



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2015년11월11일

(11) 등록번호 10-1568204

(24) 등록일자 2015년11월05일

(51) 국제특허분류(Int. Cl.)

G06F 9/50 (2006.01)

(21) 출원번호 10-2014-0089014

(22) 출원일자 2014년07월15일

심사청구일자 2014년07월15일

(56) 선행기술조사문헌

KR1020130122657 A

KR1020130036213 A

JP2012016043 A

Hong, S 외 2명, "Efficient Parallel Graph Exploration on Multi-Core CPU and GPU", 2011년 10월, Parallel Architectures and Compilation Techniques (PACT) 국제 컨퍼런스, pp. 78-88, DOI: 10.1109/PACT.2011.14

(73) 특허권자

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

최동휘

경기도 시흥시 정왕대로27번길 9, 208동 601 (정왕동, 신동아아파트)

정용화

대전광역시 유성구 송림로53번길 71-29, B동 404호 (하기동, 브리젠힐스)

(뒷면에 계속)

(74) 대리인

특허법인엠에이피에스

전체 청구항 수 : 총 7 항

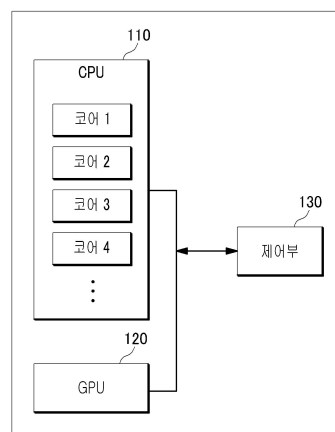
심사관 : 박미정

(54) 발명의 명칭 CPU/GPU 기반 매니코어 시스템 및 CPU/GPU의 동시처리를 위한 멀티코어 CPU의 코어 개수 결정 방법

(57) 요약

본 발명에 따른 CPU/GPU 기반 매니코어 시스템은 복수의 코어를 포함하는 CPU; GPU 및 상기 CPU와 GPU를 제어하는 제어부를 포함하되, 상기 제어부는, 상기 CPU에서 GPU로의 데이터 복사 시간에 기초하여 상기 CPU에 포함된 코어 중 작업을 수행할 코어를 선택하고, 상기 선택된 코어의 개수에 따라, 대응되는 코어에 작업을 할당하되, 상기 데이터 복사 시간이 기 설정된 임계값보다 작은 경우 상기 CPU에 포함된 전체 코어에 작업을 할당하고, 상기 기 설정된 임계값보다 큰 경우 상기 CPU에 포함된 전체 코어 중 상기 GPU로 데이터를 복사하기 위한 코어를 제외한 나머지 코어에 작업을 할당한다.

대표도 - 도1



100

(72) 발명자

김희곤

인천광역시 서구 청라에메랄드로 30, 114동 403호
(연희동, 청라자이)

이성주

서울특별시 마포구 월드컵북로22길 20, 302호 (성
산동, 한맥아이홈)

박대희

서울특별시 강남구 언주로30길 56, A동 1601호 (도
곡동, 타워팰리스)

이 발명을 지원한 국가연구개발사업

과제고유번호 1345214994

부처명 교육부

연구관리전문기관 한국연구재단

연구사업명 BK21플러스사업(0.5)

연구과제명 과학벨트 연계형 융합 SW 고급 인재 양성 사업팀

기 여 율 1/1

주관기관 고려대학교 세종캠퍼스

연구기간 2013.09.01 ~ 2014.02.28

명세서

청구범위

청구항 1

CPU/GPU 기반 매니코어 시스템에 있어서,

복수의 코어를 포함하는 CPU;

GPU 및

상기 CPU와 GPU를 제어하는 제어부를 포함하되,

상기 제어부는,

상기 CPU에서 GPU로의 데이터 복사 시간에 기초하여 상기 CPU에 포함된 코어 중 작업을 수행할 코어를 선택하고,

상기 선택된 코어의 개수에 따라, 대응되는 코어에 작업을 할당하되, 상기 데이터 복사 시간이 기 설정된 임계값보다 작은 경우 상기 CPU에 포함된 전체 코어에 작업을 할당하고, 상기 기 설정된 임계값보다 큰 경우 상기 CPU에 포함된 전체 코어 중 상기 GPU로 데이터를 복사하기 위한 코어를 제외한 나머지 코어에 작업을 할당하는 것인 매니코어 시스템.

청구항 2

제 1 항에 있어서,

상기 기 설정된 임계값은 상기 코어의 수행시간에 기초하여 설정되는 것인 매니코어 시스템.

청구항 3

제 1 항에 있어서,

상기 제어부는, 상기 CPU에서 처리하기 위한 전체 작업량을 상기 CPU에 포함된 코어의 개수인 N 으로 나눈 값과 상기 데이터 복사 시간의 합이, 상기 전체 작업량을 상기 GPU로 데이터를 복사하기 위한 코어를 제외한 나머지 코어의 개수인 $N-1$ 로 나눈 값보다 큰 경우, $N-1$ 개의 코어를 작업을 수행할 코어로 선택하는 매니코어 시스템.

