

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.

[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.

[0004] These and other implementations can each optionally include one or more of the following features. For example, the address of the particular element may be an address offset from another element of the N -dimensional tensor. To determine the address of the particular element, for each tensor index element of the one or more tensor index elements, a value stored in the respective tensor index element may be multiplied by a value stored in the corresponding tensor index element of the one or more multiple tensor index elements to generate a respective multiplied value. The address offset may be determined by determining a sum of the multiplied values.

[0005] The multiple tensor index elements may be arranged into M groups of tensor index elements, M being a positive integer greater than one. Each group of tensor index elements may be assigned to a distinct and different multi-dimensional tensor.

[0006] The apparatus may include multiple tensor bound elements, where each of the multiple tensor bound elements may have a corresponding tensor index element in the multiple tensor index elements. The instruction may represent 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, where the first loop is iterated using a first index variable, and where the second loop is iterated using a second index variable.

[0007] The one or more processors may be configured to set values stored in the multiple tensor index elements, the multiple dimension multiplier elements, and the multiple tensor bound elements to an initialization value. The one or more processors may be configured to set a value stored in 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; set a value stored in 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; set a value stored in a first dimension multiplier element of the multiple dimension multiplier elements to a predetermined value; and set a value stored in a second dimension multiplier element of the multiple dimension multiplier elements to the upper bound value of the first index variable.

[0008] To determine the address of the element of the N -dimensional tensor, the one or more processors may be configured to set a value stored in a first tensor index

element of the multiple tensor index elements to a current value of the first index variable; set a value stored in a second tensor index element of the multiple tensor index elements to a current value of the second index variable; multiply the value stored in the first dimension multiplier element by the value stored in the first tensor index element to obtain a first multiplied value; multiply the value stored in the second dimension multiplier element by the value stored in the second tensor index element to obtain a second multiplied value; and determine the address by at least summing the first multiplied value and the second multiplied value.

[0009] The one or more processors may be configured to receive an instruction to update the value stored in the first tensor index element; after receiving the instruction to update the value stored in the first tensor index element, determining that a difference between the 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 value stored in the first tensor index element and the value stored in the first tensor bound element satisfies the threshold, updating the value stored in the first tensor index element. The first index variable may be incremented by a first incremental value each time the first loop completes. To update the value stored in the first tensor index element, the value stored in the first tensor index element may be incremented by the first incremental value.

[0010] The one or more processors may be configured to receive an instruction to update the value stored in the first tensor index element; after receiving the instruction to update the value stored in the first tensor index element, determining that a difference between the value stored in the first tensor index element and the value stored in the first tensor bound element does not satisfy a threshold. In response to determining that the difference between the value stored in the first tensor index element and the value stored in the first tensor bound element does not satisfy the threshold, the one or more processors may be configured to reset the value stored in the first tensor index element, and update the value stored in the second tensor index element. The second index variable may be incremented by a second incremental value each time the second loop completes. To update the value stored in the second tensor index element, the value

stored in the second tensor index element may be incremented by the second incremental value.

[0011] The one or more processors may include one or more arithmetic logic units. Each of the tensor index elements and the dimension multiplier elements may be a register. The apparatus may include multiple branch target elements, where each of the multiple branch target elements has a corresponding tensor index element in the multiple tensor index elements, and where 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 respective branch target element.

[0012] The subject matter described in this specification can be implemented in particular embodiments so as to realize one or more of the following advantages. Using registers to keep track of memory address values allow a program to iterate deeply-nested loops with one instruction. Memory address values can be quickly determined by applying simple arithmetic calculations based on the values stored in the registers. Determining memory address values using arithmetic logic units allows the number of computational cycles at the processor to be reduced, and increases processor bandwidth for other computation tasks. Tensors can be traversed with a reduced number of instructions. A two-dimension register architecture allows multi-dimensional tensors each having different dimensions to be tracked concurrently. By having branch target tables, branch instruction-related overhead can be reduced.

[0013] Other implementations of this and other aspects include corresponding systems, apparatus, and computer programs, configured to perform the actions of the methods, encoded on computer storage devices. A system of one or more computers can be so configured by virtue of software, firmware, hardware, or a combination of them installed on the system that in operation cause the system to perform the actions. One or more computer programs can be so configured by virtue of having instructions that, when executed by data processing apparatus, cause the apparatus to perform the actions.

