



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0091206
(43) 공개일자 2020년07월30일

(51) 국제특허분류(Int. Cl.)
G06T 1/60 (2006.01) G11C 8/08 (2006.01)
G11C 8/10 (2006.01)
(52) CPC특허분류
G06T 1/60 (2013.01)
G11C 8/08 (2013.01)
(21) 출원번호 10-2019-0008189
(22) 출원일자 2019년01월22일
심사청구일자 2019년01월22일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
정성우
서울특별시 강남구 도곡로28길 8, 104동 1802호 (도곡동, 도곡1차아이파크아파트)
두 쿡 뚜안
서울특별시 동대문구 안암로 118-1, 402호(제기동)
이영서
서울특별시 동대문구 무학로49길 40, 502호(용두동)
(74) 대리인
송인호, 윤형근, 최영중, 최관락

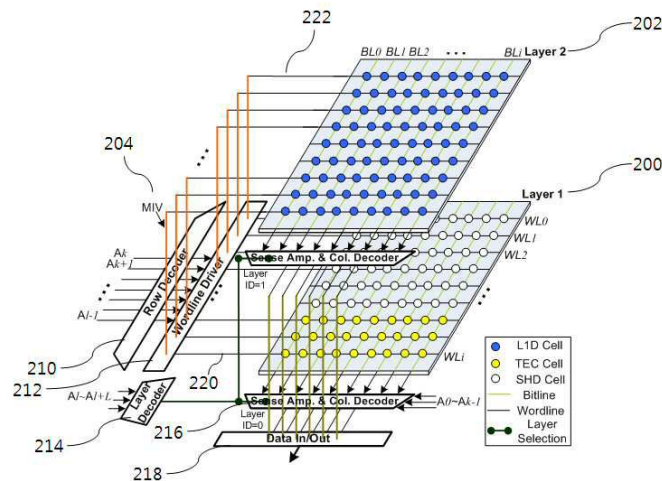
전체 청구항 수 : 총 5 항

(54) 발명의 명칭 모놀리식 3D 집적 구조 기반 GPU L1 캐시 메모리

(57) 요약

본 발명은 모놀리식 3D 집적 구조 기반 GPU L1 캐시 메모리를 개시한다. 본 발명에 따르면, 데이터 캐시, 텍스처 캐시, 상수 캐시 및 공유 메모리를 포함하는 GPU L1 캐시 메모리 장치로서, 텍스처 캐시와 상수 캐시가 통합된 셀과 공유 메모리의 비트 셀이 배치되는 제1 레이어; 상기 제1 레이어의 상부에 위치하며 데이터 캐시의 비트 셀이 배치되는 제2 레이어; 및 상기 제1 레이어 및 제2 레이어를 수직으로 연결하는 모놀리식 3D 비아(Monolithic 3D Via)를 포함하는 GPU L1 캐시 메모리 장치가 제공된다.

대표도 - 도2



(52) CPC특허분류

G11C 8/10 (2013.01)

G06T 2200/28 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711071490
부처명 과학기술정보통신부
연구관리전문기관 한국연구재단
연구사업명 나노·소재기술개발(R&D)
연구과제명 모노리틱 3D 기반 3차원 고집적 뉴로모픽 인터커넥트 설계/분석 플랫폼
기 여 율 30/100
주관기관 동국대학교 산학협력단
연구기간 2018.03.01 ~ 2019.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호 1711075684
부처명 과학기술정보통신부
연구관리전문기관 한국연구재단
연구사업명 나노·소재기술개발(R&D)
연구과제명 초절전 IoT 소자 결합 3차원 플랫폼 집적 공정과 최적 아키텍처 개발, 검증
기 여 율 30/100
주관기관 인하대학교 산학협력단
연구기간 2018.05.01 ~ 2019.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호 1711069761
부처명 과학기술정보통신부
연구관리전문기관 한국연구재단
연구사업명 개인기초연구(과기정통부)(R&D)
연구과제명 모노리틱 3D에 기반한 CPU/GPU 컴포넌트의 아키텍처적 설계
기 여 율 40/100
주관기관 고려대학교 산학협력단
연구기간 2018.03.01 ~ 2019.02.28

