(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0137356
(43) 공개일자 2023년10월04일

(51) 국제특허분류(Int. Cl.)
G06F 7/483 (2006.01) *G06N 3/063* (2023.01)
(52) CPC특허분류
G06F 7/483 (2021.08)
G06N 3/063 (2013.01)
(21) 출원번호 10-2023-7027167
(22) 출원일자(국제) 2022년01월20일
심사청구일자 없음
(85) 번역문제출일자 2023년08월09일
(86) 국제출원번호 PCT/US2022/013086
(87) 국제공개번호 WO 2022/173572
국제공개일자 2022년08월18일
(30) 우선권주장
63/148,086 2021년02월10일 미국(US)
17/361,263 2021년06월28일 미국(US)

(71) 출원인
마이크로소프트 테크놀로지 라이센싱, 엘엘씨
미국 워싱턴주 (우편번호 : 98052) 레드몬드 원 마이크로소프트 웨이
(72) 발명자
다르비쉬 로우하니 비타
미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이마이크로소프트 테크놀로지 라이센싱, 엘엘씨
엘란고 벤무길
미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이마이크로소프트 테크놀로지 라이센싱, 엘엘씨
(뒷면에 계속)
(74) 대리인
제일특허법인(유)

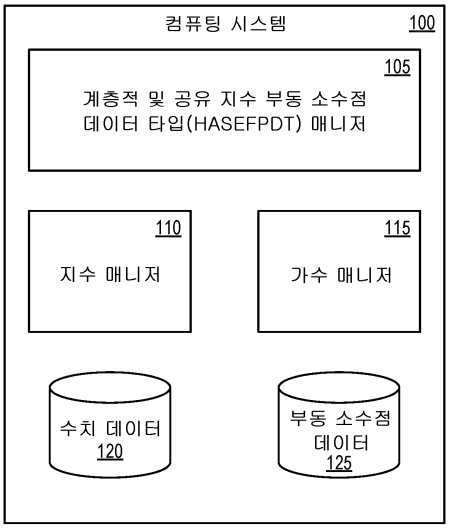
전체 청구항 수 : 총 15 항

(54) 발명의 명칭 **계층적 및 공유 지수 부동 소수점 데이터 타입**

(57) 요 약

본 개시의 실시예는 계층적 및 공유 지수 부동 소수점 데이터 타입을 제공하기 위한 시스템 및 방법을 포함한다. 제1 및 제2 공유 지수 값은 복수의 부동 소수점 값의 지수 값에 기초하여 결정된다. 제1 공유 지수값과 제2 공유 지수값에 기초하여 제3 공유 지수값이 결정된다. 제1 공유 지수값, 제2 공유 지수값, 제3 공유 지수값에 기초하여 제1 차이 값과 제2 차이 값이 결정된다. 복수의 부동 소수점 값에 대해 부호 값 및 가수 값이 결정된다. 복수의 부동 소수점 값 내의 각각의 부동 소수점 값에 대한 부호 값 및 가수 값, 제3 공유 지수 값, 제1 차이 값 및 제2 차이 값은 공유 지수 부동 소수점 데이터 타입에 대한 데이터 구조에 저장된다.

대 표 도 - 도1



(72) 발명자

샤피포어 라소울

미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이마이크로소프트 테크놀로지 라이센싱, 엘엘씨

포워스 제레미

미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이마이크로소프트 테크놀로지 라이센싱, 엘엘씨

리우 밍 강

미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이마이크로소프트 테크놀로지 라이센싱, 엘엘씨

시 진웬

미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이마이크로소프트 테크놀로지 라이센싱, 엘엘씨

버거 더글라스 씨

미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이마이크로소프트 테크놀로지 라이센싱, 엘엘씨

충 에릭 에스

미국 워싱턴주 98052-6399 레드몬드 원 마이크로소프트 웨이마이크로소프트 테크놀로지 라이센싱, 엘엘씨

명 세 서

청구범위

청구항 1

디바이스의 적어도 하나의 처리 유닛에 의해 실행 가능한 프로그램을 저장하는 머신 판독 가능 매체로서,

상기 프로그램은,

복수의 부동 소수점 값의 제1 서브세트에 있는 각 부동 소수점 값의 지수 값에 기초하여 제1 공유 지수 값을 결정하는 것과,

복수의 부동 소수점 값의 제2 서브세트에 있는 각 부동 소수점 값의 지수 값에 기초하여 제2 공유 지수 값을 결정하는 것과,

상기 제1 공유 지수 값 및 상기 제2 공유 지수 값에 기초하여 제3 공유 지수 값을 결정하는 것과,

상기 제1 공유 지수 값 및 상기 제3 공유 지수 값에 기초하여 제1 차이 값을 결정하는 것과,

상기 제2 공유 지수 값 및 상기 제3 공유 지수 값에 기초하여 제2 차이 값을 결정하는 것과,

상기 복수의 부동 소수점 값의 각 부동 소수점 값에 대한 부호 값 및 가수 값을 결정하는 것과,

공유 지수 부동 소수점 데이터 타입에 대한 데이터 구조에, 상기 복수의 부동 소수점 값의 각 부동 소수점 값에 대한 상기 부호 값 및 상기 가수 값, 상기 제3 공유 지수 값, 상기 제1 차이 값, 및 상기 제2 차이 값을 저장하는 것을 위한 명령어 세트를 포함하되,

신경망에 상기 공유 지수 부동 소수점 데이터를 사용하면 상기 신경망의 효율성이 향상되는,

머신 판독 가능 매체.

청구항 2

제1항에 있어서,

상기 제1 공유 지수 값, 상기 제2 공유 지수 값 및 상기 제3 공유 지수 값은 공유 지수 값의 계층 구조를 형성하는,

머신 판독 가능 매체.

청구항 3

제2항에 있어서,

상기 제1 공유 지수 값 및 상기 제2 공유 지수 값은 상기 공유 지수 값의 계층 구조에서 제1 레벨 지수 값이고, 상기 제3 공유 지수 값은 상기 공유 지수 값의 계층 구조에서 제2 레벨 지수 값인,

머신 판독 가능 매체.

청구항 4

제2항에 있어서,

상기 공유 지수 값의 계층 구조는 상기 제3 공유 지수 값, 상기 제1 차이 값 및 상기 제2 차이 값으로 표현되는,

머신 판독 가능 매체.

청구항 5

제1항에 있어서,

상기 프로그램은,

상기 제1 공유 지수 값 및 상기 복수의 부동 소수점 값의 상기 제1 서브세트에 있는 각 부동 소수점 값의 지수 값에 기초하여, 제1 차이 값 세트를 결정하는 것과,

상기 제2 공유 지수 값 및 상기 복수의 부동 소수점 값의 상기 제2 서브세트에 있는 각 부동 소수점 값의 지수 값에 기초하여, 제2 차이 값 세트를 결정하는 것과,

상기 공유 지수 부동 소수점 데이터 타입에 대한 데이터 구조에, 상기 제1 차이 값 세트 및 상기 제2 차이 값 세트를 저장하는 것을 위한 명령어 세트를 더 포함하는,

머신 판독 가능 매체.

청구항 6

제1항에 있어서,

상기 제1 공유 지수 값을 결정하는 것은 상기 복수의 부동 소수점 값의 상기 제1 서브세트에서 가장 높은 값을 갖는 부동 소수점 값의 지수 값을 결정하고, 상기 지수 값을 상기 제1 공유 지수 값으로 사용하는 것을 포함하는,

머신 판독 가능 매체.

청구항 7

제1항에 있어서,

상기 제2 공유 지수 값을 결정하는 것은 상기 복수의 부동 소수점 값의 상기 제2 서브세트에서 가장 높은 값을 갖는 부동 소수점 값의 지수 값을 결정하고, 상기 지수 값을 상기 제2 공유 지수 값으로 사용하는 것을 포함하는,

머신 판독 가능 매체.

청구항 8

제1항에 있어서,

상기 제3 공유 지수 값을 결정하는 것은, 상기 제1 공유 지수 값이 상기 제2 공유 지수 값보다 크거나 같다고 판단되면, 상기 제1 공유 지수 값을 상기 제3 지수 값으로 사용하는 것과, 상기 제1 공유 지수 값이 상기 제2 공유 지수 값보다 작다고 판단되면, 상기 제2 공유 지수 값을 상기 제3 지수 값으로 사용하는 것을 포함하는,

머신 판독 가능 매체.

청구항 9

제1항에 있어서,

상기 공유 지수 부동 소수점 데이터 타입에 대한 데이터 구조에 저장된 상기 가수 값을 나타내기 위해 사용되는 비트의 수는 상기 제3 지수 값을 나타내기 위해 사용되는 비트의 수보다 작은,

머신 판독 가능 매체.

청구항 10

방법으로서,

복수의 부동 소수점 값의 제1 서브세트에 있는 각 부동 소수점 값의 지수 값에 기초하여 제1 공유 지수 값을 결정하는 단계와,

복수의 부동 소수점 값의 제2 서브세트에 있는 각 부동 소수점 값의 지수 값에 기초하여 제2 공유 지수 값을 결정하는 단계와,

상기 제1 공유 지수 값 및 상기 제2 공유 지수 값에 기초하여 제3 공유 지수 값을 결정하는 단계와,

상기 제1 공유 지수 값 및 상기 제3 공유 지수 값에 기초하여 제1 차이 값을 결정하는 단계와,

상기 제2 공유 지수 값 및 상기 제3 공유 지수 값에 기초하여 제2 차이 값을 결정하는 단계와,

상기 복수의 부동 소수점 값의 각 부동 소수점 값에 대한 부호 값 및 가수 값을 결정하는 단계와,

공유 지수 부동 소수점 데이터 타입에 대한 데이터 구조에, 상기 복수의 부동 소수점 값의 각 부동 소수점 값에 대한 상기 부호 값 및 상기 가수 값, 상기 제3 공유 지수 값, 상기 제1 차이 값, 및 상기 제2 차이 값을 저장하는 단계를 포함하되,

신경망에 상기 공유 지수 부동 소수점 데이터를 사용하면 상기 신경망의 효율성이 향상되는,

방법.

청구항 11

제10항에 있어서,

상기 제1 공유 지수 값, 상기 제2 공유 지수 값 및 상기 제3 공유 지수 값은 공유 지수 값의 계층 구조를 형성하는,

방법.

청구항 12

제11항에 있어서,

상기 제1 공유 지수 값 및 상기 제2 공유 지수 값은 상기 공유 지수 값의 계층 구조에서 제1 레벨 지수 값이고, 상기 제3 공유 지수 값은 상기 공유 지수 값의 계층 구조에서 제2 레벨 지수 값인,

방법.

청구항 13

제11항에 있어서,

상기 공유 지수 값의 계층 구조는 상기 제3 공유 지수 값, 상기 제1 차이 값 및 상기 제2 차이 값으로 표현되는,

방법.

