



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0040109
(43) 공개일자 2020년04월17일

- | | |
|--|---|
| <p>(51) 국제특허분류(Int. Cl.)
G06F 12/02 (2018.01) G06F 12/0871 (2016.01)
G06T 1/60 (2006.01)</p> <p>(52) CPC특허분류
G06F 12/023 (2013.01)
G06F 12/0871 (2013.01)</p> <p>(21) 출원번호 10-2018-0119977</p> <p>(22) 출원일자 2018년10월08일
심사청구일자 2018년10월08일</p> | <p>(71) 출원인
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50</p> <p>(72) 발명자
백웅기
울산광역시 울주군 언양읍 유니스트길 50</p> <p>김규연
울산광역시 울주군 언양읍 유니스트길 50</p> <p>(74) 대리인
특허법인 무한</p> |
|--|---|

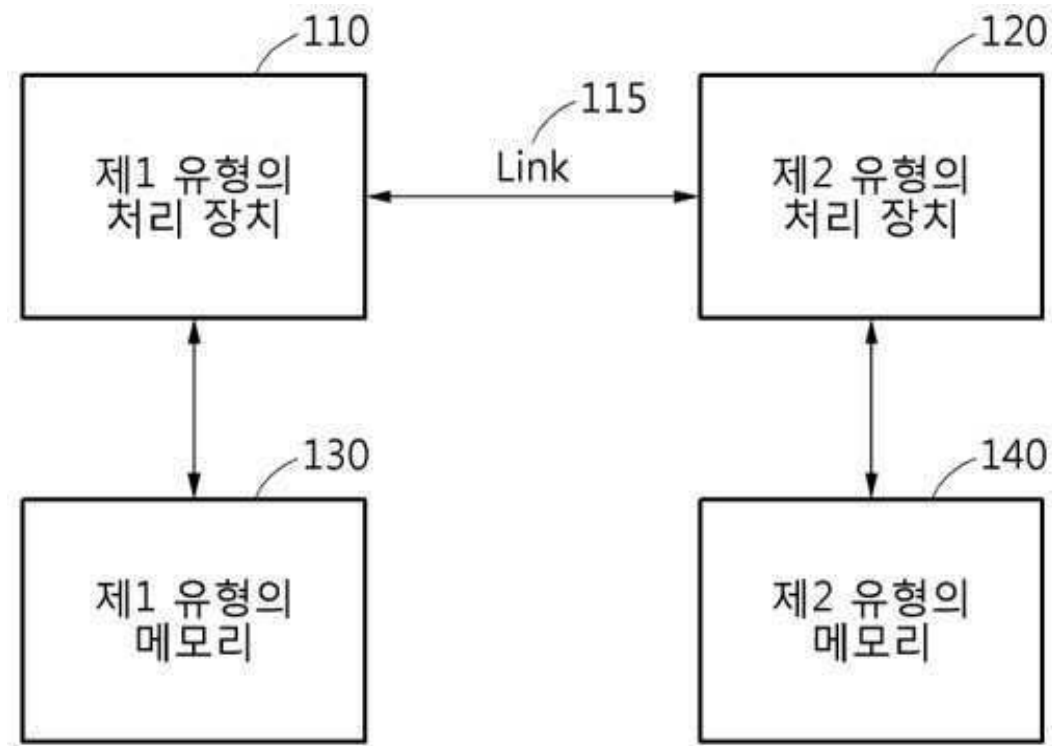
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 이중 메모리를 활용한 범용 GPU에서의 페이지징 방법 및 장치

(57) 요약

이중의 메모리들을 이용하는 페이지징 방법 및 장치가 개시된다. 일실시예에 따른 페이지징 장치는 제1 유형의 메모리의 동적 특성, 상기 제1 유형의 메모리의 정적 특성 및 제2 유형의 메모리의 정적 특성을 포함하는 정보를 획득하고, 상기 제1 유형의 메모리의 동적 특성, 상기 제1 유형의 메모리의 정적 특성에 포함된 제1 지연시간과 관(뒷면에 계속)

대표도 - 도1



런된 특성 및 상기 제2 유형의 메모리의 정적 특성에 포함된 제2 지연시간과 관련된 특성에 기초하여, 상기 제1 유형의 메모리와 상기 제2 유형의 메모리 간의 페이지 할당 정책을 결정하고, 상기 결정된 정책에 따라, 상기 제1 유형의 메모리의 정적 특성에 포함된 제1 대역폭과 관련된 특성 및 상기 제2 유형의 메모리의 정적 특성에 포함된 제2 대역폭과 관련된 특성에 기초하여, 상기 제1 유형의 메모리와 상기 제2 유형의 메모리 간의 페이지 할당 비율을 결정하고, 상기 결정된 비율에 기초하여, 상기 제1 유형의 메모리와 상기 제2 유형의 메모리에 페이지를 할당할 수 있다.

(52) CPC특허분류

G06T 1/60 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711075079

부처명 과학기술정보통신부

연구관리전문기관 한국연구재단

연구사업명 개인기초연구(과기정통부)

연구과제명 대규모 확장 가능한 고성능, 고보안, QoS 보장 빅데이터 처리를 위한 시스템 소프트웨어

기법 연구 및 개발

기 여 율 0.3/1

주관기관 울산과학기술원

연구기간 2018.03.01 ~ 2019.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호 1711074286

부처명 과학기술정보통신부

연구관리전문기관 한국연구재단

연구사업명 차세대정보·컴퓨팅기술개발

연구과제명 Peta Flops 스케일의 기계학습 프레임워크 개발

기 여 율 0.4/1

주관기관 울산과학기술원

연구기간 2018.06.01 ~ 2019.03.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711073161

부처명 과학기술정보통신부

연구관리전문기관 정보통신기술진흥센터

연구사업명 SW컴퓨팅산업원천기술개발(정보화)

연구과제명 (빅데이터-총괄) 빅데이터 처리 고도화 핵심 기술개발 사업 총괄 및 고성능 컴퓨팅 기술을

활용한 성능 가속화 기술 개발

기 여 율 0.3/1

주관기관 울산과학기술원

연구기간 2018.04.01 ~ 2019.01.31

명세서

청구범위

청구항 1

이종의 메모리들을 이용하는 페이지징 방법에 있어서,

제1 유형의 메모리의 동적 특성, 상기 제1 유형의 메모리의 정적 특성 및 제2 유형의 메모리의 정적 특성을 포함하는 정보를 획득하는 단계;

상기 제1 유형의 메모리의 동적 특성, 상기 제1 유형의 메모리의 정적 특성에 포함된 제1 지연시간과 관련된 특성 및 상기 제2 유형의 메모리의 정적 특성에 포함된 제2 지연시간과 관련된 특성에 기초하여, 상기 제1 유형의 메모리와 상기 제2 유형의 메모리 간의 페이지 할당 정책을 결정하는 단계;

상기 결정된 정책에 따라, 상기 제1 유형의 메모리의 정적 특성에 포함된 제1 대역폭과 관련된 특성 및 상기 제2 유형의 메모리의 정적 특성에 포함된 제2 대역폭과 관련된 특성에 기초하여, 상기 제1 유형의 메모리와 상기 제2 유형의 메모리 간의 페이지 할당 비율을 결정하는 단계; 및

상기 결정된 비율에 기초하여, 상기 제1 유형의 메모리와 상기 제2 유형의 메모리에 페이지를 할당하는 단계를 포함하는, 페이지징 방법.

청구항 2

제1항에 있어서,

상기 제1 지연시간은 상기 제2 지연시간보다 짧고, 상기 제1 대역폭은 상기 제2 대역폭보다 크며, 상기 제1 유형의 메모리의 용량은 상기 제2 유형의 메모리의 용량보다 작은, 페이지징 방법.

