ACCESSING DATA IN MULTI-DIMENSIONAL TENSORS

BACKGROUND

[0001] This specification generally relates to using registers for computing memory address values. Registers may be used to handle branch conditions.

SUMMARY

[0002] According to one innovative aspect of the subject matter described in this specification, registers can be used to iterate nested loop structures, to compute corresponding memory addresses, and to allow programmers to handle loop bound conditions. In general, registers may be arranged to maintain several sets of architectural status, including loop bound values, a current loop index variable in each loop, dimension multipliers for computing a memory address value, and program counter values for handling branch loop bounds. The arrangement may be good for processing one or more multi-dimensional data structures. With these architectural status sets, a processor may execute instructions to move index variables to new positions, retrieve memory address values, and to jump to a new branch target.

[0002a] According to an aspect of the present invention, there is provided an apparatus for processing an instruction for accessing an *N*-dimensional tensor, the apparatus comprising: multiple tensor index elements, wherein each tensor index element includes hardware storage circuitry configured to store an index value for a respective nested loop that is used to traverse a respective dimension of the *N*-dimensional tensor, and wherein the index value for each nested loop is adjusted after each iteration of the nested loop; multiple dimension multiplier elements, wherein each of the dimension multiplier elements has a corresponding tensor index element and includes hardware storage circuitry configured to store a multiplier value for the corresponding tensor index element, wherein each multiplier value is a constant value for the corresponding tensor index element and remains constant for each iteration of the nested loop for the corresponding tensor index element; and one or more hardware processors configured to execute one or more instructions of an instruction set executable by the one or more hardware processors, wherein execution of the one or more instructions causes the one

or more hardware processors to perform operations comprising: obtaining an instruction to access a particular element of a *N*-dimensional tensor, wherein the *N*-dimensional tensor has multiple elements arranged across each of the *N* dimensions, and wherein *N* is an integer that is equal to or greater than one; obtaining a current index value stored by each of one or more tensor index elements of the multiple tensor index elements; determining, using the current index value stored by each of the one or more tensor index elements of the multiple tensor index elements and the multiplier value stored by one or more dimension multiplier elements of the multiple dimension multiplier elements that correspond to the one or more tensor index elements, a memory address at which the particular element is stored in memory, the determining including: determining, for each dimension of the *N*-dimensional tensor, a product of (i) the current index value for the nested loop that is used to traverse the dimension and (ii) the multiplier value for the corresponding tensor index element that stores the current index value for the nested loop that is used to traverse the dimension; and determining the memory address of the particular element based on a sum of the product for each dimension of the *N*-dimensional tensor; and outputting data indicating the determined memory address to a processing unit that is configured to access the particular element of the *N*-dimensional tensor using the determined memory address.

[0002b] According to another aspect of the present invention, there is provided a system comprising: one or more hardware processors configured to perform linear algebra operations on an N-dimensional tensor, wherein the N-dimensional tensor has multiple elements arranged across each of the N dimensions, and wherein N is an integer that is equal to or greater than one; multiple tensor index elements, wherein each tensor index element includes hardware storage circuitry configured to store an index value for a respective nested loop that is used to traverse a respective dimension of the N-dimensional tensor, and wherein the index value for each nested loop is adjusted after each iteration of the nested loop; multiple dimension multiplier elements, wherein each of the dimension multiplier elements has a corresponding tensor index element and includes hardware storage circuitry configured to store a multiplier value for the corresponding tensor index element, wherein each multiplier value is a constant value for the corresponding tensor index element and remains constant for each

iteration of the nested loop for the corresponding tensor index element; and hardware circuitry configured to: obtain an instruction to access a particular element of the N-dimensional tensor; obtain a current index value stored by each of one or more tensor index elements of the multiple tensor index elements; determine, using the current index value stored by each of the one or more tensor index elements of the multiple tensor index elements and the multiplier value stored by one or more dimension multiplier elements of the multiple dimension multiplier elements that correspond to the one or more tensor index elements, a memory address at which the particular element is stored in memory, the determining comprising: determining, for each dimension of the N-dimensional tensor, a product of (i) the current index value for the nested loop that is used to traverse the dimension and (ii) the multiplier value for the corresponding tensor index element that stores the current index value for the nested loop that is used to traverse the dimension; and determining the memory address of the particular element based on a sum of the product for each dimension of the N-dimensional tensor; and output, to the one or more processors, data indicating the determined memory address for accessing the particular element of the N-dimensional tensor, wherein the one or more processors are configured to access the particular element using the determined memory address.