명세서

청구범위

청구항 1

데이터 캐시, 텍스처 캐시, 상수 캐시 및 공유 메모리를 포함하는 GPU L1 캐시 메모리 장치로서,
 텍스처 캐시와 상수 캐시가 통합된 셀과 공유 메모리의 비트 셀이 배치되는 제1 레이어;
 상기 제1 레이어의 상부에 위치하며 데이터 캐시의 비트 셀이 배치되는 제2 레이어; 및
 상기 제1 레이어 및 제2 레이어를 수직으로 연결하는 모놀리식 3D 비아(Monolithic 3D Via)를 포함하는 GPU L1 캐시 메모리 장치.

청구항 2

제1항에 있어서,
 상기 제1 레이어의 제1 워드라인 와이어와 연결되는 워드라인 드라이버를 더 포함하되,
 상기 모놀리식 3D 비아는 상기 워드라인 드라이버와 상기 제2 레이어의 제2 워드라인 와이어를 수직으로 연결하는 GPU L1 캐시 메모리 장치.

청구항 3

제2항에 있어서,
 상기 워드라인 드라이버는 로우 디코더의 출력과 연결되며,
 상기 로우 디코더에는 캐시 주소 변환기에 의해 변환된 캐시 주소가 입력되는 GPU L1 캐시 메모리 장치.

청구항 4

제3항에 있어서,
 상기 캐시 주소 변환기는, 데이터 캐시, 텍스처 캐시, 상수 캐시 및 공유 메모리 중 하나를 식별하기 위한 코드를 이용하여 기존 캐시 주소를 3차원 레이어 배치에 따른 새로운 캐시 주소로 변환하여 출력하는 GPU L1 캐시 메모리 장치.

청구항 5

제3항에 있어서,
 요구되는 비트 셀의 레이어 아이디를 식별하는 레이어 디코더;
 비트라인을 통해 전달된 비트 셀로부터의 데이터를 이용하여 요구된 데이터를 선택하는 컬럼 디코더; 및
 상기 선택된 데이터 신호를 증폭하는 센스 앰프를 더 포함하는 GPU L1 캐시 메모리 장치.

발명의 설명

기술 분야

[0001] 본 발명은 모놀리식 3D 집적 구조 기반 GPU L1 캐시 메모리에 관한 것이다.

배경 기술

[0002] GPU(Graphics Processing Unit)은 수천 개의 코어(프로세서)를 사용하여 많은 수의 스레드가 병렬로 실행되도록 하는 기기이다.

[0003] 그래픽 전용 연산을 수행하는 GPU 시스템은 CUDA(Compute Unified Device Architecture), OpenCL(Open

Computing Language), C++AMP (C++Accelerated Massive Parallelism)과 같은 언어를 이용하여 GPGPU(General-Purpose computing on Graphics Processing Units)로 수행이 가능하다. GPU 시스템은 많은 수의 프로세서를 가지고 있으며 SIMD(Single Instruction Multiple Data)와 유사한 방법으로 데이터를 처리를 한다. 때문에 2차원 및 3차원으로 구성된 대용량 데이터를 처리하는데 적합하다.