청구항 3

제1항에 있어서,

상기 페이지 할당 정책을 결정하는 단계는

상기 제1 유형의 메모리의 동적 특성, 상기 제1 지연시간과 관련된 특성, 및 상기 제2 지연시간과 관련된 특성에 기초하여, 메모리 지연 비율을 결정하는 단계; 및

상기 메모리 지연 비율이 미리 정해진 기준값 미만의 값을 가지는지 판단하는 단계

를 포함하는, 페이지징 방법.

청구항 4

제3항에 있어서,

상기 페이지 할당 정책을 결정하는 단계는

상기 메모리 지연 비율이 상기 기준값 미만의 값을 가지는 경우, 상기 제1 유형의 메모리 및 상기 제2 유형의 메모리에 페이지를 할당하는 제1 정책을 선택하는 단계; 및

상기 메모리 지연 비율이 상기 기준값 이상의 값을 가지는 경우, 상기 제1 유형의 메모리에만 페이지를 할당하는 제2 정책을 선택하는 단계

를 더 포함하는, 페이징 방법.

청구항 5

제1항에 있어서,
상기 제1 유형의 메모리는
제1 유형의 프로세서를 위한 물리적 메모리(physical memory);
상기 제1 유형의 프로세서를 위한 L1 캐시; 및
상기 제1 유형의 프로세서를 위한 L2 캐시
를 포함하는, 페이징 방법.

청구항 6

제5항에 있어서,
상기 제1 지연시간과 관련된 특성은
상기 L1 캐시의 지연시간, 상기 L2 캐시의 지연시간 및 상기 물리적 메모리의 지연시간을 포함하고,
상기 제1 유형의 메모리의 동적 특성은
상기 L1 캐시의 실패율, 및 상기 L2 캐시의 실패율
을 더 포함하는, 페이징 방법.

청구항 7

제3항에 있어서,
상기 페이지 할당 정책을 결정하는 단계는
데이터 업무량이 미리 정해진 기준값을 초과하는 값을 가지는지 판단하는 단계
를 더 포함하는, 페이징 방법.

청구항 8

제1항에 있어서,
상기 제1 유형의 메모리 및 상기 제2 유형의 메모리에 페이지를 할당하는 제1 정책이 선택된 경우,
상기 페이지 할당 비율을 결정하는 단계는
상기 제1 대역폭과 상기 제2 대역폭을 합산한 결과와 상기 제1 대역폭 사이의 비율에 기초하여, 상기 페이지 할당 비율을 결정하는 단계
를 포함하는, 페이징 방법.

청구항 9

제1항에 있어서,
상기 새로운 페이지를 할당하는 단계는

0과 1사이의 난수를 생성하는 단계;

상기 난수와 상기 제1 유형의 메모리에 할당할 페이지의 비율을 비교하는 단계; 및

상기 비교 결과에 따라, 상기 제1 유형의 메모리 및 상기 제2 유형의 메모리 중 어느 하나를 목표 노드로 선택하는 단계

를 포함하는, 페이지징 방법.

청구항 10

제9항에 있어서,

상기 새로운 페이지를 할당하는 단계는

상기 목표 노드에 빈 공간이 있는 경우, 상기 목표 노드에 페이지를 할당하는 단계; 및

상기 선택된 목표 노드에 빈 공간이 없고 선택되지 않은 메모리 노드에 빈 공간이 있는 경우, 상기 선택되지 않은 메모리 노드에 페이지를 할당하는 단계

를 더 포함하는, 페이지징 방법.

청구항 11

제1항에 있어서,

상기 정보를 획득하는 단계는

오프라인 환경에서 기 실행된 애플리케이션을 위한 메모리 할당 이력에 기초하여, 상기 정보를 획득하는 단계; 및

온라인 환경에서 실행 중인 애플리케이션을 위한 메모리 할당에 기초하여, 주기적으로 상기 정보를 획득하는 단계

중 적어도 하나를 포함하는, 페이지징 방법.

청구항 12

하드웨어와 결합되어 제1항 내지 제11항 중 어느 하나의 항의 방법을 실행시키기 위하여 매체에 저장된 컴퓨터 프로그램.

청구항 13

이종의 메모리들을 이용하는 페이지징 장치에 있어서,

제1 유형의 메모리의 동적 특성, 상기 제1 유형의 메모리의 정적 특성 및 제2 유형의 메모리의 정적 특성을 포함하는 정보를 획득하고,

상기 제1 유형의 메모리의 동적 특성, 상기 제1 유형의 메모리의 정적 특성에 포함된 제1 지연시간과 관련된 특성 및 상기 제2 유형의 메모리의 정적 특성에 포함된 제2 지연시간과 관련된 특성에 기초하여, 상기 제1 유형의 메모리와 상기 제2 유형의 메모리 간의 페이지 할당 정책을 결정하고,

상기 결정된 정책에 따라, 상기 제1 유형의 메모리의 정적 특성에 포함된 제1 대역폭과 관련된 특성 및 상기 제2 유형의 메모리의 정적 특성에 포함된 제2 대역폭과 관련된 특성에 기초하여, 상기 제1 유형의 메모리와 상기 제2 유형의 메모리 간의 페이지 할당 비율을 결정하고,

상기 결정된 비율에 기초하여, 상기 제1 유형의 메모리와 상기 제2 유형의 메모리에 페이지를 할당하는 프로세

서

를 포함하는, 페이지징 장치.

청구항 14

제13항에 있어서,

상기 제1 지연시간은 상기 제2 지연시간보다 짧고, 상기 제1 대역폭은 상기 제2 대역폭보다 크며, 상기 제1 유형의 메모리의 용량은 상기 제2 유형의 메모리의 용량보다 작은, 페이지징 장치.

청구항 15

제13항에 있어서,

상기 페이지 할당 정책을 결정하는 프로세서는

상기 제1 유형의 메모리의 동적 특성, 상기 제1 지연시간과 관련된 특성, 및 상기 제2 지연시간과 관련된 특성에 기초하여, 메모리 지연 비율을 결정하고,

상기 메모리 지연 비율이 미리 정해진 기준값 미만의 값을 가지는지 판단하는 프로세서

를 포함하는, 페이지징 장치.

청구항 16

제15항에 있어서,

상기 페이지 할당 정책을 결정하는 프로세서는

상기 메모리 지연 비율이 상기 기준값 미만의 값을 가지는 경우, 상기 제1 유형의 메모리 및 상기 제2 유형의 메모리에 페이지를 할당하는 제1 정책을 선택하고,

상기 메모리 지연 비율이 상기 기준값 이상의 값을 가지는 경우, 상기 제1 유형의 메모리에만 페이지를 할당하는 제2 정책을 선택하는 프로세서

를 더 포함하는, 페이지징 장치.

청구항 17

제13항에 있어서,

상기 제1 유형의 메모리는

제1 유형의 프로세서를 위한 물리적 메모리(physical memory);

상기 제1 유형의 프로세서를 위한 L1 캐시; 및

상기 제1 유형의 프로세서를 위한 L2 캐시

를 포함하는, 페이지징 장치.

청구항 18

제13항에 있어서,

상기 제1 유형의 메모리 및 상기 제2 유형의 메모리에 페이지를 할당하는 제1 정책이 선택된 경우,

상기 페이지 할당 비율을 결정하는 프로세서는

상기 제1 대역폭과 상기 제2 대역폭을 합산한 결과와 상기 제1 대역폭 사이의 비율에 기초하여, 상기 페이지 할당 비율을 결정하는 프로세서

를 포함하는, 페이지징 장치.