청구항 14

제10항에 있어서,

상기 제1 공유 지수 값 및 상기 복수의 부동 소수점 값의 상기 제1 서브세트에 있는 각 부동 소수점 값의 지수 값에 기초하여, 제1 차이 값 세트를 결정하는 단계와,

상기 제2 공유 지수 값 및 상기 복수의 부동 소수점 값의 상기 제2 서브세트에 있는 각 부동 소수점 값의 지수 값에 기초하여, 제2 차이 값 세트를 결정하는 단계와,

상기 공유 지수 부동 소수점 데이터 타입에 대한 데이터 구조에, 상기 제1 차이 값 세트 및 상기 제2 차이 값 세트를 저장하는 단계를 더 포함하는,

방법.

청구항 15

시스템으로서,

처리 유닛 세트와,

명령어를 저장하는 머신 판독가능 매체를 포함하되,

상기 명령어는 상기 처리 유닛 세트 내의 적어도 하나의 처리 유닛에 의해 실행되는 경우, 상기 적어도 하나의

처리 유닛으로 하여금,

복수의 부동 소수점 값의 제1 서브세트에 있는 각 부동 소수점 값의 지수 값에 기초하여 제1 공유 지수 값을 결정하는 것과,

복수의 부동 소수점 값의 제2 서브세트에 있는 각 부동 소수점 값의 지수 값에 기초하여 제2 공유 지수 값을 결정하는 것과,

상기 제1 공유 지수 값 및 상기 제2 공유 지수 값에 기초하여 제3 공유 지수 값을 결정하는 것과,

상기 제1 공유 지수 값 및 상기 제3 공유 지수 값에 기초하여 제1 차이 값을 결정하는 것과,

상기 제2 공유 지수 값 및 상기 제3 공유 지수 값에 기초하여 제2 차이 값을 결정하는 것과,

상기 복수의 부동 소수점 값의 각 부동 소수점 값에 대한 부호 값 및 가수 값을 결정하는 것과,

공유 지수 부동 소수점 데이터 타입에 대한 데이터 구조에, 상기 복수의 부동 소수점 값의 각 부동 소수점 값에 대한 상기 부호 값 및 상기 가수 값, 상기 제3 공유 지수 값, 상기 제1 차이 값, 및 상기 제2 차이 값을 저장하는 것을 수행하게 하고,

신경망에 상기 공유 지수 부동 소수점 데이터를 사용하면 상기 신경망의 효율성이 향상되는,

시스템.

발명의 설명

기 술 분 야

배 경 기 술

[0001] 본 개시는 컴퓨팅 하드웨어에 관한 것이다. 보다 상세하게, 본 개시는 부동 소수점 데이터 타입에 관한 것이다.

[0002] 신경망은 다양한 상이한 애플리케이션(예컨대, 이미지 분류, 컴퓨터 비전, 자연어 처리, 음성 인식, 필기 인식 등)에 사용되는 머신 러닝 모델이다. 신경망은 신경망을 통해 데이터세트를 실행하고, 신경망으로부터의 결과를 알려진 결과와 비교하고, 그 차이에 기초하여 신경망을 업데이트함으로써 특정 목적을 위해 훈련될 수 있다.

[0003] 신경망을 효율적으로 훈련하고, 충실도가 낮은 데이터 타입에서 추론을 위해 신경망을 사용하려면, 컴퓨팅 비용을 최소화하면서 각 비트의 충실도를 최대화하는 데이터 타입을 개발해야 할 수 있다. 이는 하드웨어 내적 유닛(hardware dot-product units)의 면적 오버헤드를 최소화하면서 양자화 신호 대 잡음비(QSNR) 메트릭을 최대화하는 것을 목표로 하는 최적화 문제로 공식화될 수 있다.

도면의 간단한 설명

[0004] 본 개시의 다양한 실시예들이 첨부된 도면에서 제한이 아닌 예시로 설명되어 있다.

도 1은 일부 실시예에 따른 컴퓨팅 시스템을 도시한 도면이다.

도 2는 일부 실시예에 따라 부동 소수점 값을 계층적 및 공유 지수 부동 소수점 데이터 타입(hierarchical and shared exponent floating point data type)으로 변환하는 일 실시예를 예시한다.

도 3은 일부 실시예에 따라 도 2에 도시된 계층적 및 공유 지수 부동 소수점 데이터 타입을 저장하기 위한 예시적인 데이터 구조를 도시한다.

도 4는 일부 실시예에 따라 부동 소수점 값을 계층적 및 공유 지수 부동 소수점 데이터 타입으로 변환하는 또 다른 예를 도시한다.

도 5는 일부 실시예에 따라 도 4에 도시된 계층적 및 공유 지수 부동 소수점 데이터 타입을 저장하기 위한 예시적인 데이터 구조를 도시한다.

도 6은 일부 실시예에 따른 계층적 및 공유 지수 부동 소수점 데이터 타입 변수를 곱하는 예를 도시한다.

도 7은 일부 실시예에 따른 계층적 및 공유 지수 부동 소수점 데이터 타입 변수를 생성하는 프로세스를 도시한다.

도 8은 일부 실시예에 따른 예시적인 컴퓨터 시스템의 단순화된 블록도이다.

도 9는 일부 실시예에 따른 신경망 처리 시스템을 도시한다.

발명을 실시하기 위한 구체적인 내용

[0005] 이하의 설명에서는, 본 개시의 완전한 이해를 제공하기 위해, 설명을 목적으로, 수 많은 실시예 및 구체적인 세부 사항들이 제시된다. 이러한 실시예 및 상세한 설명은 청구범위의 요소 또는 청구범위 전체를 부당하게 제한하는 것으로 해석되어서는 안 된다. 다양한 청구항의 언어에 기초하여, 당업자에게는 청구된 주제가 이들 실시예에 기재된 특징의 일부 또는 전부를 단독으로 또는 조합하여 포함할 수 있고, 본 명세서에 기재된 특징 및 기술의 수정 및 균등물을 더 포함할 수 있음이 명백할 것이다.

[0006] 본 명세서에서는 계층적 및 공유 지수 부동 소수점 데이터 타입을 제공하기 위한 기술들이 설명된다. 일부 실시예에서, 컴퓨팅 시스템은 계층적 및 공유 지수 부동 소수점 데이터 타입 데이터를 생성 및 관리하도록 구성될 수 있다. 예를 들어, 컴퓨팅 시스템은 필요한 경우 부동 소수점 숫자(예컨대, 반정밀도(half precision) 부동 소수점 숫자, 단정밀도 부동 소수점 숫자, 배정밀도 부동 소수점 숫자 등)로 변환된 여러 숫자로부터 계층적 및 공유 지수 부동 소수점 데이터 타입 데이터를 생성할 수 있다. 이를 위해, 컴퓨팅 시스템은 먼저 각 부동 소수점 숫자의 지수 값을 결정한다. 그런 다음, 컴퓨팅 시스템은 공유 지수 값의 계층 구조를 기반으로 이러한 지수 값을 양자화한다. 다음으로, 컴퓨팅 시스템은 양자화된 지수 값을 기반으로 각 부동 소수점 숫자의 가수(mantissa)를 양자화한다. 마지막으로, 컴퓨팅 시스템은 공유 지수 값의 계층 구조와 양자화된 가수를 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 데이터 구조에 저장한다.

[0007] 일부 실시예에서, 매우 좁은 비트 폭(예컨대, 가수의 1-4비트)에서 효율적인 훈련/추론을 수행하려면 컴퓨팅 비용을 최소화하면서 각 비트의 충실도를 최대화하는 데이터 타입을 개발해야 할 수 있다. 이는 하드웨어 내적 유닛의 면적 오버헤드를 최소화하면서 양자화 신호 대 잡음비(QSNR)를 최대화하는 것을 목표로 하는 최적화 문제로 공식화될 수 있다. 앞으로 설명하는 바와 같이, 본 명세서에 설명된 계층적 및 공유 지수 부동 소수점 데이터 타입은 다른 부동 소수점 데이터 타입에 비해, QSNR 대 곱셈 누적(MAC) 영역 파레토 프론티어를 더 넓히고 더 많은 효율을 끌어낼 수 있다. 이러한 초협소 비트폭 영역에서, 이러한 유효 QSNR의 증가는 신경망을 훈련하고 이를 추론에 사용하는 것에서 더 나은 달성 가능한 정확도로 직접적으로 해석될 수 있다.

[0008] 일부 실시예들에서, 본 명세서에 설명된 데이터 타입은 공유 지수 데이터 타입이다. 즉, 지수는 몇몇 요소들 사이에서 공유되고 있다. 요소(element)의 개수는 타일 크기 또는 블록 크기로 지칭될 수 있다. 작은 타일 크기를 사용하는 것이 하드웨어 관점에서 최선의 선택이 아닐 수 있는데, 이는 가수 비트의 수가 줄어들면 정수 곱셈기의 비용이 대부분 제거되고 누산기가 곱셈기-누산기(MAC) 유닛의 총 면적을 차지하기 시작하기 때문이다. 이와 같이, 상대적으로 큰 타일 크기를 사용하면, 타일의 개인 지수 및 타일의 공유 지수 사이의 차이로 인해 타일 내의 요소들이 0에 과도하게 매핑되는 것을 피하면서 누산기 실리콘 비용을 상각할 수 있다.

[0009] 일부 실시예에서, 본 명세서에 설명된 데이터 타입은 초협소 비트 폭 영역에서 허용 가능한 타일 크기와 인코딩 효율 사이의 균형을 제공하는 것을 목표로 한다 . 인코딩 효율은 다음 방정식에 의해 정의될 수 있다.

[0010] $$QSNR_{dB} = 20log_{10}\frac{||x||}{||x-Q(x)||} = 10log_{10}\frac{||x||^2}{||x-Q(x)||^2} = 10log_{10}\frac{\text{신호 에너지}}{\text{잡음 에너지}}$$

[0011] 여기서, $Q(\cdot): \mathfrak{R}^n \rightarrow \mathfrak{R}^n$ 은 n 차원 벡터(x)에 대해 제시된 양자화 방법이고, ||x||는 유클리드 규범이다. n은 계층적 및 공유 지수 부동 소수점 데이터 타입에서 타일 크기를 나타낸다. QSNR은 다수의 벡터(x)에 대해 전술한 비율을 평균함으로써 측정될 수 있다. 일부 실시예에서, 정의는 x가 $x=Q(y)*Q(z^T)$로 대체되는 MAC 유닛에 대한 신호 대 잡음비를 측정하도록 수정될 수 있으며, 여기서 y 및 z는 행렬 곱셈 유닛에 대한 입력을 나타낸다.

[0012] 일부 실시예에서, 계층적 및 공유 지수 부동 소수점 데이터 타입은 각 타일을 중첩된 서브 타일로 세분화하고, 각 서브 타일의 공유 지수가 그의 다음 레벨 서브 타일과 스케일 값만큼 다를 수 있게 한다. 이러한 스케일은 내적 유닛에 배치된 가산기 트리에서 오른쪽 이동으로 구현될 수 있다.