[0002c] According to another aspect of the present invention, there is provided a computer-implemented method, comprising: obtaining, by one or more hardware processors, an instruction to access a particular element of an N-dimensional tensor, wherein the N-dimensional tensor has multiple elements arranged across each of the N dimensions, and wherein N is an integer that is equal to or greater than one; obtaining, by the one or more hardware processors, a current index value stored by each of one or more tensor index elements of multiple tensor index elements, wherein: each tensor index element includes hardware storage circuitry configured to store an index value for a respective nested loop that is used to traverse a respective dimension of the N-dimensional tensor and wherein the index value for each nested loop is adjusted after each iteration of the nested loop; and each tensor index element corresponds to a respective dimension multiplier element of multiple dimension multiplier elements, wherein each dimension multiplier element includes hardware storage circuitry

configured to store a multiplier value for the corresponding tensor index element, wherein each multiplier value is a constant value for the corresponding tensor index element and remains constant for each iteration of the nested loop for the corresponding tensor index element; determining, by the one or more hardware processors and using the current index value stored by each of the one or more tensor index elements of the multiple tensor index elements and a multiplier value stored by one or more dimension multiplier elements of multiple dimension multiplier elements that correspond to the one or more tensor index elements, a memory address at which the particular element is stored in memory, the determining including: determining, for each dimension of the N-dimensional tensor, a product of (i) the current index value for the nested loop that is used to traverse the dimension and (ii) the multiplier value for the corresponding tensor index element that stores the current index value for the nested loop that is used to traverse the dimension; and determining the memory address of the particular element based on a sum of the product for each dimension of the N-dimensional tensor; and outputting, by the one or more hardware processors, data indicating the determined memory address to a processing unit that is configured to access the particular element of the N-dimensional tensor using the determined memory address.

[0003] In general, one innovative aspect of the subject matter described in this specification can be embodied in an apparatus for processing an instruction for accessing a *N*-dimensional tensor. The apparatus includes multiple tensor index elements and multiple dimension multiplier elements, where each of the dimension multiplier elements has a corresponding tensor index element. The apparatus includes one or more processors configured to obtain an instruction to access a particular element of a *N*-dimensional tensor, where the *N*-dimensional tensor has multiple elements arranged across each of the *N* dimensions, and where *N* is an integer that is equal to or greater than one; determine, using one or more tensor index elements of the multiple tensor index elements and one or more dimension multiplier elements of the multiple dimension multiplier elements, an address of the particular element; and output data indicating the determined address for accessing the particular element of the *N*-dimensional tensor.

CLAIMS

1. An apparatus for processing an instruction for accessing an *N*-dimensional tensor, the apparatus comprising:

multiple tensor index elements, wherein each tensor index element includes hardware storage circuitry configured to store an index value for a respective nested loop that is used to traverse a respective dimension of the *N*-dimensional tensor, and wherein the index value for each nested loop is adjusted after each iteration of the nested loop;

multiple dimension multiplier elements, wherein each of the dimension multiplier elements has a corresponding tensor index element and includes hardware storage circuitry configured to store a multiplier value for the corresponding tensor index element, wherein each multiplier value is a constant value for the corresponding tensor index element and remains constant for each iteration of the nested loop for the corresponding tensor index element; and

one or more hardware processors configured to execute one or more instructions of an instruction set executable by the one or more hardware processors, wherein execution of the one or more instructions causes the one or more hardware processors to perform operations comprising:

obtaining an instruction to access a particular element of a *N*-dimensional tensor, wherein the *N*-dimensional tensor has multiple elements arranged across each of the *N* dimensions, and wherein *N* is an integer that is equal to or greater than one;

obtaining a current index value stored by each of one or more tensor index elements of the multiple tensor index elements;

determining, using the current index value stored by each of the one or more tensor index elements of the multiple tensor index elements and the multiplier value stored by one or more dimension multiplier elements of the multiple dimension multiplier elements that correspond to the one or more tensor index elements, a memory address at which the particular element is stored in memory, the determining including:

determining, for each dimension of the *N*-dimensional tensor, a product of (i) the current index value for the nested loop that is used to traverse the dimension and (ii) the multiplier value for the corresponding tensor index element that stores the current index value for the nested loop that is used to traverse the dimension; and

determining the memory address of the particular element based on a sum of the product for each dimension of the *N*-dimensional tensor; and

outputting data indicating the determined memory address to a processing unit that is configured to access the particular element of the *N*-dimensional tensor using the determined memory address.