청구항 19

제13항에 있어서,

상기 새로운 페이지를 할당하는 프로세서는

0과 1사이의 난수를 생성하고,

상기 난수와 상기 제1 유형의 메모리에 할당할 페이지의 비율을 비교하고,

상기 비교 결과에 따라, 상기 제1 유형의 메모리 및 상기 제2 유형의 메모리 중 어느 하나를 목표 노드로 선택하는 프로세서

를 포함하는, 페이지징 장치.

청구항 20

제13항에 있어서,

상기 정보를 획득하는 프로세서는

오프라인 환경에서 기 실행된 애플리케이션을 위한 메모리 할당 이력에 기초하여, 상기 정보를 획득하고,

온라인 환경에서 실행 중인 애플리케이션을 위한 메모리 할당에 기초하여, 주기적으로 상기 정보를 획득하는 프로세서

중 적어도 하나를 포함하는, 페이지징 장치.

발명의 설명

기술 분야

[0001] 아래 실시예들은 범용 GPU에서의 메모리 관리 기술에 관한 것이다.

배경 기술

[0002] 최근에 그래픽 프로세서는 알려진 그래픽 관련 연산들 중 상당 부분이 파이프라인화될 수 있다는 점과 반복적인 스트림 연산들로 프로그래밍될 수 있다는 점을 이용하여, 수많은 스트림 프로세서 코어들을 클러스터링한 아키텍처에 기반하는 범용 GPU(General Purpose Graphic Processor Unit, GPGPU)로 발전하고 있다.

[0003] 범용 GPU는 그래픽 어플리케이션에 가장 적합하지만, 파이프라인 및 스트림 연산의 특징을 가지도록 프로그래밍할 수 있는 물리 효과 시뮬레이션이나 계산 생물학, 암호학, 영상 처리, 입자 핵물리 시뮬레이션, 천체 시뮬레이션 등 다양한 과학 및 공학 분야에서의 연산 속도 향상에도 활용될 수 있으며, 기계 학습 및 데이터 분석과 같은 영역에서도 고성능 및 에너지 효율적인 컴퓨팅을 위한 유망한 솔루션으로 부각되고 있다.

[0004] 최근까지 그 컴퓨팅 효율성에도 불구하고, 가상 메모리는 범용 GPU에 대해 널리 지원되지 않았다. 가상 메모리를 지원하지 않는 경우 프로그래머는 GPU와 CPU 물리적 메모리 간의 데이터 전송을 수동으로 조정하여야 하고, 이는 범용 GPU 컴퓨팅의 프로그램 작동 가능성과 유연성을 크게 저하시키게 된다.

[0005] 선행 연구들은 GPU 메모리 노드와 CPU 메모리 노드로 구성된 이중 처리 장치를 활용한 범용 GPU에 지원되는 가상 메모리를 설계 및 구현하였다.

- [0006] 이중 처리 장치를 활용한 범용 GPU에서 최상의 성능을 얻으려면 대역폭 및 지연 시간에 대한 성능 차이를 고려하여 이중 처리 장치 메모리 노드에 페이지를 할당하는 것이 중요하다. 선행 연구들은 주소 변환(address translation), 캐시 계층(cache hierarchy) 및 워프 스케줄링(wrap scheduling)의 효율적인 설계와 구현을 연구하였으나, 이중 처리 장치 메모리 관리 기능을 고려하지 않았다.

발명의 내용

해결하려는 과제

과제의 해결 수단

- [0007] 일실시에에 따른 이중의 메모리들을 이용하는 페이지징 방법에 의하면, 지연 시간 및 데이터 업무량에 기초하여 결정된 할당 정책에 따라 페이지를 할당할 메모리를 결정함으로써, 상황에 따라 유동적으로 하나 또는 복수 개의 메모리에 페이지가 할당될 수 있다. 일실시에에 따른 이중의 메모리들을 이용하는 페이지징 방법에 의하면, 결정된 정책에 따라 이중의 메모리의 대역폭과 관련된 특성에 기초하여 페이지 할당 비율을 결정함으로써, 정보를 처리하는 데 소요되는 시간이 감소될 수 있다.
- [0008] 일실시에에 따른 이중의 메모리들을 이용하는 페이지징 방법은 제1 유형의 메모리의 동적 특성, 상기 제1 유형의 메모리의 정적 특성 및 제2 유형의 메모리의 정적 특성을 포함하는 정보를 획득하는 단계; 상기 제1 유형의 메모리의 동적 특성, 상기 제1 유형의 메모리의 정적 특성에 포함된 제1 지연시간과 관련된 특성 및 상기 제2 유형의 메모리의 정적 특성에 포함된 제2 지연시간과 관련된 특성에 기초하여, 상기 제1 유형의 메모리와 상기 제2 유형의 메모리 간의 페이지 할당 정책을 결정하는 단계; 상기 결정된 정책에 따라, 상기 제1 유형의 메모리의 정적 특성에 포함된 제1 대역폭과 관련된 특성 및 상기 제2 유형의 메모리의 정적 특성에 포함된 제2 대역폭과 관련된 특성에 기초하여, 상기 제1 유형의 메모리와 상기 제2 유형의 메모리 간의 페이지 할당 비율을 결정하는 단계; 및 상기 결정된 비율에 기초하여, 상기 제1 유형의 메모리와 상기 제2 유형의 메모리에 페이지를 할당하는 단계를 포함한다.
- [0009] 일실시에에 따르면, 상기 제1 지연시간은 상기 제2 지연시간보다 짧고, 상기 제1 대역폭은 상기 제2 대역폭보다 크며, 상기 제1 유형의 메모리의 용량은 상기 제2 유형의 메모리의 용량보다 작을 수 있다.
- [0010] 일실시에에 따르면, 상기 페이지 할당 정책을 결정하는 단계는 상기 제1 유형의 메모리의 동적 특성, 상기 제1 지연시간과 관련된 특성, 및 상기 제2 지연시간과 관련된 특성에 기초하여, 메모리 지연 비율을 결정하는 단계; 및 상기 메모리 지연 비율이 미리 정해진 기준값 이하의 값을 가지는지 판단하는 단계를 포함할 수 있다.
- [0011] 일실시에에 따르면, 상기 페이지 할당 정책을 결정하는 단계는 상기 메모리 지연 비율이 상기 기준값 이하의 값을 가지는 경우, 상기 제1 유형의 메모리 및 상기 제2 유형의 메모리에 페이지를 할당하는 제1 정책을 선택하는 단계; 및 상기 메모리 지연 비율이 상기 기준값을 초과하는 값을 가지는 경우, 상기 제1 유형의 메모리에만 페이지를 할당하는 제2 정책을 선택하는 단계를 더 포함할 수 있다.
- [0012] 일실시에에 따르면, 상기 제1 유형의 메모리는 제1 유형의 프로세서를 위한 물리적 메모리(physical memory); 상기 제1 유형의 프로세서를 위한 L1 캐시; 및 상기 제1 유형의 프로세서를 위한 L2 캐시를 포함할 수 있다.
- [0013] 일실시에에 따르면, 상기 제1 지연시간과 관련된 특성은 상기 L1 캐시의 지연시간, 상기 L2 캐시의 지연시간 및 상기 물리 메모리의 지연시간을 포함하고, 상기 제1 유형의 메모리의 동적 특성은 상기 L1 캐시의 실패율, 및 상기 L2 캐시의 실패율을 더 포함할 수 있다.
- [0014] 일실시에에 따르면, 상기 페이지 할당 정책을 결정하는 단계는 데이터 업무량이 미리 정해진 기준값 이상의 값을 가지는지 판단하는 단계를 더 포함할 수 있다.
- [0015] 일실시에에 따르면, 상기 제1 유형의 메모리 및 상기 제2 유형의 메모리에 페이지를 할당하는 제1 정책이 선택된 경우, 상기 페이지 할당 비율을 결정하는 단계는 상기 제1 대역폭과 상기 제2 대역폭을 합산한 결과와 상기 제1 대역폭 사이의 비율에 기초하여, 상기 페이지 할당 비율을 결정하는 단계를 포함할 수 있다.
- [0016] 일실시에에 따르면, 상기 새로운 페이지를 할당하는 단계는 0과 1사이의 난수를 생성하는 단계; 상기 난수와 상기 제1 유형의 메모리에 할당할 페이지의 비율을 비교하는 단계; 및 상기 비교 결과에 따라, 상기 제1 유형의

메모리 및 상기 제2 유형의 메모리 중 어느 하나를 목표 노드로 선택하는 단계를 포함할 수 있다.