[0013] 일부 실시예에서, 본 명세서에 설명된 데이터 타입은 조정 가능한 파라미터를 다수 포함한다. 하나의 파라미터

는 타일 크기이며, 이는 전역 지수를 공유하는 요소의 수이다. 또 다른 파라미터는 가수 비트의 수이다. 일부 실시예에서, 가수 값은 2의 보수를 사용하여 표현되는 반면, 다른 실시예에서, 가수 값은 부호 크기 형식을 사용하여 표현된다. 가수 비트의 수는 1보다 크거나 동일한 임의의 값을 취할 수 있다.

[0014] 또 다른 파라미터는 공유 지수의 표현이다. 일부 실시예에서, 공유 지수는 단일 레벨을 사용하여 표현될 수 있다. 단일 레벨 표현의 경우, 하나의 파라미터는 서브 타일 크기이며, 이는 전역 지수를 조정하기 위해 로컬 스케일을 공유하는 요소의 수이다. 각 타일은 동일한 크기(m)의 여러 서브 타일로 분할될 수 있으며, 각 서브 타일은 공통 서브 스케일을 공유한다. m=1이면, 오른쪽 시프터가 내적 유닛에서 가산기 트리의 잎(leaves)에 추가될 수 있다. m 값이 클수록 오른쪽 시프터가 가산기 트리의 루트 쪽으로 이동한다. 예를 들어, m=16이면, 오른쪽 시프터가 가산기 트리의 $Log_2(16) = 4$번째 레벨에 추가된다. 단일 레벨 표현을 위한 또 다른 파라미터는 요소당 허용되는 오른쪽 시프트의 양을 지정하는 서브 타일 스케일이다. 스케일은 전역 공유 지수와 서브 타일 내에서의 선택된 지수 간에 허용되는 최대 차이를 인코딩한다. 예를 들어, 가산기 트리의 잎에서 스케일이 1이라는 것은 각 입력 요소가 0 또는 1만큼 시프트되어 2비트 조합 멀티플렉스를 생성할 수 있다는 것을 의미한다. 경우에 따라, 타일당 최대 지수를 전역 공유 지수로 사용할 수 있다. 따라서, 스케일은 항상 0보다 크거나 같은 양수이다. 일부 실시예에서는 다른 지수 선택 기법이 사용될 수 있다. 전역 공유 지수가 타일의 최대 지수가 아닌 경우, 선택된 스케일은 양수 값 또는 음수 값일 수 있다.

[0015] 다른 실시예에서, 공유 지수는 계층적 방식으로 표현된다. 공유 지수의 스케일을 조정하기 위해 큰 단일 레벨 시프트를 사용하는 대신, 시프트는 가산기 트리의 여러 레벨로 분산하고 대신 시프트당 스케일을 제한한다. 계층적 표현에서의 하나의 파라미터는 계층적 레벨이며, 이는 공유 지수를 조정하는 데 사용할 가산기 트리의 계층적 레벨의 수이다. 인덱싱은 트리의 잎에서부터 시작된다. 예를 들어 타일 크기가 32인 경우, 가산기 트리에는 최대 5개의 레벨이 있다. 3의 계층적 레벨은 잎에서부터 시작하는 가산기 트리의 처음 3개의 레벨 내에서 오른쪽 시프트를 사용한다는 의미한다. 계층적 표현을 위한 또 다른 파라미터는 각 레벨에서 허용되는 요소당 최대 스케일인 서브 타일 스케일이다. 스케일은 다음 레벨 공유 지수(반드시 전역 지수일 필요는 없음)와의 차이를 인코딩한다. 공유 지수는 최대 지수일 수도 있고 아닐 수도 있다. 예를 들어, 타일의 최대 및 최소 지수가 계산되는 경우, 최대와 최소의 차이가 가장 큰 스케일 차이를 제공한다. 계층적 레벨이 5로 설정된 경우 이(최대-최소) 지수 차이는 5로 제한될 수 있으며(서브 타일링의 레벨이 5개이므로) 전역 지수는 (최대-최소).클립(5)/2로 설정될 수 있다. 이는 타일을 계층적으로 서브 타일로 분할하고, 각 레벨에서 각 서브 타일에 대해, 서브 타일의 지수에 따라 -1/0/1의 스케일이 설정되어야 하는지 여부를 확인한다.

[0016] 누산기 비트 폭 파라미터는 가산기 트리의 끝에서 누산에 사용되는 데이터 타입을 지정한다. 이러한 데이터 타입의 예는 반정밀도 부동 소수점 데이터 타입, 단정밀도 부동 소수점 데이터 타입, 배정밀도 부동 소수점 데이터 타입 등을 포함한다. 일부 실시예에서, 기본 데이터 타입은 단정밀도 부동 소수점 값일 수 있다. 절삭 비트 폭은 또 다른 파라미터일 수 있다. 이 파라미터는 영역 오버헤드를 효과적으로 줄일 수 있는, 각 정수 추가 이후의 절삭의 양을 지정한다. 효과적인 절삭 체계의 예는 가산기 트리의 잎에 있는 정수 곱셈기의 출력을 8비트 가수로 잘라내고, 가산기 트리의 각 레벨에서 정수 곱셈기 출력의 비트 폭을 1씩 증가시키는 방식이 있다. 예를 들어, 이러한 접근 방식은 달성 가능한 QSNR(즉, 정확도)에 최소한의 영향을 미치면서 면적 오버헤드를 감소시킬 수 있다.

[0017] 도 1은 일부 실시예에 따른 컴퓨팅 시스템(100)을 예시한다. 도시된 바와 같이, 컴퓨팅 시스템(100)은 계층적 및 공유 지수 부동 소수점 데이터 타입(HASEFPDT) 매니저(105), 지수 매니저(110), 가수 매니저(115), 수치 데이터 저장소(120) 및 부동 소수점 데이터 저장소(125)를 포함한다. 수치 데이터 저장소(120)는 계층적 및 공유 지수 부동 소수점 데이터 타입으로 변환될 수 있는 수치 데이터를 저장한다. 수치 데이터의 예는 부동 소수점 숫자, 정수 숫자 등이 있다. 부동 소수점 데이터 저장소(125)는 계층적 및 공유 지수 부동 소수점 데이터 타입을 저장하도록 구성된다. 이러한 데이터 타입의 예는 아래에서 상세히 설명될 것이다. 또한, 부동 소수점 데이터 저장소(125)는 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 정의를 저장할 수 있다.

[0018] HASEFPDT 매니저(105)는 계층적 및 공유 지수 부동 소수점 데이터 타입과 이러한 데이터 타입에 따라 저장된 데이터를 관리할 수 있다. 예를 들어, HASEFPDT 매니저(105)는 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 정의를 부동 소수점 데이터 저장소(125)로부터 검색하여 계층적 및 공유 지수 부동 소수점 데이터 타입으로 저장된 데이터를 생성할 수 있다. 다음으로, HASEFPDT 매니저(105)는 수치 데이터 저장소(120)로부터 수치 데이터를 검색하여 이 수치 데이터를 부동 소수점 숫자(예컨대, 반정밀도 부동 소수점 숫자, 단정밀도 부동 소수점 숫자, 배정밀도 부동 소수점 숫자 등)로 변환할 수 있다. HASEFPDT 매니저(105)가 검색하는 수치 데이터

의 수는 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 정의에 명시된 것에 기초할 수 있다. 그런 다음, HASEFPDT 매니저(105)는 지수 매니저(110)에 변환된 수치 데이터, 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 정의, 및 수치 데이터에 대한 공유 지수 값을 결정하기 위한 요청을 전송한다. HASEFPDT 매니저(105)가 지수 매니저(110)로부터 공유 지수 값을 수신하면, HASEFPDT 매니저(105)는 가수 매니저(115)에 수치 데이터, 공유 지수 값, 계층 및 공유 지수 부동 소수점 데이터 타입에 대한 정의, 및 수치 데이터에 대한 가수 값을 결정하기 위한 요청을 전송한다. HASFPDT 매니저(105)가 가수 매니저(115)로부터 가수 값을 수신하면, HASEFPDT 매니저(105)는 가수 값과 공유 지수 값을 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 데이터를 저장하도록 구성된 데이터 구조에 저장한다.

[0019] 또한, HASEFPDT 매니저(105)는 계층적 및 공유 지수 부동 소수점 데이터 타입 데이터로부터 값을 판독할 수도 있다. 예를 들어, HASEFPDT 매니저(105)는 계층적 및 공유 지수 부동 소수점 데이터 타입에 따라 (예를 들어, 계층적 및 공유 지수 부동 소수점 데이터 타입에 따라 데이터를 저장하기 위한 데이터 구조에) 저장된 데이터로부터 특정 값을 읽기 위한 요청을 (예를 들어, 컴퓨팅 시스템(100)에서 동작하는 컴포넌트로부터, 다른 컴퓨팅 시스템에서 동작하는 컴포넌트 등으로부터) 수신할 수 있다. 이에 대응하여, HASEFPDT 매니저(105)는 계층적 및 공유 지수 부동 소수점 데이터 타입 데이터로부터 요청된 값의 가수 값을 검색하고, 계층적 및 공유 지수 부동 소수점 데이터 타입 데이터에 저장된 공유 지수 값에 기초하여 요청된 값의 지수 값을 결정하고, 가수 값과 지수 값에 기초하여 요청된 값을 생성한다.

[0020] 지수 매니저(110)는 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 지수 값을 결정하도록 구성된다. 예를 들어, 지수 매니저(110)는 HASEFPDT 매니저(105)로부터 수치 데이터, 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 정의, 및 수치 데이터에 대한 지수 값을 결정하기 위한 요청을 수신할 수 있다. 이에 대한 응답으로, 지수 매니저(110)는 수치 데이터 및 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 정의에 기초하여 지수 값의 계층 구조를 결정한다. 그런 다음, 지수 매니저(110)는 지수 값의 계층 구조를 HASEFPDT 매니저(105)로 전송한다.

[0021] 가수 매니저(115)는 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 가수 값의 결정을 처리한다. 예를 들어, 가수 매니저(115)는 HASEFPDT 매니저(105)로부터 수치 데이터, 공유 지수 값, 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 정의, 및 수치 데이터에 대한 가수 값을 결정하기 위한 요청을 수신할 수 있다. 공유 지수 값과 계층 및 공유 지수 부동 소수점 데이터 타입에 대한 정의를 기반으로, 가수 관리자(115)는 수치 데이터에 대한 가수 값을 결정한다. 그런 다음, 가수 관리자(115)는 가수 값을 HASEFPDT 관리자(105)로 전송한다.