- [0004] 하지만 많은 프로세서가 하나의 전역 메모리에 동시접근하기 위해 경쟁을 함으로써 메모리 접근 지연으로 인한 성능 감소가 발생한다.
- [0005] 이를 해결하기 위해 공유 메모리를 사용하는 기법이 제안되었으나, 구현 복잡도가 높아지는 문제점이 있었다.
- [0006] 이에, 근래에는 캐시 메모리를 제공하여 CPU에서처럼 전역 메모리로부터 읽어온 값을 캐싱한다.
- [0007] 캐시 메모리와 공유 메모리는 같은 하드웨어에 위치하고 있으며 실행 환경에 따라서 메모리 공간들이 구분이 된다. 그러므로 캐시 메모리를 활용한다면 공유 메모리를 사용하지 않아도 동일한 성능을 볼 수 있을 것이며 공유 메모리 사용으로 나타나는 구현 복잡도는 피할 수 있을 것이다.
- [0008] GPU에서 워프 스케줄러(warp scheduler)는 GPU 하드웨어 자원을 효율적으로 사용하여 GPU 성능을 향상시키는데 중요한 역할을 한다.
- [0009] GPU 성능 향상을 위해 다수의 워프 스케줄러가 제안되었지만 GPU 하드웨어 자원을 아직 완전히 활용하고 있지 못하다.
- [0010] 이는 워프 스케줄러의 효율성 및 GPU 성능이 워프 스케줄러 알고리즘 자체보다는 L1 캐시 메모리 용량에 크게 의존하기 때문이다.
- [0011] 특히 작은 용량의 L1 캐시 메모리는 GPU(GPGPU)의 범용 컴퓨팅에서 자주 발생하는 대규모 작업 세트를 수용할 수 없고 이는 캐시 스래싱(cache thrashing)으로 이어질뿐만 아니라 메모리 시스템 리소스를 포화상태로 만든다.

선행기술문헌

특허문헌

- [0012] (특허문헌 0001) 등록특허 10-1639943

발명의 내용

해결하려는 과제

- [0013] 상기한 종래기술의 문제점을 해결하기 위해, 캐시 액세스 시간은 악화시키지 않으면서 L1 캐시 용량을 증가시켜 GPU 성능을 향상시킬 수 있는 모듈리식 3D 집적 구조 기반 GPU L1 캐시메모리를 제안하고자 한다.

과제의 해결 수단

- [0014] 상기한 바와 같은 목적을 달성하기 위하여, 본 발명의 일 실시예에 따르면, 데이터 캐시, 텍스처 캐시, 상수 캐시 및 공유 메모리를 포함하는 GPU L1 캐시 메모리 장치로서, 텍스처 캐시와 상수 캐시가 통합된 셀과 공유 메모리의 비트 셀이 배치되는 제1 레이어; 상기 제1 레이어의 상부에 위치하며 데이터 캐시의 비트 셀이 배치되는 제2 레이어; 및 상기 제1 레이어 및 제2 레이어를 수직으로 연결하는 모듈리식 3D 비아(Monolithic 3D Via)를 포함하는 GPU L1 캐시 메모리 장치가 제공된다.
- [0015] 상기 제1 레이어의 제1 워드라인 와이어와 연결되는 워드라인 드라이버를 더 포함되, 상기 모듈리식 3D 비아는 상기 워드라인 드라이버와 상기 제2 레이어의 제2 워드라인 와이어를 수직으로 연결할 수 있다.
- [0016] 상기 워드라인 드라이버는 로우 디코더의 출력과 연결되며, 상기 로우 디코더에는 캐시 주소 변환기에 의해 변환된 캐시 주소가 입력될 수 있다.
- [0017] 상기 캐시 주소 변환기는, 데이터 캐시, 텍스처 캐시, 상수 캐시 및 공유 메모리 중 하나를 식별하기 위한 코드를 이용하여 기존 캐시 주소를 3차원 레이어 배치에 따른 새로운 캐시 주소로 변환하여 출력할 수 있다.

[0018] 요구되는 비트 셀의 레이어 아이디를 식별하는 레이어 디코더; 비트라인을 통해 전달된 비트 셀로부터의 데이터를 이용하여 요구된 데이터를 선택하는 컬럼 디코더; 및 상기 선택된 데이터 신호를 증폭하는 센스 앰프를 더 포함할 수 있다.

발명의 효과

[0019] 본 발명에 따르면, 데이터 캐시와 텍스처 캐시/상수 캐시 및 공유 메모리를 서로 다른 레이어로 분리하고, 이들을 모놀리식 3D 비아로 연결하여 액세스 시간을 한층 단축시킬 수 있는 장점이 있다.

도면의 간단한 설명

[0020] 도 1은 일반적인 GPU의 SM 구조를 도시한 도면이다.

도 2는 본 발명의 바람직한 일 실시예에 따른 GPU L1 캐시 메모리 구조를 도시한 도면이다.