청구항 4

제 1 항에 있어서,

상기 데이터 복사 시간은 비디오 스트리밍 서비스에 제공되는 영상의 크기에 기초하여 결정되는 것인 매니코어 시스템.

청구항 5

CPU/GPU 기반 매니코어 시스템에 의해 수행되는 CPU/GPU의 동시처리를 위한 멀티코어 CPU의 코어 개수 결정 방법에 있어서,

상기 매니코어 시스템의 제어부가, CPU에서 GPU로의 데이터 복사 시간에 기초하여 작업을 수행할 코어를 선택하는 단계 및

상기 제어부가 상기 선택된 코어의 개수에 따라, 대응되는 코어에 작업을 할당하는 단계를 포함하되,

상기 작업을 수행할 코어를 선택하는 단계는,

상기 데이터 복사 시간이 기 설정된 임계값보다 작은 경우 상기 CPU에 포함된 전체 코어를 작업을 수행할 코어로 선택하고,

상기 기 설정된 임계값보다 큰 경우 상기 CPU에 포함된 코어 중 상기 GPU로 데이터를 복사하기 위한 코어를 제

외한 나머지 코어를 작업을 수행할 코어로 선택하는 것인 코어 개수 결정 방법.

청구항 6

제 5 항에 있어서,

상기 기 설정된 임계값은 상기 코어의 수행시간에 기초하여 설정되는 것인 코어 개수 결정 방법.

청구항 7

제 5 항에 있어서,

상기 작업을 수행할 코어를 선택하는 단계는,

상기 멀티코어 CPU가 처리하기 위한 전체 작업량을 상기 멀티코어 CPU에 포함된 코어 개수인 N 으로 나눈 값과 상기 데이터 복사 시간의 합이, 상기 전체 작업량을 상기 GPU로 데이터를 복사하기 위한 코어를 제외한 나머지 코어의 개수인 $N-1$ 로 나눈 값보다 큰 경우, $N-1$ 개의 코어를 작업을 수행할 코어로 선택하는 것인 코어 개수 결정 방법.

발명의 설명

기술 분야

[0001] 본 발명은 CPU/GPU 기반 매니코어 시스템 및 CPU/GPU의 동시처리를 위한 멀티코어 CPU의 코어 개수 결정 방법에 관한 것이다.

배경 기술

[0002] 최근 컴퓨터 구조의 발달로 인해 일반적인 PC나 모바일 컴퓨팅 환경도 멀티코어 CPU/GPU 구조를 가지고 있다. 이에 따라, 멀티코어 CPU 또는 GPU를 활용하여 속도를 개선하고자 하는 병렬 처리 기법에 대한 연구가 중요한 이슈로 떠오르고 있으며, 이와 더불어 클라우드 컴퓨팅이나 빅데이터 처리를 위한 분산처리 시스템 분야에 있어서 여러 종류로 이루어진(heterogeneous) 컴퓨팅 환경이 발달함에 따라 CPU 및 GPU를 동시에 활용하는 연구가 활발히 진행되고 있다.

[0003] 한편, CPU와 GPU를 동시에 활용할 경우, GPU 가속기는 데이터 복사시간이 큰 오버헤드로 작용하고, CPU와 GPU를 동시에 처리할 때에는 CPU에서 GPU로 데이터를 복사해야 하며, 이러한 데이터 복사시간은 상대적으로 작은 태스크를 병렬화할 때 매우 큰 병렬 오버헤드로 작용한다는 특징을 가지고 있다.

[0004] 따라서, 멀티코어 CPU와 GPU를 동시에 이용할 경우, 태스크를 적절하게 나누어주고, 이에 대한 태스크를 처리하기 위해 CPU의 모든 코어를 이용하는 것이 일반적이다. 그러나, 태스크의 크기가 작은 경우에는 데이터 복사시간이 태스크를 처리하기 위한 시간보다 상대적으로 길다. 특히, 비디오 스트리밍의 경우, 비디오 영상의 크기에 따라 데이터 복사 비율이 달라지게 되며, 이에 따라 태스크의 크기를 고려하지 않고, 모든 코어를 이용할 경우 작업 수행시간이 길어지게 되는 문제가 있다.

[0005] 이와 같은 오브젝트 태스킹은, 하나의 작업량은 작지만 스트리밍 서비스에 적용할 경우에는 많은 작업량을 갖게 되므로, 수행시간을 단축시킬 필요가 있다. 또한, 오브젝트 태스킹은 전 영상의 특징을 참조하기 때문에 각 프레임에서 데이터 종속성을 갖게 되므로, 반드시 작은 작업량을 갖는 태스크 레벨에서 병렬화를 해야 한다. 이와 같이 작은 태스크에 대하여, CPU 및 GPU를 동시에 활용할 경우, 효율적으로 병렬 처리하는 방법이 필요하다.