[0022] 도 2는 일부 실시예에 따라 부동 소수점 값을 계층적 및 공유 지수 부동 소수점 데이터 타입으로 변환하는 일례를 도시한다. 구체적으로, 도 2는 부동 소수점 값을 계층적 및 공유 지수 부동 소수점 데이터 타입으로 변환하는 과정 전반에 걸쳐 다양한 값을 포함하는 테이블(200)을 예시한다. 이 예는 도 1을 참조하여 설명한다. 이 예에서는, 행(205)에 있는 4개의 값은 3개의 레벨을 갖는 공유 지수 값의 계층 구조를 사용하여 공유 지수 값을 저장하는 계층적 및 공유 지수 부동 소수점 데이터 타입으로 변환될 것이다. 표시된 바와 같이, 4개의 값은 십진수 값으로 표시된다. HASEFPDT 매니저(105)는 부동 소수점 데이터 저장소(125)로부터 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 정의를 검색하는 것으로 시작한다. 다음으로, HASEFPDT 매니저(105)는 수치 데이터 저장소(120)로부터 행(205) 내의 값을 검색하여 이들을 부동 소수점 값(예컨대, 반정밀도 부동 소수점 값, 단정밀도 부동 소수점 값, 배정밀도 부동 소수점 값 등)으로 변환한다. 그런 다음, HASEFPDT 매니저(105)는 지수 매니저(110)에게 부동 소수점 값, 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 정의, 및 부동 소수점 값에 대한 지수 값을 결정하기 위한 요청을 전송한다.

[0023] 부동 소수점 값, 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 정의, 및 요청을 수신하면, 지수 매니저(110)는 표(200)의 행(210)에 도시된 바와 같이 각 부동 소수점 값의 지수 값을 결정한다. 이러한 지수 값은 행(215)에 도시된 바와 같이, 이 예에서 레벨 0 지수 값으로 간주된다. 그런 다음, 지수 매니저(110)는 각 지수 값 쌍에 대해 레벨 1 공유 지수 값을 결정한다. 이 예에서, 지수 매니저(110)는 각 지수 값 쌍에서 가장 높은 값을 갖는 지수 값을 결정하고 이를 레벨 1 공유 지수 값으로 사용하는 방식으로 레벨 1 공유 지수 값을 결정한다. 행(220)에 도시된 바와 같이, 제1 값 쌍에 대한 지수 값들 사이의 가장 높은 지수 값은 -3이다. 제2 값 쌍에 대한 지수 값들 사이의 가장 높은 지수 값은 -2이다. 다음으로, 지수 매니저(110)는 레벨 0 지수 값과 대응하는 레벨 1 공유 지수 값의 각 쌍에 대해 레벨 1 스케일 값을 결정한다. 여기서, 지수 매니저(110)는 레벨 0 지수 값과 대응하는 레벨 1 공유 지수 값 사이의 차이를 결정하여 레벨 1 스케일 값을 결정한다. 그 차이가 1보다 크면, 지수 매니저(110)는 레벨 1 스케일 값을 1로 근사화한다.

[0024] 레벨 0 지수 값과 대응하는 레벨 1 공유 지수 값의 제1 및 제2 쌍, 및 제2 값에 대해, 지수 매니저(110)는 레벨 1 스케일 값을 각각 0으로 결정한다(즉, -3과 -3 사이의 차이는 0이다). 레벨 0 지수 값과 대응하는 레벨 1 공유 지수 값의 제3 쌍에 대해, 지수 매니저(110)는 레벨 0 지수 값과 레벨 1 공유 지수 값 사이의 차이를 2인 것으로 결정한다. 이 값은 1보다 크므로, 지수 매니저(110)는 레벨 1 스케일 값을 1로 근사화한다. 레벨 0 지수 값과 대응하는 레벨 1 공유 지수 값의 제4 쌍에 대해, 지수 매니저(110)는 레벨 1 스케일 값을 0으로 결정한다(즉, -2와 -2 사이의 차이는 0이다). 레벨 1 스케일 값은 표(200)의 행(225)에 도시되어 있다.

[0025] 레벨 1 공유 지수 값 및 스케일 값을 결정한 후, 지수 매니저(110)는 레벨 2 공유 지수 값을 결정한다. 이 예에서, 지수 매니저(110)는 레벨 1 공유 지수 값에 기초하여 레벨 2 공유 지수 값을 결정한다. 특히, 지수 매니저(110)는 레벨 1 공유 지수 값 중에서 가장 높은 값을 갖는 지수 값을 결정하여 레벨 2 공유 지수 값으로 사용하는 방식으로 레벨 2 공유 지수 값을 결정한다. 여기서, 지수 매니저(110)는 행(230)에 도시된 바와 같이 레벨 2 공유 지수 값을 -2로 결정한다.

[0026] 다음으로, 지수 매니저(110)는 레벨 1 공유 지수 값과 대응하는 레벨 2 공유 지수 값의 각 쌍에 대해 레벨 2 스케일 값을 결정한다. 이 예에서, 지수 매니저(110)는 레벨 1 공유 지수 값과 대응하는 레벨 2 공유 지수 값의 차이를 결정하여 레벨 2 스케일 값을 결정한다. 차이가 1보다 큰 경우, 지수 매니저(110)는 레벨 2 스케일 값을 1로 근사화한다. 레벨 1 공유 지수 값과 대응하는 레벨 2 공유 지수 값의 제1 쌍에 대해, 지수 매니저(110)는 레벨 2 스케일 값을 1인 것로 결정한다(즉, -2와 -3의 차이는 1이다). 레벨 1 공유 지수 값 및 대응하는 레벨 2 공유 지수 값의 제2 쌍에 대해, 지수 매니저(110)는 레벨 2 스케일 값을 0인 것으로 결정한다(즉, -2와 -2 사이의 차이는 0이다). 레벨 2 스케일 값은 행(235)에 표시된다. 행(240)은 레벨 1 스케일 값과 레벨 2 스케일 값의 합인 각 부동 소수점 값에 대한 총 스케일 값을 나타낸다. 행(245)은 각 부동 소수점 값에 대한 총 공유 지수 스케일 값을 나타내며, 이는 레벨 2 공유 지수 값에서 각각의 총 스케일 값을 뺌으로써 부동 소수점 값에 대해 결정된 지수 값이다. 행(250)은 각 부동 소수점 값에 대해 결정된 지수 값과 원(original) 지수 값 사이의 차이를 나타낸다. 행(255)은 부동 소수점 값에 대한 가수의 양자화된 크기를 보여준다. 이 예에서는, 2비트 마인티사가 사용되는데 제1 비트는 부호용으로 사용되고(표(200)에는 표시되지 않음) 제2 비트는 크기용으로 사용된다. 행(260)은 행(245)의 공유 지수 스케일 값과 동일한 각 부동 소수점 값에 대한 양자화된 지수 값을 보여준다. 끝으로, 행(265)은 원 부동 소수점 값을 계층적 및 공유 지수 부동 소수점 데이터 타입으로 변환한 후 양자화된 부동 소수점 값의 10진수 표현을 나타낸다.

[0027] 지수 매니저(110)가 레벨 2 공유 지수, 레벨 2 스케일 값 및 레벨 1 스케일 값을 결정한 후, 지수 매니저(110)는 이들을 HASEFPDT 매니저(105)로 전송한다. 그런 다음, HASEFPDT 매니저(105)는 행(205)의 부동 소수점 값, 레벨 2 공유 지수, 레벨 2 스케일 값, 레벨 1 스케일 값, 계층 및 공유 지수 부동 소수점 데이터 타입에 대한 정의, 부동 소수점 값의 가수 값을 결정하기 위한 요청을 가수 매니저(115)로 전송한다. 가수 매니저(115)가 데이터 및 요청을 수신하면, 가수 매니저(115)는 각 부동 소수점 값에 대한 가수 값을 결정한다. 지수 차이(행(250)에 도시됨)가 0보다 큰 경우, 가수 관리자(115)는 각 부동 소수점 값에 대한 가수의 크기를 0으로 양자화한다. 그렇지 않으면, 가수 관리자(115)는 사용 가능한 비트(이 예에서는 부호에 1비트, 크기에 1비트)와 양자화된 각 지수 값(행(260)에 도시됨)을 사용하여 원 부동 소수점 값에 가장 근사하게 접근한다. 여기서, 가수 매니저(115)는 각 부동 소수점 값에 대한 양자화된 크기 값을 행(255)에 도시된 값으로 결정한다. 다음으로, 가수 매니저(115)는 결정된 가수 값을 HASEFPDT 매니저(105)로 전송한다.

[0028] HASEFPDT 매니저(105)가 가수 매니저(115)로부터 가수 값을 수신하면, HASEFPDT 매니저(105)는 가수 값, 레벨 2 공유 지수, 레벨 2 스케일 값 및 레벨 1 스케일 값을, 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 데이터를 저장하도록 구성된 데이터 구조에 저장한다. 도 3은 일부 실시예에 따라 도 2에 도시된 계층적 및 공유 지수 부동 소수점 데이터 타입을 저장하기 위한 데이터 구조(300)를 예시적으로 도시한 것이다. 데이터 구조(300)는 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 정의에 기초하여 생성될 수 있다. 여기서, 데이터 타입에 대한 정의는 4개의 부동 소수점 값에 대한 4개의 2비트 가수 값, 4개의 1비트 레벨 1 스케일 값, 2개의 1비트 레벨 2 스케일 값, 및 8비트 레벨 2 공유 지수 값을 지정할 것이다. 도시된 바와 같이, 데이터 구조(300)는 4개의 속성, 즉 가수 속성, 레벨 1 스케일 속성, 레벨 2 스케일 속성 및 공유 지수 속성을 포함한다. 가수 속성은 각 부동 소수점 값에 대해 결정된 2비트 가수 값을 저장한다. 레벨 1 스케일 속성은 각 부동 소수점 값에 대해 결정된 1비트 레벨 1 스케일 값을 저장한다. 레벨 2 스케일 값 속성은 각 부동 소수점 값 쌍에 대해 결정된 1비트 레벨 2 스케일 값을 저장한다. 공유 지수 속성 값은 부동 소수점 값에 대해 결정된 8비트 레벨 2 공유 지수 값을 저장한다. 이 예에서, 레벨 2 공유 지수 값의 값은 127의 바이어스를 사용하여 표시된다. 따라서, 125(즉, 127 + (-2))의 값이 이 예의 공유 지수 속성에 저장된다.

[0029] 도 4는 일부 실시예에 따라 부동 소수점 값을 계층적 및 공유 지수 부동 소수점 데이터 타입으로 변환하는 또 다른 예를 예시한다. 특히, 도 4는 부동 소수점 값을 다른 계층적 및 공유 지수 부동 소수점 데이터 타입으로 변환하는 과정 전반에 걸쳐 다양한 값을 포함하는 테이블(400)을 예시한다. 이 예는 또한 도 1을 참조하여 설명될 것이다. 이 예제에서는 위의 예제에서 사용된 행(405)에 표시된 것과 동일한 4개의 값이 2개의 레벨을 갖는 공유 지수 값의 계층 구조를 사용하여 공유 지수 값을 저장하는 계층적 공유 지수 부동 소수점 데이터 타입으로 변환될 것이다. 이 예제는 HASEFPDT 매니저(105)가 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 정의를 부동 소수점 데이터 저장소(125)에서 검색하는 것으로 시작된다. 그런 다음, HASEFPDT 매니저(105)는 수치 데이터 저장소(120)에서 행(405)의 값을 불러와서 이들을 부동 소수점 값(예컨대, 반정밀도 부동 소수점 값, 단정밀도 부동 소수점 값, 배정밀도 부동 소수점 값 등)으로 변환한다. 다음으로, HASEFPDT 매니저(105) 지수 매니저(110)에게, 부동 소수점 값, 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 정의, 및 부동 소수점 값에 대한 지수 값을 결정하기 위한 요청을 전송한다.

