상기 제2 주기는 서로 다른 메모리 뱅크 그룹들을 번갈아 접근하는 경우의 지연시간(tCCD_S)에 대응되는, 방법.

청구항 13

제 11항에 있어서,

상기 제1 주기는 상기 제2 주기의 n배(n은 2 이상의 자연수)인, 방법.

청구항 14

제 13항에 있어서,

상기 제2 주기마다 상기 내부 데이터 및 상기 외부 데이터에 대한 PIM 연산을 수행하는 단계는,

상기 내부 데이터 및 상기 외부 데이터에 대한 연산을 위해 상기 내부 데이터를 n번 재사용하는 단계를 더 포함하는, 방법.

청구항 15

제 11항에 있어서,

상기 외부 데이터는 상기 반도체 메모리 장치에 포함되는 메모리 채널들 중 상기 메모리 뱅크 그룹을 포함하는 메모리 채널 외의 다른 메모리 채널에 저장된 데이터, 상기 반도체 메모리 장치의 외부의 다른 메모리 장치에 저장된 데이터 및 상기 반도체 메모리 장치 외부의 호스트 프로세서에서 처리되는 데이터 중 적어도 하나에 대응되는, 방법.

청구항 16

제 11항에 있어서,

상기 방법은,

상기 내부 데이터 및 상기 외부 데이터에 대한 연산의 결과를 상기 제1 주기를 상기 제2 주기로 나눈 몫인 n에 대응되는 개수의 복수의 버퍼들 각각에 한번씩 번갈아 가며 저장하는 단계를 더 포함하는, 방법.

청구항 17

제 16항에 있어서,

상기 내부 데이터 및 상기 외부 데이터에 대한 연산의 결과를 상기 n개의 버퍼들 각각에 한번씩 번갈아 가며 저장하는 단계는,

상기 n개의 버퍼들 중 상기 내부 데이터 및 상기 외부 데이터에 대한 연산의 결과를 저장할 버퍼를 선택하기 위한 제어 신호를 출력하는 단계를 포함하는, 방법.

청구항 18

반도체 메모리 장치에 있어서,

각각이 병렬적으로 접근 가능한 복수의 메모리 뱅크들;

상기 복수의 메모리 뱅크들의 외부로부터 외부 데이터를 수신하는 내부 메모리 버스; 및

제1 주기마다 상기 복수의 메모리 뱅크들 중 제1 메모리 뱅크로부터 내부 데이터를 수신하고, 상기 제1 주기보다 짧은 제2 주기마다 상기 내부 메모리 버스를 통해 상기 외부 데이터를 수신하며, 상기 제2 주기마다 상기 내부 데이터 및 상기 외부 데이터에 대한 연산을 수행하는 제1 연산기를 포함하는, 반도체 메모리 장치.

청구항 19

제 18항에 있어서,

상기 외부 데이터는 상기 반도체 메모리 장치에 포함되는 메모리 채널들 중 상기 메모리 뱅크를 포함하는 메모리 채널 외의 다른 메모리 채널에 저장된 데이터, 상기 반도체 메모리 장치의 외부의 다른 메모리 장치에 저장된 데이터 및 상기 반도체 메모리 장치 외부의 호스트 프로세서에서 처리되는 데이터 중 적어도 하나에 대응되는, 반도체 메모리 장치.

청구항 20

제 18항에 있어서,

상기 제1 연산기는 상기 복수의 메모리 뱅크들 중 제2 메모리 뱅크와 연결되는 제2 연산기와 병렬적으로 동작하는, 반도체 메모리 장치.

발명의 설명

기 술 분 야

[0001] 본 개시는 PIM을 채용하는 반도체 메모리 장치 및 그 동작 방법에 관한 것이다.

배 경 기 술

[0002] 종래의 반도체 메모리 장치는 연산 작업을 수행하는 프로세서와 기능이 완전히 분리되어 있었다. 따라서, 많은 양의 데이터에 대한 연산이 요구되는 뉴럴 네트워크, 빅 데이터, 사물 인터넷 등과 같은 응용들을 구현하는 시스템에서 반도체 메모리 장치와 프로세서 간에 많은 양의 데이터가 송수신됨에 따라 병목 현상이 자주 발생되는 문제가 있었다. 이와 같은 문제를 해결하기 위해 메모리 기능에 연산 작업을 수행하는 프로세서의 기능을 합친 반도체 메모리 장치로서, 프로세싱 인 메모리(Processing in memory: PIM)에 대한 연구가 진행되고 있다.

발명의 내용

해결하려는 과제

[0003] 다양한 실시예들은 PIM을 채용하는 반도체 메모리 장치 및 그 동작 방법을 제공하는데 있다. 본 개시가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제들로 한정되지 않으며, 이하의 실시예들로부터 또 다른 기술적 과제들이 유추될 수 있다.

과제의 해결 수단

[0004] 상술한 기술적 과제를 해결하기 위한 수단으로서, 일 측면에 따른 반도체 메모리 장치는, 각각이 병렬적으로 접근 가능한 복수의 메모리 뱅크 그룹들; 상기 복수의 메모리 뱅크 그룹들의 외부로부터 외부 데이터를 수신하는 내부 메모리 버스; 및 제1 주기마다 상기 복수의 메모리 뱅크 그룹들 중 제1 메모리 뱅크 그룹으로부터 내부 데이터를 수신하고, 상기 제1 주기보다 짧은 제2 주기마다 상기 내부 메모리 버스를 통해 상기 외부 데이터를 수신하며, 상기 제2 주기마다 상기 내부 데이터 및 상기 외부 데이터에 대한 연산을 수행하는 제1 연산기를 포함할 수 있다.

[0005] 또한, 다른 측면에 따른 반도체 메모리 장치에 포함되는 연산기의 동작 방법은, 제1 주기마다 메모리 뱅크 그룹으로부터 상기 메모리 뱅크 그룹에 저장된 내부 데이터를 수신하는 단계; 상기 제1 주기보다 짧은 제2 주기마다 내부 메모리 버스를 통해 상기 메모리 뱅크 그룹의 외부에 저장되거나 상기 메모리 뱅크 그룹의 외부에서 처리되는 외부 데이터를 수신하는 단계; 및 상기 제2 주기마다 상기 내부 데이터 및 상기 외부 데이터에 대한 PIM(Processing in memory) 연산을 수행하는 단계를 포함할 수 있다.

[0006] 또한, 또 다른 측면에 따른 반도체 메모리 장치는, 각각이 병렬적으로 접근 가능한 복수의 메모리 뱅크들; 상기 복수의 메모리 뱅크들의 외부로부터 외부 데이터를 수신하는 내부 메모리 버스; 및 제1 주기마다 상기 복수의 메모리 뱅크들 중 제1 메모리 뱅크로부터 내부 데이터를 수신하고, 상기 제1 주기보다 짧은 제2 주기마다 상기 내부 메모리 버스를 통해 상기 외부 데이터를 수신하며, 상기 제2 주기마다 상기 내부 데이터 및 상기 외부 데이터에 대한 연산을 수행하는 제1 연산기를 포함할 수 있다.

도면의 간단한 설명

[0007] 도 1은 일부 실시예에 따른 반도체 메모리 장치의 구성을 나타내는 도면이다.

도 2는 일부 실시예에 따른 제1 주기 및 제2 주기를 설명하기 위한 도면이다.

도 3 및 도 4는 일부 실시예에 따른 반도체 메모리 장치의 동작 방법과 종래 기술에 따른 반도체 메모리 장치의 동작 방법을 비교하는 도면이다.

도 5는 일부 실시예에 따른 외부 데이터의 일 예를 설명하기 위한 도면이다.

도 6은 일부 실시예에 따른 연산기의 구성을 나타내는 도면이다.

도 7은 일부 실시예에 따른 반도체 메모리 장치가 행렬 곱 연산을 수행하는 과정의 예시를 설명하기 위한 도면이다.

도 8은 일부 실시예에 따른 반도체 메모리 장치의 동작 방법과 종래 기술에 따른 반도체 메모리 장치의 동작 방법의 성능을 비교하기 위한 시뮬레이션 결과를 나타내는 도면이다.

도 9는 다른 실시예에 따른 반도체 메모리 장치의 구성을 나타내는 도면이다.

도 10은 일부 실시예에 따른 반도체 메모리 장치에 포함되는 연산기의 동작 방법을 나타내는 흐름도이다.

도 11은 일부 실시예에 따른 전자 시스템의 구성을 나타내는 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0008] 본 실시예들에서 사용되는 용어는 본 실시예들에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 기술분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 임의로 선정된 용어도 있으며, 이 경우 해당 실시예의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서, 본 실시예들에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 실시예들의 전반에 걸친 내용을 토대로 정의되어야 한다.

[0009] 실시예들에 대한 설명들에서, 어떤 부분이 다른 부분과 연결되어 있다고 할 때, 이는 직접적으로 연결되어 있는 경우뿐 아니라, 그 중간에 다른 구성요소를 사이에 두고 전기적으로 연결되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 포함한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0010] 본 실시예들에서 사용되는 "구성된다" 또는 "포함한다" 등의 용어는 명세서 상에 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.

[0011] 또한, 본 명세서에서 사용되는 '제 1' 또는 '제 2' 등과 같이 서수를 포함하는 용어는 다양한 구성 요소들을 설명하는데 사용할 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만 사용된다.

[0012] 하기 실시예들에 대한 설명은 권리범위를 제한하는 것으로 해석되지 말아야 하며, 해당 기술분야의 당업자가 용이하게 유추할 수 있는 것은 실시예들의 권리범위에 속하는 것으로 해석되어야 할 것이다. 이하 첨부된 도면들을 참조하면서 오로지 예시를 위한 실시예들을 상세히 설명하기로 한다.