[0006] 이와 관련하여, 한국공개특허공보 제2012-0091550호(발명의 명칭: 범용 그래픽 처리장치의 병렬 프로세서를 이용한 고속 영상 처리 장치 및 방법)는 GPU가 탑재된 범용 비디오카드를 이용하여 병렬 처리함으로써 대용량의 데이터 변환에 소요되는 연산속도를 향상시키기 위한 기술을 개시하고 있다.

발명의 내용

해결하려는 과제

[0007] 본 발명은 전술한 종래 기술의 문제점을 해결하기 위한 것으로서, 본 발명의 일부 실시예는 CPU와 GPU를 동시에 활용하여 작은 태스크를 효율적으로 병렬처리하기 위한 CPU/GPU 기반 매니코어 시스템 및 CPU/GPU의 동시처리를

위한 멀티코어 CPU의 코어 개수 결정 방법을 제공하는 것을 그 목적으로 한다.

과제의 해결 수단

[0008] 상술한 기술적 과제를 달성하기 위한 기술적 수단으로서, 본 발명의 제 1 측면에 따른 CPU/GPU 기반 매니코어 시스템은 복수의 코어를 포함하는 CPU; GPU 및 상기 CPU와 GPU를 제어하는 제어부를 포함하되, 상기 제어부는, 상기 CPU에서 GPU로의 데이터 복사 시간에 기초하여 상기 CPU에 포함된 코어 중 작업을 수행할 코어를 선택하고, 상기 선택된 코어의 개수에 따라, 대응되는 코어에 작업을 할당하되, 상기 데이터 복사 시간이 기 설정된 임계값보다 작은 경우 상기 CPU에 포함된 전체 코어에 작업을 할당하고, 상기 기 설정된 임계값보다 큰 경우 상기 CPU에 포함된 전체 코어 중 상기 GPU로 데이터를 복사하기 위한 코어를 제외한 나머지 코어에 작업을 할당한다.

[0009] 또한, 본 발명의 제 2 측면에 따른 CPU/GPU의 동시처리를 위한 멀티코어 CPU의 코어 개수 결정 방법은 CPU에서 GPU로의 데이터 복사 시간에 기초하여 작업을 수행할 코어를 선택하는 단계 및 상기 선택된 코어의 개수에 따라, 대응되는 코어에 작업을 할당하는 단계를 포함하되, 상기 작업을 수행할 코어를 선택하는 단계는, 상기 데이터 복사 시간이 기 설정된 임계값보다 작은 경우 상기 CPU에 포함된 전체 코어를 작업을 수행할 코어로 선택하고, 상기 기 설정된 임계값보다 큰 경우 상기 CPU에 포함된 코어 중 상기 GPU로 데이터를 복사하기 위한 코어를 제외한 나머지 코어를 작업을 수행할 코어로 선택한다.

발명의 효과

[0010] 전술한 본 발명의 과제 해결 수단의 어느 실시예에 의하면, GPU로 데이터를 복사하는데 소요되는 시간에 기초하여 작업을 수행할 CPU의 코어 개수를 결정함으로써 CPU의 작업 처리 속도를 향상시킬 수 있다.

도면의 간단한 설명

[0011] 도 1은 본 발명의 일 실시예에 따른 매니코어 시스템의 블록도이다.
 도 2는 데이터 복사시간에 따라 코어의 개수를 결정하는 방법의 일 예시를 도시한 도면이다.
 도 3은 본 발명의 일 실시예에 따른 CPU/GPU의 동시처리를 위한 멀티코어 CPU의 코어 개수 결정 방법의 순서도이다.
 도 4는 이미지 프레임 크기에 따른 수행시간을 나타낸 그래프이다.
 도 5는 CPU와 GPU의 동시처리 수행시 속도 향상 정도를 비교한 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0012] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0013] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "전기적으로 연결"되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0014] 도 1은 본 발명의 일 실시예에 따른 CPU/GPU 기반 매니코어 시스템(100)의 블록도이다.

[0015] 본 발명의 일 실시예에 따른 매니코어 시스템(100)은 CPU(110), GPU(120) 및 CPU(110)와 GPU(120)를 제어하는 제어부(130)를 포함한다. 이때, CPU(110)는 작업을 수행하기 위한 복수의 코어를 포함하고 있다.

[0016] 참고로, 본 발명의 실시예에 따른 도 1에 도시된 구성 요소들은 소프트웨어 또는 FPGA(Field Programmable Gate Array) 또는 ASIC(Application Specific Integrated Circuit)와 같은 하드웨어 구성 요소를 의미하며, 소정의 역할들을 수행한다.

[0017] 그렇지만 '구성 요소들'은 소프트웨어 또는 하드웨어에 한정되는 의미는 아니며, 각 구성 요소는 어드레싱할 수

있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다.

[0018] 따라서, 일 예로서 구성 요소는 소프트웨어 구성 요소들, 객체지향 소프트웨어 구성 요소들, 클래스 구성 요소들 및 태스크 구성 요소들과 같은 구성 요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 및 변수들을 포함한다.

