(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2012-0059590
(43) 공개일자 2012년06월08일

(51) 국제특허분류(Int. Cl.)
G06F 9/38 (2006.01) *G06F 9/46* (2006.01)
(21) 출원번호 10-2012-7008022
(22) 출원일자(국제) 2010년09월03일
심사청구일자 없음
(85) 번역문제출일자 2012년03월28일
(86) 국제출원번호 PCT/US2010/047784
(87) 국제공개번호 WO 2011/028984
국제공개일자 2011년03월10일
(30) 우선권주장
12/616,636 2009년11월11일 미국(US)
61/239,730 2009년09월03일 미국(US)

(71) 출원인
어드밴스드 마이크로 디바이시즈, 인코포레이티드
미국 캘리포니아 94088-3453 서니베일 원 에이엠디 플레이스 메일 스톱68
(72) 발명자
사도우스키 그레그
미국 매사추세츠 02139 캠브리지 321 하바드 스트리트 #303
이오우르챠 콘스탄틴
미국 캘리포니아 95120 산 호세 우디드 레이크 드라이브 7186
브라더스 존
미국 캘리포니아 94085 서니베일 아파트먼트 1226 레이크 사이드 드라이브 1257
(74) 대리인
박장원

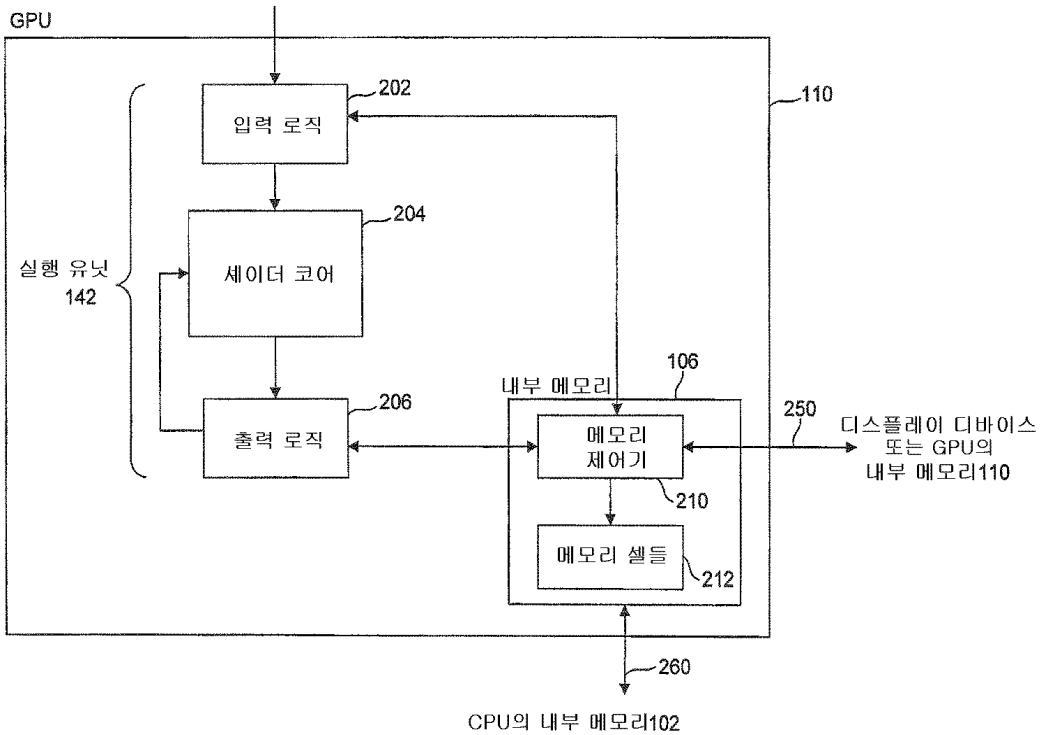
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **범용 사용을 위한 내부의, 처리-유닛 메모리**

(57) 요 약

여기에서는 범용 사용을 위한 내부 메모리를 구비한 그래픽 처리 유닛(GPU) 및 그 응용이 개시된다. 그러한 GPU는 제1 내부 메모리, 상기 제1 내부 메모리에 연결된 실행 유닛, 및 상기 제1 내부 메모리를 다른 처리 유닛의 제2 내부 메모리에 연결하도록 된 인터페이스를 포함한다. 제1 내부 메모리는 적층 동적 랜덤 액세스 메모리(DRAM) 또는 임베디드 DRAM을 포함할 수 있다. 인터페이스는 또한 디스플레이 디바이스에 제1 내부 메모리를 연결하도록 되어 있다. GPU는 또한 제1 내부 메모리를 중앙 처리 유닛에 연결하도록 되어 있다. 추가로, GPU는 소프트웨어에 구현될 수 있고 그리고/또는 컴퓨팅 시스템에 포함될 수 있다.

대 표 도



특허청구의 범위

청구항 1

그래픽 처리 유닛(GPU)에 있어서,

제1 내부 메모리와;

상기 제1 내부 메모리에 연결된 실행 유닛과; 그리고

상기 제1 내부 메모리를 다른 처리 유닛의 제2 내부 메모리에 연결하도록 된 인터페이스를 포함하는 것을 특징으로 하는 그래픽 처리 유닛.

청구항 2

제1 항에 있어서,

상기 다른 처리 유닛은 GPU를 포함하는 것을 특징으로 하는 그래픽 처리 유닛.

청구항 3

제1 항에 있어서,

상기 다른 처리 유닛은 중앙 처리 유닛을 포함하는 것을 특징으로 하는 그래픽 처리 유닛.

청구항 4

제1 항에 있어서,

상기 제1 내부 메모리는 적층(stacked) 동적 랜덤 액세스 메모리를 포함하는 것을 특징으로 하는 그래픽 처리 유닛.

청구항 5

제1 항에 잇어서,

상기 제1 내부 메모리는 임베디드(embedded) 동적 랜덤 액세스 메모리를 포함하는 것을 특징으로 하는 그래픽 처리 유닛.

청구항 6

제1 항에 있어서,

상기 인터페이스는 상기 제1 내부 메모리를 디스플레이 디바이스에 연결하도록 된 것을 특징으로 하는 그래픽 처리 유닛.

청구항 7

컴퓨팅 디바이스 상에서 실행시 그래픽-처리 유닛(GPU)을 정의(define)하는 명령들이 내포된 컴퓨터-판독가능 저장 매체를 포함하는 컴퓨터-프로그램 물(computer-program product)로서, 상기 GPU는,

제1 내부 메모리와;

상기 제1 내부 메모리에 연결된 실행 유닛과; 그리고

상기 제1 내부 메모리를 다른 처리 유닛의 제2 내부 메모리에 연결하도록 된 인터페이스를 포함하는 것을 특징으로 하는 컴퓨터-프로그램 물.

청구항 8

제7 항에 있어서,

상기 다른 처리 유닛은 GPU를 포함하는 것을 특징으로 하는 컴퓨터-프로그램 물.

청구항 9

제7 항에 있어서,

상기 다른 처리 유닛은 중앙 처리 유닛을 포함하는 것을 특징으로 하는 컴퓨터-프로그램 물.

청구항 10

제7 항에 있어서,

상기 GPU의 제1 내부 메모리는 적층(stacked) 동적 랜덤-액세스 메모리를 포함하는 것을 특징으로 하는 컴퓨터-프로그램 물.

청구항 11

제7 항에 있어서,

상기 GPU의 제1 내부 메모리는 임베디드(embedded) 동적 랜덤-액세스 메모리를 포함하는 것을 특징으로 하는 컴퓨터-프로그램 물.

청구항 12

제7 항에 있어서,

상기 GPU는 하드웨어 기술 언어 소프트웨어로 구현된 것을 특징으로 하는 컴퓨터-프로그램 물.

청구항 13

제7 항에 있어서,

상기 GPU는 베릴로그 하드웨어 기술 언어 소프트웨어, 베릴로그-A 하드웨어 기술 언어 소프트웨어, 및 VHDL 하드웨어 기술 언어 소프트웨어 중 하나로 구현된 것을 특징으로 하는 컴퓨터-프로그램 물.