[0013] 도 1은 일부 실시예에 따른 반도체 메모리 장치의 구성을 나타내는 도면이다.

[0014] 도 1을 참조하면, 반도체 메모리 장치(10)는 제1 메모리 뱅크 그룹(110), 제2 메모리 뱅크 그룹(115), 내부 메모리 버스(120), 제1 연산기(130) 및 제2 연산기(135)를 포함할 수 있다. 다만, 도 1에 도시된 반도체 메모리 장치(10)에는 본 실시예들와 관련된 구성요소들만이 도시되어 있다. 따라서, 반도체 메모리 장치(10)에는 도 1에 도시된 구성요소들 외에 다른 범용적인 구성요소들이 더 포함될 수 있음은 당업자에게 자명하다. 예를 들어, 반도체 메모리 장치(10)는 메모리 컨트롤러(미도시)를 더 포함할 수 있다. 메모리 컨트롤러는 반도체 메모리 장치(10)를 제어하기 위한 전반적인 기능을 수행하는 역할을 한다. 메모리 컨트롤러는 다수의 논리 게이트들의 어레이로 구현될 수 있고, 범용적인 마이크로 프로세서와 마이크로 프로세서에서 실행될 수 있는 프로그램이 저장된 메모리의 조합으로 구현될 수도 있다.

[0015] 한편, 도 1에는 반도체 메모리 장치(10)에 포함되는 하나의 메모리 채널만이 도시된 것이고, 반도체 메모리 장치(10)는 도 1에 도시된 메모리 채널 외에 다른 메모리 채널들을 더 포함할 수 있다. 또한, 도 1에는 설명의 편의를 위해 2개의 메모리 뱅크 그룹들 및 2개의 연산기들만이 도시되어 있으나, 반도체 메모리 장치(10)가 더 많은 수의 메모리 뱅크 그룹들 및 연산기들을 포함할 수 있음은 해당 기술분야의 통상의 기술자라면 쉽게 이해할 것이다.

[0016] 제1 메모리 뱅크 그룹(110) 및 제2 메모리 뱅크 그룹(115) 각각은 병렬적으로 접근 가능한 메모리 영역을 의미할 수 있다. 제1 메모리 뱅크 그룹(110) 및 제2 메모리 뱅크 그룹(115)은 내부 메모리 버스(120)를 공유하므로, 일반적인 메모리 독출/기입 동작 시에는 제1 메모리 뱅크 그룹(110) 및 제2 메모리 뱅크 그룹(115) 중 어느 하나만 내부 메모리 버스(120)를 이용할 수 있다. 다만, 제1 메모리 뱅크 그룹(110) 및 제2 메모리 뱅크 그룹(115) 각각은 독립적으로 접근 가능하므로, 제1 메모리 뱅크 그룹(110) 및 제2 메모리 뱅크 그룹(115)에 대해 개별적인 독출 동작이 동시에 수행될 수도 있다. 예를 들어, 제1 메모리 뱅크 그룹(110)으로부터 데이터가 독출될 때, 제2 메모리 뱅크 그룹(115)에서도 데이터가 독출될 수 있다. 이 경우 반도체 메모리 장치(10)는 뱅크 그룹 단위 병렬성을 갖는다고 표현될 수 있다. 한편, 제1 메모리 뱅크 그룹(110) 및 제2 메모리 뱅크 그룹(115) 각각은 복수의 메모리 뱅크들을 포함할 수 있다. 복수의 메모리 뱅크들 각각은 병렬적으로 접근 가능한 메모리 영역의 최소 단위를 의미할 수 있다.

[0017] 내부 메모리 버스(120)는 반도체 메모리 장치(10)에 포함되는 구성요소들 간에 데이터가 송수신될 수 있도록 형

성되는 데이터 전송 경로를 의미할 수 있다. 예를 들어, 내부 메모리 버스(120)는 제1 메모리 뱅크 그룹(110) 및 제2 메모리 뱅크 그룹(115) 간에 공유될 수 있다. 또한, 내부 메모리 버스(120)는 반도체 메모리 장치(10)에 포함되는 구성요소들 및 반도체 메모리 장치(10) 외부의 구성요소들 간의 연결 경로를 형성할 수 있다. 예를 들어, 내부 메모리 버스(120)는 반도체 메모리 장치(10)에 포함되는 복수의 메모리 뱅크 그룹들의 외부로부터 외부 데이터를 수신할 수 있다. 내부 메모리 버스(120)는 외부로부터 수신된 외부 데이터를 제1 연산기(130) 및 제2 연산기(135) 중 적어도 하나로 전달할 수 있다. 이하 도 5를 참조하여 복수의 메모리 뱅크 그룹들의 외부의 의미에 대해서 보다 상세히 설명할 것이다.

[0018] 제1 연산기(130)는 반도체 메모리 장치(10) 내부에서 연산 작업을 수행하는 하드웨어를 의미할 수 있다. 예를 들어, 제1 연산기(130)는 ALU(Arithmetic Logic Unit)를 포함할 수 있고, 연산 작업은 덧셈, 뺄셈, 적분, 가산 등을 포함하는 함수 연산을 포함할 수 있다. 다만, 이에 제한되는 것은 아니다. 제1 연산기(130)는 다수의 논리 게이트들의 어레이로 구현될 수 있고, 논리 게이트들의 어레이와 데이터를 일시적으로 저장하는 버퍼의 조합으로 구현될 수도 있다. 한편, 반도체 메모리 장치(10)는 내부에 연산 작업을 수행하는 제1 연산기(130)를 포함하는바, PIM에 해당할 수 있다.

[0019] 제1 연산기(130)는 반도체 메모리 장치(10) 외부의 CPU(Central Processing Unit), GPU(Graphics Processing Unit), DSP(Digital Signal Processor) 등과 같은 호스트 프로세서와 유사하게, 프로세싱 기능을 갖는 하드웨어로서, 반도체 메모리 장치(10)의 메모리 패키지(memory package)의 칩 내에서 복수의 메모리 뱅크 그룹들과 함께 패키징된 프로세서일 수 있다. 제1 연산기(130)는 반도체 메모리 장치(10)의 내부에 존재하므로 내부 프로세서로 지칭될 수 있고, 호스트 프로세서는 반도체 메모리 장치(10)의 외부에 존재하므로 외부 프로세서로 지칭될 수 있다. PIM 아키텍처는 내부 프로세서와 메모리가 온칩(On-chip)으로 구현되기 때문에, 낮은 레이턴시(low latency)의 빠른 메모리 액세스가 가능하다. 또한, PIM 아키텍처가 뱅크 단위 또는 뱅크 그룹 단위의 병렬성을 이용하는 경우 일반적인 메모리 접근과 비교하여 수 배 내지 수십 배의 메모리 대역폭을 활용할 수 있다. PIM 아키텍처를 갖는 반도체 메모리 장치(10)는 intelligent RAM(Random Access Memory), computational RAM, 또는 smart memory 등과 같은 용어로도 지칭될 수 있다.

[0020] 한편, 제1 연산기(130)는 반도체 메모리 장치(10)에 포함되는 복수의 메모리 뱅크 그룹들 중 제2 메모리 뱅크 그룹(115)과 연결되는 제2 연산기(135)와 병렬적으로 동작할 수 있다. 제1 연산기(130) 및 제2 연산기(135) 각각은 제1 메모리 뱅크 그룹(110) 및 제2 메모리 뱅크 그룹(115)과 개별적으로 연결되고, 서로 독립적으로 연산 작업을 수행할 수 있다. 예를 들어, 제1 연산기(130)가 제1 메모리 뱅크 그룹(110)으로부터 수신된 데이터를 이용하여 연산 작업을 수행할 때, 제2 연산기(135) 또한 제2 메모리 뱅크 그룹(115)으로부터 수신된 데이터를 이용하여 연산 작업을 수행할 수 있다.

[0021] 제1 연산기(130)는 제1 주기마다 반도체 메모리 장치(10)에 포함되는 복수의 메모리 뱅크 그룹들 중 제1 메모리 뱅크 그룹(110)으로부터 내부 데이터를 수신할 수 있다. 내부 데이터는 복수의 메모리 뱅크 그룹들 각각에 저장된 데이터를 의미하는 것으로서, 복수의 메모리 뱅크 그룹들 각각에 할당된 복수의 메모리 어드레스들 중 적어도 일부에 대응될 수 있다. 또한, 제1 연산기(130)는 제1 주기보다 짧은 제2 주기마다 내부 메모리 버스(120)를 통해 외부 데이터를 수신할 수 있다. 외부 데이터는 복수의 메모리 뱅크 그룹들의 외부에 저장되거나 복수의 메모리 뱅크 그룹들의 외부에서 처리되는 데이터를 지칭하는 것으로서, 이하 도 5를 참조하여 외부 데이터의 의미에 대해 보다 상세히 설명할 것이다. 제1 연산기(130)는 제2 주기마다 내부 데이터 및 외부 데이터에 대한 연산을 수행할 수 있다. 이하 도 2를 참조하여 제1 주기 및 제2 주기에 대해 상세히 설명하고, 도 3 및 도 4를 참조하여 전술한 제1 연산기(130)의 동작에 의해 얻어지는 효과에 대해 보다 상세히 설명한다.

[0022] 도 2는 일부 실시예에 따른 제1 주기 및 제2 주기를 설명하기 위한 도면이다.