[0017] 일실시예에 따르면, 상기 새로운 페이지를 할당하는 단계는 상기 목표 노드에 빈 공간이 있는 경우, 상기 목표 노드에 페이지를 할당하는 단계; 및 상기 선택된 목표 노드에 빈 공간이 없고 선택되지 않은 메모리 노드에 빈 공간이 있는 경우, 상기 선택되지 않은 메모리 노드에 페이지를 할당하는 단계를 더 포함할 수 있다.

[0018] 일실시예에 따르면, 상기 정보를 획득하는 단계는 오프라인 환경에서 기 실행된 애플리케이션을 위한 메모리 할당 이력에 기초하여, 상기 정보를 획득하는 단계; 및 온라인 환경에서 실행 중인 애플리케이션을 위한 메모리 할당에 기초하여, 주기적으로 상기 정보를 획득하는 단계 중 적어도 하나를 포함할 수 있다.

[0019] 일실시예에 따른 이종의 메모리들을 이용하는 페이지징 장치는 제1 유형의 메모리의 동적 특성, 상기 제1 유형의 메모리의 정적 특성 및 제2 유형의 메모리의 정적 특성을 포함하는 정보를 획득하고, 상기 제1 유형의 메모리의 동적 특성, 상기 제1 유형의 메모리의 정적 특성에 포함된 제1 지연시간과 관련된 특성 및 상기 제2 유형의 메모리의 정적 특성에 포함된 제2 지연시간과 관련된 특성에 기초하여, 상기 제1 유형의 메모리와 상기 제2 유형의 메모리 간의 페이지 할당 정책을 결정하고, 상기 결정된 정책에 따라, 상기 제1 유형의 메모리의 정적 특성에 포함된 제1 대역폭과 관련된 특성 및 상기 제2 유형의 메모리의 정적 특성에 포함된 제2 대역폭과 관련된 특성에 기초하여, 상기 제1 유형의 메모리와 상기 제2 유형의 메모리 간의 페이지 할당 비율을 결정하고, 상기 결정된 비율에 기초하여, 상기 제1 유형의 메모리와 상기 제2 유형의 메모리에 페이지를 할당하는 프로세서를 포함한다.

도면의 간단한 설명

[0020] 도 1은 일실시예에 따른 이종 처리 장치를 설명하기 위한 도면이다.

도 2는 일실시예에 따른 이종 처리 장치의 예시도이다.

도 3은 일실시예에 따른 페이지징 장치를 설명하기 위한 도면이다.

도 4는 일실시예에 따른 페이지징 방법을 설명하기 위한 동작 흐름도이다.

도 5는 일실시예에 따른 페이지 할당 정책을 결정하는 동작을 설명하기 위한 도면이다.

도 6은 일실시예에 따른 페이지징 동작을 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0021] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형태로 변경되어 실시될 수 있다. 따라서, 실시예들은 특정한 개시형태로 한정되는 것이 아니며, 본 명세서의 범위는 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.

[0022] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.

[0023] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.

[0024] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로써 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0025] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 해당 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0026] 이하, 실시예들을 첨부된 도면을 참조하여 상세하게 설명한다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

- [0028] 도 1은 일실시예에 따른 이중 처리 장치를 설명하기 위한 도면이다.
- [0029] 도 1을 참조하면, 일실시예에 따른 이중 처리 장치는 제1 유형의 처리 장치(110), 제2 유형의 처리 장치(120), 제1 유형의 메모리(130), 및 제2 유형의 메모리(140)를 포함한다. 이중 처리 장치는 서로 다른 유형의 처리 장치가 연동하여 데이터를 처리하는 장치일 수 있다.
- [0030] 제1 유형의 처리 장치 및 제2 유형의 처리 장치는 데이터를 처리하는 장치로서, 사용자로부터 입력 받은 명령어를 해석, 연산한 후 그 결과를 출력할 수 있다. 제1 유형의 처리 장치의 스펙(spec)과 제2 유형의 처리 장치의 스펙은 상이할 수 있다. 예를 들어, 제1 유형의 처리 장치는 미리 정해진 특정 태스크(task)들을 처리하기에 적합한 스펙을 가지고, 제2 유형의 처리 장치는 범용 태스크들을 처리하기에 적합한 스펙을 가질 수 있다. 도 2를 참조하면, GPU(210)는 제1 유형의 처리 장치(110)에 해당할 수 있고, CPU(220)는 제2 유형의 처리 장치(120)에 해당할 수 있다.
- [0031] 제1 유형의 메모리는 제1 유형의 처리 장치에 대응하여, 정보를 읽고 쓰는 메모리일 수 있다. 제2 유형의 메모리는 제2 유형의 처리 장치에 대응하여, 정보를 읽고 쓰는 메모리일 수 있다. 제1 유형의 처리 장치 및 제2 유형의 처리 장치는 링크(115)를 통하여 서로 연동할 수 있다. 이중 처리 장치는 제1 유형의 메모리 및 제2 유형의 메모리에 적응적으로 페이지를 할당할 수 있다.
- [0032] 제1 유형의 메모리의 정적 특성과 제2 유형의 메모리의 정적 특성은 상이할 수 있다. 예를 들어, 제1 유형의 메모리의 제1 지연시간은 제2 유형의 메모리의 제2 지연시간보다 짧을 수 있다. 제1 유형의 메모리의 제1 대역폭은 제2 유형의 메모리의 제2 대역폭보다 클 수 있다. 제1 유형의 메모리의 용량은 제2 유형의 메모리의 용량보다 작을 수 있다. 도 2를 참조하면, 제1 유형의 메모리(130)는 GPU 메모리(230)에 해당할 수 있고, 제2 유형의 메모리(140)는 CPU 메모리(240)에 해당할 수 있다. GPU 메모리(230)는 CPU 메모리(240)에 비하여 지연시간이 짧고, 대역폭은 큰 반면, GPU 메모리(230)의 용량은 CPU 메모리(240)의 용량보다 작을 수 있다.
- [0033] 아래에서 설명하겠으나, 정적 특성은 해당 메모리에서 수행되는 태스크와 무관하게 설계 단계에서 정해지는 특성을 포함할 수 있다. 또한, 동적 특성은 해당 메모리에서 수행되는 태스크에 의존하여 결정되는 특성을 포함할 수 있다.
- [0034] 이중 처리 장치는 이중의 메모리들의 특성을 고려하여, 이중의 메모리들에 페이지를 적응적으로 할당하는 기술을 제공할 수 있다.
- [0036] 도 3은 일실시예에 따른 페이지징 장치를 설명하기 위한 도면이다.
- [0037] 도 3을 참조하면, 페이지징 장치는 애플리케이션 모니터부(310), 할당 비율 조절부(320), 및 메모리 할당부(330)를 포함할 수 있다. 페이지징 장치는 페이지를 할당하는 장치로, 예를 들어 하나 또는 그 이상의 소프트웨어 모듈, 하나 또는 그 이상의 하드웨어 모듈, 또는 이들의 다양한 조합으로 구현될 수 있다. 페이지징 장치는 이중 처리 장치에 포함된 이중 메모리들에 페이지를 할당할 수 있다. 페이지징 장치는 이중 처리 장치에 탑재되거나, 이중 처리 장치와 독립적으로 구현될 수 있다.
- [0038] 애플리케이션 모니터부(310)는 이중 메모리의 동적 특성을 획득한다. 이중 메모리의 동적 특성은 애플리케이션이 수행되는 동안 변경될 수 있다. 애플리케이션 모니터부(310)는 이중 메모리의 동적 특성을 모니터링하여 할당 비율 조절부(320)에 전달할 수 있다. 일 실시예에 따르면, 애플리케이션 모니터부(310)는 이중 메모리 중 미리 정해진 하나 또는 그 이상의 메모리의 동적 특성을 모니터링 할 수 있다. 예를 들어, 애플리케이션 모니터부(310)는 제1 유형의 메모리의 동적 특성을 획득할 수 있다. 애플리케이션 모니터부(310)는 제1 유형의 프로세서를 위한 L1 캐시의 실패율(miss rate), 제1 유형의 프로세서를 위한 L2 캐시의 실패율 및 데이터 업무량(data traffic)에 관한 정보를 획득할 수 있다.
- [0039] 할당 비율 조절부(320)는 애플리케이션 모니터부(310)에서 이중 메모리의 동적 특성을 획득할 수 있다. 할당 비율 조절부(320)는 이중 메모리의 정적 특성을 획득할 수 있다. 이중 메모리의 정적 특성은 동적 실행 환경에 따라 변하지 않고 미리 정해질 수 있다. 할당 비율 조절부(320)는 이중 메모리의 정적 특성과 관련하여 기 저장된 정보를 획득하고, 획득한 정보에 기초하여 이중 메모리의 정적 특성(예를 들어, 제1 유형의 메모리의 제1 지연시간과 제2 유형의 메모리의 제2 지연시간 등)을 결정할 수 있다.