청구항 14

제1 내부 메모리, 상기 제1 내부 메모리에 연결된 제1 실행 유닛, 및 상기 제1 내부 메모리를 다른 GPU의 내부 메모리에 연결하도록 된 제1 인터페이스를 포함하는 제1 그래픽 처리 유닛(GPU)과;

제2 내부 메모리, 상기 제2 내부 메모리에 연결된 제2 실행 유닛, 및 상기 제2 내부 메모리를 다른 GPU의 내부 메모리에 연결하도록 된 제2 인터페이스를 포함하는 제2 GPU를 포함하며,

여기서 상기 제1 내부 메모리와 상기 제2 내부 메모리는 서로 연결되어, 상기 제1 GPU의 제1 실행 유닛이 상기 제2 GPU의 제2 내부 메모리에 액세스할 수 있게 하고, 상기 제2 GPU의 제2 실행 유닛이 상기 제1 GPU의 제1 내부 메모리에 액세스할 수 있게 하는 것을 특징으로 하는 시스템.

청구항 15

제14 항에 있어서,

상기 제1 내부 메모리는 적층 동작 랜덤 액세스 메모리를 포함하는 것을 특징으로 하는 시스템.

청구항 16

제14 항에 있어서,

상기 제1 내부 메모리는 임베디드 동적 랜덤-액세스 메모리를 포함하는 것을 특징으로 하는 시스템.

청구항 17

제16 항에 있어서,

상기 제1 인터페이스는 상기 제1 내부 메모리를 상기 디스플레이 디바이스에 연결하도록 되어있고; 그리고

상기 제2 인터페이스는 상기 제2 내부 메모리를 상기 디스플레이 디바이스에 연결하도록 된 것을 특징으로 하는 시스템.

청구항 18

제14 항에 있어서,

외부 메모리와;

캐시 메모리를 포함하는 중앙 처리 유닛(CPU)과; 그리고

상기 외부 메모리와 CPU 사이에 연결된 버스를 포함하는 것을 특징으로 하는 시스템.

청구항 19

제18 항에 있어서,

상기 제1 GPU는 상기 제1 내부 메모리를 상기 CPU의 캐시 메모리에 연결하도록 된 것을 특징으로 하는 시스템.

청구항 20

제18 항에 있어서,

상기 제2 GPU는 상기 제2 내부 메모리를 상기 CPU의 캐시 메모리에 연결하도록 된 것을 특징으로 하는 시스템.

명 세 서

기 술 분 야

[0001] 본 발명은 일반적으로 컴퓨팅 디바이스(예를 들어, 컴퓨터, 임베디드 디바이스, 휴대형 디바이스 등)에 관한 것이다. 보다 구체적으로, 본 발명은 각각의 컴퓨팅 디바이스들의 처리 유닛들에 의해 사용되는 메모리에 관한 것이다.

배 경 기 술

[0002] 컴퓨팅 디바이스는 일반적으로 중앙-처리 유닛(CPU) 및 그래픽-처리 유닛(GPU)과 같은 하나 이상의 처리 유닛들을 포함한다. CPU는 정밀한 명령어들의 세트를 따름으로써 컴퓨팅 디바이스의 동작들을 조정한다. GPU는, 최종 사용자 애플리케이션(예를 들어, 비디오-게임 애플리케이션)에 의해 요구될 수 있는 그래픽-처리 태스크들 및/또는 물리 시뮬레이션들과 같은 데이터-병렬 컴퓨팅 태스크들을 수행함으로써 CPU를 보조한다. GPU 및 CPU는 개별 디바이스들 및/또는 패키지들의 일부이거나 동일한 디바이스 및/또는 패키지에 포함될 수 있다. 또한, 각각의 프로세싱 유닛은 또 다른 더 큰 디바이스에 포함될 수 있다. 예를 들어, GPU들은 종종 예를 들어, 노스브리지 디바이스들과 같은 라우팅 또는 브리지 디바이스들에 통합된다.

[0003] 최종-사용자 애플리케이션과 GPU 사이에는 몇개의 소프트웨어 계층들이 존재한다. 최종-사용자 애플리케이션은 애플리케이션-프로그래밍 인터페이스(API)와 통신한다. API는 최종 사용자 애플리케이션으로 하여금 그래픽 데이터 및 커맨드들을, GPU에 의존적인 포맷이 아닌, 표준화된 포맷으로 출력할 수 있게 한다. Redmond, Washington의 Microsoft Corporation에 의해 개발된 DirectX®

, Khronos Group에 의해 발표된 OpenGL®

을 포함하여 몇몇 타입의 API들이 상용화되어 있다. API는 드라이버와 통신한다. 드라이버는 API로부터 수신된 표준 코드를 GPU에 의해 이해되는 네이티브 포맷의 명령들로 변환한다. 드라이버는 일반적으로 GPU의 제조사에 의해 작성된다. GPU는 드라이버로부터의 명령들을 실행한다.

[0004] 종래의 시스템에서, CPU와 GPU는 각각 일반적으로 외부 메모리에 연결된다. 외부 메모리는 CPU 및/또는 GPU에 의해 실행될 명령들 및/또는 사용될 데이터를 포함할 수 있다. 외부 메모리는, 예를 들어, 다이나믹 랜덤-액세스 메모리(DRAM)일 수 있다. 외부 메모리는 상당히 크게 구성될 수 있으며, 그럼으로써 그것이 연결된 각각의 처리 유닛에 충분한 저장 용량을 제공할 수 있다. 불행하게도, 외부 메모리에 액세스하는 것은 수백 클럭 사이클들이 소요될 수 있다. 따라서, 외부 메모리는 메모리에 고급(high-end) GPU들을 위한 충분한 대역폭 또는 빠른 메모리 액세스를 제공할 수 없다.

[0005] GPU에 충분한 메모리 대역폭을 제공하기 위한 한가지 가능한 해법은 GPU에 내부 메모리(internal memory)를 제공하는 것이다. 내부 메모리는, 예를 들어, 임베디드 또는 적층(stacked) DRAM 일 수 있다. 외부 메모리와 비교하여, 내부 메모리는 더 높은 대역폭, 더 빠른 메모리 액세스를 제공하며, 전력을 덜 소비한다. 그러나, 내부 메모리의 용량(capacity)은 고급 GPU들을 위한 저장 요구들을 충족하도록 쉽게 스케일링될 수 없다. 예를 들어, 고급 GPU는 GPU의 내부 메모리에 포함될 수 있는 것보다 더 많은 메모리를 요구할 수 있다.

[0006] 상기를 고려하여, (외부 메모리와 같은) 충분한 메모리 용량 및 (임베디드 메모리와 같은) 높은 대역폭 둘 모두를 제공하는 메모리 및 그 응용이 필요하다.

발명의 내용

과제의 해결 수단

[0007] 본 발명의 실시예들은 내부의, 범용 사용을 위한 처리-유닛 메모리 및 그 응용들을 제공함으로써 위에서 기술된 필요들에 부응한다. 본 발명의 실시예들의 내부의 처리-유닛 메모리는, 처리 유닛에 임베딩되기 때문에 높은 대역폭을 제공한다. 내부의 처리-유닛 메모리는 또한 복수의 처리-유닛 메모리들이 충분히 큰 메모리 풀로 결합될 수 있기 때문에 충분한 저장 용량을 제공한다.

[0008] 예를 들어, 본 발명의 일 실시예는 GPU를 제공한다. GPU는 제1 내부 메모리, 상기 제1 내부 메모리에 연결된 실행 유닛, 그리고 상기 제1 내부 메모리를 또 다른 처리 유닛의 제2 내부 메모리에 연결하도록 된 인터페이스를 포함한다. 일 실시예에서, GPU는 소프트웨어에서 구현된다. 또 다른 실시예에서, GPU는 시스템 내에 포함된다. 시스템은, 예를 들어, 슈퍼컴퓨터, 데스크탑 컴퓨터, 랩탑 컴퓨터, 비디오-게임 콘솔, 임베디드 디바이스, 휴대용 디바이스(예를 들어, 모바일 텔레폰, 스마트폰, MP3 플레이어, 카메라, GPS 디바이스, 등), 또는 GPU를 포함하거나 GPU를 포함하도록 된 또 다른 시스템을 포함할 수 있다.