[0023] 도 2를 참조하면, 클럭(CLK)에 따라 메모리 뱅크 그룹에 대한 독출 동작이 수행되는 타이밍을 나타낸 도면이 도시되어 있다. 메모리 뱅크 그룹에 대한 독출 동작을 수행하라는 명령(Column CMD)이 메모리 뱅크 그룹에 전송되는 시점을 살펴보면, 서로 다른 뱅크 그룹들을 번갈아 접근하는 경우의 지연시간(tCCD_S)보다 동일한 메모리 뱅크 그룹에 연속하여 접근하는 경우의 지연시간(tCCD_L)이 더 긴 것을 확인할 수 있다. 도 1의 제1 주기는 동일한 메모리 뱅크 그룹을 연속으로 접근하는 경우의 지연시간(tCCD_L)에 대응되고, 제2 주기는 서로 다른 메모리 뱅크 그룹들을 번갈아 접근하는 경우의 지연시간(tCCD_S)에 대응될 수 있다. tCCD는 Column to column delay를 나타내고, L은 Long을 나타내며, S는 Short을 나타낸다.

[0024] 한편, PIM은 DDR(Double Data Rate) DRAM (Dynamic RAM) 또는 DDR SDRAM (Synchronous DRAM)로 구현될 수 있는데, 최신의 DDR DRAM 또는 DDR SDRAM를 기반으로 하는 PIM은 일반적으로 뱅크 단위의 병렬성 또는 뱅크 그룹 단

위의 병렬성을 이용한다. 예를 들어, 뱅크 그룹 단위의 병렬성을 이용하는 DDR4 기반의 PIM은 동일한 메모리 뱅크 그룹에 연속하여 접근해야 하기 때문에 지연시간이 항상 tCCD_L이 되는 문제를 갖는다. 복수의 메모리 뱅크 그룹들로부터 동시에 데이터가 독출되므로, 한 번에 독출되는 데이터의 양은 증가될 수 있으나, 데이터가 독출되는 속도는 항상 tCCD_L 주기로 제한된다. 이에 따라, PIM의 전체적인 연산 속도에 제한이 생길 수 있다.

[0025] 본 개시의 일부 실시예에 따르면, 메모리 뱅크 그룹으로부터 데이터가 독출되는 지연시간(tCCD_L)이 일반적인 메모리 사용시 데이터가 독출되는 지연시간(tCCD_S)보다 길기 때문에 발생되는 대기 시간에 외부 데이터를 tCCD_S 주기로 제1 연산기(130)에 공급해줌으로써 PIM의 전체적인 연산 속도를 증가시킬 수 있다.

[0026] 도 3 및 도 4는 일부 실시예에 따른 반도체 메모리 장치의 동작 방법과 종래 기술에 따른 반도체 메모리 장치의 동작 방법을 비교하는 도면이다.

[0027] 도 3을 참조하면, 종래 기술에 따른 반도체 메모리 장치의 동작 방법을 나타내는 개념도(310) 및 일부 실시예에 따른 반도체 메모리 장치의 동작 방법을 나타내는 개념도(320)가 도시되어 있다. 도 3에서 Bank/Bankgroup은 메모리 뱅크 또는 메모리 뱅크 그룹을 나타내고, ALU, MUX 및 buf 각각은 산술 논리 장치(Arithmetic Logic Unit), 멀티플렉서(Multiplexer) 및 버퍼(buffer)를 의미하는 것으로서, 연산기에 포함되는 구성들을 나타낸다. 또한, bus는 내부 메모리 버스를 나타낸다.

[0028] 개념도(310)를 참조하면, 메모리 뱅크 그룹으로부터 제1 주기마다 내부 데이터가 독출되므로, 내부 데이터는 제1 주기마다 연산기로 공급될 수 있다. 또한, 내부 메모리 버스를 통해 외부 데이터가 제1 주기마다 연산기로 공급될 수 있다. 연산기는 내부 데이터 및 외부 데이터를 이용하여 연산을 수행하므로, 제1 주기마다 연산을 수행할 수 있다. 이와 같이, 종래 기술에 따른 반도체 메모리 장치의 동작 방법에 따르면, 연산기의 연산 속도는 제1 주기로 제한될 수 있다.

[0029] 개념도(320)를 참조하면, 개념도(310)와 동일하게 메모리 뱅크 그룹으로부터 제1 주기마다 내부 데이터가 독출되므로, 내부 데이터는 제1 주기마다 연산기로 공급될 수 있다. 다만, 개념도(310)와는 달리, 일부 실시예에 따른 반도체 메모리 장치에서는 내부 메모리 버스를 통해 외부 데이터가 제2 주기마다 공급되므로, 연산기가 제2 주기마다 연산을 수행할 수 있다. 제2 주기는 제1 주기보다 짧으므로, 일부 실시예에 따른 반도체 메모리 장치의 동작 방법에 따르면, 연산기의 연산 속도가 증가될 수 있다.

[0030] 한편, 제1 주기는 제2 주기의 n배(n은 2 이상의 자연수)일 수 있다. 다만, n은 반드시 2 이상의 자연수일 필요는 없으며, n은 2 이상이기만 하면 어떤 값을 갖더라도 상관 없다. 예를 들어, n은 2 이상의 임의의 실수일 수 있다. 표(330)에는 일부 실시예에 따른 반도체 메모리 장치에서 제1 주기가 4ns이고, 제2 주기가 2ns이며, n=2일 때, 시간에 따른 메모리 뱅크 그룹으로부터 공급되는 데이터 및 외부로부터 공급되는 데이터가 도시되어 있다. 이하 도 5를 참조하여 외부 또는 외부 데이터의 의미에 대해 보다 상세히 설명한다.

[0031] 도 5는 일부 실시예에 따른 외부 데이터의 일 예를 설명하기 위한 도면이다.

[0032] 도 5를 참조하면, 반도체 메모리 장치의 구성을 나타내는 도면이 도시되어 있다. 반도체 메모리 장치는 복수의 메모리 채널들을 포함할 수 있다. 예를 들어, 반도체 메모리 장치는 제1 메모리 뱅크 그룹(110), 내부 메모리 버스(120) 및 제1 연산기(130)를 포함하는 메모리 채널(510) 외에 다른 메모리 채널(520)을 더 포함할 수 있다. 메모리 채널(520)은 메모리 채널(510)과 동일한 구조를 가질 수 있다. 한편, 도 5에는 2개의 메모리 채널들만이 도시되어 있으나, 반도체 메모리 장치가 더 많은 수의 메모리 채널들을 포함할 수 있음은 해당 기술분야의 통상의 기술자라면 쉽게 이해할 것이다.

[0033] 외부 데이터는 반도체 메모리 장치에 포함되는 메모리 채널들 중 제1 메모리 뱅크 그룹(110)을 포함하는 메모리 채널(510) 외의 다른 메모리 채널(520)에 저장된 데이터일 수 있다. 이 때, 외부는 메모리 채널(520)을 의미할 수 있다. 다만, 이에 제한되는 것은 아니며, 외부는 메모리 채널(510)의 외부이기만 하면 어떤 소스이든 상관없다. 예를 들어, 외부 데이터는 반도체 메모리 장치 외부의 다른 메모리 장치에 저장된 데이터일 수 있고, 반도체 메모리 장치 외부의 호스트 프로세서에서 처리되는 데이터일 수도 있다. 호스트 프로세서는 CPU, GPU, DSP 등을 포함할 수 있으나, 이에 제한되는 것은 아니다.

[0034] 다시 도 3으로 돌아와서, 표(330)를 참조하면, 연산기는 0ns에 메모리 뱅크 그룹으로부터 내부 데이터 W0를 수신할 수 있고, 0ns에 외부로부터 내부 메모리 버스를 통해 외부 데이터 A0를 수신할 수 있다. 연산기는 수신된 내부 데이터 W0 및 외부 데이터 A0에 대한 연산을 수행할 수 있다.

[0035] 이후 메모리 뱅크 그룹으로부터 새로운 내부 데이터가 공급되지 않은 시점인 2ns에 연산기는 외부로부터 내부

메모리 버스를 통해 새로운 외부 데이터 B0를 수신할 수 있다. 이에 따라, 연산기는 2ns에 기존의 내부 데이터 W0 및 새로운 외부 데이터 B0에 대한 연산을 수행할 수 있다. 이와 같이, 연산기는 메모리 뱅크 그룹으로부터 새로운 데이터가 공급되지 않았더라도, 외부로부터 공급된 새로운 외부 데이터를 이용하여 추가적인 연산을 수행할 수 있다. 내부 데이터는 내부 데이터 및 외부 데이터에 대한 연산을 위해 n번 재사용될 수 있고, 반도체 메모리 장치의 전체적인 연산 속도는 약 n배 증가될 수 있다.

[0036] 도 4를 참조하면, 종래 기술에 따른 반도체 메모리 장치의 동작 방법을 나타내는 타이밍도(410) 및 일부 실시예에 따른 반도체 메모리 장치의 동작 방법을 나타내는 타이밍도(420)가 도시되어 있다.

[0037] 타이밍도(410) 및 타이밍도(420)에서 행들 각각은 메모리 뱅크 그룹의 시간에 따른 동작을 나타내고, R(Read)은 메모리 뱅크 그룹으로부터 내부 데이터가 독출됨을 의미하며, C(Compute)는 연산기에 의해 연산 작업이 수행됨을 의미할 수 있다.