2. The apparatus of claim 1,

wherein the multiple tensor index elements are arranged into *M* groups of tensor index elements, *M* being a positive integer greater than one, and

wherein each group of tensor index elements is assigned to a distinct and different multi-dimensional tensor.

3. The apparatus of claim 1, further comprising multiple tensor bound elements, wherein each of the multiple tensor bound elements has a corresponding tensor index element in the multiple tensor index elements.

4. The apparatus of claim 3, wherein the obtained instruction represents an instruction for processing a nested loop that includes a first loop and a second loop that is an outer loop of the first loop, wherein the first loop is iterated using a first index variable, and wherein the second loop is iterated using a second index variable.

5. The apparatus of claim 4, wherein the one or more hardware processors are further configured to perform operations comprising:

setting values stored in the multiple tensor index elements, the multiple dimension multiplier elements, and the multiple tensor bound elements to an initialization value, including setting at least one of the dimension multiplier values to a

value based on an upper bound value of one or more of the nested loops.

6. The apparatus of claim 4, wherein the one or more hardware processors are further configured to perform operations comprising:

setting a value stored by a first tensor bound element of the multiple tensor bound elements to an upper bound value of the first index variable for iterating the first loop;

setting a value stored by a second tensor bound element of the multiple tensor bound elements to an upper bound value of the second index variable for iterating the second loop;

setting a first multiplier value stored by a first dimension multiplier element of the multiple dimension multiplier elements to a predetermined value; and

setting a second multiplier value stored by a second dimension multiplier element of the multiple dimension multiplier elements to the upper bound value of the first index variable.

7. The apparatus of claim 6, wherein determining the memory address of the particular element of the *N*-dimensional tensor comprises:

setting a first index value stored in a first tensor index element of the multiple tensor index elements to a current value of the first index variable;

setting a second index value stored in a second tensor index element of the multiple tensor index elements to a current value of the second index variable;

multiplying the first multiplier value stored in the first dimension multiplier element by the first index value stored in the first tensor index element to obtain a first multiplied value;

multiplying the second multiplier value stored in the second dimension multiplier element by the second index value stored in the second tensor index element to obtain a second multiplied value; and

determining the address by at least summing the first multiplied value and the second multiplied value.

8. The apparatus of claim 7, wherein the one or more hardware processors are further configured to perform operations comprising:

receive an instruction to update the first index value stored in the first tensor index element;

after receiving the instruction to update the first index value stored in the first tensor index element, determining that a difference between the first index value stored in the first tensor index element and the value stored in the first tensor bound element satisfies a threshold; and

in response to determining that the difference between the first index value stored in the first tensor index element and the value stored in the first tensor bound element satisfies the threshold, updating the first index value stored in the first tensor index element.

9. The apparatus of claim 8,

wherein the first index variable is incremented by a first incremental value each time the first loop completes, and

wherein updating the first index value stored in the first tensor index element comprises incrementing the first index value stored in the first tensor index element by the first incremental value.

10. The apparatus of claim 7, wherein the one or more hardware processors are further configured to perform operations comprising:

receiving an instruction to update the first index value stored in the first tensor index element;

after receiving the instruction to update the first index value stored in the first tensor index element, determining that a difference between the first index value stored in the first tensor index element and the value stored in the first tensor bound element does not satisfy a threshold; and

in response to determining that the difference between the first index value stored in the first tensor index element and the value stored in the first tensor bound element does not satisfy the threshold:

resetting the first index value stored in the first tensor index element; and

updating the second index value stored in the second tensor index element.

11. The apparatus of claim 10,

wherein the second index variable is incremented by a second incremental value each time the second loop completes, and

wherein updating the second index value stored in the second tensor index element comprises incrementing the second index value stored in the second tensor index element by the second incremental value.