[0030] 부동 소수점 값, 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 정의, 및 요청을 수신한 후, 지수 매니저(110)는 행(410)에 도시된 바와 같이 각 부동 소수점 값의 지수 값을 결정한다. 이 예에서, 지수 매니저(110)는 원 지수 값의 각 쌍에 대한 지수 값을 레벨 0 공유 지수로 공유한다. 이 예에서, 지수 매니저(110)는 원 지수 값의 각각의 쌍에서 가장 높은 값을 갖는 지수 값을 결정하고 이를 레벨 0 공유 지수 값으로 사용함으로써 레벨 0 공유 지수 값을 결정한다. 행(415)에 도시된 바와 같이, 제1 값 쌍의 지수 값들 사이에서 가장 높은 지수 값은 -3이고, 제2 값 쌍의 지수 값들 사이에서 가장 높은 지수 값은 -2이다. 그런 다음, 지수 매니저(110)는 레벨 0 공유 지수 값에 기초하여 레벨 1 공유 지수를 결정한다. 이 예에서, 지수 매니저(110)는 가장 높은 값을 갖는 레벨 0 공유 지수 값을 결정하고 이를 레벨 1 공유 지수 값으로 사용함으로써 레벨 1 공유 지수를 결정한다. 여기서, 지수 매니저(110)는 행(420)에 도시된 바와 같이 레벨 1 공유 지수 값이 -2임을 결정한다. 다음으로, 지수 매니저(110)는 레벨 0 공유 지수 값과 대응하는 레벨 1 공유 지수 값의 각 쌍에 대해 레벨 1 스케일 값을 결정한다. 이 예에서, 지수 매니저(110)는 레벨 0 공유 지수 값과 대응하는 레벨 1 공유 지수 값의 차이를 결정하여 레벨 1 스케일 값을 결정한다. 차이가 1보다 큰 경우, 지수 매니저(110)는 레벨 1 스케일 값을 1로 근사화한다.

[0031] 레벨 0 공유 지수 값과 대응하는 레벨 1 공유 지수 값의 제1 쌍에 대해, 지수 매니저(110)는 레벨 1 스케일 값을 1인 것으로 결정한다(즉, -2와 -3의 차이는 1이다). 레벨 0 공유 지수 값 및 대응하는 레벨 1 공유 지수 값의 제2 쌍에 대해, 지수 매니저(110)는 레벨 1 스케일 값을 0인 것으로 결정한다(즉, -2와 -2 사이의 차이는 0이다). 레벨 1 스케일 값은 행(425)에 표시된다. 행(430)은 레벨 1 스케일 값과 동일한 총 스케일 값을 나타낸다. 행(435)은 각 부동 소수점 값에 대한 총 공유 지수 스케일 값을 나타내며, 이는 레벨 1 공유 지수 값에서 제각기의 총 스케일 값을 뺌으로써 부동 소수점 값에 대해 결정된 지수 값이다. 행(440)은 각 부동 소수점 값에 대해 결정된 지수 값과 원 지수 값 사이의 차이를 나타낸다. 행(445)은 부동 소수점 값에 대한 가수의 양자화된 크기를 나타낸다. 이 예에서, 3비트 가수가 사용되는데, 제1 비트는 부호(표(400)에 표시되지 않음)에 대해 사용되고, 제2 비트는 크기에 대해 사용된다. 행(450)은 행(435)의 공유 지수 스케일 값과 동일한 각 부동 소수점 값에 대한 양자화된 지수 값을 표시한다. 그런 다음, 행(455)은 원 부동 소수점 값을 계층적 및 공유 지수 부동 소수점 데이터 타입으로 변환한 후 양자화된 부동 소수점 값의 10진수 표현을 나타낸다.

[0032] 레벨 1 공유 지수 및 레벨 1 스케일 값이 결정되면, 지수 매니저(110)는 이를 HASEFPDT 매니저(105)로 전송한다. HASEFPDT 매니저(105)는 행(405)의 부동 소수점 값, 레벨 1 공유 지수, 레벨 1 스케일 값, 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 정의, 및 부동 소수점 값에 대한 가수 값을 결정하기 위한 요청을 가수 매니저(115)로 전송한다. 가수 매니저(115)가 데이터 및 요청을 수신하면, 가수 매니저(115)는 각 부동 소수점 값에 대한 가수 값을 결정한다. 지수 차이(행(440)에 도시됨)가 1보다 큰 경우, 가수 매니저(115)는 각 부동 소수점 값에 대한 가수의 크기를 0으로 양자화한다. 그렇지 않으면, 가수 매니저(115)는 사용 가능한 비트(이 예에서는 부호에 1비트, 크기에 1비트)와 각각의 양자화된 지수 값(행(450)에 도시됨)을 사용하여 원 부동 소수점 값에 가장 근접한다. 가수 매니저(115)는 각 부동 소수점 값에 대한 양자화된 크기 값을 행(445)에 표시된 값인 것으로 결정한다. 그런 다음, 가수 매니저(115)는 결정된 가수 값을 HASEFPDT 매니저(105)로 전송한다.

[0033] HASEFPDT 매니저(105)가 가수 매니저(115)로부터 가수 값을 수신하면, HASEFPDT 매니저(105)는 계층적 및 공유 지수 부동 소수점 데이터 타입에 대한 데이터를 저장하도록 구성된 데이터 구조에 가수 값, 레벨 1 공유 지수, 및 레벨 1 스케일 값을 저장한다. 도 5는 일부 실시예에 따라 도 4에 도시된 계층적 및 공유 지수 부동 소수점 데이터 타입을 저장하기 위한 예시적인 데이터 구조(500)를 도시한이다. 데이터 구조(500)는 계층적 및 공유

지수 부동 소수점 데이터 타입에 대한 정의에 기초하여 생성될 수 있다. 이 예에서, 데이터 타입에 대한 정의는 4개의 부동 소수점 값에 대한 4개의 2비트 가수 값, 2개의 2비트 레벨 1 스케일 값, 및 8비트 레벨 1 공유 지수 값을 지정할 것이다. 도시된 바와 같이, 데이터 구조(500)는 3개의 속성, 즉 가수 속성, 레벨 1 스케일 속성, 및 공유 지수 속성을 포함한다. 가수 속성은 각 부동 소수점 값에 대해 결정된 2비트 가수 값을 저장한다. 레벨 1 스케일 속성은 각 부동 소수점 값에 대해 결정된 2비트 레벨 1 스케일 값을 저장한다. 공유 지수 속성 값은 부동 소수점 값에 대해 결정된 8비트 레벨 1 공유 지수 값을 저장한다. 이 예에서, 레벨 1 공유 지수 값의 값은 127의 바이어스를 사용하여 표시된다. 따라서, 125(즉, 127 + (-2))의 값은 이 예의 공유 지수 속성에 저장된다.

[0034] 도 1 내지 도 5를 참조하여 상술한 실시예들은 부동 소수점 숫자가 어떻게 두 개의 상이한 계층적 및 공유 지수 부동 소수점 데이터 타입으로 변환될 수 있는지를 예시한다. 당업자는 일부 실시예들에서 추가의 및/또는 다른 계층적 및 공유 지수 부동 소수점 데이터 타입이 사용될 수 있음을 인식할 것이다. 예를 들어, 계층적 및 공유 지수 부동 소수점 데이터 타입은 지수를 공유하는 상이한 수의 부동 소수점 값들을 나타낼 수 있고(상기 예들은 각각 4개의 부동 소수점 값들을 나타낸다), 상이한 수의 비트를 사용하여 가수 값, 레벨 스케일 값 및 공유 지수 값들을 나타내거나, 지수 값들의 계층 구조에서 상이한 수의 레벨(예를 들어, 스케일 값의 상이한 수의 레벨)을 사용할 수 있다.

[0035] 또한, 도 2-5를 참조하여 상술한 예들은 공유 지수 값의 계층 구조가 최고 레벨의 공유 지수 및 레벨 스케일 값으로 어떻게 표현될 수 있는지를 보여준다. 예를 들어, 도 2에 도시된 예는 레벨 0 지수 값, 레벨 1 공유 지수 값 및 레벨 2 공유 지수 값의 3가지 레벨을 가진 공유 지수의 계층 구조를 사용한다. 도 3에 표시된 것처럼, 공유 지수 값의 계층 구조는 레벨 2 공유 지수 값, 레벨 2 스케일 값, 및 레벨 1 스케일 값으로 표시된다. 부동 소수점 값의 양자화된 지수 값(행(260)의 값)은 레벨 2 공유 지수 값, 레벨 2 스케일 값, 및 레벨 1 스케일 값으로부터 결정될 수 있다. 다른 예로서, 도 4에 도시된 예는 레벨 0 지수 값과 레벨 1 공유 지수 값이라는 두 개의 서로 다른 레벨을 갖는 공유 지수의 계층 구조를 사용한다. 도 5에서 볼 수 있듯이, 공유 지수 값의 계층 구조는 레벨 1 공유 지수 값과 레벨 1 스케일 값으로 표현된다. 부동 소수점 값의 양자화된 지수 값(행(450)의 값)은 레벨 1 공유 지수 값 및 레벨 1 스케일 값으로부터 결정될 수 있다. 계층적 및 공유 지수 부동 소수점 데이터 타입에 따라 저장된 부동 소수점 값을 판독할 때(예를 들어, 행(265 및 455)에 도시된 것과 같은 양자화된 값을 결정할 때), HASEFPDT 매니저(105)는 이러한 방식으로 양자화된 지수 값을 결정한다.

[0036] 또한, 도 2-5를 참조하여 상술한 실시예들은 하위 레벨 지수/공유 지수들 중에서 가장 높은 값을 선택함으로써 공유 지수 값을 결정한다. 당업자는 특정 레벨에 대한 공유 지수 값을 결정하기 위해 추가의 및/또는 다른 통계적 메트릭이 사용될 수 있음을 이해할 것이다. 이러한 메트릭의 예로는 가중 평균, 평균에 정의된 개수의 표준 편차(예컨대, 하나의 표준 편차, 2개의 표준 편차, 3개의 표준 편차 등)를 더한 값 등이 있다.

[0037] 계층적 및 공유 지수 부동 소수점 데이터 타입의 예시적 적용

[0038] 상술한 예들은 계층적 및 공유 지수 부동 소수점 데이터 타입을 생성하는 방법을 설명한다. 다음 예는 이러한 데이터 타입의 많은 응용 중 하나를 예시한다. 도 6은 일부 실시예에 따른 계층적 및 공유 지수 부동 소수점 데이터 타입 변수를 곱하는 예시를 도시한다. 구체적으로, 도 6은 계층적 및 공유 지수 부동 소수점 데이터 타입에 따라 저장된 두 변수 사이의 내적(dot product)을 수행하기 위한 하드웨어 아키텍처(600)를 예시한다. 일부 실시예에서, 하드웨어 아키텍처(600)는 신경망을 훈련시키고 추론을 위해 신경망을 사용하는 데 사용되는 인공지능(AI) 가속기를 구현하는 데 사용될 수 있다. 도시된 바와 같이, 하드웨어 아키텍처(600)는 곱셈기(605-1-605-n), 시프터(610-1-610-n 및 620-1-620-m), 가산기(615-1-615-m, 625-1-625-k 및 630) 및 절삭/반올림 모듈(635)을 포함한다.