[0038] 타이밍도(410)를 참조하면, 0ns에 모든 메모리 뱅크 그룹들 각각에서 동시에 내부 데이터가 독출됨에 따라 연산기에 의한 연산 작업이 수행된 후 4ns에 모든 메모리 뱅크 그룹들 각각에서 새로운 내부 데이터가 독출되어야만 연산기에 의한 새로운 연산 작업이 수행됨을 알 수 있다. 이와 같이, 종래 기술에 따른 반도체 메모리 장치의 동작 방법에 따르면, 연산기의 연산 작업은 메모리 뱅크 그룹으로부터 내부 데이터가 독출되는 속도에 종속되는 바, 연산기의 연산 속도가 4ns 주기로 제한됨을 알 수 있다.

[0039] 이와 달리, 타이밍도(420)를 참조하면, 0ns에 모든 메모리 뱅크 그룹들 각각에서 동시에 내부 데이터가 독출됨에 따라 연산기에 의한 연산 작업이 수행된 이후 2ns에 외부로부터 새로운 외부 데이터가 연산기로 공급됨에 따라 연산기에 의한 연산 작업이 다시 수행됨을 알 수 있다. 일부 실시예에 따른 반도체 메모리 장치의 동작 방법에 따르면, 연산기의 연산 속도가 2ns 주기로 증가되는 것이다. 연산기의 연산 속도가 증가됨에 따라 반도체 메모리 장치의 전체적인 연산 속도가 증가될 수 있다. 이하 도 6을 참조하여 연산기에 대해 보다 상세히 설명한다.

[0040] 도 6은 일부 실시예에 따른 연산기의 구성을 나타내는 도면이다.

[0041] 도 6을 참조하면, 연산기는 제1 주기마다 메모리 뱅크 그룹으로부터 공급되는 내부 데이터를 저장하는 피연산자 버퍼(610) 및 제2 주기마다 내부 메모리 버스로부터 공급되는 외부 데이터를 저장하는 피연산자 버퍼(615)를 포함할 수 있다. 또한, 연산기는 피연산자 버퍼(610)에 저장된 내부 데이터 및 피연산자 버퍼(615)에 저장된 외부 데이터에 대한 연산을 수행하는 ALU(620)를 더 포함할 수 있다. ALU(620)는 곱셈기(multiplier), 덧셈기(adder), 적분기(integrator) 등일 수 있으나, 이에 제한되는 것은 아니다. ALU(620)는 PIM에서 수행될 연산의 종류에 따라 임의의 적절한 방식으로 구현될 수 있다.

[0042] 한편, 연산기는 내부 데이터 및 외부 데이터에 대한 연산의 결과를 저장하는 복수의 버퍼들을 포함할 수 있다. 복수의 버퍼들의 개수는 제1 주기를 제2 주기로 나눈 몫에 대응되는 개수일 수 있다. 예를 들어, 제1 주기가 제2 주기의 n배일 때, 연산기는 n개의 버퍼를 포함할 수 있다. 도 3 및 도 4를 참조하여 설명한 예시에서 n=2를 가정하였으므로, 도 6에는 연산기가 2개의 버퍼들(630 및 635)을 포함하는 것으로 도시하였다. 제1 주기 및 제2 주기 간의 관계가 달라짐에 따라 연산기에 포함된 버퍼들의 개수가 달라질 수 있음은 해당 기술분야의 통상의 기술자라면 쉽게 이해할 것이다.

[0043] 연산기에 포함되는 제1 버퍼(630) 및 제2 버퍼(635) 각각은 내부 데이터 및 외부 데이터에 대한 연산의 결과를 한번씩 번갈아 가며 저장할 수 있다. 예를 들어, 도 3의 표(330)을 참조하여 설명하면, 0ns에 제1 버퍼(630)가 내부 데이터 W0 및 외부 데이터 A0에 대한 연산의 결과를 저장하고, 이후 2ns에 제2 버퍼(635)가 내부 데이터 W0 및 외부 데이터 B0에 대한 연산의 결과를 저장한다. 이후 4ns에 제1 버퍼(630)가 내부 데이터 W1 및 외부 데이터 A1에 대한 연산의 결과를 저장하고, 이후 6ns에 제2 버퍼(635)가 내부 데이터 W1 및 외부 데이터 B1에 대한 연산의 결과를 저장한다.

[0044] 연산기는 n개의 버퍼들 중 연산의 결과를 저장하는 버퍼를 선택하기 위해 셀렉터(selector)(640), 디멀티플렉서(demultiplexer)(645) 및 멀티플렉서(multiplexer)(650) 중 적어도 하나를 더 포함할 수 있다. 셀렉터(640)는 제1 버퍼(630) 및 제2 버퍼(635) 중 내부 데이터 및 외부 데이터에 대한 연산의 결과를 저장할 버퍼를 선택하기 위한 제어 신호를 출력할 수 있다.

[0045] 디멀티플렉서(645)는 하나의 입력을 통해 수신된 데이터를 여러 개의 출력선 중 하나로 출력하는 조합 회로를 의미하고, 멀티플렉서(650)는 여러 개의 입력선 중에서 하나를 선택하여 단일 출력선으로 연결하는 조합 회로를 의미할 수 있다. 디멀티플렉서(645) 및 멀티플렉서(650)는 셀렉터(640)에 의해 출력되는 제어 신호에 기초하여

제1 버퍼(630) 및 제2 버퍼(635) 중 연산기와 연결될 버퍼를 결정할 수 있다.

[0046] 한편, 셀렉터(640)는 제2 주기마다 계수하는 카운터를 포함할 수 있고, 셀렉터(640)에 의해 출력되는 제어 신호는 카운터의 출력 신호에 대응될 수 있다. 예를 들어, 셀렉터(640)는 제2 주기마다 카운트 값이 변경되는 카운터를 포함할 수 있고, 제2 주기마다 변경되는 카운트 값이 카운터로부터 디멀티플렉서(645) 및 멀티플렉서(650)의 제어단으로 출력됨에 따라 제1 버퍼(630) 및 제2 버퍼(635) 중 연산기와 연결될 버퍼가 주기적으로 변경될 수 있다.

[0047] 도 7은 일부 실시예에 따른 반도체 메모리 장치가 행렬 곱 연산을 수행하는 과정의 예시를 설명하기 위한 도면이다.

[0048] 도 7을 참조하면, 일부 실시예에 따른 반도체 메모리 장치가 행렬 $\begin{pmatrix} A0 & A1 \\ A2 & A3 \end{pmatrix}$ 및 행렬 $\begin{pmatrix} B0 & B1 \\ B2 & B3 \end{pmatrix}$간의 곱셈 연산을 수행하는 과정의 예시가 도시되어 있다. 도 7의 예시는 제1 주기가 4ns이고, 제2 주기가 2ns이며, n=2일 때를 가정한 경우에 대응된다.

[0049] 0ns에 메모리 뱅크 그룹으로부터 행렬 $\begin{pmatrix} B0 & B1 \\ B2 & B3 \end{pmatrix}$의 (1,1) 성분인 B0에 대응되는 내부 데이터가 연산기로 공급되고, 외부로부터 행렬 $\begin{pmatrix} A0 & A1 \\ A2 & A3 \end{pmatrix}$의 (1,1) 성분인 A0에 대응되는 외부 데이터가 연산기로 공급된다. 연산기는 행렬 곱 연산을 수행하기 위해 곱셈기로 구성될 수 있고, A0 및 B0에 대한 곱셈 연산을 수행할 수 있다. 곱셈 연산의 결과인 A0*B0는 제1 버퍼에 저장될 수 있다.

[0050] 이후 2ns에 외부로부터 행렬 $\begin{pmatrix} A0 & A1 \\ A2 & A3 \end{pmatrix}$의 (2,1) 성분인 A2에 대응되는 외부 데이터가 연산기로 공급됨에 따라 연산기는 새로운 외부 데이터에 대응되는 값인 A2 및 기존의 내부 데이터에 대응되는 값인 B0에 대한 곱셈 연산을 수행할 수 있다. 곱셈 연산의 결과인 A2*B0는 제2 버퍼에 저장될 수 있다.

[0051] 이후 4ns에 메모리 뱅크 그룹으로부터 행렬 $\begin{pmatrix} B0 & B1 \\ B2 & B3 \end{pmatrix}$의 (2,1) 성분인 B2에 대응되는 내부 데이터가 연산기로 새롭게 공급되고, 외부로부터 행렬 $\begin{pmatrix} A0 & A1 \\ A2 & A3 \end{pmatrix}$의 (1,2) 성분인 A1에 대응되는 외부 데이터가 연산기로 새롭게 공급될 수 있다. 연산기는 새로운 내부 데이터 및 외부 데이터가 수신됨에 따라 A1 및 B2에 대한 곱셈 연산을 수행할 수 있다. 곱셈 연산의 결과인 A1*B2는 제1 버퍼에 누적하여 저장될 수 있다. 이와 같이, 연산기에 포함되는 버퍼들 각각의 전단 및 후단 중 적어도 하나에는 연산 결과를 누적하여 저장하기 위해 누산기(accumulator)가 추가될 수 있다. 곱셈 연산의 결과인 A1*B2가 제1 버퍼에 누적됨에 따라 제1 버퍼는 A0*B0+A1*B2라는 값을 저장할 수 있다. A0*B0+A1*B2는 행렬 $\begin{pmatrix} A0 & A1 \\ A2 & A3 \end{pmatrix}$ 및 행렬 $\begin{pmatrix} B0 & B1 \\ B2 & B3 \end{pmatrix}$간의 곱셈 연산이 수행됨에 따라 생성되는 행렬의 (1,1) 성분에 대응되는 값일 수 있다.