[0019] 구성 요소들과 해당 구성 요소들 안에서 제공되는 기능은 더 작은 수의 구성 요소들로 결합되거나 추가적인 구성 요소들로 더 분리될 수 있다.

[0020] 제어부(130)는 CPU(110)에서 GPU(120)로의 데이터 복사 시간에 기초하여 CPU(110)에 포함된 코어 중 작업을 수행할 코어를 선택한다. 이때, 데이터 복사 시간은 비디오 스트리밍 서비스에 제공되는 영상의 크기에 기초하여 결정될 수 있다.

[0021] 제어부(130)는, 선택된 코어의 개수에 따라, 대응되는 코어에 작업을 할당한다. 이때, 데이터 복사 시간이 기 설정된 임계값보다 작은 경우, CPU(110)에 포함된 전체 코어에 작업을 할당한다.

[0022] 반대로, 데이터 복사 시간이 기 설정된 임계값보다 큰 경우, CPU(110)에 포함된 전체 코어 중 GPU(120)로 데이터를 복사하기 위한 코어를 제외한 나머지 코어에 작업을 할당한다.

[0023] 이때, 기 설정된 임계값은 코어의 수행시간에 기초하여 설정될 수 있다. 예를 들어, GPU(120)로 데이터를 복사하기 위한 코어에서 처리하는 일반 작업 수행시간을 임계값으로 설정하여 데이터 복사 시간과 비교할 수 있다. 비교 결과, 데이터 복사 시간이 일반 작업 수행시간보다 작은 경우, CPU의 수행 속도에 거의 영향을 미치지 않으므로, 전체 코어에 작업을 할당한다.

[0024] 제어부(130)는 데이터 복사시간의 비율에 따라 CPU(110)의 N개의 코어를 모두 이용할지, N-1개의 코어를 이용할지 결정해야 하며, 하기의 수학적 식 1을 이용하여 이를 결정할 수 있다.

[0025] [수학적 식 1]

$$D(n) = \begin{cases} \text{using } n - 1 \text{ cores} : \frac{\text{workload}}{n} + \text{data copy time} > \frac{\text{workload}}{n-1} \\ \text{using } n \text{ cores} : \text{else} \end{cases}$$

[0026]

[0027] 즉, 제어부(130)는 위 수학적 식, 즉 판별식을 통해 CPU(110)에서 처리하기 위한 전체 작업량을 CPU(110)에 포함된 코어의 개수인 N으로 나눈 값과 데이터 복사시간의 합이, 전체 작업량을 GPU(120)로 데이터를 복사하기 위한 코어를 제외한 나머지 코어의 개수인 N-1로 나눈 값보다 큰 경우, N-1개의 코어를 작업을 수행할 코어로 선택할 수 있다. 이하에서는 도 2를 참조하여, 코어의 개수를 선택하는 방법에 대하여 설명하도록 한다.

[0028] 도 2는 데이터 복사시간에 따라 코어의 개수를 결정하는 방법의 일 예시를 도시한 도면이다.

[0029] 도 2를 참조하면, 수행시간은 데이터 복사시간(CPU-GPU data copy)에 따라 민감하게 영향을 받는 것을 확인할 수 있다. 예를 들어, 데이터 복사시간이 긴 경우, 도 2의 (a)보다 도 2의 (b)가 더 빠른 수행시간을 얻을 수 있다. 즉, 데이터 복사시간이 긴 경우, 도 2의 (a)처럼 CPU(110)의 코어 4개 각각에 동일한 작업량을 할당하면, CPU(110)의 수행시간은 ‘코어 4’가 수행하는 시간만큼 오래 걸리게 된다. 따라서, 도 2의 (b)의 경우처럼 ‘코어 4’에는 작업을 할당하지 않고, ‘코어 1’ 내지 ‘코어 3’에 ‘코어 4’가 수행할 작업을 분배할 경우, CPU(110)의 전체 수행시간을 단축할 수 있게 된다.

[0030] 반면, 데이터 복사시간이 짧은 경우, 도 2의 (c)가 도 2의 (d)보다 더 빠른 수행시간을 얻을 수 있다. 즉, 데이터 복사 시간이 짧은 경우, 데이터를 복사하기 위한 코어에 작업을 할당하여도 전체 수행시간에 큰 영향을 미치지 않게 되므로 도 2의 (c)와 같이 전체 코어에 작업을 분배함으로써, 전체 수행시간을 단축할 수 있게 된다.

[0031] 한편, 작업을 수행할 코어를 선택하면, CPU(110) 및 GPU(120)에 최적의 작업량을 할당함으로써 전체 수행시간을 단축할 수 있는바, 이하에서 구체적으로 설명하도록 한다.

[0032] 먼저, 하나의 작업에 대해 CPU(110)와 GPU(120)에 할당하는 작업량을 W_{CPU} 및 W_{GPU} 라 할 때, 전체 작업량(W_{TOTAL})을 아래의 수학적 식 2와 같이 나타낼 수 있다.