- [0040] 할당 비율 조절부(320)는 이중 메모리의 특성을 고려하여 메모리 지연 비율을 결정한 뒤, 메모리 지연 비율에 기초하여 페이지 할당 정책을 결정할 수 있다. 예를 들어, 할당 비율 조절부(320)는 제1 지연시간, 제2 지연시간, 제1 유형의 메모리의 대역폭, 및 제2 유형의 메모리의 대역폭에 기초하여 메모리 지연 비율을 결정할 수 있다. 할당 비율 조절부(320)는 메모리 지연 비율과 데이터 업무량에 기초하여, 페이지 할당 정책을 결정할 수 있다.
- [0041] 메모리의 메모리 지연 비율을 결정하는 것과 관련된 보다 상세한 사항은 도 5를 통하여 후술한다.
- [0042] 할당 비율 조절부(320)는 제1 유형의 메모리 및 제2 유형의 메모리 간의 페이지 할당 정책을 결정하고, 제1 유형의 메모리 및 제2 유형의 메모리 간의 페이지 할당 비율을 결정한다. 페이지 할당 정책은 메모리 지연 비율과 데이터 업무량에 기초하여 결정될 수 있다. 예를 들어, 메모리 지연 비율이 1.4 미만이고, 데이터 업무량이 140MB/s를 초과하는 경우, 할당 비율 조절부(320)는 제1 유형의 메모리 및 제2 유형의 메모리에 페이지를 할당하는 제1 정책을 선택할 수 있으며, 조건을 만족하지 못하는 경우 제1 유형의 메모리에만 페이지를 할당하는 제2 정책을 선택할 수 있다.
- [0043] 할당 비율 조절부(320)는 제1 정책이 선택된 경우, 제1 대역폭과 제2 대역폭을 합산한 결과와 제1 대역폭 사이의 비율에 기초하여 페이지 할당 비율을 결정할 수 있다. 예를 들어, 제1 대역폭과 제2 대역폭의 비율이 2:1 인 경우, 페이지 할당 비율은 2/3으로 결정될 수 있다.
- [0044] 페이지 할당 정책 및 페이지 할당 비율을 결정하는 것과 관련된 보다 상세한 사항은 도 5를 통하여 후술한다.
- [0045] 메모리 할당부(330)는 할당 비율 조절부(320)에 의하여 결정된 페이지 할당 비율에 기초하여 이중 메모리에 페이지를 할당한다. 예를 들어, 메모리 할당부(330)는 0과 1 사이의 난수를 생성하여 난수와 할당 비율 조절부(320)에서 결정된 할당 비율을 비교함으로써 제1 유형의 메모리 및 제2 유형의 메모리에 확률적으로 페이지를 할당할 수 있다. 예를 들어, 페이지 할당 비율이 2/3으로 결정된 경우, 신규한 페이지는 2/3 확률로 제1 유형의 메모리에 할당될 수 있다.
- [0046] 확률적으로 페이지를 할당하는 것과 관련된 보다 상세한 사항은 도 6을 통하여 후술한다.
- [0047] 일 실시예에 따르면, 애플리케이션 모니터부(310)에서 정보를 획득하는 단계는 오프라인 환경에서 기 실행된 애플리케이션을 위한 메모리 할당 이력에 기초하여 정보를 획득하는 단계 및 온라인 환경에서 실행 중인 애플리케이션을 위한 메모리 할당에 기초하여 주기적으로 정보를 획득하는 단계 중 적어도 하나를 포함할 수 있다. 오프라인 환경에서 기 실행된 애플리케이션을 위한 메모리 할당 이력에 기초하여 정보를 획득하는 단계는 정적 페이지징 단계에 해당할 수 있다. 온라인 환경에서 실행 중인 애플리케이션을 위한 메모리 할당에 기초하여 주기적으로 정보를 획득하는 단계는 동적 페이지징 단계에 해당할 수 있다.
- [0048] 동적 페이지징 단계는 정적 페이지징 단계에 비하여 오프라인 환경에서의 정보 획득이 요구되지 않을 수 있다. 정적 페이지징 단계는 동적 페이지징 단계에 비하여 최적의 할당 비율이 계산되기 전이라도 페이지를 할당할 수 있다.
- [0050] 도 4는 일실시예에 따른 페이지징 방법을 설명하기 위한 동작 흐름도이다.
- [0051] 도 4를 참조하면, 정보 획득 단계(410)에서 제1 유형의 메모리의 동적 특성, 제1 유형의 메모리의 정적 특성 및 제2 유형의 메모리의 정적 특성을 포함하는 정보가 획득된다. 정보 획득 단계(410)는 도 3의 애플리케이션 모니터부(310)에 의하여 수행될 수 있다.
- [0052] 페이지 할당 정책 결정 단계(420)에서 제1 유형의 메모리의 동적 특성, 제1 유형의 메모리의 정적 특성에 포함된 제1 지연시간과 관련된 특성 및 제2 유형의 메모리의 정적 특성에 포함된 제2 지연시간과 관련된 특성에 기초하여, 제1 유형의 메모리와 제2 유형의 메모리 간의 페이지 할당 정책이 결정된다.
- [0053] 페이지 할당 비율 결정 단계(430)에서 제1 유형의 메모리의 정적 특성에 포함된 제1 대역폭과 관련된 특성 및 제2 유형의 메모리의 정적 특성에 포함된 제2 대역폭과 관련된 특성에 기초하여, 제1 유형의 메모리와 제2 유형의 메모리 간의 페이지 할당 비율이 결정된다.
- [0054] 페이지 할당 정책 결정 단계(420) 및 페이지 할당 비율 결정 단계(430)는 도 3의 할당 비율 조절부(320)에 의하여 수행될 수 있다.
- [0055] 페이지징 단계(440)에서 제1 유형의 메모리와 제2 유형의 메모리에 페이지가 할당된다. 페이지징 단계(440)는 도 3

의 메모리 할당부(330)에 의하여 수행될 수 있다.