[0052] 이후 6ns에 외부로부터 행렬의 $\begin{pmatrix} A0 & A1 \\ A2 & A3 \end{pmatrix}$의 (2,2) 성분인 A3에 대응되는 외부 데이터가 연산기로 공급됨에 따라 연산기는 새로운 외부 데이터에 대응되는 값인 A3 및 기존의 내부 데이터에 대응되는 값인 B2에 대한 곱셈 연산을 수행할 수 있다. 곱셈 연산의 결과인 A3*B2는 제2 버퍼에 누적하여 저장될 수 있다. 곱셈 연산의 결과인 A3*B2가 제2 버퍼에 누적됨에 따라 제2 버퍼는 A2*B0+A3*B2라는 값을 저장할 수 있다. A2*B0+A3*B2는 행렬 $\begin{pmatrix} A0 & A1 \\ A2 & A3 \end{pmatrix}$ 및 행렬 $\begin{pmatrix} B0 & B1 \\ B2 & B3 \end{pmatrix}$간의 곱셈 연산이 수행됨에 따라 생성되는 행렬의 (2,1) 성분에 대응되는 값일 수 있다.

[0053] 일부 실시예에 따른 반도체 메모리 장치는 전술한 과정을 반복함으로써 행렬 $\begin{pmatrix} A0 & A1 \\ A2 & A3 \end{pmatrix}$및 행렬 $\begin{pmatrix} B0 & B1 \\ B2 & B3 \end{pmatrix}$간의 곱셈 연산을 수행할 수 있다. 메모리 뱅크 그룹으로부터 데이터가 독출되는 제1 주기보다 빠른 주기인 제2 주기로 외부 데이터가 공급됨에 따라, 반도체 메모리 장치는 행렬 $\begin{pmatrix} A0 & A1 \\ A2 & A3 \end{pmatrix}$ 및 행렬 $\begin{pmatrix} B0 & B1 \\ B2 & B3 \end{pmatrix}$간의 곱셈 연산을 종래보다 훨씬 빠르게 수행할 수 있다. 한편, 반도체 메모리 장치가 수행할 수 있는 연산은 행렬 곱 연산에 제

한되지 않는다. 반도체 메모리 장치가 다양한 연산들을 수행할 수 있고, 임의의 PIM 연산 명령에 의해 반도체 메모리 장치에 포함되는 구성들이 다양한 방식으로 제어될 수 있음은 해당 기술분야의 통상의 기술자라면 쉽게 이해할 것이다.

[0054] 도 8은 일부 실시예에 따른 반도체 메모리 장치의 동작 방법과 종래 기술에 따른 반도체 메모리 장치의 동작 방법의 성능을 비교하기 위한 시뮬레이션 결과를 나타내는 도면이다.

[0055] 도 8을 참조하면, 일부 실시예에 따른 반도체 메모리 장치 및 종래 기술에 따른 반도체 메모리 장치 각각이 32비트 부동 소수점 데이터의 곱셈 연산을 204,800번 수행하기 위해 소요되는 지연시간(사이클)을 비교한 도면이다. 지연시간은 메모리 시뮬레이터 Ramulator를 이용하여 측정되었다.

[0056] 그래프(810)를 참조하면, 종래 기술에 따른 반도체 메모리 장치가 32비트 부동 소수점 데이터의 곱셈 연산을 204,800번 수행하기 위해 소요되는 지연시간은 약 730 사이클임을 알 수 있고, 그래프(820)를 참조하면, 일부 실시예에 따른 반도체 메모리 장치가 32비트 부동 소수점 데이터의 곱셈 연산을 204,800번 수행하기 위해 소요되는 지연시간은 약 390 사이클임을 알 수 있다.

[0057] 일부 실시예에 따른 반도체 메모리 장치의 동작 방법에 따르면, 동일한 수의 곱셈 연산을 수행하기 위해 종래 기술보다 훨씬 적은 사이클이 소요되는바, 일부 실시예에 따른 반도체 메모리 장치의 연산 속도가 종래 기술에 따른 반도체 메모리 장치의 연산 속도보다 크게 증가됨을 확인할 수 있다.

[0058] 도 9는 다른 실시예에 따른 반도체 메모리 장치의 구성을 나타내는 도면이다.

[0059] 도 9를 참조하면, 뱅크 그룹 단위의 병렬성을 이용하는 반도체 메모리 장치(10)를 도시한 도 1과는 달리, 뱅크 단위의 병렬성을 이용하는 반도체 메모리 장치(90)가 도시되어 있다. 반도체 메모리 장치(90)는 반도체 메모리 장치(10)와 기본적인 구조는 동일하나, 메모리 뱅크 그룹들 각각마다 연산기가 할당되는 것이 아니라 메모리 뱅크 그룹에 포함되는 메모리 뱅크들 각각마다 연산기가 할당되는 점에서 차이가 있다.

[0060] 예를 들어, 반도체 메모리 장치(90)는 메모리 뱅크(910), 메모리 뱅크(915), 내부 메모리 버스(920), 연산기(930) 및 연산기(935)를 포함할 수 있다. 메모리 뱅크(910)는 연산기(930)와 연결되고, 메모리 뱅크(915)는 연산기(935)와 연결된다. 연산기(930) 및 연산기(935)는 서로 독립적으로 연산 작업을 수행할 수 있다. 예를 들어, 연산기(930)가 메모리 뱅크(910)으로부터 수신된 데이터를 이용하여 연산 작업을 수행할 때, 연산기(935) 또한 메모리 뱅크(915)으로부터 수신된 데이터를 이용하여 연산 작업을 수행할 수 있다. 연산기(930) 및 연산기(935) 각각이 도 1을 참조하여 설명한 제1 연산기(130)와 동일 또는 유사한 방식으로 동작할 수 있음은 해당 기술분야의 통상의 기술자에게 자명할 것이므로, 중복되는 설명은 생략한다.

[0061] 도 10은 일부 실시예에 따른 반도체 메모리 장치에 포함되는 연산기의 동작 방법을 나타내는 흐름도이다.

[0062] 도 10을 참조하면, 반도체 메모리 장치에 포함되는 연산기의 동작 방법은 도 1 내지 도 9에 도시된 연산기에서 시계열적으로 처리되는 단계들로 구성된다. 따라서, 이하에서 생략된 내용이라고 하더라도 도 1 내지 도 9의 연산기에 관하여 이상에서 기술된 내용은 도 10의 연산기의 동작 방법에도 적용됨을 알 수 있다.

[0063] 단계 1010에서, 연산기는 제1 주기마다 메모리 뱅크 그룹으로부터 메모리 뱅크 그룹에 저장된 내부 데이터를 수신할 수 있다. 동일한 메모리 뱅크 그룹으로부터 내부 데이터가 독출되는 주기는 물리적인 제약으로 인해 제1 주기로 제한될 수 있다.

[0064] 단계 1020에서, 연산기는 제1 주기보다 짧은 제2 주기마다 내부 메모리 버스를 통해 메모리 뱅크 그룹의 외부에 저장되거나 메모리 뱅크 그룹의 외부에서 처리되는 외부 데이터를 수신할 수 있다. 외부 데이터는 반도체 메모리 장치에 포함되는 메모리 채널들 중 메모리 뱅크 그룹을 포함하는 메모리 채널 외의 다른 메모리 채널에 저장된 데이터, 반도체 메모리 장치의 외부의 다른 메모리 장치에 저장된 데이터 및 반도체 메모리 장치 외부의 호스트 프로세서에서 처리되는 데이터일 수 있다. 반도체 메모리 장치 외부의 호스트 프로세서는 CPU, GPU, DSP 등일 수 있으나, 이에 제한되는 것은 아니다.

[0065] 제1 주기는 동일한 메모리 뱅크 그룹을 연속으로 접근하는 경우의 지연시간(tCCD_L)에 대응되고, 제2 주기는 서로 다른 메모리 뱅크 그룹들을 번갈아 접근하는 경우의 지연시간(tCCD_S)에 대응될 수 있다. 일 예에서, 제1 주기는 제2 주기의 n배(n은 2 이상의 자연수)일 수 있다. 다만, n은 반드시 2 이상의 자연수일 필요는 없으며, n은 2 이상이기만 하면 어떤 값을 갖더라도 상관 없다.

[0066] 단계 1030에서, 연산기는 제2 주기마다 내부 데이터 및 외부 데이터에 대한 PIM 연산을 수행할 수 있다. PIM 연

산은 반도체 메모리 장치 내부에서 프로세싱 기능을 갖는 연산기에 의해 수행되는 연산을 의미할 수 있다. 이와 같이, 본 개시의 일부 실시예에 따르면, 메모리 뱅크 그룹으로부터 데이터가 독출되는 지연시간(tCCD_L)이 일반적인 메모리 사용시 데이터가 독출되는 지연시간(tCCD_S)보다 길기 때문에 발생되는 대기 시간에 외부 데이터가 제2 주기로 연산기에 공급될 수 있고, 이에 따라, 연산기는 제2 주기로 연산을 수행할 수 있다. 연산기는 내부 데이터 및 외부 데이터에 대한 연산을 위해 내부 데이터를 n번 재사용할 수 있다. 제2 주기는 제1 주기보다 짧으므로, 일부 실시예에 따른 반도체 메모리 장치의 동작 방법에 따르면, 연산기의 연산 속도가 증가될 수 있다.

[0067] 연산기는 내부 데이터 및 외부 데이터에 대한 연산의 결과를 제1 주기를 제2 주기로 나눈 몫인 n에 대응되는 개수의 복수의 버퍼들 각각에 한번씩 번갈아 가며 저장할 수 있다. 예를 들어, 연산기는 n개의 버퍼들 중 내부 데이터 및 외부 데이터에 대한 연산의 결과를 저장할 버퍼를 선택하기 위한 제어 신호를 출력함으로써 n개의 버퍼들 중 연산기와 연결될 하나의 버퍼를 선택할 수 있다.