도 3은 본 발명의 일 실시예에 따른 캐시 주소 변환기의 구성을 도시한 도면이다.

도 4는 본 실시예에 따른 캐시 주소 변환기에 의한 캐시 주소 변환 전후 결과를 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0021] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다.

[0022] 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0024] 도 1은 일반적인 GPU의 SM 구조를 도시한 도면이다.

[0025] GPU는 SM(Streaming Multiprocessor)의 집합으로 구성되며 각각의 SM은 복수의 SP(Scalar Processor) 및 복수의 스레드를 포함한다.

[0026] 도 1에 도시된 바와 같이, SM은 8개의 SP 및 32개의 스레드를 포함할 수 있다.

[0027] 여기서, 32개의 스레드를 워프(warp)라는 단위로 부르며 16개의 스레드로 메모리참조를 하는 경우에는 하프워프라고 한다.

[0028] GPU는 1SM 단위(32스레드)로 동작(메모리참조)한다.

[0029] SP는 4개의 스레드로 구성되고 실행된다.

[0030] 여기서, 스레드는 최소 명령어 처리 단위이다.

[0031] 이와 같은 SM은 데이터 캐시(L1D), 텍스처 캐시(TEX), 상수 캐시(CONST) 및 공유 메모리(SHD)를 포함하는 특정 용도의 캐시들을 포함한다.

[0032] 이처럼 특정 용도의 캐시들을 분리하는 이유는 빠른 액세스 시간을 위한 것이다.

[0033] SM 내에서 다른 SP로부터의 메모리 요청은 큐에 저장되어 특정 용도 캐시에 대한 메모리 요청 순서를 유지한다.

[0034] 본 발명의 바람직한 일 실시예에 따르면, 빠른 액세스를 보장하기 위해, 즉 데이터 캐시에서 간섭을 피하기 위해 최하위 계층에 텍스처 캐시와 상수 캐시를 통합하고, 공유 메모리 역시 최하위 계층에 배치한다.

[0035] 또한, 데이터 캐시는 모놀리식 3D 집적 구조 기반으로 SRAM의 상위 계층에 배치한다.

[0036] 도 2는 본 발명의 바람직한 일 실시예에 따른 GPU L1 캐시 메모리 구조를 도시한 도면이다.

[0037] 도 2에 도시된 바와 같이, GPU L1 캐시 메모리는 하위 계층인 제1 레이어(Layer1, 200) 및 상위 계층인 제2 레이어(Layer2, 202)를 포함할 수 있다.

[0038] 제1 레이어(200)에는 텍스처 캐시와 상수 캐시가 통합된 TEC 비트 셀과 공유 메모리의 SHD 비트 셀이 배치된다.

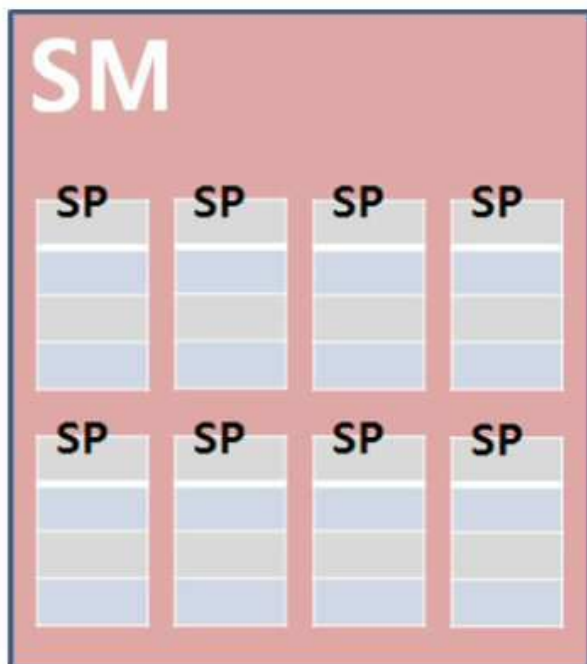
[0039] 또한, 제2 레이어(202)에는 데이터 캐시인 L1D 비트 셀이 배치된다.