[0009] 본 발명의 추가적인 피쳐들 및 이점들은, 본 발명의 다양한 실시예들의 구조 및 동작과 함께, 첨부의 도면들을 참조로 하기에서 자세히 설명된다. 본 발명은 여기에 기술된 특정 실시예들로 제한된 것이 아님에 주목하여야 한다. 그러한 실시예들은 여기에서 단지 설명의 목적으로 제시된 것이다. 본 관련 기술분야의 통상의 기술자들에게는 여기에 포함된 내용에 근거하여 추가적인 실시예들이 자명할 것이다.

도면의 간단한 설명

[0010] 여기에 포함되며 본 명세서의 일부를 이루는 첨부의 도면들은 본 발명을 예시하며, 상세한 설명과 함께, 본 발명의 원리들을 추가로 설명하고 관련 기술 분야의 기술자들로 하여금 본 발명을 만들고 사용할 수 있도록 더 기능한다.

도 1a 및 1b는 본 발명의 실시예에 따라 내부의, 범용 사용을 위한 처리-유닛 메모리들을 포함하는 예시적인 시스템들을 도시한다.

도 2는 본 발명의 실시예에 따라 범용 사용을 위한 내부 메모리를 구비한 예시적인 GPU의 세부사항을 도시한다.

도 3은 본 발명의 실시예에 따라 처리 소자에 포함될 수 있는 예시적인 적층 메모리(stacked memory)를 도시한다.

도 4는 본 발명의 실시예에 따라 도 2의 GPU에 의해 구현되는 예시적인 방법을 도시한다.

본 발명의 피쳐들 및 이점들은 도면들과 함께 고려될 때, 하기에서 설명된 상세한 설명으로부터 명확해질 것이며, 도면들에서, 유사한 참조 부호들은 도면들 전체에 걸쳐 대응하는 구성요소들을 식별한다. 도면들에서, 유사한 참조 부호들은 일반적으로, 동일한, 기능적으로 유사한, 그리고/또는 구조적으로 유사한 구성요소들을 표시한다. 구성요소가 처음으로 나타나는 도면은 대응하는 참조 숫자의 최좌측 번호(들)에 의해 표시된다.

발명을 실시하기 위한 구체적인 내용

[0011] I. 개괄(Overview)

[0012] 본 발명은 범용 사용을 위한 내부의 GPU 메모리 및 그 응용들을 제공한다. 하기의 상세한 설명에서, "일 실시예", "실시예", "예시적인 실시예" 등에 대한 참조는 기술된 실시예가 특정한 피쳐, 구조, 또는 특징을 포함할 수 있음을 나타내나, 모든 실시예가 반드시 그 특정한 피쳐, 구조, 또는 특징을 포함할 필요는 없다.

또한, 그러한 표현이 반드시 동일한 실시예를 참조하는 것일 필요는 없다. 또한, 특정한 피쳐, 구조, 또는 특징이 임의의 실시예와 관련하여 기술될 때, 명시적으로 설명되든 아니든, 본 기술분야의 기술자의 지식의 범위내에서, 다른 실시예들과 관련되어 그러한 피쳐, 구조, 또는 특징에 영향을 줄 수 있는 것으로 고려된다.

[0013] 실시예에 따르면, GPU는 하나 이상의 다른 처리 유닛들에 의해 사용되도록 된 내부 메모리(예를 들어, 임베디드 또는 적층 DRAM)을 포함한다. GPU는 인터페이스를 포함하며 프로토콜을 실시하여, 하나 이상의 다른 GPU들이 그것의 내부 메모리에 액세스할 수 있게 한다. 인터페이스는 각각의 다른 GPU들에 내부 메모리에 대한 전용 액세스(dedicated access)를 제공하거나 다른 GPU들에 내부 메모리에 대한 공유된 액세스를 제공할 수 있다. GPU의 내부 메모리에 대한 액세스는 그 GPU 자체에 의해 또는 각각의 다른 GPU들에 의해 제어될 수 있다.

[0014] 실시예에서, 인터페이스 및 프로토콜은 내부 메모리를 외부 메모리들과 결합될 수 있게 하여, GPU에 의해 액세스가능한 더 큰 메모리 풀을 형성한다. 외부 메모리들은 다른 GPU들에 포함될 수 있다. 일 실시예에서, 예를 들어, 컴퓨팅 디바이스는 복수의 GPU들을 포함하며, 여기서 각각의 GPU는 다른 GPU들과 공유되도록 된 내부 메모리를 포함한다. 이 예에서, 각각의 GPU의 내부 메모리는 통합 메모리 풀로 결합된다. 메모리 풀의 사이즈는 참여 GPU들의 수로 스케일링된다. 임의의 참여 GPU는 그것의 저장 필요를 위해 메모리 풀을 사용할 수 있다.

[0015] 본 발명의 실시예에 따른 예시적인 GPU의 추가의 세부사항들이 하기에 기술된다. 그러나, 이 세부사항들을 제공하기 전에, 그러한 GPU들이 구현될 수 있는 예시적인 컴퓨팅 디바이스를 설명하는 것이 유용할 것이다.

[0016] II. 예시적인 컴퓨팅 디바이스

[0017] 도 1a 및 1b는 복수의 GPU들을 구비한 예시적인 컴퓨팅 시스템(100)을 도시하며, 여기서, 본 발명의 실시예들에 따라, 각각의 GPU가 범용 사용을 위해 구성된 내부 메모리를 포함한다. 외부 메모리들에 비하여, 내부 메모리들은 각각의 GPU들에 데이터에 대한 더 높은 대역폭의 액세스를 제공한다. 또한, 각각의 GPU의 내부 메모리들은 각각의 GPU에 의해 액세스가능한 더 큰 메모리 풀로 결합될 수 있으며, 그럼으로써 각각의 GPU에 충분한 저장 용량을 제공한다.

[0018] 도 1a의 실시예에서, 각각의 GPU에는 또 다른 GPU의 내부 메모리에 대한 전용 액세스가 주어진다. 도 1b의 실시예에서, 각각의 GPU는 공유 인터페이스를 통해 다른 GPU의 내부 메모리들에 대한 공유된 액세스를 가진다. 실시예들에서, 컴퓨팅 시스템(100)은 슈퍼컴퓨터, 데스크탑 컴퓨터, 랩탑 컴퓨터, 비디오 게임 콘솔, 임베디드 디바이스, 휴대형 디바이스(예를 들어, 모바일 전화기, 스마트폰, MP3 플레이어, 카메라, GPS 디바이스 등), 또는 GPU 및/또는 GPU를 포함하거나 포함하도록 된 어떤 다른 디바이스를 포함할 수 있다.

[0019] 도 1a 및 1b를 참조하면, 컴퓨팅 디바이스(100)는 CPU(102), 제1 GPU(110a), 및 제2 GPU(110b)를 포함한다. CPU(102)는 컴퓨팅 디바이스(100)의 기능을 제어하기 위한 명령어들을 실행한다. GPU들(110)은 데이터-병렬 처리 태스크들(예를 들어, 그래픽-처리 태스크들 및/또는 일반-컴퓨팅 태스크들)을 수행함으로써 CPU(102)를 보조한다. GPU들(110)의 설계에 근거하여, GPU들(110)은 일반적으로 CPU가 소프트웨어에서 수행할 수 있는 것보다 빠르게 데이터-병렬 처리 태스크들을 수행할 수 있다.