[0068] 한편, 전술한 연산기의 동작 방법은 그 방법을 실행하는 명령어들을 포함하는 하나 이상의 프로그램이 기록된 컴퓨터로 읽을 수 있는 기록 매체에 기록될 수 있다. 컴퓨터로 읽을 수 있는 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(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 명령어의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0069] 도 11은 일부 실시예에 따른 전자 시스템의 구성을 나타내는 블록도이다.

[0070] 도 11을 참조하면, 전자 시스템(1100)은 뉴럴 네트워크를 기초로 입력 데이터를 실시간으로 분석하여 유효한 정보를 추출하고, 추출된 정보를 기초로 상황 판단을 하거나 또는 전자 시스템(1100)이 탑재되는 전자 디바이스의 구성들을 제어할 수 있다. 예컨대 전자 시스템(1100)은 드론(drone), 첨단 운전자 보조 시스템(Advanced Drivers Assistance System; ADAS) 등과 같은 로봇 장치, 스마트 TV, 스마트폰, 의료 디바이스, 모바일 디바이스, 영상 표시 디바이스, 계측 디바이스, IoT 디바이스 등에 적용될 수 있으며, 이 외에도 다양한 종류의 전자 디바이스들 중 적어도 하나에 탑재될 수 있다. 예를 들어, 전자 시스템(1100)은 서버일 수도 있다.

[0071] 전자 시스템(1100)은 프로세서(1110), 반도체 메모리 장치(1120), 뉴럴 네트워크 장치(1130), 시스템 메모리(1140), 센서 모듈(1150) 및 통신 모듈(1160)을 포함할 수 있다. 전자 시스템(1100)은 입출력 모듈, 보안 모듈, 전력 제어 장치 등을 더 포함할 수 있다. 전자 시스템(1100)의 하드웨어 구성들 중 일부는 적어도 하나의 반도체 칩에 탑재될 수 있다.

[0072] 프로세서(1110)는 전자 시스템(1100)의 전반적인 동작을 제어한다. 프로세서(1110)는 하나의 프로세서 코어(Single Core)를 포함하거나, 복수의 프로세서 코어들(Multi-Core)을 포함할 수 있다. 프로세서(1110)는 시스템 메모리(1140)에 저장된 프로그램들 및/또는 데이터를 처리 또는 실행할 수 있다. 일부 실시예에 있어서, 프로세서(1110)는 시스템 메모리(1140)에 저장된 프로그램들을 실행함으로써, 뉴럴 네트워크 장치(1130)의 기능을 제어할 수 있다. 프로세서(1110)는 CPU, GPU, AP(Application Processor) 등으로 구현될 수 있다.

[0073] 반도체 메모리 장치(1120)는 프로그램들, 데이터, 또는 명령들(instructions)을 일시적으로 저장할 수 있다. 예컨대 시스템 메모리(1140)에 저장된 프로그램들 및/또는 데이터는 프로세서(1110)의 제어 또는 부팅 코드에 따라 반도체 메모리 장치(1120)에 일시적으로 저장될 수 있다. 반도체 메모리 장치(1120)은 DRAM 등의 메모리로 구현될 수 있다. 한편, 반도체 메모리 장치(1120)는 도 1 내지 도 10을 참조하여 설명한 반도체 메모리 장치에 대응될 수 있다. 반도체 메모리 장치(1120)는 PIM에 해당하므로, 데이터를 저장하는 메모리 기능뿐만 아니라 연산 작업을 수행하는 프로세서의 기능을 수행할 수 있다. 이에 따라, 프로세서(1110)에 연산 작업이 과도하게 할당되지 않을 수 있고, 전자 시스템(1100)의 전체적인 성능이 증가될 수 있다.

[0074] 뉴럴 네트워크 장치(1130)는 수신되는 입력 데이터를 기초로 뉴럴 네트워크의 연산을 수행하고, 수행 결과를 기초로 정보 신호를 생성할 수 있다. 뉴럴 네트워크는 CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), FNN(Feedforward Neural Network), Deep Belief Networks, Restricted Boltzman Machines 등을 포함할 수 있으나 이에 제한되지 않는다.

[0075] 정보 신호는 음성 인식 신호, 사물 인식 신호, 영상 인식 신호, 생체 정보 인식 신호 등과 같은 다양한 종류의 인식 신호 중 하나를 포함할 수 있다. 예를 들어, 뉴럴 네트워크 장치(1130)는 비디오 스트림에 포함되는 프레임 데이터를 입력 데이터로서 수신하고, 프레임 데이터로부터 프레임 데이터가 나타내는 이미지에 포함된 사물

에 대한 인식 신호를 생성할 수 있다. 그러나, 이에 제한되는 것은 아니며, 전자 시스템(1100)이 탑재된 전자 장치의 종류 또는 기능에 따라 뉴럴 네트워크 장치(1130)는 다양한 종류의 입력 데이터를 수신할 수 있고, 입력 데이터에 따른 인식 신호를 생성할 수 있다.

[0076] 시스템 메모리(1140)는 데이터를 저장하기 위한 저장 장소로서, OS(Operating System), 각종 프로그램들, 및 각종 데이터를 저장할 수 있다. 실시예에 있어서, 시스템 메모리(1140)는 뉴럴 네트워크 장치(1130)의 연산 수행 과정에서 생성되는 중간 결과들을 저장할 수 있다.

[0077] 시스템 메모리(1140)는 휘발성 메모리 또는 불휘발성 메모리 중 적어도 하나를 포함할 수 있다. 불휘발성 메모리는 ROM, PROM, EPROM, EEPROM, 플래시 메모리, PRAM, MRAM, RRAM, FRAM 등을 포함한다. 휘발성 메모리는 DRAM, SRAM, SDRAM, PRAM, MRAM, RRAM, FeRAM 등을 포함한다. 실시예에 있어서, 시스템 메모리(1140)는 HDD, SSD, CF, SD, Micro-SD, Mini-SD, xD 또는 Memory Stick 중 적어도 하나를 포함할 수 있다.

[0078] 센서 모듈(1150)은 전자 시스템(1100)이 탑재되는 전자 장치 주변의 정보를 수집할 수 있다. 센서 모듈(1150)은 전자 장치의 외부로부터 신호(예컨대 영상 신호, 음성 신호, 자기 신호, 생체 신호, 터치 신호 등)를 센싱 또는 수신하고, 센싱 또는 수신된 신호를 데이터로 변환할 수 있다. 이를 위해, 센서 모듈(1150)은 센싱 장치, 예컨대 마이크, 촬상 장치, 이미지 센서, 라이더(LIDAR; light detection and ranging) 센서, 초음파 센서, 적외선 센서, 바이오 센서, 및 터치 센서 등 다양한 종류의 센싱 장치 중 적어도 하나를 포함할 수 있다.

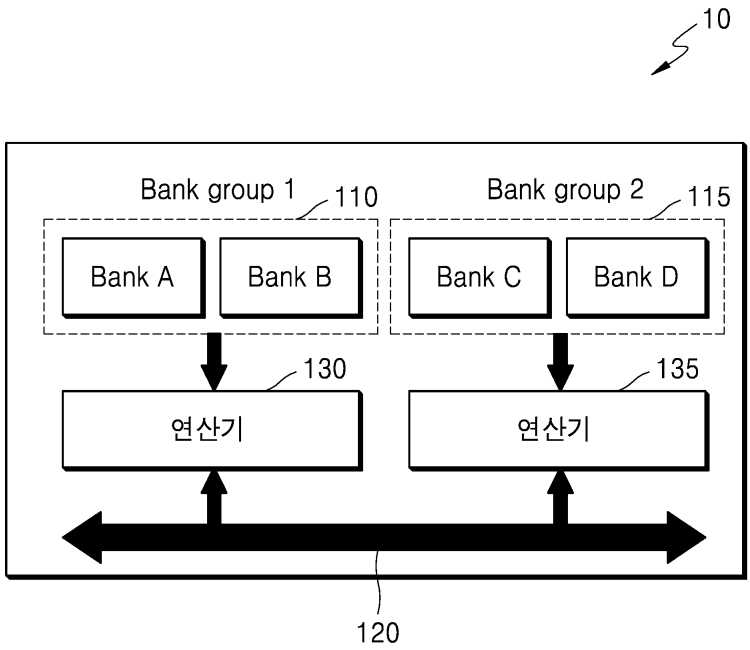
[0079] 센서 모듈(1150)은 변환된 데이터를 뉴럴 네트워크 장치(1130)에 입력 데이터로서 제공할 수 있다. 예를 들어, 센서 모듈(1150)은 이미지 센서를 포함할 수 있으며, 전자 장치의 외부 환경을 촬영하여 비디오 스트림을 생성하고, 비디오 스트림의 연속하는 데이터 프레임을 뉴럴 네트워크 장치(1130)에 입력 데이터로서 순서대로 제공할 수 있다. 그러나 이에 제한되는 것은 아니며 센서 모듈(1150)은 다양한 종류의 데이터를 뉴럴 네트워크 장치(1130)에 제공할 수 있다.

[0080] 통신 모듈(1160)은 외부 디바이스와 통신할 수 있는 다양한 유선 또는 무선 인터페이스를 구비할 수 있다. 예컨대 통신 모듈(1160)은 유선 근거리통신망(Local Area Network; LAN), Wi-fi(Wireless Fidelity)와 같은 무선 근거리 통신망 (Wireless Local Area Network; WLAN), 블루투스(Bluetooth)와 같은 무선 개인 통신망(Wireless Personal Area Network; WPAN), 무선 USB (Wireless Universal Serial Bus), Zigbee, NFC (Near Field Communication), RFID (Radio-frequency identification), PLC(Power Line communication), 또는 3G (3rd Generation), 4G (4th Generation), LTE (Long Term Evolution) 등 이동 통신망(mobile cellular network)에 접속 가능한 통신 인터페이스 등을 포함할 수 있다.