[수학식 2]

$$W_{TOTAL} = W_{CPU} + W_{GPU}$$

이때, W_{CPU} 및 W_{GPU} 는 전체 작업량 대비 차지하는 비율로 표현될 수 있다. 예를 들어, 전체 작업량(W_{TOTAL})을 100%라 할 경우, W_{CPU} 및 W_{GPU} 는 각각 80% 및 20%로 표현될 수 있다.

다음으로, CPU(110)와 GPU(120)의 동시처리 수행시간(T_{TOTAL})은 수학식 3과 같이 나타낼 수 있다.

[수학식 3]

$$T_{TOTAL}(W_{CPU}, W_{GPU}) = \max[T_{CPU}(W_{CPU}), T_{GPU}(W_{GPU})]$$

이때, $T_{CPU}(W)$ 는 CPU(110)의 수행시간을 나타내고, $T_{GPU}(W)$ 는 GPU(120)의 수행시간을 나타낸다. CPU(110)와 GPU(120)의 동시처리 수행시간(T_{TOTAL})은 산출된 CPU(110)의 수행시간($T_{CPU}(W_{CPU})$) 및 산출된 GPU(120)의 수행시간($T_{GPU}(W_{GPU})$) 중 더 긴 시간으로 결정된다.

한편, CPU(110)의 수행시간($T_{CPU}(W_{CPU})$)은 CPU(110)의 순차처리 수행시간($T_{SEQ}(W_{CPU})$)에 기초하여 산출된다. 그리고 GPU(120)의 수행시간($T_{GPU}(W_{GPU})$)은 GPU(120)로 데이터를 복사하는 코어의 순차처리 수행시간($T_{SEQ}(W_{GPU})$) 및 GPU(120)의 데이터 복사시간($CP(W_{GPU})$)에 기초하여 산출된다.

이때, CPU(110)의 수행시간($T_{CPU}(W_{CPU})$) 및 GPU(120)의 수행시간($T_{GPU}(W_{GPU})$)은 암달의 법칙에 의해 수학식 4 및 수학식 5와 같이 나타낼 수 있다. 한편, 제어부(130)는 CPU(110)의 순차처리 대비 상대적인 속도향상 정도(ES_{CPU} , ES_{GPU})를 더 고려하여 CPU(110)의 수행시간($T_{CPU}(W_{CPU})$) 및 GPU(120)의 수행시간($T_{GPU}(W_{GPU})$)을 산출할 수 있다. 수학식 4 및 5에서 ES_{CPU} 및 ES_{GPU} 는 CPU(110)와 GPU(120) 각각의 CPU(110) 순차처리 대비 속도향상 정도를 의미한다.

[수학식 4]

$$T_{CPU}(W_{CPU}) = \frac{T_{SEQ}(W_{CPU})}{ES_{CPU}}$$

[수학식 5]

$$T_{GPU}(W_{GPU}) = \frac{T_{SEQ}(W_{GPU})}{ES_{GPU}} + CP(W_{GPU})$$

CPU(110)의 순차처리 수행시간($T_{SEQ}(W)$)은 선형 보간법을 이용하여 수학식 6과 같이 나타낼 수 있으며, GPU(120)의 데이터 복사시간($CP(W)$)은 수학식 7과 같이 나타낼 수 있다.

[수학식 6]

$$T_{SEQ}(W) = \alpha W + \beta$$

[수학식 7]

$$CP(W_{GPU}) = HtoD(W_{GPU}) + DtoH(W'_{GPU}) \approx W_{GPU} + \delta$$

이때, GPU(120)의 데이터 복사시간($CP(W)$)은 CPU(110)에서 GPU(120)로의 데이터 전송시간($HtoD(W_{GPU})$) 및

GPU(120)에서 CPU(110)로의 데이터 전송시간($D_{toH}(W'_{GPU})$)에 기초하여 산출될 수 있다. 그리고, GPU(120)로의 데이터 전송시간 및 CPU(110)로의 데이터 전송시간은 각각 GPU(120)의 입력 데이터의 작업량(W_{GPU}) 및 출력 데이터의 작업량(W'_{GPU})에 기초하여 산출될 수 있다.

한편, 각각의 파라미터인 $\alpha, \beta, \gamma, \delta$ 는 실험적으로 얻을 수 있으며, 이때, 각 파라미터는 상수에 해당한다. 이와 같이, 실험에 의해 구해진 파라미터를 통하여 CPU(110) 및 GPU(120)의 동시활용을 위한 최적의 작업량을 계산할 수 있다. 각각의 파라미터에 대한 일 예시로 아래의 표와 같은 수치를 얻을 수 있다.

[표]

ES_{CPU}	ES_{GPU}	α	β	γ	δ
2.4	40.0	1.55×10^{-8}	1.05×10^{-3}	9.87×10^{-10}	5.61×10^{-4}

다음으로, 제어부(130)는 동시처리 수행시간(T_{TOTAL})이 최소가 되는 때의 작업량을 CPU(110) 및 GPU(120)에 각각 할당한다. 이때, 동시처리 수행시간이 최소가 되는 때는 CPU(110)와 GPU(120)의 수행시간이 같을 때를 의미하며, 이를 아래의 수학적 식 8과 같이 나타낼 수 있다.