[0014] The details of one or more implementations of the subject matter described in this specification are set forth in the accompanying drawings and the description below.

Other potential features, aspects, and advantages of the subject matter will become apparent from the description, the drawings, and the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] FIG. 1 is a block diagram of an example computation system.

[0016] FIGS. 2A-2H illustrate an example tensor traversal unit.

[0017] FIG. 3 illustrates an example tensor traversal unit.

[0018] FIG. 4 is a flow diagram that illustrates an example of a process for determining an address of a multi-dimensional tensor variable.

[0019] Like reference numbers and designations in the various drawings indicate like elements.

DETAILED DESCRIPTION

[0020] In general, when a software algorithm processes an N -dimensional tensor, a nested loop may be used, where each loop is responsible for traversing each dimension of the N -dimensional tensor. A multi-dimensional tensor may be a matrix, or multi-dimensional matrices. Each dimension of the N -dimensional tensor may include one or more elements, where each element may store a respective data value. For example, a tensor may be a variable in a program, where the variable may have three dimensions. The first dimension may have a length of three hundred elements, the second dimension may have a length of thousand elements, and the third dimension may have a length of twenty elements. Traversing the tensor in a nested loop requires a computation of a memory address value of an element to load or store the corresponding data value of the element. For example, a for-loop is a nested loop, where three loops tracked by three loop index variables can be nested to traverse through the three-dimensional tensor. In some cases, a processor may need to execute a loop bound condition, such as setting a loop bound of an inner loop with an outer loop index variable. For example, in determining whether to exit the inner-most loop of a nested loop, the program may compare the current value of the loop index variable of

the inner-most loop with the current value of the loop index variable of the outer-most loop of the nested loop.

[0021] These tasks may require a significant number of instructions such as branch instructions and integer arithmetic instructions. When each loop bound is small and the number of loops is large, the computation may take a significant portion of the overall execution time, and seriously degrade overall performance. For example, a processor may be designed to process N -dimensional tensor input. The processor may be implemented on a 2-dimensional architecture, which leaves $(N-2)$ dimensions to be processed by software. When N is large, the handling of the $N-2$ dimensions may be computationally heavy for the processor because the computation requires very high scalar computation bandwidth. A hardware tensor traversal unit for a processor may increase the computation bandwidth of the processor by reducing the number of dimensions that the processor is required to process when traversing a tensor.

[0022] FIG. 1 shows a block diagram of an example computing system 100 for traversing a tensor. In general, the computing system 100 processes an input 104 to generate an output 116. The computing system 100 may be configured to perform linear algebra computations. The input 104 may be any suitable data that can be processed by the computing system 100. The computing system 100 includes a processing unit 102, a storage medium 104, and a tensor traversal unit 106.

[0023] In general, when the processing unit 102 executes an instruction for accessing a particular element of a tensor, the tensor traversal unit 106 determines the address of the particular element of the tensor, such that the processing unit 102 may access the storage medium 104 to read data 114 representing the value of the particular element. For example, a program may include a nested loop and the processing unit 102 may execute an instruction to access an element of a two-dimensional array variable within the nested loop according to current index variable values associated with the nested loop. Based on the current index variable values associated with the nested loop, the tensor traversal unit 106 may determine an offset value that represents an offset from a first element of the two-dimensional array variable. The processing unit 102 may then

access, using the offset value and from the storage medium, the particular element of the two-dimensional array variable.

[0024] The processing unit 102 is configured to process instructions for execution within the computing system 100, including instructions 112 stored in the storage medium 104 or other instructions stored in another storage device. The processing unit 102 may include one or more processors. The storage medium 104 stores information within the computing system 100. In some implementations, the storage medium 104 is a volatile memory unit or units. In some other implementations, the storage medium 104 is a non-volatile memory unit or units. The storage medium 104 may also be another form of computer-readable medium, such as a floppy disk device, a hard disk device, an optical disk device, or a tape device, a flash memory or other similar solid state memory device, or an array of devices, including devices in a storage area network or other configurations. The instructions, when executed by the processing unit 102, cause the processing unit 102 to perform one or more tasks.