[0039] 이 예에서, 하드웨어 아키텍처(600)는 두 개의 입력 A 및 B를 수신하도록 구성된다. 입력(A 및 B) 각각은 n 개의 부동 소수점 값(예컨대, 데이터 구조(300), 데이터 구조(500) 등)을 저장하는 계층적 및 공유 지수 부동 소수점 데이터 타입 변수이다. 각 변수에 있는 부동 소수점 값의 가수 값은 대응하는 곱셈기(605)에 입력된다. 예를 들어, 변수 A의 제1 부동 소수점 값의 가수 값과 변수 B의 제1 부동 소수점 값의 가수 값이 곱셈기(605-1)에 입력되고, 변수 A의 제2 부동 소수점 값의 가수 값과 변수 B의 제2 부동 소수점 값의 가수 값이 곱셈기(605-2)에 입력되고, 변수 A의 제3 부동 소수점 값의 가수 값과 변수 B의 제3 부동 소수점 값의 가수 값이 곱셈기(605-3)에 입력되는 등의 예가 있을 수 있다. 도시된 바와 같이, 각 곱셈기(605)에 입력되는 입력은 비트 폭이 dw이다. 각 곱셈기(605)는 입력된 두 개의 가수 값을 곱하여 비트 폭이 2dw인 산출물을 출력한다. 곱셈기(605)의 출력은 시프터(610)의 상단에 도시된 제어 입력에 기초하여 오른쪽 시프트를 수행하도록 구성된 대응하

는 시프터(610)에 입력된다. 시프터(610)에 대한 제어 입력은 곱셈기(605)에 대한 각 입력의 레벨 1 스케일 값의 합이다. 시프터(610)는 제어 입력의 값과 동일한 다수의 오른쪽 시프트 작업을 수행한다. 시프터(610)의 출력은 2dw+3의 비트 폭을 갖는다. 두 개의 시프터(610)의 출력은 두 입력을 더하여 비트 폭이 2dw+4인 합계를 생성하는 대응하는 가산기(615)에 입력된다. 이 프로세스는 대응하는 레벨 스케일 값을 사용하여 시프터를 제어하도록 계속 사용된다(예컨대, 레벨 2 스케일 값의 합은 시프터(620-1-620-m)를 제어하는 데 사용된다). 레벨 수는 $\log_2(n)$에 의해 결정된다. 위에서 언급했듯이, n은 변수에 저장된 부동 소수점 값의 개수이다. 마지막 레벨의 출력은 $2dw + 4\log_2(n)$의 비트 폭을 갖는다.

[0040] 절삭/반올림 모듈(635)이 마지막 레벨로부터 출력을 수신하면, 절삭/반올림 모듈(635)은 각 변수의 공유 지수 값의 합에 기초하여 출력 값에 대한 시프트 작업을 수행한다. 예를 들어, 공유 지수 값의 합이 양수 값인 경우, 절삭/반올림 모듈(635)은 공유 지수 값의 합과 동일한 값에 대해 여러 번의 왼쪽 시프트 작업을 수행한다. 공유 지수 값의 합이 음수 값이면, 절삭/반올림 모듈(635)은 공유 지수 값의 합의 절대값과 같은 값에 대해 여러 번의 오른쪽 시프트 작업을 수행한다. 그런 다음, 절삭/반올림 모듈(635)은 값을 정의된 비트 수로 잘라내어 출력 값(C)을 생성한다.

[0041] 도 7은 일부 실시예에 따라 계층적 및 공유 지수 부동 소수점 데이터 타입 변수를 생성하기 위한 프로세스(700)를 도시한다. 일부 실시예들에서, 컴퓨팅 시스템(100)은 프로세스(700)를 수행한다. 프로세스(700)는 복수의 부동 소수점 값의 제1 서브세트에 있는 각 부동 소수점 값의 지수 값에 기초하여, 제1 공유 지수 값을 결정함으로써 시작된다(710). 도 1 및 도 4를 예시적으로 참조하면, 지수 매니저(110)는 제1 부동 소수점 값 쌍의 원 지수 값에 기초하여 레벨 0 공유 지수 값을 결정할 수 있다.

[0042] 다음으로, 프로세스(700)는 720에서, 복수의 부동 소수점 값의 제2 서브세트에 있는 각 부동 소수점 값의 지수 값에 기초하여 제2 공유 지수 값을 결정한다. 도 1 및 도 4를 예로서 참조하면, 지수 매니저(110)는 제2 부동 소수점 값 쌍의 원 지수 값에 기초하여 레벨 0 공유 지수 값을 결정할 수 있다.

[0043] 이어서, 프로세스(700)는 제1 공유 지수 값 및 제 2 공유 지수 값에 기초하여 제3 공유 지수 값을 결정한다(730). 도 1 및 도 4를 예로서 참조하면, 지수 매니저(110)는 레벨 0 공유 지수 값에 기초하여 레벨 1 공유 지수 값을 결정할 수 있다.

[0044] 740에서, 프로세스(700)는 제1 공유 지수 값 및 제3 공유 지수 값에 기초하여 제1 차이 값을 결정한다. 도 1 및 도 4를 예로서 참조하면, 지수 매니저(110)는 레벨 0 공유 지수 값 및 대응하는 레벨 1 공유 지수 값의 제1 쌍에 대한 레벨 1 스케일 값을 결정할 수 있다. 그런 다음, 750에서, 프로세스(700)는 제2 공유 지수 값 및 제3 공유 지수 값에 기초하여 제2 차이 값을 결정한다. 도 1 및 도 4를 예로서 참조하면, 지수 매니저(110)는 레벨 0 공유 지수 값 및 대응하는 레벨 1 공유 지수 값의 제2 쌍에 대한 레벨 1 스케일 값을 결정할 수 있다.

[0045] 다음으로, 프로세스(700)는, 760에서, 복수의 부동 소수점 값 내의 각 부동 소수점 값에 대한 부호 값 및 가수 값을 결정한다. 도 1 및 도 4를 예로서 참조하면, 가수 관리자(115)는 표(400)의 행(445)에 도시된 바와 같이 가수에 대한 양자화된 크기 값을 결정할 수 있다. 770에서, 프로세스(700)는 최종적으로, 공유 지수 부동 소수점 데이터 타입에 대한 데이터 구조에, 복수의 부동 소수점 값 내의 각 부동 소수점 값에 대한 부호 값 및 가수 값, 제3 공유 지수 값, 제1 차이 값 및 제2 차이 값을 저장한다. 도 1 및 도 5를 예시적으로 참조하면, HASEFPDT 매니저(105)는 가수 값, 레벨 1 공유 지수, 레벨 1 스케일 값을 데이터 구조(500)에 저장한다.

[0046] 상술한 기법들은 신경망을 처리하도록 구성된 광범위한 컴퓨터 시스템에서 구현될 수 있다. 도 8은 전술한 개시에서 설명된 기술들을 구현하는 데 사용될 수 있는 예시적인 컴퓨터 시스템(800)(예컨대, 컴퓨팅 시스템(100))의 단순화된 블록 다이어그램을 도시한다. 도 8에 도시된 바와 같이, 컴퓨터 시스템(800)은 버스 서브시스템(804)을 통해 다수의 주변 장치와 통신하는 하나 이상의 프로세서(802)를 포함한다. 이러한 주변 장치는 저장 서브시스템(806)(예를 들어, 메모리 서브시스템(808) 및 파일 저장 서브시스템(810)을 포함함) 및 네트워크 인터페이스 서브시스템(816)을 포함할 수 있다. 일부 컴퓨터 시스템은 사용자 인터페이스 입력 디바이스(812) 및/또는 사용자 인터페이스 출력 디바이스(814)를 더 포함할 수 있다.

[0047] 버스 서브시스템(804)은 컴퓨터 시스템(800)의 다양한 컴포넌트 및 서브시스템이 의도된 바와 같이 서로 통신할 수 있도록 하는 메커니즘을 제공할 수 있다. 버스 서브시스템(804)은 단일 버스로 개략적으로 도시되어 있지만, 버스 서브시스템의 다른 실시예는 다수의 버스를 활용할 수 있다.

[0048] 네트워크 인터페이스 서브시스템(816)은 컴퓨터 시스템(800)과 다른 컴퓨터 시스템 또는 네트워크 사이에서 데

이터를 통신하기 위한 인터페이스로서 작용할 수 있다. 네트워크 인터페이스 서브시스템(816)의 실시예는 예를 들어, 이더넷, 와이파이 및/또는 셀룰러 어댑터, 모뎀(전화, 위성, 케이블, ISDN 등), 디지털 가입자 회선(DSL) 유닛, 및/또는 이와 같은 것을 포함할 수 있다.

[0049] 저장 서브시스템(806)은 메모리 서브시스템(808) 및 파일/디스크 저장 서브시스템(810)을 포함한다. 서브시스템(808, 810) 및 본 명세서에 설명된 다른 메모리들은 본 개시의 실시예의 기능을 제공하는 실행 가능한 프로그램 코드 및/또는 데이터를 저장할 수 있는 비일시적 컴퓨터 판독 가능 저장 매체의 예이다.

[0050] 메모리 서브시스템(808)은 프로그램 실행 중 명령어 및 데이터의 저장을 위한 메인 랜덤 액세스 메모리(RAM)(818) 및 고정 명령어들이 저장된 읽기 전용 메모리(ROM)(820)를 포함하는 다수의 메모리를 포함한다. 파일 저장 서브시스템(810)은 프로그램 및 데이터 파일을 위한 영구(예컨대, 비휘발성) 저장소를 제공할 수 있으며, 자기 또는 솔리드 스테이트 하드 디스크 드라이브, 광학 드라이브 및 관련 이동식 매체(예컨대, CD-ROM, DVD, 블루레이 등), 이동식 플래시 메모리 기반 드라이브 또는 카드 및/또는 당업자에게 알려진 다른 유형의 저장 매체를 포함할 수 있다.

[0051] 컴퓨터 시스템(800)은 예시적인 것이며, 시스템(800)보다 더 많거나 더 적은 컴포넌트를 갖는 다른 많은 구성이 가능하다는 것을 이해해야 한다.