[수학적 식 8]

$$\frac{\alpha W_{TOTAL} - \alpha W_{GPU} + \beta}{ES_{CPU}} - \left(\frac{\alpha W_{GPU} + \beta}{ES_{GPU}} + \gamma W_{GPU} + \delta \right) = 0$$

수학적 식 8을 통해 GPU(120)의 최적의 작업량 OW_{GPU} 와 CPU(110)의 최적의 작업량 OW_{CPU} 를 구할 수 있다.

$$OW_{GPU} = \frac{\alpha ES_{GPU} W_{TOTAL} + \beta (ES_{GPU} - ES_{CPU}) - \delta ES_{CPU} ES_{GPU}}{(\alpha (ES_{GPU} + ES_{CPU}) + \gamma ES_{GPU} ES_{CPU})}$$

$$OW_{CPU} = W_{TOTAL} - OW_{GPU}$$

최적의 작업량을 산출하여 CPU(110) 및 GPU(120)에 할당할 경우, CPU(110) 또는 GPU(120)의 단일 병렬화 방법에 비해 수행시간을 약 50%정도 단축시킬 수 있다는 효과를 기대할 수 있다.

도 3은 본 발명의 일 실시예에 따른 CPU/GPU의 동시처리를 위한 멀티코어 CPU의 코어 개수 결정 방법의 순서도이다.

본 발명의 일 실시예에 따른 CPU/GPU의 동시처리를 위한 멀티코어 CPU의 코어 개수 결정 방법은 먼저, CPU(110)에서 GPU(120)로의 데이터 복사 시간에 기초하여 작업을 수행할 코어를 선택한다(S110).

다음으로, 데이터 복사시간에 기초하여 작업을 수행할 코어를 선택하게 된다(S120). 즉, 데이터 복사시간이 기 설정된 임계값보다 작은 경우, CPU(110)에 포함된 전체 코어를 작업을 수행할 코어로 선택한다. 반면, 데이터 복사시간이 기 설정된 임계값보다 큰 경우 CPU(110)에 포함된 코어 중 GPU(120)로 데이터를 복사하기 위한 코어를 제외한 나머지 코어를 작업을 수행할 코어로 선택한다. 한편, 기 설정된 임계값은 코어의 수행시간에 기초하여 설정될 수 있으며, 데이터 복사시간은 비디오 스트리밍 서비스에 제공되는 영상의 크기에 기초하여 결정될 수 있다.

구체적으로, 작업을 수행할 코어를 선택할 때, 전체 코어를 선택하거나 전체 코어에서 1개 적은 코어를 선택하여 작업을 수행하게 할 수 있다. 즉, 멀티코어 CPU(110)가 처리하기 위한 전체 작업량을 CPU(110)에 포함된 코어 개수인 N으로 나눈 값과 데이터 복사 시간의 합이, 전체 작업량을 GPU(120)로 데이터를 복사하기 위한 코어를 제외한 나머지 코어의 개수인 N-1로 나눈 값보다 큰 경우, 1개 더 적은 N-1개의 코어를 작업을 수행할 코어로 선택할 수 있다.

[0066] 이와 같은 방법에 의해 선택된 코어의 개수에 따라, 대응되는 코어에 작업을 할당한다.

[0067] 이하에서는 도 4 및 도 5를 참조하여, 본 발명에 따른 CPU/GPU 기반 매니코어 시스템(100) 및 CPU/GPU의 동시처리를 위한 멀티코어 CPU의 코어 개수 결정 방법에서의 수행시간 및 속도 향상 정도를 설명하도록 한다.

[0068] 도 4는 이미지 프레임 크기에 따른 수행시간을 나타낸 그래프이며, 도 5는 CPU(110)와 GPU(120)의 동시처리 수행시 속도 향상 정도를 비교한 그래프이다.

[0069] CPU(110)와 GPU(120)의 동시 처리에 따른 수행시간 및 속도 향상 정도를 알아보기 위해 4개의 코어를 가지고 있는 인텔 코어 i5-3570 CPU와 GeForce GTX 660 GPU를 이용하였다. 그리고 3000×2000 , 1440×1080 , 1280×720 , 800×600 , 720×480 및 640×480 의 이미지 크기를 갖는 120개의 이미지 프레임을 이용하였다. 또한, CPU(110)와 GPU(120)의 병렬적 특성을 추출하기 위하여 OpenMP와 OpenCL을 이용하였다.

[0070] 본 발명에 따른 특성을 예측하기 위하여, 수행시간의 예상치(Estimated)와 측정치(Measured) 비교하여 도 4에 도시하였다. 도 4를 참조하면, CPU(110)와 GPU(120)의 최적의 작업량 비율에 따른 수행시간의 예상치가 측정치와 근접함을 확인할 수 있다. CPU(110)와 GPU(120)의 최적의 작업량 비율은 3000×2000 크기의 이미지 프레임에서는 80%:20%이고, 640×480 크기의 이미지 프레임에서는 75%:25%임을 확인할 수 있다.