[0020] 제1 GPU(110A) 및 제2 GPU(110B) 각각은 그들 고유의 내부 메모리 및 실행 유닛을 포함한다. 구체적으로, 제1 GPU(106A)는 내부 메모리(106A) 및 실행 유닛(142A)을 포함하고, 제2 GPU(106B)는 내부 메모리(106B) 및 실행 유닛(142B)을 포함한다. 마찬가지로, CPU(102)는 캐시 메모리(130) 및 실행 유닛(132)을 포함한다. 내부 메모리(106)(및 선택적으로 캐시 메모리(130))는 특정 데이터에 대해,상기 데이터가 외부에 저장된 경우(예를 들어, 데이터가 시스템 메모리(104)에 저장된 경우)에 가능했을 것보다 더 빠른 액세스 및 더 높은 대역폭을 제공하도록 GPU들(110)에 의해 사용가능하다. 내부 메모리들(106)은, 예를 들어, 임베디드 또는 적층 DRAM을 포함할 수 있다.

[0021] 내부 메모리들(106A, 106B)(및 선택적으로 캐시 메모리(130))은 더 큰 메모리 풀로 결합되어, 빠르고 높은 대역폭의 메모리 액세스를 제공함과 아울러, 상당한 저장 용량(예를 들어, 4GB 보다 큰 저장 용량)을 제공할 수 있다. 비록 종래의 외부 메모리들이 (예를 들어, 4GB 보다 큰) 충분한 저장 용량을 제공할 수 있으나, 종래의 외부 메모리들은 특정한 고급 사용에 대해 불충분한 대역폭을 제공한다. 마찬가지로, 종래의 임베디드 메모리들이 이러한 고급 사용들에 대해 충분한 대역폭을 제공할 수 있으나, 종래의 임베디드 메모리들은 이 고급 사용들에 대해 불충분한 저장 용량(예를 들어, 4GB 미만의 저장 용량)을 제공한다. 종래의 외부 메모리들 및/또는 종래의 임베디드 메모리들과는 달리, 본 발명의 실시예들은 (예를 들어, 4GB 보다 큰) 충분한 저장 용량을

제공할 뿐만아니라, 또한 범용 사용을 위해 다른 GPU들에 의해 사용가능한 내부 메모리들을 포함하는 GPU들을 제공함으로써 높은 대역폭을 제공한다.

[0022] 예를 들어, 고급 GPU의 프레임 버퍼(즉, 디스플레이 디바이스 상에 디스플레이될 데이터의 완전한 프레임(complete frame)을 저장하는 버퍼)는 실질적으로 큰 메모리(예를 들어, 4 기가바이트(GB)보다 큰 메모리)에 대한 고 대역폭 액세스를 요구할 수 있다. 실시예들에서, 제1 GPU(110A)는 제1 GPU(110A)의 프레임 버퍼를 정의하기 위하여 내부 메모리들(106A,B) 및 선택적으로 CPU(102)의 캐시 메모리(130)를 사용할 수 있다. 마찬가지로, 제2 GPU(110B)는 또한 제2 GPU(110B)의 프레임 버퍼를 정의하기 위하여 내부 메모리들(106A,B) 및 선택적으로 CPU(102)의 캐시 메모리(130)를 사용할 수 있다. 이러한 방식으로, 종래의 외부 메모리 또는 임베디드 메모리와 다르게, 본 발명의 실시예들에 따라 정의된 프레임 버퍼는 실질적으로 큰 메모리(예를 들어, 4GB보다 큰 메모리)에 대해 고 대역폭 액세스를 제공한다.

[0023] 도 1a의 실시예에서, 각각의 GPU(110)에는, 위에서 암시된 바와 같이, 또 다른 처리 유닛의 내부 메모리(106)에 대한 전용 액세스가 주어진다. 구체적으로, 제1 인터페이스(101)는 제1 GPU(110A)에 제2 GPU(110B)의 내부 메모리(106B)에 대한 전용 액세스를 제공하고 제2 GPU(110B)에 제1 GPU(110A)의 내부 메모리(106A)에 대한 전용 액세스를 제공한다. 데이터는 그 데이터의 어드레스 범위에 근거하여 내부 메모리(106A) 또는 내부 메모리(106B)에 기록될 수 있다(또는 내부 메모리(106A) 또는 내부 메모리(106B)로부터 검색될 수 있다). 예를 들어, 내부 메모리(106A)에는 (예를 들어, 제1 소정의 어드레스 A보다 작고 제2 소정의 어드레스 B 이상인) 제1 어드레스 범위가 할당될 수 있고, 그리고 내부 메모리(106B)에는 제2 어드레스 범위(예를 들어, 제1 어드레스 범위 내에 있지 않은 모든 어드레스)가 할당될 수 있다. 그러나, 제1 GPU(110A) 및 제2 GPU(110B)가 각각 제1 GPU(110A)의 내부 메모리(106A) 및 제2 GPU(110B)의 내부 메모리(106B)에 대한 액세스를 가질 수 있다면, 본 발명의 범주 및 정신으로부터 벗어남이 없이 내부 메모리(106A) 및/또는 내부 메모리(106B)에 데이터를 기록하고 그리고 내부 메모리(106A) 및/또는 내부 메모리(106B)로부터 데이터를 검색하기 위한 다른 기법들이 구현될 수 있다.

[0024] 실시예에서, 제1 인터페이스(101)는 디스플레이 제어기 인터페이스를 포함한다. 디스플레이 제어기 인터페이스는 디스플레이 디바이스(140)에 GPU의 프레임 버퍼에 대한 액세스를 제공한다. 디스플레이 제어기 인터페이스를 제1 인터페이스(101)에 포함시킴으로써, 제1 인터페이스(101)는 종래의 GPU 디자인에 이미 포함된 표준 핀(standard pin) 상에 제공될 수 있다.

[0025] 제1 인터페이스(101)에 부가하여, 제2 인터페이스(103)는 CPU(102)에 제2 GPU(110B)의 내부 메모리(106B)에 대한 전용 액세스를 제공하고 제2 GPU(110B)에 CPU(102)의 캐시 메모리(130)에 대한 전용 액세스를 제공한다. 이러한 식으로, 제2 GPU(110B) 및 CPU(102)는 각각 제2 GPU(110B)의 내부 메모리(106B) 및 CPU(102)의 캐시 메모리(130)에 대한 액세스를 가질 수 있다. 마찬가지로, 제3 인터페이스(105)는 제1 GPU(110A)에 CPU(102)의 캐시 메모리(130)에 대한 전용 액세스를 제공하고 CPU(102)에 제1 GPU(110A)의 내부 메모리(106A)에 대한 전용 액세스를 제공한다. 이러한 식으로, CPU(102)의 제1 GPU(110A)는 각각 제1 GPU(110A)의 내부 메모리 및 CPU(102)의 캐시 메모리(130)에 대한 액세스를 가질 수 있다.

[0026] 도 1b의 실시예에서, 각각의 처리 유닛은 공유 인터페이스(164)를 통해 다른 처리 유닛들의 내부 메모리들에 대한 공유된 액세스를 가진다. 공유 인터페이스(164)는 각각의 처리 유닛(예를 들어, 제1 GPU(110A), 제2 GPU(110B), 및 CPU(102))에 다른 처리 유닛들의 내부 메모리에 대한 고 대역폭 액세스를 제공한다. 데이터는 상기 데이터의 어드레스 범위에 근거하여 내부 메모리(106A), 내부 메모리(106B), 또는 캐시 메모리(130)에 기록되거나 내부 메모리(106A), 내부 메모리(106B), 또는 캐시 메모리(130)로부터 검색될 수 있다. 예를 들어, 내부 메모리(106A)에는 제1 어드레스 범위가 할당될 수 있고, 내부 메모리(106B)에는 제2 어드레스 범위가 할당될 수 있고, 그리고 캐시 메모리(130)에는 제3 어드레스 범위가 할당될 수 있다. 그러나, 제1 GPU(110A), 제2 GPU(110B), 및 CPU(102)는 각각 제1 GPU(110A)의 내부 메모리(106A), 제2 GPU(110B)의 내부 메모리(106B), 및 CPU(102)의 캐시 메모리(130)에 대한 액세스를 가질 수 있다면, 본 발명의 정신 및 범주로부터 벗어남이 없이 데이터를 내부 메모리(106A), 내부 메모리(106B), 및/또는 캐시 메모리(130)에 기록하고 내부 메모리(106A), 내부 메모리(106B), 및/또는 캐시 메모리(130)로부터 검색하기 위한 다른 기법들이 구현될 수 있음이 이해될 것이다.