[0057] 도 5는 일실시예에 따른 페이지 할당 정책을 결정하는 동작을 설명하기 위한 도면이다. 도 5에 도시된 단계들은 도 4의 페이지 할당 정책 결정 단계(420)에 포함될 수 있고, 도 3의 할당 비율 조절부(320)에 의하여 수행될 수 있다.

[0058] 도 5를 참조하면, 데이터 업무량이 미리 정해진 기준값을 초과하고(510), 메모리 지연 비율이 미리 정해진 기준값 미만의 값을 가지는 경우(520), 제1 유형의 메모리(130) 내지 제2 유형의 메모리(140)에 페이지를 할당하는 제1 정책(530)이 선택될 수 있다. 데이터 업무량이 미리 정해진 기준값 이하의 값을 가지거나, 메모리 지연 비율이 상기 기준값 이상의 값을 가지는 경우 제1 유형의 메모리에만 페이지를 할당하는 제2 정책이 선택될 수 있다(540).

[0059] 데이터 업무량에 관한 정보는 도 3의 애플리케이션 모니터부(310)에서 획득될 수 있다.

[0060] 메모리 지연 비율을 결정하는 방법을 설명하기에 앞서, 실시예들에 따른 이중 메모리의 구조를 간략하게 설명한다. 설명의 편의를 위하여, 일 실시예에 따른 이중 메모리가 도 1에 도시된 것과 같이 제1 유형의 메모리와 제2 유형의 메모리를 포함하는 것으로 설명하였으나, 실시예들은 이중 메모리가 3개 이상의 서로 다른 유형의 메모리들을 포함하는 경우로 확장될 수 있다.

[0061] 일실시예에 따르면, 3개 이상의 이중 메모리들을 사용하는 이중 처리 장치에 있어서, 제k-1 유형의 메모리의 제k-1 지연시간은 제k 유형의 메모리의 제k 지연시간보다 짧을 수 있고, 제k-1 유형의 메모리의 제k-1 대역폭은 제k 유형의 메모리의 제k 대역폭보다 클 수 있으며, 제k-1 유형의 메모리의 용량은 제k 유형의 메모리의 용량보다 작을 수 있다. 여기서, k=2, 3, ... n일 수 있다.

[0063] 메모리 지연 비율에 따른 페이지 할당 정책

[0064] 메모리 지연 비율을 결정하기 위해서는 우선 제1 지연시간 및 제2 지연시간을 결정해야 한다. 제1 지연시간은 수학적 식 1을 이용하여 결정될 수 있다.

수학적 식 1

$$T_1 = T_{1,L1D} + m_{1,L1D} \times (T_{1,L2} + m_{1,L2} \times T_{1,M})$$

[0066] 여기서, T_1 은 제1 지연시간이고, $T_{1,L1D}$ 는 제1 유형의 프로세서를 위한 L1 캐시의 지연시간이며, $T_{1,L2D}$ 는 제1 유형의 프로세서를 위한 L2 캐시의 지연시간이고, $T_{1,M}$ 은 제1 유형의 프로세서를 위한 물리적 메모리의 지연시간이며, $m_{1,L1D}$ 는 제1 유형의 프로세서를 위한 L1 캐시의 실패율이고, $m_{1,L2D}$ 는 제1 유형의 프로세서를 위한 L2 캐시의 실패율이다.

[0067] 제2 지연시간은 수학적 식 2를 이용하여 결정될 수 있다.

수학적 식 2

$$T_x = T_{x,M}$$

[0069] 여기서, x는 2 이상 n 이하의 자연수이고, T_x 는 제x 지연시간이고, $T_{x,M}$ 은 제x 유형의 프로세서를 위한 물리적 메모리 및 제1 유형의 프로세서와 제x 유형의 프로세서 사이의 물리적 연결에 의한 지연시간이다.

[0070] 일 실시예에 따르면, 이중 처리 장치에서, 캐시 일관성(cache coherence)이 지원되지 않을 수 있다. 예를 들어, 제x 유형의 메모리에 속하는 데이터를 제1 유형의 프로세서를 위한 L1 캐시에 쓰는 경우, 해당 데이터는 L1 캐시에만 반영되고, 제x 유형의 프로세서는 해당 데이터가 쓰여진 사실을 즉각적으로 인지하지 못할 수 있다. 때

문에, 제1 유형의 프로세서가 데이터를 읽거나 쓸 때, 해당 데이터가 제1 유형의 메모리에 속할 때만 L1 캐시에 데이터를 저장하고, 해당 데이터가 제x 유형의 프로세서의 메모리에 속하면 L1 캐시에 저장하지 않도록 하는 선택적 캐싱(selective caching)을 실시할 수 있다.

[0071] 일 실시예에 따르면, 이중 처리 장치에서, 메모리 요청(memory request)의 경로는 이중의 프로세서를 위한 캐시를 거치지 않을 수 있다. 예를 들어, 제1 유형의 프로세서가 제x 유형의 프로세서의 메모리를 읽거나 쓰기 위한 메모리 요청을 하는 경우, 해당 요청은 제x 유형의 프로세서를 위한 Lx 캐시를 거치지 않고 제x 유형 프로세서를 위한 메모리 컨트롤러만 거치기 때문에, 제1 유형의 프로세서에 의한 메모리 요청은 Lx 캐시에 저장되지 않을 수 있다.

[0072] 이로 인하여, 제x 지연 시간은 제x 유형의 프로세서를 위한 물리적 메모리 및 제1 유형의 프로세서와 제x 유형의 프로세서 사이의 물리적 연결에 의해서만 결정될 수 있다.

[0073] 메모리 지연 비율은 수학식 3을 이용하여 결정될 수 있다.

수학식 3

$$r_L = \frac{\sum_{j=1}^n \frac{B_j}{\sum_{i=1}^n B_i} \times T_j}{T_1}$$

[0074]

[0075] 여기서, r_L 은 메모리 지연 비율이고, y는 1 이상 n 이하의 자연수일 때 B_y 는 제y 유형의 메모리의 제y 대역폭이다.

[0076] 메모리 지연 비율이 결정되면 이에 기초하여 페이지 할당 정책을 결정할 수 있다. 예를 들어, 단계(520)에서 메모리 지연 비율과 미리 정해진 기준값을 비교함으로써, 단계(530)에서 제1 정책이 선택되거나 단계(540)에서 제2 정책이 선택될 수 있다.

[0078] 페이지 할당 비율

[0079] 페이지 할당 정책이 결정되면 이에 기초하여 제1 유형의 메모리 내지 제n 유형의 메모리 간의 페이지 할당 비율을 결정할 수 있다.

[0080] 제1 정책을 선택한 경우, 제y 유형의 메모리에 할당할 페이지의 비율을 p_y 로 설정할 수 있다. 이 경우, 제y 유형의 프로세서에서 정보를 처리하는 시간은 수학식 4를 이용하여 결정될 수 있다.

수학식 4

$$t_y = \frac{p_y \times D}{B_y}, \sum_{i=1}^n p_i = 1$$

[0081]

[0082] 여기서, t_y 는 제y 유형의 프로세서에서 정보를 처리하는 시간이고, D는 총 정보이다.

[0083] 총 실행 시간은 수학식 5를 이용하여 결정될 수 있다.

수학식 5

$$t_T = \max(t_1, t_2, \dots, t_n)$$

[0084]

[0085] 여기서, t_T 는 총 실행 시간이다.

[0086] t_1 내지 t_n 이 모두 같은 값을 가질 때, 총 실행 시간이 최소화될 수 있다.