- [0040] 도 2에서는 제2 레이어(202)가 하나로 도시되어 있으나, 이는 일예에 불과하며 L1 데이터 캐시의 경우 적중률이 다른 캐시들에 비해 낮기 때문에 용량을 키우기 위해 제2 레이어(202)는 복수개로 구성될 수 있다.
- [0041] 이러한 경우, 제1 레이어(200) 상에 수직으로 하나 이상의 제2 레이어(202)가 배치될 수 있고, 복수의 제2 레이어(202) 각각도 수직으로 배치될 수 있다.
- [0042] 본 실시예에 따르면, 제1 레이어(200)와 제2 레이어(202)는 모놀리식 3D 비아(Monolithic 3D Via, 204)를 통해 연결된다.
- [0043] 여기서 모놀리식 3D 집적 구조는 3차원 시퀀셜 집적구조로 정의된다.
- [0044] 모놀리식 3D 집적 소자 형성 기술은 하나의 웨이퍼에 소자 형성 후 epi 성장, recrystallization 또는 절연체 층을 증착한 후 얇은 웨이퍼 위에 bonding을 통해 새로운 substrate layer를 형성 후 그 위에 소자를 시퀀셜하게 형성하는 공정 기술이다.
- [0045] 모놀리식 3D 집적 소자 형성 기술은 각 레이어의 소자를 매우 얇은 채널 (50~500nm) 형성으로 Via specification의 변경이 가능 [(Via diameter (10~50nm), Via pitch (100~300nm))] 하여 TSV(Through Si Via)의 한계점으로 지적되는 웨이퍼의 두께에 따른 문제점들을 해결 가능한 새로운 대안 제조 기술이다.
- [0046] 이와 같은 모놀리식 3D 집적 구조에서는 각 레이어의 소자를 연결하는 비아를 인터-레이어 비아(inter-layer via) 또는 인터-타이어 비아(inter-tier via)로 정의하며, 본 명세서에서는 상위 계층인 L1D 비트 셀과 하위 계층인 TEX 비트 셀 및 SHD 비트 셀이 모놀리식 3D 집적 구조를 이루어진다는 점에서 모놀리식 3D 인터-비아(MIV)로 정의한다.
- [0047] 또한, 본 실시예에 따른 GPU L1 캐시 메모리는 로우 디코더(Row Decoder, 210), 워드라인 드라이버(Wordline Driver, 212), 레이어 디코더(214), 센스 앰프 및 컬럼 디코더(216) 및 데이터 입출력부(218)를 포함할 수 있다.
- [0048] 본 실시예에 따른 M3D 기반 GPU L1 캐시 메모리(SRAM)의 동작을 설명하기 위해 수직 워드라인 개념이 사용된다.
- [0049] 전술한 바와 같이, M3D 기반 SRAM에서 MIV(M3D Via, 204)는 상하위 레이어(200,202)를 수직으로 연결하는데 사용된다.
- [0050] 여기서, 제1 레이어(200)의 워드라인이 글로벌 워드라인이 된다.
- [0051] 2D H-트리 네트워크를 통한 뱅크 라우팅은 수직 뱅크 라우팅 경로가 필요없는 제1 레이어(200)에서만 수행된다.
- [0052] 모든 캐시들이 M3D 기반 SRAM에 구현되므로 M3D 기반 SRAM에 올바르게 액세스하기 위해서는 캐시 주소 지정의 간단히 수정이 요구된다.
- [0053] 도 3은 본 발명의 일 실시예에 따른 캐시 주소 변환기의 구성을 도시한 도면이고, 도 4는 본 실시예에 따른 캐시 주소 변환기에 의한 캐시 주소 변환 전후 결과를 도시한 도면이다.
- [0054] 도 3을 참조하면, 본 실시예에 따른 캐시 주소 변환기는 믹스(Mux, 300), 포맷부(Format, 302) 및 풀 애더(Full Adder, 304)를 포함할 수 있다.
- [0055] 즉, 캐시 주소 변환기는 기존 주소를 모놀리식 3D 비아를 통한 3차원 레이어 배치에 따른 새로운 캐시 주소로 변환한다.
- [0056] 믹스(300)는 SHD, TEC, L1D 캐시 중 하나에 대한 코드를 출력한다.
- [0057] 여기서, SHD, TEC, L1D 코드는 각각 0000, 0100 및 1000일 수 있다.
- [0058] 여기서, SHD, TEC 코드의 최상위 비트인 0은 하위 계층인 제1 레이어(200)에 대응되는 비트이고, L1D 코드의 최상위 비트인 1은 상위 계층인 제2 레이어(202)에 대응되는 비트이다.
- [0059] 예를 들어, L1D 캐시에 대해 010이라는 캐시 주소가 입력되는 경우, 포맷부(300)는 최상위 비트에 0을 추가하여 0010으로 변환하여 출력한다.
- [0060] 상기한 캐시 주소는 L1D 캐시에 대한 주소이므로 믹스(300)는 L1D 코드인 1000을 출력하고, 풀 애더(304)는 최종적으로 1010을 출력한다.
- [0061] 상기한 구성을 통해 기존 캐시 주소가 본 실시예에 따른 M3D 기반 GPU L1 캐시 메모리 구조에 맞게 수정된다.