[0027] 실시예들에서, 컴퓨팅 디바이스(100)는 또한 시스템 메모리(104), 제2 메모리(120), 입출력(I/O) 인터페이스(116), 및/또는 디스플레이 디바이스(140)를 포함한다. 시스템 메모리(104)는 CPU(102) 상에서 구동되는 프로그램들에 의해 빈번하게 액세스되는 정보를 저장한다. 시스템 메모리(104)는 일반적으로 휘발성 메모리를 포함하는바, 이는 컴퓨팅 디바이스(100)의 파워가 턴오프될 때 시스템(104)에 저장된 데이터가 손실된다는 것을

의미한다. 제2 메모리(120)는 컴퓨팅 디바이스(00)에 의해 사용된 데이터 및/또는 애플리케이션들을 저장한다. 제2 메모리(120)는 일반적으로 시스템 메모리(104)에 비해 더 큰 저장 용량을 가지며 일반적으로 비휘발성(영구) 메모리를 포함하는바, 이는 제2 메모리(120)에 저장된 데이터가 컴퓨팅 디바이스(100)의 파워가 턴오프될 때 조차도 유지됨을 의미한다. I/O 인터페이스(116)는 외부 디바이스(116)(외부 디스플레이 디바이스, 외부 저장 디바이스(예를 들어, 비디오-게임 카트리지, CD, DVD, 플래시 드라이브 등), 네트워크 카드, 또는 어떤 다른 타입의 외부 디바이스)에 연결될 수 있다. 디스플레이 디바이스(140)는 컴퓨팅 디바이스(100)의 내용(content)을 디스플레이한다. 디스플레이 디바이스는, 캐쏘드 레이 튜브, 액정 디스플레이(LCD), 플라즈마 스크린, 또는 현재 알려져 있거나 나중에 개발될 어떤 다른 타입의 디스플레이 디바이스를 포함할 수 있다.

[0028] GPU(110)와 CPU(102)는 버스(1145)를 통해 서로, 그리고 시스템 메모리(104), 제2 메모리(120), 및 I/O 인터페이스(116)와 통신한다. 버스(114)는 컴퓨팅 디바이스들에서 사용되는 임의의 타입의 버스일 수 있으며, 주변 컴포넌트 인터페이스(PCI) 버스, 가속 그래픽 포트(AGP) 버스, PCI 익스프레스(PCIE) 버스, 또는 현재 사용가능하거나 나중에 개발될 또 다른 타입의 버스를 포함한다.

[0029] 실시예들에서, 컴퓨팅 디바이스(100)는 GPU(110) 대신에 또는 GPU(110)에 부가하여, 비디오 처리 유닛(VPU)을 포함할 수 있다. 예를 들어, 실시예에서, 컴퓨팅 디바이스(100)는, GPU(110A), CPU(102)를 포함하고, 그리고도 1a 및 1b에 도시된 GPU(110B) 대신에, 컴퓨팅 디바이스(100)는 VPU를 포함한다. 이러한 식으로, CPU(102)는 일반적인 처리 기능들(general processing functions)을 수행할 수 있고, GPU(110A)는 그래픽 처리 기능들을 수행할 수 있고, 그리고 VPU는 비디오 처리 기능들을 수행할 수 있다.

[0030] III. 예시적인 GPU

[0031] 도 2는 내부 메모리(106)를 가진 GPU(110)의 예시적인 세부사항들을 도시한다. 본 발명의 실시예에 따라, 내부 메모리(106)는, 증가된 메모리 풋프린트 사이즈를 토대로 그래픽 처리력(graphic processing power)을 결합함으로써 전체적인 시스템 성능을 증가시키기 위하여 또 다른 GPU 또는 CPU에 의해 사용될 수 있다.

[0032] 위에서 언급된 바와 같이, GPU(110)는 실행 유닛(142) 및 내부 메모리(106)를 포함한다. 도 2를 참조로, 실행 유닛(142)은 입력 로직(202), 셰이더 코어(204), 및 출력 로직(206)을 포함한다. 내부 메모리(106)는 메모리 제어기(210) 및 메모리 셀들(212)을 포함한다. 메모리 제어기(210)는 메모리 셀들에 대한 액세스를 제어한다. 메모리 셀들(212)은 데이터를 저장한다.

[0033] 실시예에서, 내부 메모리(106)는 임베디드, 동적 랜덤 액세스 메모리(DRAM)를 포함한다. 임베디드 DRAM은 처리 유닛과 함께 공통 패키지에 캡슐화된 메모리이다. 또다른 실시예에서, 내부 메모리(106)는 도 3에 도시된 것과 같은 적층(stacked) DRAM을 포함한다. 적층 메모리는 3차원 구조로 서로의 상부에 적층된 복수의 메모리 소자들을 포함한다.

[0034] 내부 메모리(106)는 입력 로직(202)과 출력 로직(206) 둘 모두를 통해 실행 유닛(142)에 연결된다. 특히, 입력 로직(20)은 내부 메모리(106)로부터 데이터를 검색할 수 있고, 출력 로직(206)은 메모리 셀들(212)에 저장될 내부 메모리(106)에 데이터를 송신할 수 있다.

[0035] 내부 메모리(106)는 또한 제1 인터페이스(250)를 통해 또 다른 GPU의 내부 메모리에 연결될 수 있다. 또 다른 GPU의 내부 메모리에 내부 메모리(106)를 연결하는 것은 실행 유닛(142)에서 사용가능한 전체 메모리 풀을 증가시킬 수 있다. 실시예에서, 도 1a의 인터페이스(101)에 의해 도시된 바와 같이, 제1 인터페이스(250)는 GPU(110)의 내부 메모리(106)와 또 다른 GPU의 내부 메모리 사이의 전용 액세스를 제공한다. 이 실시예에서, 제1 인터페이스(250)는 종래의 GPU의 표준 핀 상에 제공된다. 예를 들어, 제1 인터페이스(250)는 디스플레이-제어기 인터페이스를 포함할 수 있으며, 이는 내부 메모리(106)에 포함된 로컬 프레임 버퍼에 디스플레이 디바이스 액세스를 제공한다. 또 다른 실시예에서, 도 1b의 인터페이스(164)에 의해 도시된 바와 같이, 제1 인터페이스(250)는 GPU(110)의 내부 메모리(106)와 다른 처리 유닛들의 내부 메모리들 사이에 공유된 액세스를 제공한다.

[0036] 내부 메모리(106)는 또한 제2 인터페이스(260)를 통해 CPU(102)의 캐시 메모리(130)에 연결될 수 있다. 따라서, 내부 메모리(106)와 캐시 메모리(130)의 조합은 GPU(110)가 사용할 수 있는 메모리 풀을 증가시킬 수 있다. 실시예에서, 제2 인터페이스(260)는 도 1a의 연결(103) 또는 연결(105)과 같은, GPU(110)의 내부 메모리(106)와 CPU(102)의 캐시 메모리(130) 사이의 전용 연결을 제공한다. 또 다른 실시예에서, 제2 인터페이스

(260)는 오직 GPU(110)와 CPU(102)에 의해 공유되는 연결, 예컨대 도 1b의 연결(164)을 제공한다. 또 다른 실시예에서, 제2 인터페이스는 도 1a 및 1b의 버스(114)와 같은 공통 버스 상에서 GPU(110)를 CPU(102)에 연결한다.