[0087] 즉, 제1 정책이 선택된 경우 최적의 페이지 할당 비율은 수학식 6을 이용하여 결정될 수 있다.

수학식 6

[0088]
$$p_{yOPT,S} = \frac{B_y}{\sum_{i=1}^n B_i}$$

[0089] 여기서, $p_{yOPT,S}$ 는 제1 정책이 선택된 경우에 제y 유형의 메모리에 할당할 페이지의 비율이다.

[0090] 제2 정책을 선택한 경우, 제1 유형의 메모리에만 페이지가 할당될 수 있다.

[0091] 따라서 제2 정책이 선택된 경우 최적의 페이지 할당 비율은 수학식 7을 이용하여 결정될 수 있다.

수학식 7

[0092]
$$p_{1OPT,I} = 1, p_{kOPT,I} = 0 \text{ for } k = 2, 3, \dots, n$$

[0093] 여기서, $p_{yOPT,I}$ 는 제2 정책이 선택된 경우에 제y 유형의 메모리에 할당할 페이지의 비율이다.

[0094] 결론적으로, 최적의 페이지 할당 비율은 수학식 8을 이용하여 결정될 수 있다.

수학식 8

[0095]
$$p_{yOPT} = \begin{cases} p_{yOPT,S} & \text{if } r_L < \theta_L \cap B_M > \theta_B \\ p_{yOPT,I} & \text{otherwise} \end{cases}$$

[0096] 여기서, p_{yOPT} 는 제y 유형의 메모리에 할당할 페이지의 비율이고, B_M 은 데이터 업무량이며, θ_L 은 메모리 지연 비율에 대하여 미리 정해진 기준값이고, θ_B 는 데이터 업무량에 대하여 미리 정해진 기준값이다.

[0097] 도 6은 일실시예에 따른 페이징 동작을 설명하기 위한 도면이다. 도 6에 도시된 단계들은 도 4의 페이징 단계(440)에 포함될 수 있고, 도 3의 메모리 할당부(330)에 의하여 수행될 수 있다.

[0098] 도 6을 참조하면, 메모리 할당부는 0과 1 사이의 난수를 생성할 수 있고(610), 난수와 페이지 할당 비율 값을 비교할 수 있다(620). 어느 메모리를 목표 노드로 선택할지는 수학식 9를 이용하여 결정될 수 있다.

수학식 9

[0099]
$$\begin{cases} \text{제1 유형의 메모리를 목표 노드로 선택} & \text{if } \alpha \leq p_{1OPT} \\ \text{제k 유형의 메모리를 목표 노드로 선택} & \text{if } \sum_{i=1}^{k-1} p_{iOPT} < \alpha \leq \sum_{i=1}^k p_{iOPT} \end{cases}$$

[0100] 여기서, α 는 생성된 난수이다.

[0101] 메모리 할당부는 난수가 제1 유형의 메모리에 할당할 페이지의 비율 값 이하인 경우, 제1 유형의 메모리를 목표 노드로 선택할 수 있다(631). 메모리 할당부는 난수가 제1 유형 내지 제k-1 유형의 메모리에 할당할 페이지의

비율을 합산한 값을 초과하고, 제1 유형 내지 제k 유형의 메모리에 할당할 페이지의 비율을 합산한 값 이하인 경우(630), 제k 유형의 메모리를 목표 노드로 선택할 수 있다(632).

[0102] 메모리 할당부는 선택된 목표 노드에 대하여, 빈 공간이 있는지를 판단할 수 있다(640). 메모리 할당부는 목표 노드에 빈 공간이 있을 경우, 목표 노드에 페이지를 할당할 수 있다(651). 메모리 할당부는 목표 노드에 빈 공간이 없을 경우, 선택되지 않은 메모리 노드에 빈 공간이 있는지를 판단할 수 있다(650). 메모리 노드에 빈 공간이 있는지를 판단하는 과정에 있어서, $z=2, 3, \dots, n$ 에 대하여, 제 $z-1$ 유형의 메모리의 제 $z-1$ 지연시간은 제 z 유형의 메모리의 제 z 지연시간보다 짧고, 제 $z-1$ 유형의 메모리의 제 $z-1$ 대역폭은 제 z 유형의 메모리의 제 z 대역폭보다 크기 때문에, 제1 유형의 메모리부터 순서대로 빈 공간이 있는지를 판단할 수 있다. 메모리 할당부는 제 m 유형의 메모리 노드에 빈 공간이 있음이 확인된 경우, 제 m 유형의 메모리 노드에 페이지를 할당할 수 있다(652). 메모리 할당부는 선택되지 않은 메모리 노드에 빈 공간이 없을 경우, 목표 노드에 빈 공간이 있는지를 다시 판단할 수 있다(660).

[0103] 페이지 할당 비율은 각각의 메모리에 할당되는 페이지의 비율일 수 있다. 따라서, 메모리 할당부는 0과 1 사이의 난수를 생성하여 난수와 할당 비율 조절부에서 결정된 할당 비율을 비교함으로써 메모리에 확률적으로 페이지를 할당할 수 있다.

[0105] 이하에서, 전술한 동작들을 도 2의 실시예에 적용하여 설명한다.

[0106] 메모리 지연 비율 r_L 은 수학식 10에 의하여 결정된다.

수학식 10

$$r_L = \frac{\frac{B_G}{B_G+B_C} \cdot L_G + \frac{B_C}{B_G+B_C} \cdot L_C}{L_G}$$

[0107]

[0108] 이 때, $L_G = L_{G,L1D} + m_{G,L1D} \cdot (L_{G,L2} + m_{G,L2} \cdot L_{G,M})$ 이고,

$L_C = L_{C,M}$ 일 수 있다. 여기서, B_G 는 제1 유형의 메모리의 제1 대역폭이고, B_C 는 제2 유형의 메모리의 제2 대역폭이며, L_G 는 제1 지연시간이고, $L_{G,L1D}$ 는 제1 유형의 프로세서를 위한 L1 캐시의 지연시간이며, $L_{G,L2D}$ 는 제1 유형의 프로세서를 위한 L2 캐시의 지연시간이고, $L_{G,M}$ 은 제1 유형의 프로세서를 위한 물리적 메모리의 지연시간이며, $m_{G,L1D}$ 는 제1 유형의 프로세서를 위한 L1 캐시의 실패율이고, $m_{G,L2D}$ 는 제1 유형의 프로세서를 위한 L2 캐시의 실패율이다.

[0110] 메모리 지연 비율이 결정되면, 페이지 할당 비율 p_{OPT} 는 수학식 11에 의하여 결정된다.

수학식 11

$$p_{OPT} = \begin{cases} p_{OPT,S} & \text{if } r_L < \theta_L \wedge B_M > \theta_B \\ p_{OPT,I} & \text{otherwise} \end{cases}$$

[0111]

$$p_{OPT,S} = \frac{B_G}{B_G + B_C} \quad \text{이고,} \quad p_{OPT,I} = 1 \quad \text{일 수 있다.}$$

[0112]

이 때,

[0113] 여기서, $p_{OPT,S}$ 는 제1 정책이 선택된 경우에 제1 유형의 메모리에 할당할 페이지의 비율이고, $p_{OPT,I}$ 는 제2 정책이 선택된 경우에 제1 유형의 메모리에 할당할 페이지의 비율이며, r_L 은 메모리 지연 비율이고, B_M 은 데이터 업무량이며, θ_L 은 메모리 지연 비율에 대하여 미리 정해진 기준값이고, θ_B 는 데이터 업무량에 대하여 미리 정해진 기준값이다.

[0115] 페이지 할당 비율이 결정되면, 이중 메모리에 확률적으로 페이지를 할당하는 동작은 표 1의 수도 코드를 통하여 수행될 수 있다.