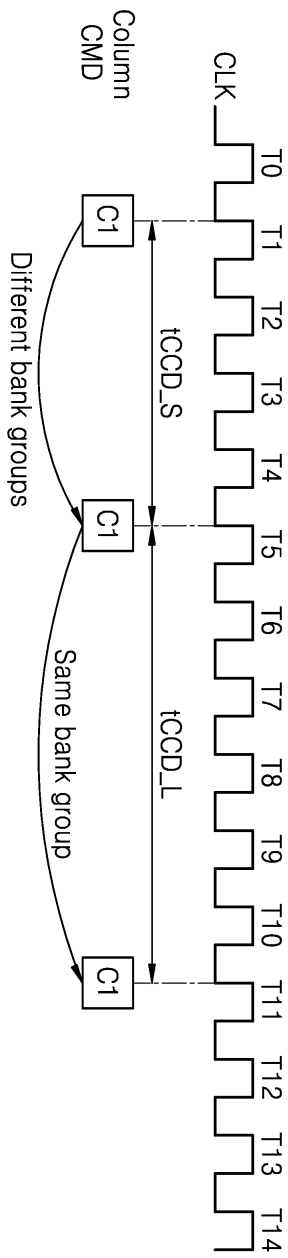
[0081] 이상에서 실시예들에 대하여 상세하게 설명하였지만 본 발명의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속한다.

도면

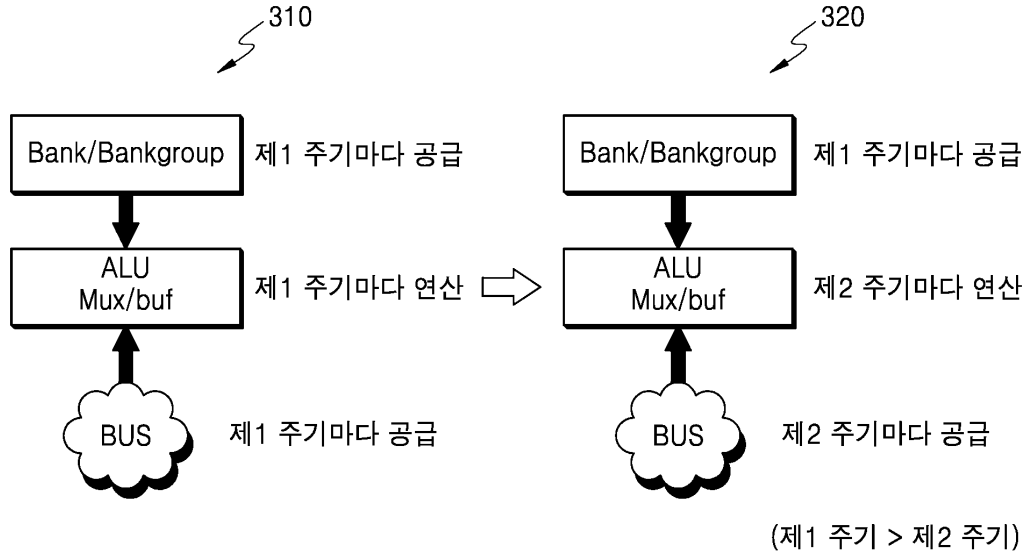
도면1

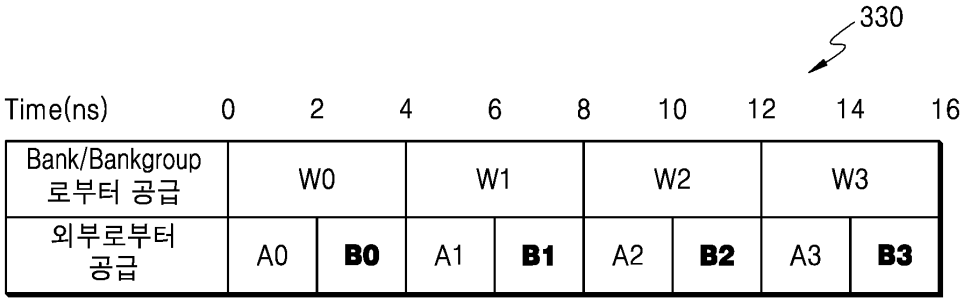

10
Bank group 1
110
Bank group 2
115
Bank A
Bank B
Bank C
Bank D
130
135
연산기
연산기
120

도면2

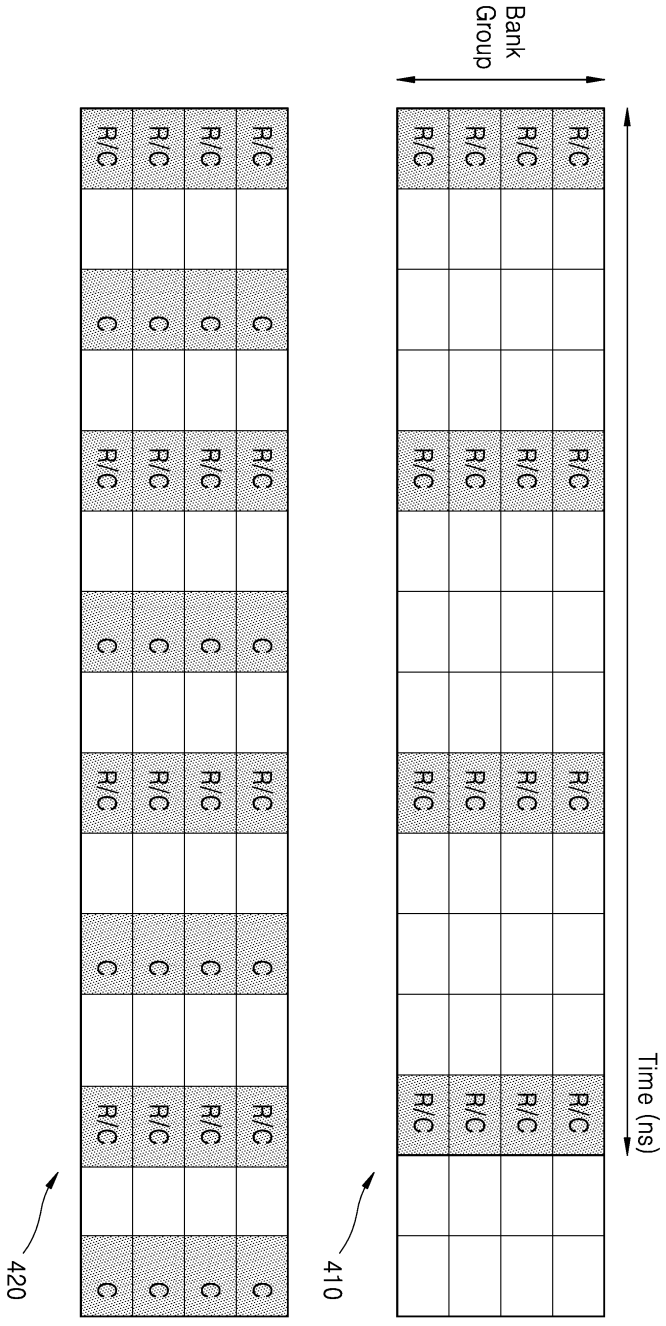

CLK
Column CMD
T0
T1
T2
T3
T4
T5
T6
T7
T8
T9
T10
T11
T12
T13
T14
C1
C1
C1
tCCD_S
tCCD_L
Different bank groups
Same bank group

도면3

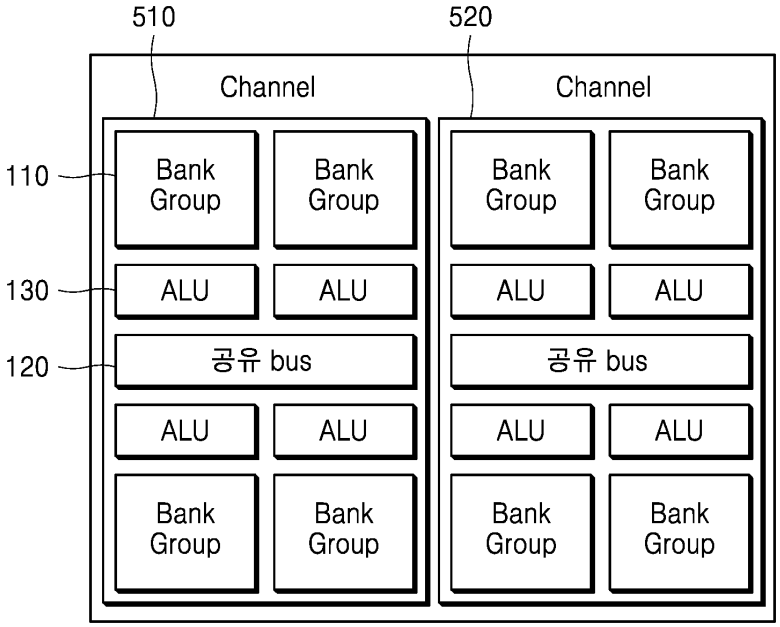

310
320
Bank/Bankgroup
제1 주기마다 공급
Bank/Bankgroup
제1 주기마다 공급
ALU
Mux/buf
제1 주기마다 연산
ALU
Mux/buf
제2 주기마다 연산
BUS
제1 주기마다 공급
BUS
제2 주기마다 공급
(제1 주기 > 제2 주기)