12. The apparatus of claim 1, wherein the one or more hardware processors include one or more arithmetic logic units.

13. The apparatus of claim 1, wherein each of the tensor index elements and the dimension multiplier elements is a hardware register.

14. The apparatus of claim 1, further comprising multiple branch target elements, wherein each of the multiple branch target elements has a corresponding tensor index element in the multiple tensor index elements, and wherein each of the multiple branch target elements is configured to store an address for an instruction to be executed upon a completion of an iteration of a loop that corresponds to the branch target element.

15. A system comprising:

one or more hardware processors configured to perform linear algebra operations on an *N*-dimensional tensor, wherein the *N*-dimensional tensor has multiple elements arranged across each of the *N* dimensions, and wherein *N* is an integer that is equal to or greater than one;

multiple tensor index elements, wherein each tensor index element includes hardware storage circuitry configured to store an index value for a respective nested loop that is used to traverse a respective dimension of the *N*-dimensional tensor, and

wherein the index value for each nested loop is adjusted after each iteration of the nested loop;

multiple dimension multiplier elements, wherein each of the dimension multiplier elements has a corresponding tensor index element and includes hardware storage circuitry configured to store a multiplier value for the corresponding tensor index element, wherein each multiplier value is a constant value for the corresponding tensor index element and remains constant for each iteration of the nested loop for the corresponding tensor index element; and

hardware circuitry configured to:

obtain an instruction to access a particular element of the *N*-dimensional tensor;

obtain a current index value stored by each of one or more tensor index elements of the multiple tensor index elements;

determine, using the current index value stored by each of the one or more tensor index elements of the multiple tensor index elements and the multiplier value stored by one or more dimension multiplier elements of the multiple dimension multiplier elements that correspond to the one or more tensor index elements, a memory address at which the particular element is stored in memory, the determining comprising:

determining, for each dimension of the *N*-dimensional tensor, a product of (i) the current index value for the nested loop that is used to traverse the dimension and (ii) the multiplier value for the corresponding tensor index element that stores the current index value for the nested loop that is used to traverse the dimension; and

determining the memory address of the particular element based on a sum of the product for each dimension of the *N*-dimensional tensor; and

output, to the one or more processors, data indicating the determined memory address for accessing the particular element of the *N*-dimensional tensor, wherein the one or more processors are configured to access the particular element using the determined memory address.

16. A computer-implemented method, comprising:

obtaining, by one or more hardware processors, an instruction to access a particular element of an *N*-dimensional tensor, wherein the *N*-dimensional tensor has multiple elements arranged across each of the *N* dimensions, and wherein *N* is an integer that is equal to or greater than one;

obtaining, by the one or more hardware processors, a current index value stored by each of one or more tensor index elements of multiple tensor index elements, wherein:

each tensor index element includes hardware storage circuitry configured to store an index value for a respective nested loop that is used to traverse a respective dimension of the N-dimensional tensor and wherein the index value for each nested loop is adjusted after each iteration of the nested loop; and

each tensor index element corresponds to a respective dimension multiplier element of multiple dimension multiplier elements, wherein each dimension multiplier element includes hardware storage circuitry configured to store a multiplier value for the corresponding tensor index element, wherein each multiplier value is a constant value for the corresponding tensor index element and remains constant for each iteration of the nested loop for the corresponding tensor index element;

determining, by the one or more hardware processors and using the current index value stored by each of the one or more tensor index elements of the multiple tensor index elements and a multiplier value stored by one or more dimension multiplier elements of multiple dimension multiplier elements that correspond to the one or more tensor index elements, a memory address at which the particular element is stored in memory, the determining including:

determining, for each dimension of the *N*-dimensional tensor, a product of (i) the current index value for the nested loop that is used to traverse the dimension and (ii) the multiplier value for the corresponding tensor index element that stores the current index value for the nested loop that is used to traverse the dimension; and

determining the memory address of the particular element based on a sum of the product for each dimension of the *N*-dimensional tensor; and

outputting, by the one or more hardware processors, data indicating the

determined memory address to a processing unit that is configured to access the particular element of the *N*-dimensional tensor using the determined memory address.

17. The method of claim 16, wherein the instruction represents an instruction for processing a nested loop that includes a first loop and a second loop that is an outer loop of the first loop, wherein the first loop is iterated using a first index variable, and wherein the second loop is iterated using a second index variable.