[0037] IV. GPU(110)의 예시적인 동작

[0038] 도 4는 본 발명의 실시예에 따라 GPU(110)에 의해 구현되는 예시적인 방법(400)을 도시한다. 방법(400)은 도 3 및 4를 참조로 하기에서 설명된다.

[0039] 방법(400)은 명령이 수신되는 단계(402)에서 시작된다. 실시예에서, 입력 로직(202)은 GPU(110)에 의해 실행될 명령들을 수신한다. 명령들은, 예를 들어, 시스템(100)의 CPU(102) 상에서 구동되는 최종 사용자 애플리케이션에 의해 제공되는 그래픽-처리 태스크 또는 데이터-병렬 처리 태스크를 포함할 수 있다.

[0040] 단계(404)에서, 명령과 관련된 데이터의 위치가 식별된다. 일 예에서, 데이터는 수신된 명령에 포함될 수 있다. 그러한 데이터는 흔히 즉시 데이터(immediate data)라고 칭해진다. 또 다른 예에서, 명령은 데이터의 위치를 제공한다. 예를 들어, 명령은 내부에 데이터가 저장된 어드레스를 포함할 수 있다. 또 다른 예에서, 명령은 정보를 포함하는바, 입력 로직(202)은 데이터가 저장된 어드레스를 상기 정보로부터 계산한다. 데이터는 내부 메모리(106), 내부 메모리(106)가 연결된 또 다른 GPU의 내부 메모리, 또는 CPU(102)의 캐시 메모리(130)에 저장될 수 있다.

[0041] 단계(406)에서, 데이터가 검색된다. 데이터가 즉시 데이터이면, 입력 로직(202)은 명령으로부터 그 즉시 데이터를 단순히 추출한다. 데이터가 내부 메모리(106)에 저장되거나 또는 내부 메모리(106)가 연결된 메모리에 저장되면, 입력 로직(202)은 데이터에 대한 액세스를 위해 메모리 제어기(210)에 요청을 송신한다. 반면에, 데이터가 메모리 셀들(212)에 저장되면, 데이터는 검색되고 입력 로직(202)에 제공된다. 반면, 데이터가 내부 메모리(106)에 연결된 또 다른 메모리에 저장되면, 입력 로직(202)으로부터의 요청이 인터페이스(250) 또는 인터페이스(260)를 통해 다른 메모리로 포워딩된다. 데이터는 그후 또 다른 메모리로부터 검색되고 입력 로직(202)에 제공된다.

[0042] 단계(408)에서, 명령이 실행된다. 셰이더 코어(204)는 단계(406)에서 입력 로직(202)에 의해 얻어진 데이터에 근거하여 명령을 실행한다.

[0043] 단계(410)에서, 명령 실행의 결과들이 출력 로직(206)에 제공된다. 출력 로직(206)은, 결정 단계(412)에 표시된 바와 같이, 이 결과들에 대해 추가의 처리가 요구되는지 여부를 결정한다. 출력 로직(206)에 제공되는 결과들은 추가적인 처리가 필요한지 여부를 표시하기 위한 플래그 또는 어떤 다른 표시(indicia)를 가질 수 있다. 결정 단계(412)에서 출력 로직(206)이 추가의 처리가 필요하다고 결정하면, 출력 로직(206)은 셰이더 코어(205)에 결과들을 다시 포워딩하고 방법(400)의 단계(408) 및 단계(410)이 반복된다. 다른 한편으로, 결정 단계(412)에서, 출력 로직(206)이 어떠한 추가의 처리도 필요하지 않음을 결정하면, 단계(414)에 표시된 바와 같이, 출력 로직(206)은 결과들을 내부 메모리(106)에 제공한다.

[0044] 결과들이 기록될 어드레스에 따라, 결과들은 내부 메모리(106)에, 또는 내부 메모리(106)에 연결된 메모리에 기록될 수 있다. 결과들이 내부 메모리(106)에 기록될 것이라면, 메모리 제어기(210)는 메모리 셀들(212)의 적절한 어드레스에 액세스를 제공하고, 결과들이 그곳에 저장된다. 반면, 결과들이 내부 메모리(106)에 연결된 메모리에 기록될 것이라면, 메모리 제어기(210)는 인터페이스(250) 또는 인터페이스(260)를 통해 다른 메모리에 결과들을 포워딩하고 결과들이 다른 메모리의 메모리 셀들에 저장된다.

[0045] V. 예시적인 소프트웨어 구현

[0046] GPU(110)의 하드웨어 구현에 부가하여, 그러한 GPU들은 또한, 예를 들어, 소프트웨어(예를 들어, 컴퓨터-판독가능 프로그램 코드)를 저장하도록 된 컴퓨터-판독가능 매체에 배치된 소프트웨어에 수록될 수 있다. 컴퓨터-판독가능 프로그램 코드는 다음의 실시예들, 즉, (i) 여기에 기술된 시스템 및 기법의 기능(예를 들어, GPU(1100)에 태스크들을 제공, GPU(110)에서 태스크들을 스케쥴링, GPU(110)에서 태스크들을 실행, 등); (ii) 여기에 개시된 시스템 및 기법들의 제작(예를 들어, GPU(110)의 제작); 또는 (iii) 여기에 개시된 시스템 및 기법의 기능과 제작의 결합을 포함하는 본 발명의 실시예들을 가능하게 한다.

[0047] 이는, 예를 들어, 일반적인 프로그래밍 언어들(예를 들어, C 또는 C++), 베릴로그 HDL, VHDL, 알테라

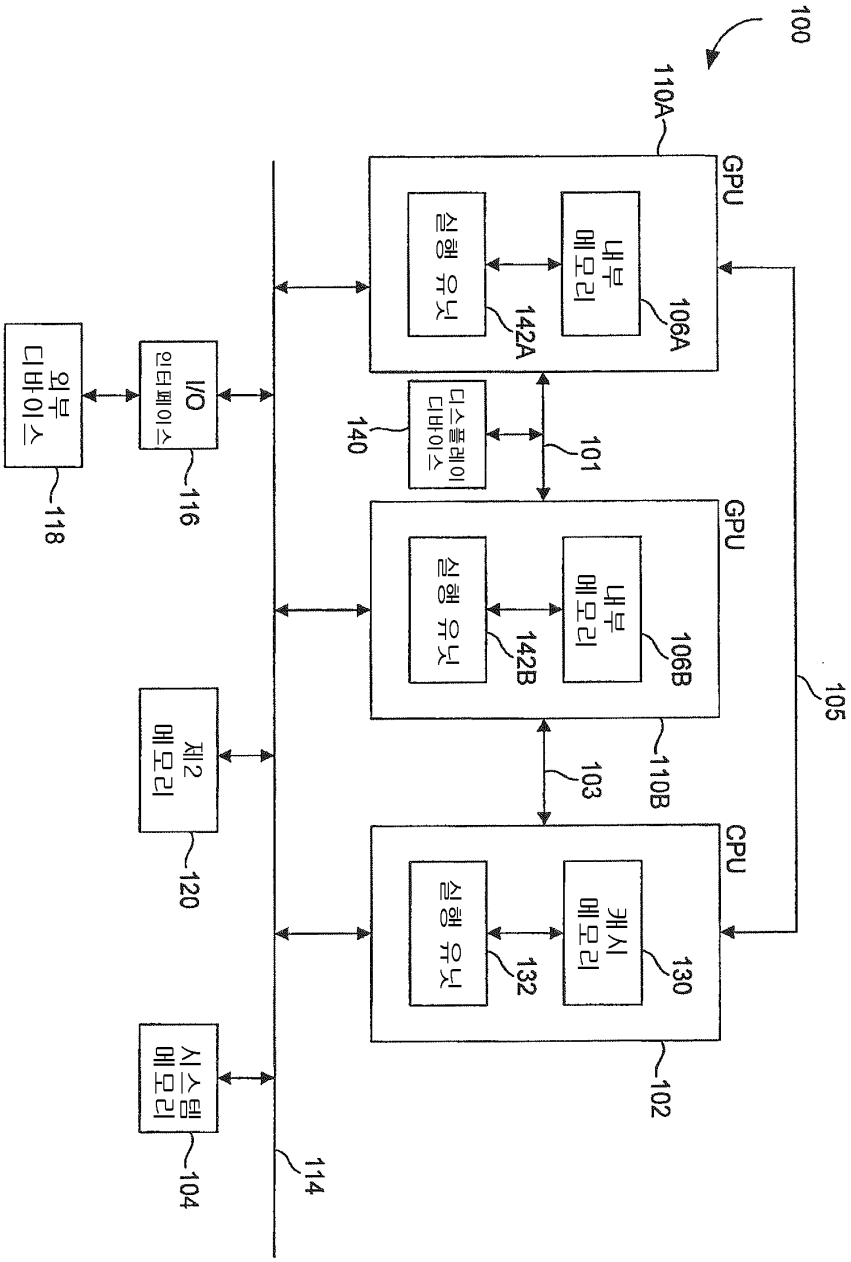
HDL(AHDL) 등의 하드웨어 기술 언어들(HDL), 또는 다른 사용가능한 프로그래밍 및/또는 스케매틱 캡쳐 툴들(예를 들어, 회로-캡쳐 툴들)의 사용을 통해 달성될 수 있다. 컴퓨터-판독가능 프로그램 코드는 반도체, 자기 디스크, 또는 광학 디스크(예를 들어, CD-ROM, DVD-ROM)를 포함하는 임의의 알려진 컴퓨터-판독가능 매체에 배치될 수 있다. 따라서, 컴퓨터-판독가능 프로그램 코드는 인터넷(Internet and internets)을 포함하는 통신 네트워크들을 통해 전송될 수 있다. 위에서 기술된 시스템 및 기법들에 의해 제공되는 구조 및/또는 달성된 기능들은 컴퓨터-판독가능 프로그램 코드에서 구현되는 코어(예를 들어, 셰이더 코어)에 나타날 수 있고 집적 회로 제품의 일부로서 하드웨어로 변환될 수 있다.