[0052] 도 9는 일부 실시예에 따른 신경망 처리 시스템을 예시한다. 다양한 실시예에서, 본 개시에 따른 신경망은 하나 이상의 신경망 프로세서를 포함하는 하드웨어 환경에서 구현 및 훈련될 수 있다. 신경망 프로세서는 다양한 그래픽 처리 장치(GPU)(예컨대, 엔비디아(Nvidia Corp®)에서 생산한 신경망 처리용 GPU), 필드 프로그래머블 게이트 어레이(FPGA)(예컨대, 자일링스(Xilinx®)에서 생산한 신경망 처리용 FPGA) 또는 신경망 계산에 최적화된 하드웨어 아키텍처를 포함하는 다양한 애플리케이션 특정 집적 회로(ASIC) 또는 신경망 프로세서를 지칭할 수 있다. 이 예시적 환경에서, 도 8에 도시된 아키텍처를 포함할 수 있는 하나 이상의 서버(902)는 통신 네트워크(901)(예컨대, 스위치, 라우터 등)를 통해 복수의 제어기(910(1)-910(M))에 결합될 수 있다. 제어기(910(1)-910(M)은 또한 도 8에 도시된 아키텍처를 포함할 수 있다. 각 제어기(910(1)-910(M)는 예를 들어, 프로세서(911(1)-911(N) 및 912(1)-912(N)) 등과 같은 하나 이상의 NN 프로세서에 결합될 수 있다. NN 프로세서(911(1)-911(N) 및 912(1)-912(N))는 훈련 또는 추론과 같은 신경망 처리에 최적화된 다양한 기능 처리 블록 및 메모리 구성을 포함할 수 있다. NN 프로세서는 신경망 연산에 최적화되어 있다. 서버(902)는 예를 들어, 병렬로 NN 프로세서(911(1)-911(N) 및 912(1)-912(N))에 의해 로드 및 실행될 수 있는 NN 모델 및 그 모델에 대한 입력 데이터를 통해 제어기(910)를 구성할 수 있다. 모델은 예를 들어 위에서 설명한 대로 계층 및 관련 가중치를 포함할 수 있다. NN 프로세서는 모델을 로드하고 입력을 적용하여 출력 결과를 생성할 수 있다. NN 프로세서는 예를 들어, 본원에 설명된 훈련 알고리즘을 구현할 수 있다.

[0053] 추가 실시예들

[0054] 다양한 실시예에서, 본 개시는 계층적 및 공유 지수 부동 소수점 데이터 타입을 제공하기 위한 시스템, 방법, 및 장치를 포함한다. 본 명세서에 기술된 기술들은 컴퓨터 시스템에 의해 실행 가능한 프로그램을 저장하는 비일시적 머신 판독 가능 매체로 구현될 수 있으며, 프로그램은 본 명세서에 기술된 기술들을 수행하기 위한 명령어 세트를 포함한다. 일부 실시예에서, 시스템은 처리 유닛 세트, 및 처리 유닛 세트의 적어도 하나의 처리 유닛에 의해 실행될 때, 적어도 하나의 처리 유닛이 상술한 기술들을 수행하도록 하는 명령어 세트를 저장하는 비일시적 머신 판독 가능 매체를 포함한다. 일부 실시예들에서, 비일시적 머신 판독 가능 매체는 예를 들어, 하나 이상의 제어기 또는 하나 이상의 인공 지능 프로세서에 결합될 수 있는 메모리일 수 있다.

[0055] 다음 기술들은 단독으로 또는 상이한 조합으로 구현될 수 있으며, 본원에 설명된 다른 기술들과 함께 추가로 구현될 수 있다.

[0056] 예를 들어, 일 실시예에서, 본 개시는 디바이스의 적어도 하나의 처리 유닛에 의해 실행 가능한 프로그램을 저장하는 비일시적 머신 판독 가능 매체를 포함한다. 프로그램은 복수의 부동 소수점 값의 제1 서브세트에 있는 각 부동 소수점 값의 지수 값에 기초하여 제1 공유 지수 값을 결정하는 것과, 복수의 부동 소수점 값의 제2 서브세트에 있는 각 부동 소수점 값의 지수 값에 기초하여 제2 공유 지수 값을 결정하는 것과, 제1 공유 지수 값 및 제2 공유 지수 값에 기초하여 제3 공유 지수 값을 결정하는 것과, 제1 공유 지수 값 및 제3 공유 지수 값에 기초하여 제1 차이 값을 결정하는 것과, 제2 공유 지수 값 및 제3 공유 지수 값에 기초하여 제2 차이 값을 결정하는 것과, 복수의 부동 소수점 값의 각 부동 소수점 값에 대한 부호 값 및 가수 값을 결정하는 것과, 공유 지수 부동 소수점 데이터 타입에 대한 데이터 구조에, 복수의 부동 소수점 값의 각 부동 소수점 값에 대한 부호

값 및 가수 값, 제3 공유 지수 값, 제1 차이 값, 및 제2 차이 값을 저장하는 것을 위한 명령어 세트를 포함한다.

[0057] 일 실시예에서, 제1 공유 지수 값, 제2 공유 지수 값 및 제3 공유 지수 값은 공유 지수 값의 계층 구조를 형성한다.

[0058] 일 실시예에서, 제1 공유 지수 값 및 제2 공유 지수 값은 공유 지수 값의 계층 구조에서 제1 레벨 지수 값이고, 제3 공유 지수 값은 공유 지수 값의 계층 구조에서 제2 레벨 지수 값이다.

[0059] 일 실시예에서, 공유 지수 값의 계층 구조는 제3 공유 지수 값, 제1 차이 값 및 제2 차이 값으로 표현된다.

[0060] 일 실시예에서, 본 개시는, 제1 공유 지수 값 및 복수의 부동 소수점 값의 제1 서브세트에 있는 각 부동 소수점 값의 지수 값에 기초하여 제1 차이 값 세트를 결정하는 것과, 제2 공유 지수 값 및 복수의 부동 소수점 값의 제2 서브세트에 있는 각 부동 소수점 값의 지수 값에 기초하여 제2 차이 값 세트를 결정하는 것과, 공유 지수 부동 소수점 데이터 타입에 대한 데이터 구조에, 제1 차이 값 세트 및 제2 차이 값 세트를 저장하는 것을 더 포함할 수 있다.

[0061] 일 실시예에서, 제1 공유 지수 값을 결정하는 것은 복수의 부동 소수점 값의 제1 서브세트에서 가장 높은 값을 갖는 부동 소수점 값의 지수 값을 결정하고, 그 지수 값을 제1 공유 지수 값으로 사용하는 것을 포함한다.

[0062] 일 실시예에서, 제2 공유 지수 값을 결정하는 것은 복수의 부동 소수점 값의 제2 서브세트에서 가장 높은 값을 갖는 부동 소수점 값의 지수 값을 결정하고, 그 지수 값을 제2 공유 지수 값으로 사용하는 것을 포함한다.

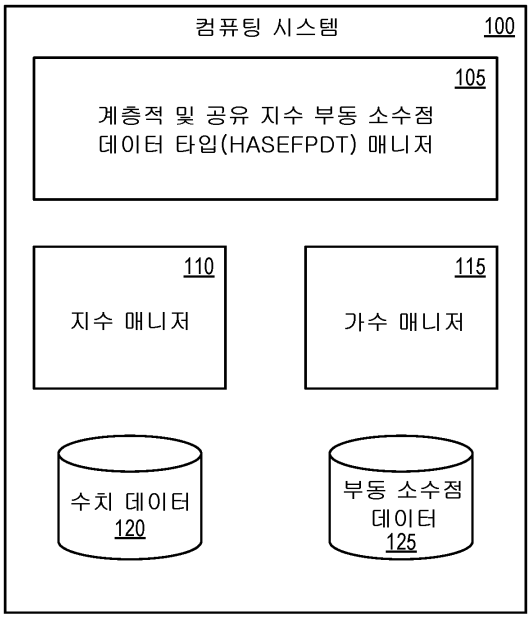
[0063] 일 실시예에서, 제3 공유 지수 값을 결정하는 것은 제1 공유 지수 값이 제2 공유 지수 값보다 크거나 같다고 판단되면, 제1 공유 지수 값을 제3 지수 값으로 사용하는 것과, 제1 공유 지수 값이 제2 공유 지수 값보다 작다고 판단되면, 제2 공유 지수 값을 제3 지수 값으로 사용하는 것을 포함한다.

[0064] 일 실시예에서, 공유 지수 부동 소수점 데이터 타입에 대한 데이터 구조에 저장된 가수 값을 나타내기 위해 사용되는 비트의 수는 제3 지수 값을 나타내기 위해 사용되는 비트의 수보다 작다.

[0065] 상기의 설명은 본 개시의 다양한 실시예와 함께 특정 실시예의 측면들이 어떻게 구현될 수 있는지에 대한 예들을 예시한다. 상기 실시예들은 유일한 실시예로 간주되어서는 안되며, 다음 청구범위에 정의된 바와 같은 특정 실시예들의 유연성 및 장점을 설명하기 위해 제시된 것이다. 상기 개시 및 다음의 청구범위에 기초하여, 청구범위에 정의된 바와 같은 본 개시의 범위를 벗어나지 않고 다른 배열, 실시예, 구현 및 균등물들이 사용될 수 있다.

도면

도면1



도면2

200

205	값	0.1946	0.1535	0.0904	0.4727
210	원 지수	-3	-3	-4	-2
215	레벨 0 지수	-3	-3	-4	-2
220	레벨 1 공유 지수	-3		-2	
225	레벨 1 스케일	0	0	1	0
230	레벨 2 공유 지수	-2			
235	레벨 2 스케일	1		0	
240	총 스케일	1	1	1	0
245	공유 지수 스케일	-3	-3	-3	-2
250	지수 차이	0	0	1	0
255	양자화된 1비트 크기	1	1	0	1
260	양자화된 지수	-3	-3	-3	-2
265	양자화된 값	0.125	0.125	0	0.25

도면3

300

가수(1비트 부호, 1비트 크기)	레벨 1 스케일	레벨 2 스케일	공유 지수
2비트	1비트	1비트	8비트
2비트	1비트		
2비트	1비트	1비트	
2비트	1비트		