[0071] 이와 같이 상대적으로 작은 크기의 이미지 프레임에서의 GPU(120)의 최적의 작업량 크기는 큰 이미지 프레임에서보다 증가하게 되며, 이는 GPU(120)의 데이터 복사 시간에 따른 오버헤드에 따른 영향을 감소시켜 더 향상된 성능을 제공하기 위함이다.

[0072] 도 5를 참조하면, GPU(120)는 데이터 복사 시간에 따른 오버헤드를 가지고 있기 때문에, 다양한 이미지 프레임 크기에서 GPU(120)만을 사용할 경우(GPU-Only) CPU(110)만을 사용한 경우(CPU-Only)보다는 더 나은 성능을 가져올 수 있다. 반면, 본 발명에 따른 CPU/GPU 기반 매니코어 시스템(100) 및 CPU/GPU의 동시처리를 위한 멀티코어 CPU의 코어 개수 결정 방법은 CPU(110)와 GPU(120)를 동시에 사용함으로써(CPU/GPU(Proposed)), GPU(120)만을 사용한 경우보다 상대적으로 118%(3000×2000 이미지 프레임 크기의 경우)의 향상된 성능을 나타내며, CPU(110)만을 사용한 경우보다 상대적으로 141%(640×480 이미지 프레임 크기의 경우) 향상된 성능을 보여줌을 확인할 수 있다.

[0073] 본 발명의 일 실시예는 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행 가능한 명령어를 포함하는 기록 매체의 형태로도 구현될 수 있다. 컴퓨터 판독 가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 및 비분리형 매체를 모두 포함한다. 또한, 컴퓨터 판독가능 매체는 컴퓨터 저장 매체 및 통신 매체를 모두 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함한다. 통신 매체는 전형적으로 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈, 또는 반송파와 같은 변조된 데이터 신호의 기타 데이터, 또는 기타 전송 매커니즘을 포함하며, 임의의 정보 전달 매체를 포함한다.

[0074] 본 발명의 방법 및 시스템은 특정 실시예와 관련하여 설명되었지만, 그것들의 구성 요소 또는 동작의 일부 또는 전부는 범용 하드웨어 아키텍처를 갖는 컴퓨터 시스템을 사용하여 구현될 수 있다.