[0048] VI. 결론

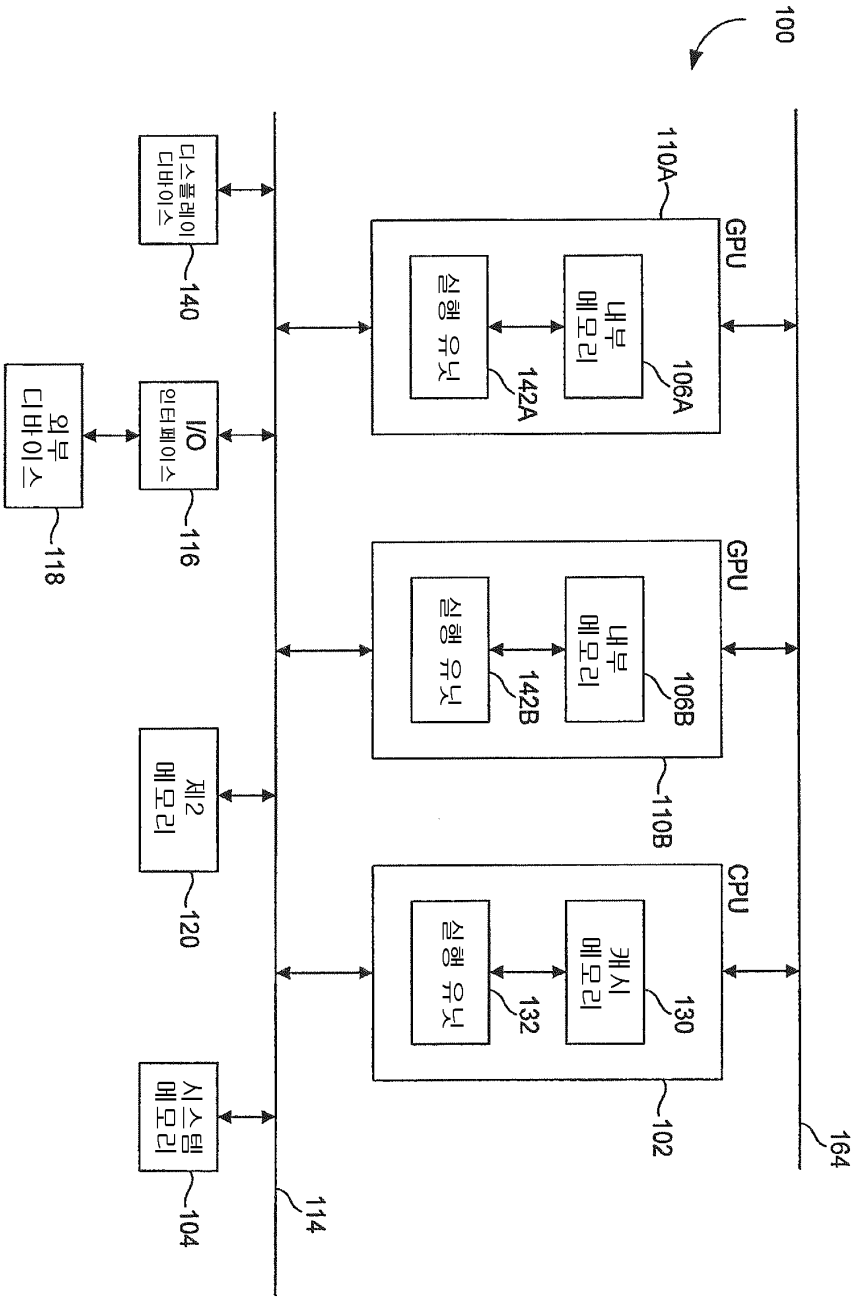
[0049] 위에서 설명된 것은 내부의, 범용 사용을 위한 GPU 메모리, 및 그 응용이다. 개요 및 요약 부분이 아닌 상세한 설명 부분은 청구항들을 해석하는데 사용되도록 의도된 것임이 이해되어야 한다. 개요 및 요약 부분들은 발명자(들)에 의해 고려되는 본 발명의 모든 예시적 실시예들이 아닌 단지 한 두가지의 실시예들을 제시하는 것으로서, 본 발명 및 첨부된 청구항들을 어떠한 식으로든 제한하려 의도된 것이 아니다.