[0025] The tensor traversal unit 106 is configured to determine a status associated with one or more tensors. The status may include loop bound values, current loop index variable values, dimension multipliers for computing a memory address value, and/or program counter values for handling branch loop bounds. The tensor traversal unit 106 includes tensor status elements 122 and an arithmetic logic unit 124. Each of the tensor status elements 122 may be a storage element, for example a register or any other suitable storage circuitry. In some implementations, the tensor status elements 122 may be physically or logically arranged into different groups, as described in more detail below with reference to FIGS. 2A-2H and FIGS. 3A-3B. In some implementations, a group of the tensor status elements 122 may be physically or logically arranged into a multi-dimension array. For example, each group of the tensor status elements 122 may be physically or logically arranged into a two-dimensional array. The arithmetic logic unit 124 includes one or more arithmetic logic units (ALUs) that are configured to perform arithmetic and logic operations. For example, the arithmetic logic unit 124 may include digital circuitry that is configured to perform addition, subtraction, multiplication, division, logic AND, logic OR, or any other suitable arithmetic and logic operations.

[0026] FIG. 2A shows an example set of tensor status elements 200 of a tensor traversal unit. The tensor status elements 200 may correspond to the tensor status elements 122 of the tensor traversal unit 106. The tensor traversal unit 200 includes a group of tensor index elements 202, a group of tensor bound elements 204, and a group of dimension multiplier elements 206. The tensor index elements 202 may be physically or logically arranged as a 2-D array having M rows and N columns, where M and N are integers greater than or equal to one. In some implementations, each row of the tensor index elements 202 may represent tensor index information for a tensor. For example, if a program defines two array variables $V1$ and $V2$, the tensor traversal unit may assign rows 202a and 202b to track the tensor index information for array variables $V1$ and $V2$, respectively. In some implementations, each column of the tensor index elements 202 may represent tensor index information for nested loop index variable values that are associated with a tensor. For example, if the program defines a nested loop having three loops for accessing the variable $V1$, where each loop of the nested loop is indexed by nested loop index variables i , j , and k , the tensor traversal unit may assign tensor index element $X_{1,1}$, $X_{1,2}$, and $X_{1,3}$ to track the nested loop index variable i , j , and k , respectively. The tensor index information is described in more detail below with reference to FIGS. 2D-2H below.

[0027] The tensor bound elements 204 may be physically or logically arranged as a 2-D array having the same dimensions as the tensor index elements 202, where each element of the tensor bound elements 204 has a corresponding element in the tensor index elements 202. In some implementations, each row of the tensor bound elements 204 may represent tensor bound information for a tensor. For example, if the program defines two array variables $V1$ and $V2$, the tensor traversal unit may assign rows 204a and 204b to track the tensor bound information for array variables $V1$ and $V2$, respectively. In some implementations, each column of the tensor bound elements 204 may represent tensor bound information for nested loop index variable values that are associated with a tensor. For example, if the program defines a nested loop having three loops for accessing the variable $V1$, where each loop of the nested loop is indexed by nested loop index variables i , j , and k , the tensor traversal unit may assign tensor bound element $Y_{1,1}$, $Y_{1,2}$, and $Y_{1,3}$ to track the tensor bound information associated with

the nested loop index variables i , j , and k , respectively. The tensor bound information is described in more detail below with reference to FIG. 2C below.

[0028] The dimension multiplier elements 206 may be physically or logically arranged as a 2-D array having the same dimensions as the tensor index elements 202, where each element of the dimension multiplier elements 206 has a corresponding element in the tensor index elements 202. In some implementations, each row of the dimension multiplier elements 206 may represent dimension multiplier information for a tensor. For example, if the program defines two array variables $V1$ and $V2$, the tensor traversal unit may assign rows 206a and 206b to track the dimension multiplier information for array variables $V1$ and $V2$, respectively. In some implementations, each column of the dimension multiplier elements 206 may represent dimension multiplier information for nested loop index variable values that are associated with a tensor. For example, if the program defines a nested loop having three loops for accessing the variable $V1$, where each loop of the nested loop is indexed by nested loop index variables i , j , and k , the tensor traversal unit may assign dimension multiplier elements $Z_{1,1}$, $Z_{1,2}$, and $Z_{1,3}$ to track the dimension multiplier information associated with the nested loop index variables i , j , and k , respectively. The dimension multiplier information is described in more detail below with reference to FIG. 2D below.