[0075] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

[0076] 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

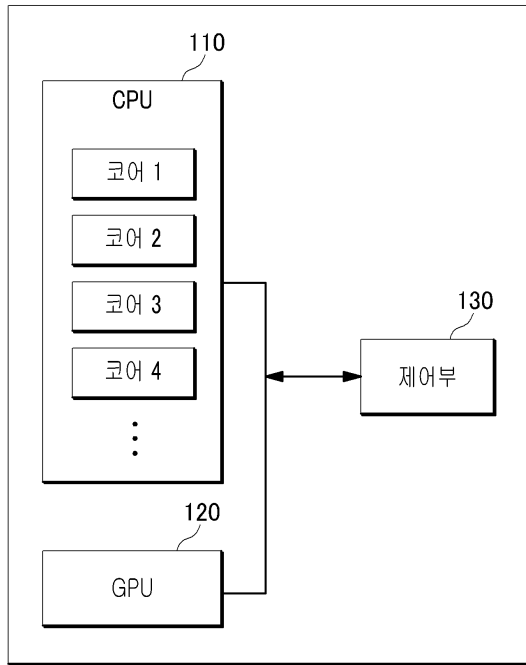
부호의 설명

[0077] 100: 매니코어 시스템 110: CPU

 120: GPU 130: 제어부

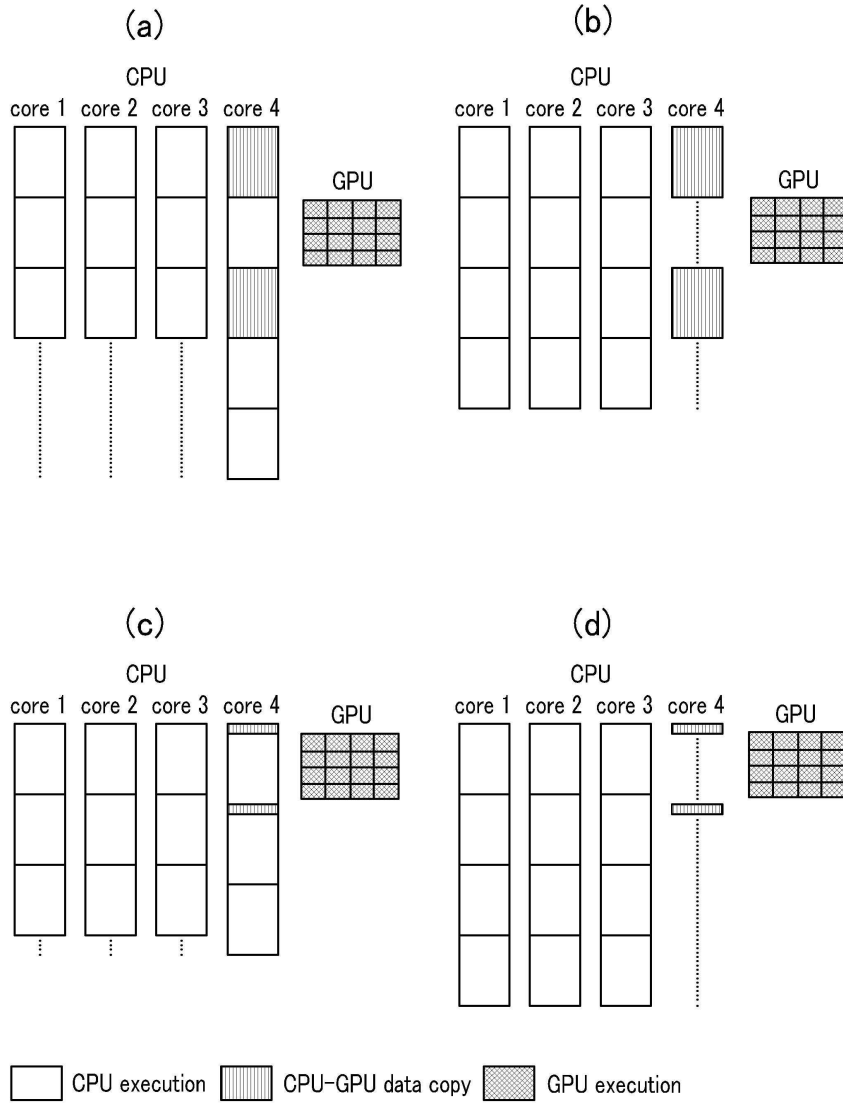
도면

도면1

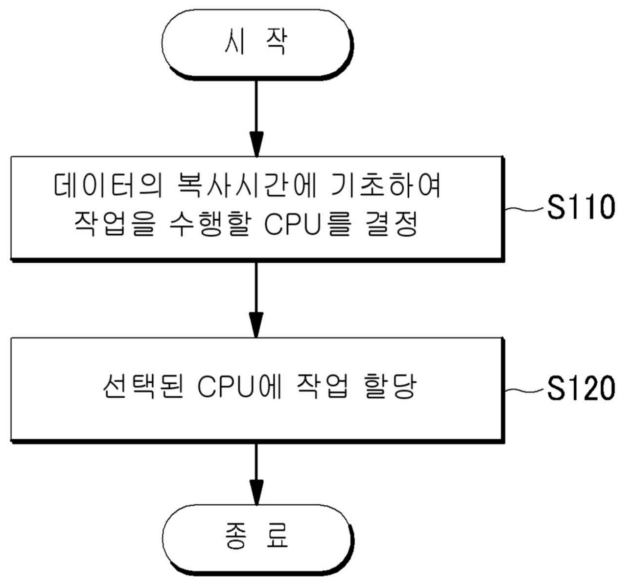


100

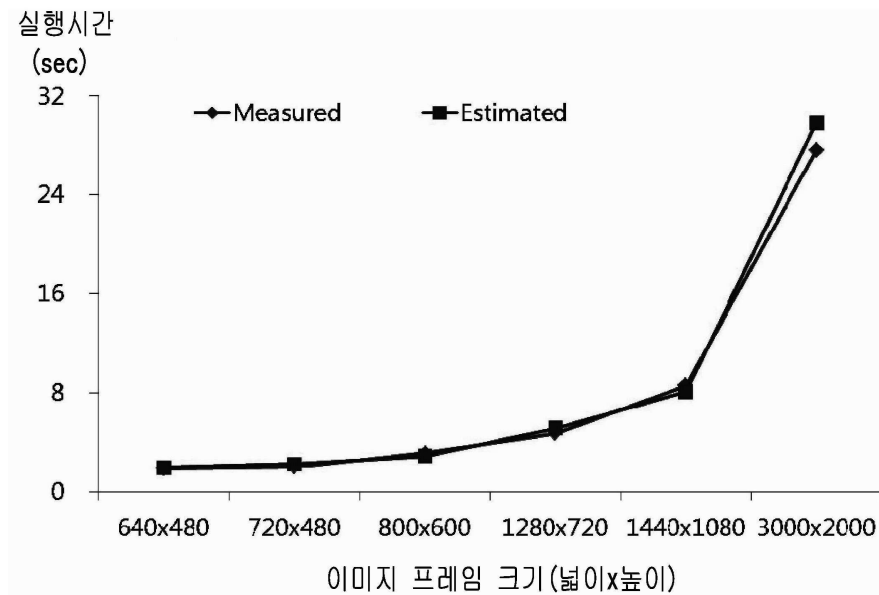
도면2



도면3



도면4



도면5