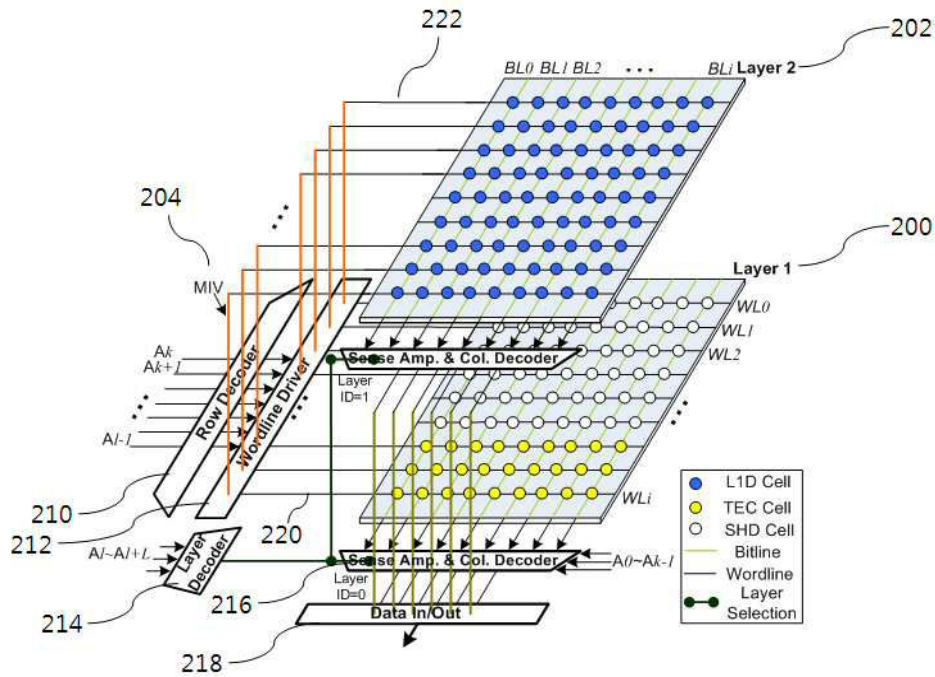
- [0062] 본 실시예에 따른 캐시 주소 변환기는 SHD, TEC 및 L1D 각각에 대해 고유 코드(0000, 0100, 1000)를 이용하여 기존 캐시 주소를 새로운 캐시 주소로 변환한다.
- [0063] 변환된 캐시 주소가 로우 디코더(210)로 입력된다.
- [0064] 새로운 캐시 주소에 기초하여 M3D 기반 SRAM의 워드라인들 및 비트라인들이 선택된다.
- [0065] 워드라인 드라이버(212)는 새로운 캐시 주소에 의해 어느 글로벌 워드라인이 선택되는지를 식별한다.
- [0066] 레이어 디코더(214)는 요구되는 비트 셀의 레이어 아이디를 식별한다.
- [0067] 전술한 바와 같이, 새로운 캐시 주소의 최상위 비트는 제1 레이어(200) 및 제2 레이어(202) 중 하나의 선택을 위해 사용된다.
- [0068] 필요한 비트 셀이 하위 계층인 제1 레이어(200)에 있다면 이들은 제1 워드라인 와이어(220)를 통해 액세스된다.
- [0069] 그렇지 않으면 MIV(204)를 통해 상위 계층인 제2 레이어(202)로 액세스한 다음 제2 레이어(202)의 제2 워드라인 와이어(222)를 통해 소정 비트 셀에 액세스한다.
- [0070] 센스 앰프 및 컬럼 디코더(216)의 컬럼 디코더는 비트라인을 통해 전달된 비트 셀로부터의 데이터를 이용하여 적절한 워드(요구된 데이터)를 선택하고, 선택된 데이터 신호는 센스 앰프에 의해 증폭된다.
- [0071] 데이터 입출력부(218)는 증폭된 신호를 출력한다.
- [0072] 상기한 본 발명의 실시예는 예시의 목적을 위해 개시된 것이고, 본 발명에 대한 통상의 지식을 가지는 당업자라면 본 발명의 사상과 범위 안에서 다양한 수정, 변경, 부가가 가능할 것이며, 이러한 수정, 변경 및 부가는 하기의 특허청구범위에 속하는 것으로 보아야 할 것이다.