330
Time(ns)
0
2
4
6
8
10
12
14
16
Bank/Bankgroup 로부터 공급
W0
W1
W2
W3
외부로부터 공급
A0
B0
A1
B1
A2
B2
A3
B3

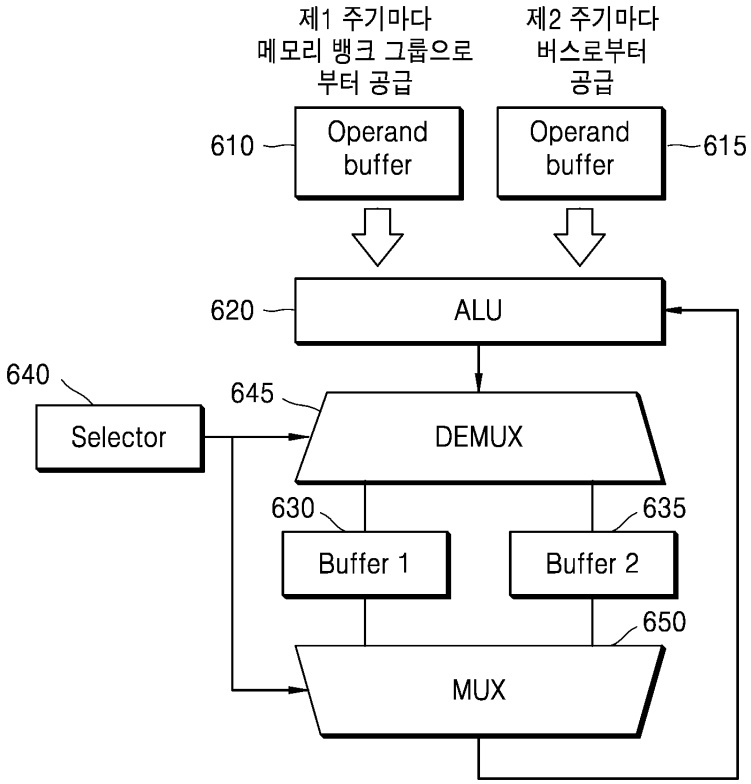
도면4



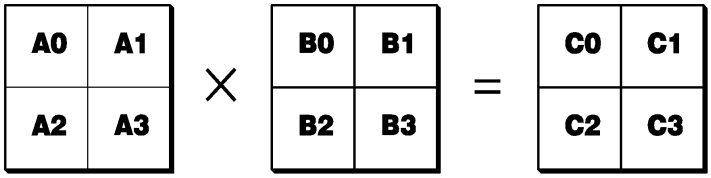
도면5


510
520
Channel
Channel
110
Bank Group
Bank Group
Bank Group
Bank Group
130
ALU
ALU
ALU
ALU
120
공유 bus
공유 bus
ALU
ALU
ALU
ALU
Bank Group
Bank Group
Bank Group
Bank Group

도면6

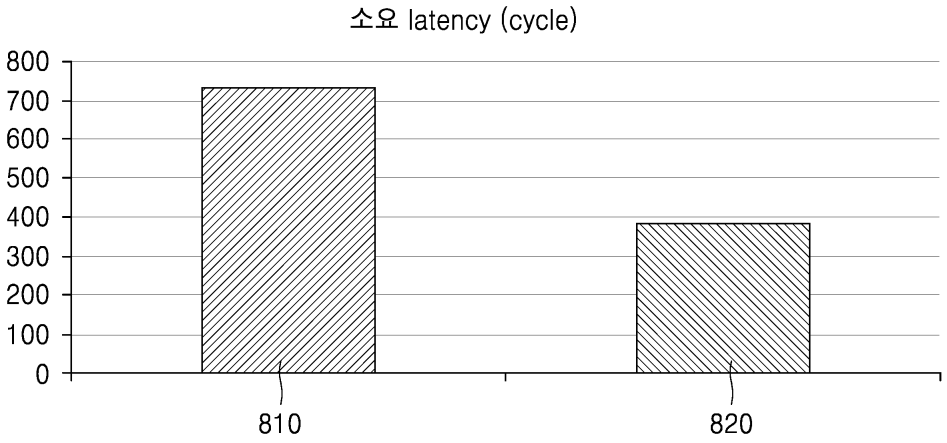

제1 주기마다 메모리 뱅크 그룹으로부터 공급
제2 주기마다 버스로부터 공급
610
Operand buffer
Operand buffer
615
620
ALU
640
Selector
645
DEMUX
630
635
Buffer 1
Buffer 2
650
MUX

도면7

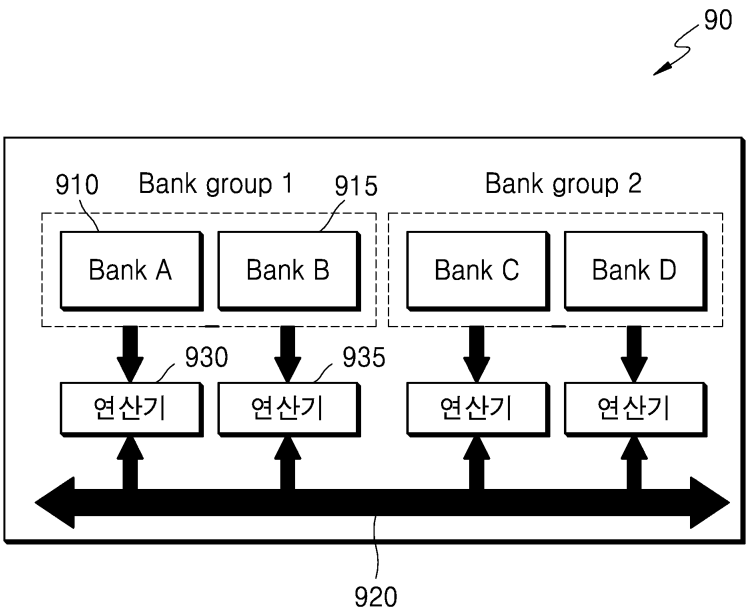


Time	0	2ns	4ns	6ns	8ns
Bank로부터 공급	B0		B2		
외부로부터 공급	A0	A2	A1	A3	
Buffer에 누적되고 있는 값	A0*B0	A2*B0	A0*B0+A1*B2	A2*B0+A3*B2	

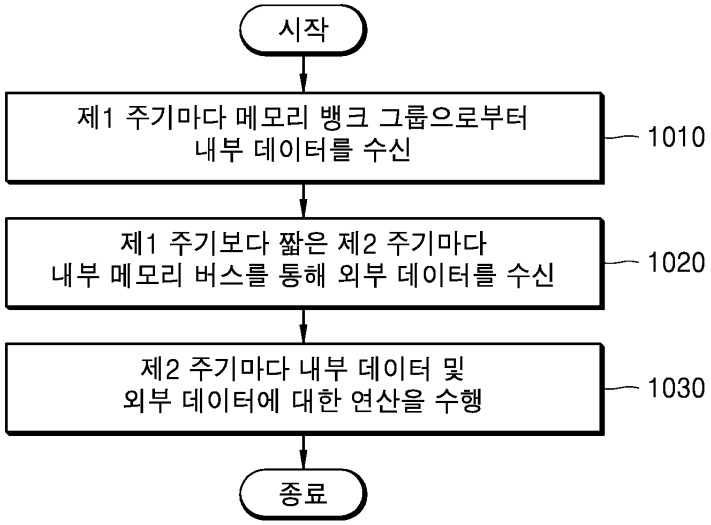
도면8



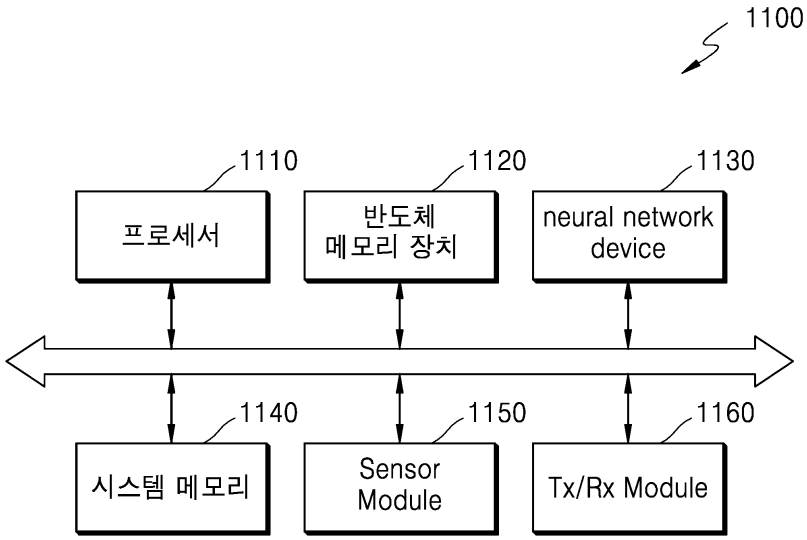
도면9



도면10


시작
제1 주기마다 메모리 뱅크 그룹으로부터
내부 데이터를 수신
1010
제1 주기보다 짧은 제2 주기마다
내부 메모리 버스를 통해 외부 데이터를 수신
1020
제2 주기마다 내부 데이터 및
외부 데이터에 대한 연산을 수행
1030
종료

도면11


1100
1110
프로세서
1120
반도체
메모리 장치
1130
neural network
device
1140
시스템 메모리
1150
Sensor
Module
1160
Tx/Rx Module