도면

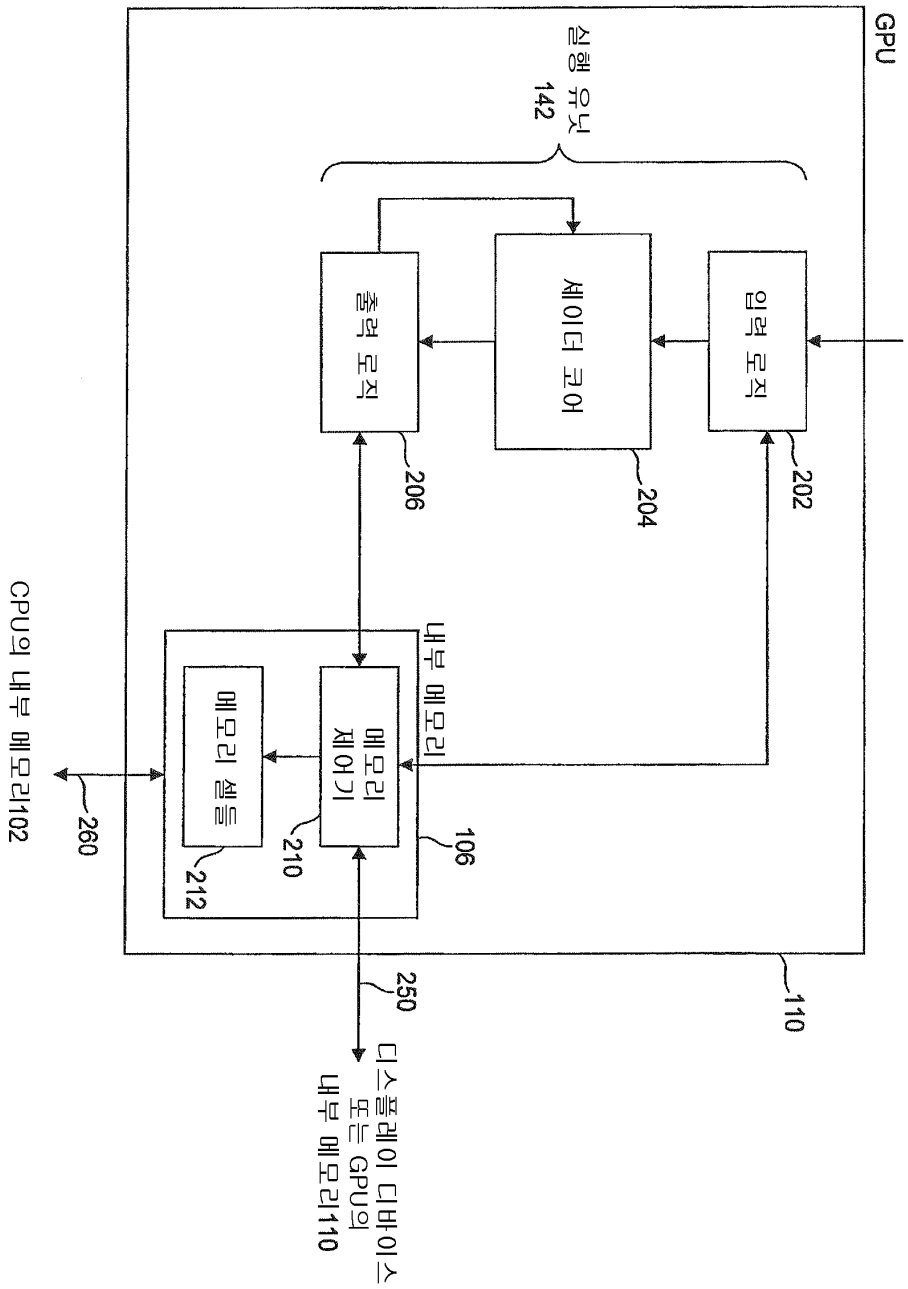
도면1a


100
110A
GPU
실행 유닛
142A
내부 메모리
106A
디스플레이 디바이스
140
101
GPU
실행 유닛
142B
내부 메모리
106B
110B
105
103
CPU
실행 유닛
132
캐시 메모리
130
102
I/O 인터페이스
116
외부 디바이스
118
제2 메모리
120
시스템 메모리
104
114

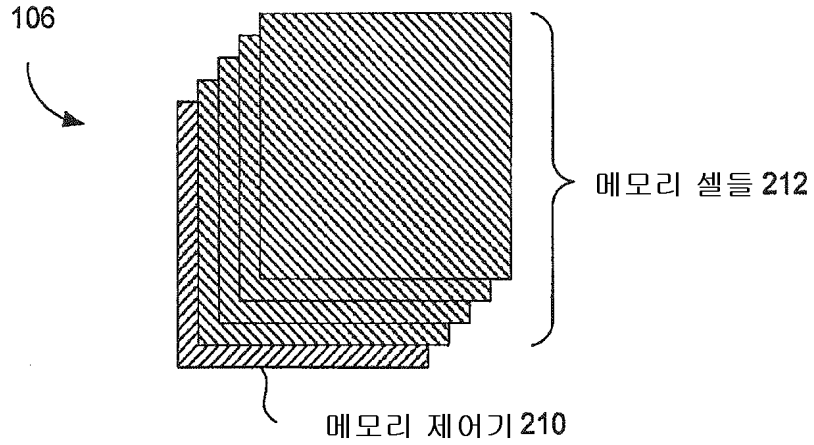
도면1b


100
110A
GPU
실행 유닛
142A
내부 메모리
106A
110B
GPU
실행 유닛
142B
내부 메모리
106B
102
CPU
실행 유닛
132
캐시 메모리
130
164
114
디스플레이 디바이스
140
I/O 인터페이스
116
외부 디바이스
118
제2 메모리
120
시스템 메모리
104

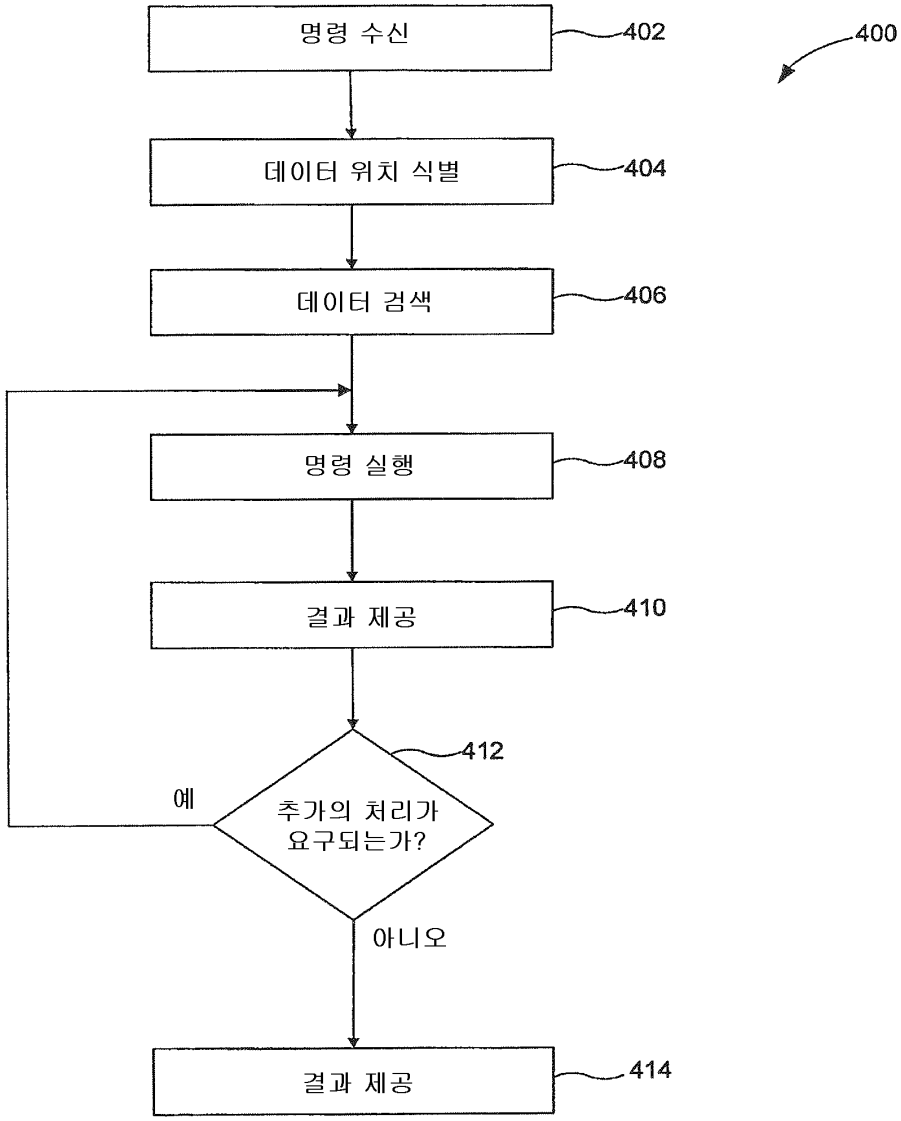
도면2


GPU
실행 유닛
142
출력 로직
셰이더 코어
입력 로직
206
204
202
내부 메모리
메모리 셀들
메모리 제어기
212
210
106
260
CPU의 내부 메모리102
250
110
디스플레이 디바이스
또는 GPU의
내부 메모리110

도면3


106
메모리 셀들 212
메모리 제어기 210

도면4


400
명령 수신
402
데이터 위치 식별
404
데이터 검색
406
명령 실행
408
결과 제공
410
412
추가의 처리가 요구되는가?
예
아니오
결과 제공
414