도면

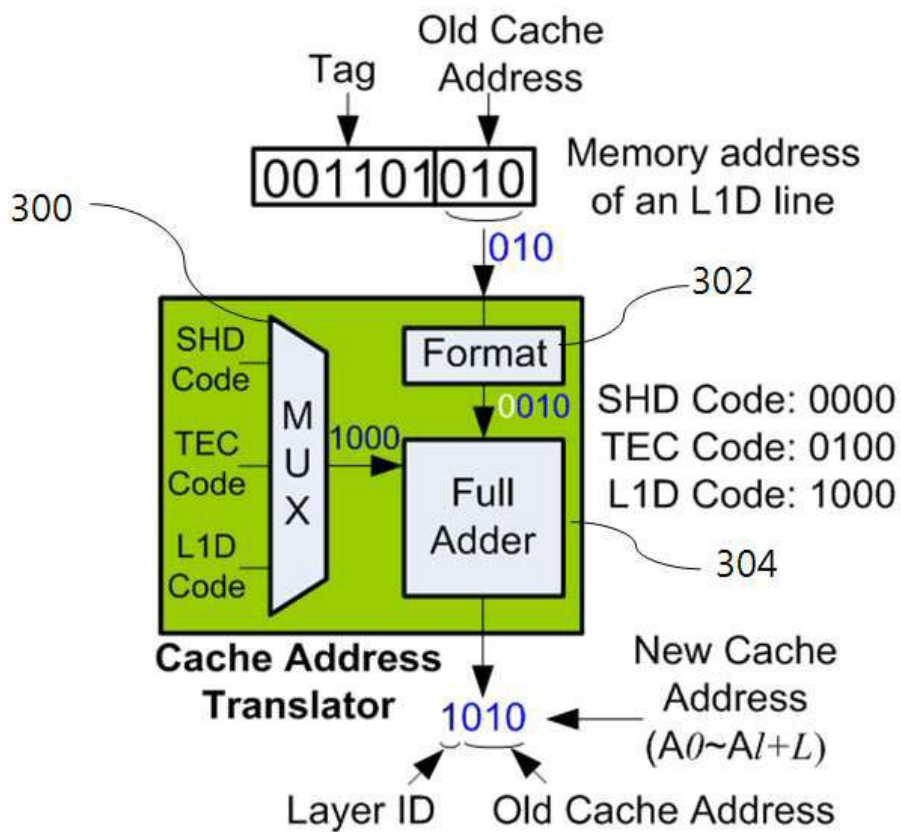
도면1



도면2



도면3



도면4