표 1

```

1: procedure BLPPALLOC
2:   node  $\leftarrow$  initialNode
3:    $r \leftarrow$  getRandom(0, 1)
4:   if  $r < p_{OPT}$  then
5:     | node  $\leftarrow$  nodeGPU
6:   else
7:     | node  $\leftarrow$  nodeCPU
8:   end if
9:   if node.isFull = true then
10:    | otherNode  $\leftarrow$  getOtherNode(node)
11:    | if otherNode.isFull = false then
12:      | node  $\leftarrow$  otherNode
13:    | end if
14:  end if
15:  page  $\leftarrow$  allocPage(node)
16:  return page
17: end procedure

```

[0117] 여기서, r 은 0과 1 사이의 난수이고, 페이지 할당 비율 p_{OPT} 는 페이지 할당 비율이다.

[0119] 일실시예에 따른 이중의 메모리들을 이용하는 페이지징 방법은 지연 시간 및 데이터 업무량에 기초하여 결정된 할당 정책에 따라 페이지를 할당할 메모리를 결정함으로써 상황에 따라 유동적으로 하나 또는 복수 개의 메모리에 페이지를 할당할 수 있다. 일실시예에 따른 이중의 메모리들을 이용하는 페이지징 방법은 결정된 정책에 따라 이중의 메모리의 대역폭과 관련된 특성에 기초하여 페이지 할당 비율을 결정함으로써 정보를 처리하는 데 소요되는 시간을 최소화할 수 있다.

[0121] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나

이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0122] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

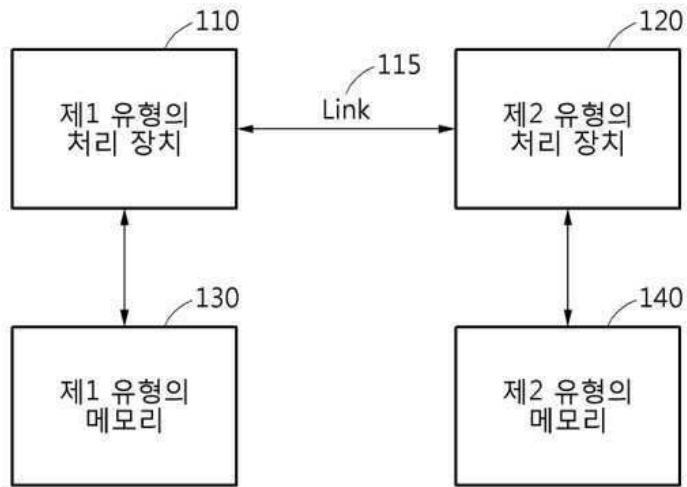
[0123] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic 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 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0124] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

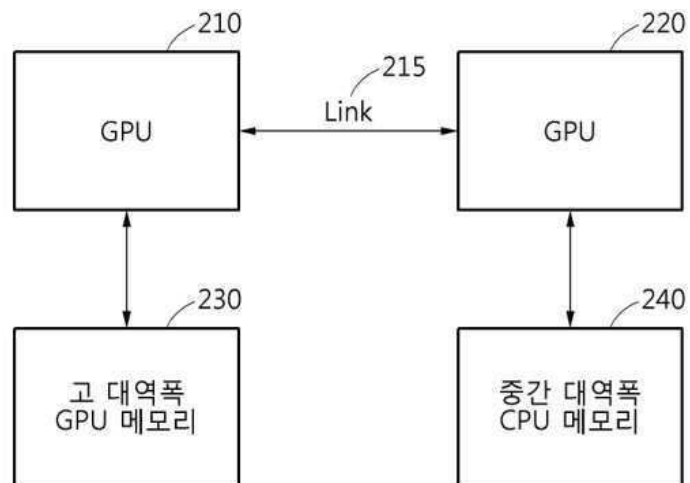
[0125] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

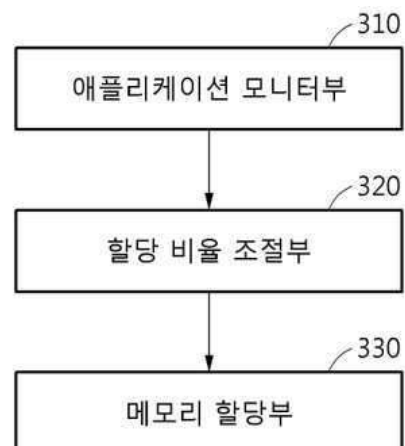
도면1



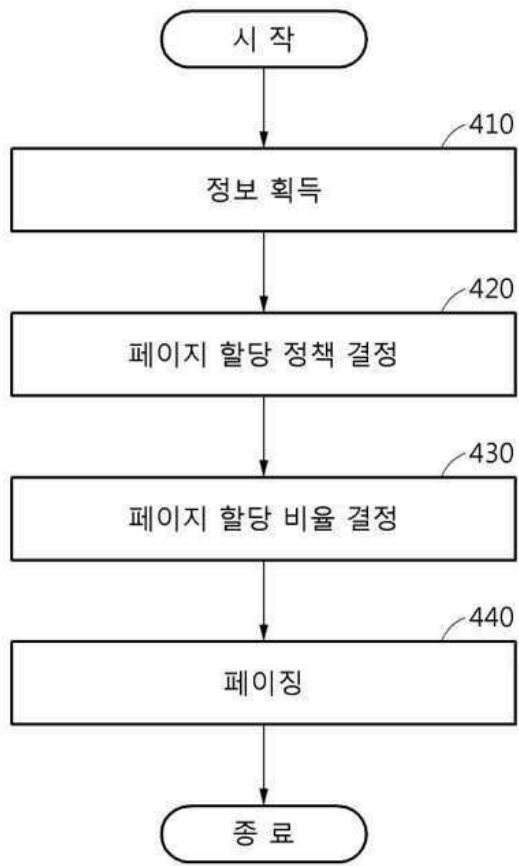
도면2



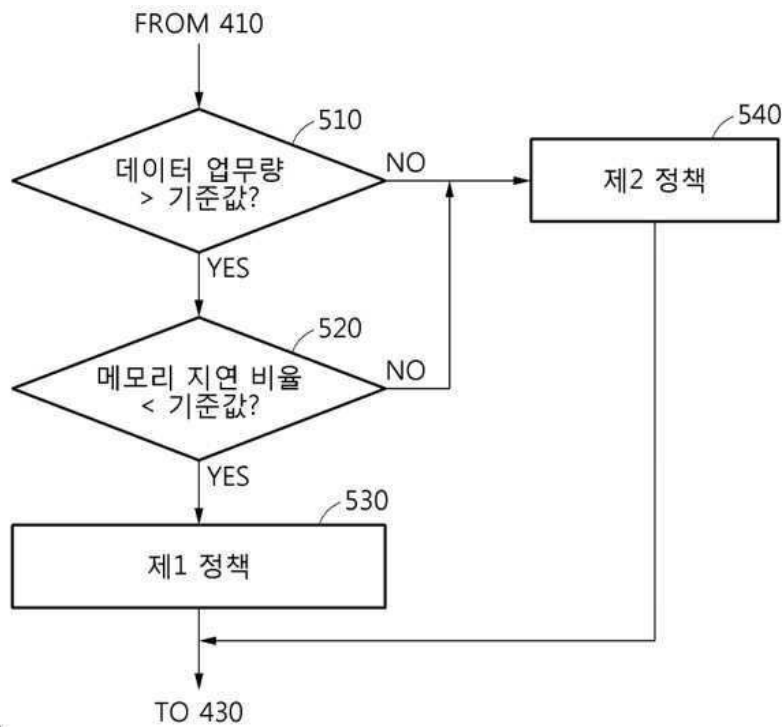
도면3



도면4



도면5



도면6