도면4

400

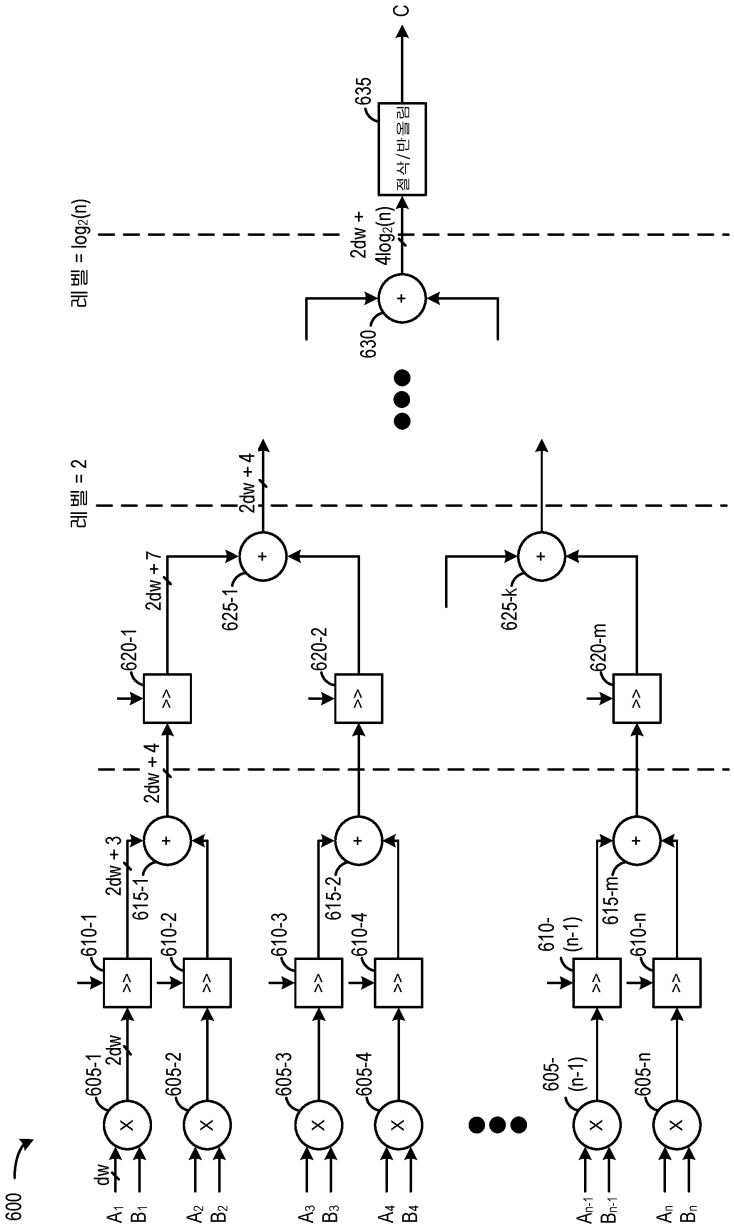
405	값	0.1946	0.1535	0.0904	0.4727
410	원 지수	-3	-3	-4	-2
415	레벨 0 공유 지수	-3		-2	
420	레벨 1 공유 지수	-2			
425	레벨 1 스케일	1		0	
430	총 스케일	1	1	0	0
435	공유 지수 스케일	-3	-3	-2	-2
440	지수 차이	0	0	2	0
445	양자화된 2비트 크기	3	2	0	3
450	양자화된 지수	-3	-3	-2	-2
455	양자화된 값	0.1875	0.125	0	0.375

도면5

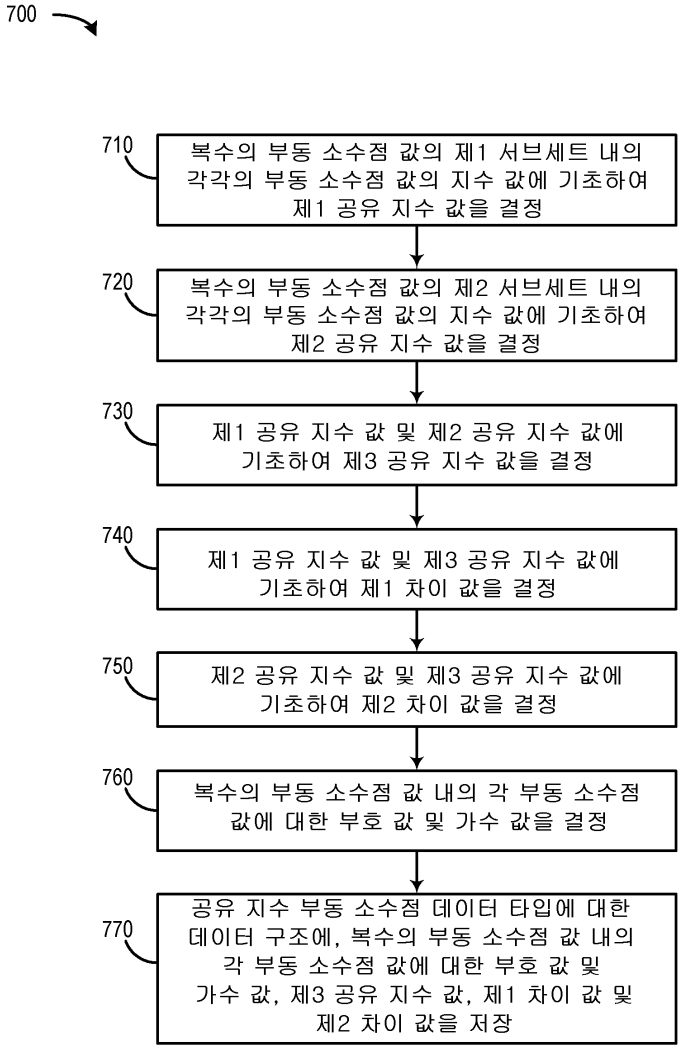
500

가수(1비트 부호, 1비트 크기)	레벨 1 스케일	공유 지수
2비트	2비트	8비트
2비트		
2비트	2비트	
2비트		

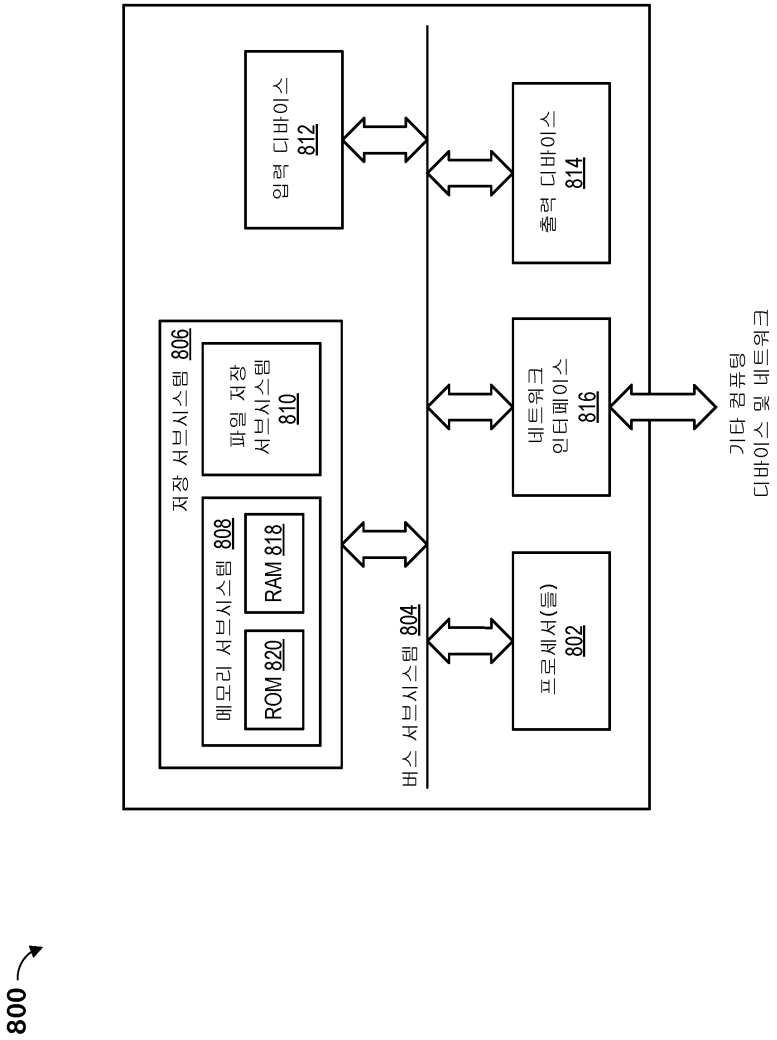
도면6


600
dw
A1
B1
A2
B2
A3
B3
A4
B4
An-1
Bn-1
An
Bn
605-1
2dw
605-2
605-3
605-4
605-(n-1)
605-n
610-1
610-2
610-3
610-4
610-(n-1)
610-n
>>
615-1
2dw + 3
615-2
615-m
2dw + 4
620-1
620-2
620-m
625-1
2dw + 7
625-k
2dw + 4
레벨 = 2
630
2dw + 4log2(n)
레벨 = log2(n)
635
절삭/반올림
C

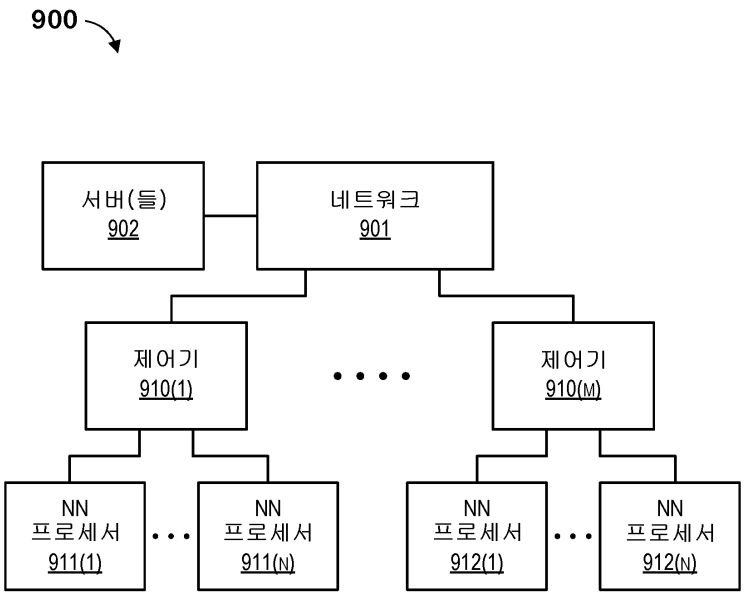
도면7


700
710
복수의 부동 소수점 값의 제1 서브세트 내의 각각의 부동 소수점 값의 지수 값에 기초하여 제1 공유 지수 값을 결정
720
복수의 부동 소수점 값의 제2 서브세트 내의 각각의 부동 소수점 값의 지수 값에 기초하여 제2 공유 지수 값을 결정
730
제1 공유 지수 값 및 제2 공유 지수 값에 기초하여 제3 공유 지수 값을 결정
740
제1 공유 지수 값 및 제3 공유 지수 값에 기초하여 제1 차이 값을 결정
750
제2 공유 지수 값 및 제3 공유 지수 값에 기초하여 제2 차이 값을 결정
760
복수의 부동 소수점 값 내의 각 부동 소수점 값에 대한 부호 값 및 가수 값을 결정
770
공유 지수 부동 소수점 데이터 타입에 대한 데이터 구조에, 복수의 부동 소수점 값 내의 각 부동 소수점 값에 대한 부호 값 및 가수 값, 제3 공유 지수 값, 제1 차이 값 및 제2 차이 값을 저장

도면8


800
입력 디바이스 812
출력 디바이스 814
저장 서브시스템 806
파일 저장 서브시스템 810
메모리 서브시스템 808
RAM 818
ROM 820
네트워크 인터페이스 816
기타 컴퓨팅 디바이스 및 네트워크
버스 서브시스템 804
프로세서(들) 802

도면9


900
서버(들) 902
네트워크 901
제어기 910(1)
제어기 910(M)
NN 프로세서 911(1)
NN 프로세서 911(N)
NN 프로세서 912(1)
NN 프로세서 912(N)