[0029] FIGS. 2B-2H show an example of how the tensor status elements 200 may be used by a tensor traversal unit to process a tensor. Referring to FIG. 2B, a program 212 may be stored in the storage medium 104 or another storage medium that can be executed by the processing unit 102. The program 212 specifies a character array variable V having a first dimension of 4 and a second dimension of 2. The program 212 specifies a nested for-loop for traversing the variable V , where the for-loop traverses the first dimension of V in an outer loop tracked by a nested loop index variable i , and traverses the second dimension of V in an inner loop tracked by a nested loop index variable j . In some implementations, the tensor status elements 200 may be initialized at the beginning of a program. For example, a processor may execute an instruction “*InitializeElements*” that initializes the tensor status elements 200. The instruction may be a hardware instruction of an instruction set executable by a processor. In some implementations, after initialization, each element of the tensor status elements 200 is

set to a predetermined value. Here, after initialization, each element of the tensor status elements 200 may be set to zero.

[0030] FIG. 2C illustrates setting tensor bound information in the tensor bound elements 204 according to the program 212. For example, a processor may execute an instruction “*SetTensorBound*” that sets tensor bound information in the tensor bound elements 204. In some implementations, the tensor bound information represents an upper bound of the corresponding nested loop index variable. Here, the inner nested loop index variable i has an upper bound of 4, and the outer nested loop index variable j has an upper bound of 2. After setting tensor bound information in the tensor bound elements 204, the element 214 (i.e., $Y_{1,2}$) is set to 4, and the element 216 (i.e., $Y_{1,1}$) is set to 2.

[0031] FIG. 2D illustrates setting dimension multiplier information in the dimension multiplier elements 206 according to the program 212. For example, the processor may execute an instruction “*SetDimensionMultiplier*” that sets dimension multiplier information in the dimension multiplier elements 206. In some implementations, the dimension multiplier information represents a multiplier that a corresponding element in the tensor index elements 202 may be multiplied by. Here, the character variable V has a first dimension of 4 and a second dimension of 2. If the element $V[0][0]$ is stored at a base memory address P and the element $V[0][1]$ is stored at a memory address $P+1$, the element $V[1][0]$ is stored at a memory address $P+2$. The corresponding memory address for an element $V[i][j]$ is therefore $P+(i \times 2)+(j \times 1)$. Thus, the dimension multiplier information for the outer loop is 2, and the dimension multiplier information for the inner loop is 1. After setting dimension multiplier information in the dimension multiplier elements 206, the element 222 (i.e., $Z_{1,2}$) is set to 2, and the element 224 (i.e., $Z_{1,1}$) is set to 1.

[0032] FIG. 2E illustrates accessing the element $V[0][0]$ according to the program 212. For example, the processor may execute an instruction “*LocateTensor*” that locates a memory address that corresponds to the element being accessed. In some implementations, the instruction may include a base memory address. For example, the instruction “*LocateTensor*” may include a memory address of the element $V[0][0]$

that is the first element of the variable V . In some implementations, the instruction may include a row number corresponding to a tensor to be accessed. For example, the instruction “*LocateTensor*” may include a row number corresponding to the variable V . Here, the row number is 1.

[0033] In some implementations, in response to receiving the instruction, an arithmetic logic unit (e.g., the arithmetic logic unit 106) determines a memory address offset by calculating a sum of the multiplicative product between values stored in each of the tensor index elements 202 and values stored in the corresponding dimension multiplier elements 206. Here, an arithmetic logic unit multiplies the value stored in element $X_{1,N}$ by the value stored in element $Z_{1,N}$, the value stored in element $X_{1,N-1}$ by the value stored in element $Z_{1,N-1}$, . . . , the value stored in element $X_{1,2}$ by the value stored in element $Z_{1,2}$, and the value stored in element $X_{1,1}$ by the value stored in element $Z_{1,1}$. The arithmetic logic unit then sums all the multiplied products together to determine the memory address that corresponds to the element being accessed. Since all values in the tensor index elements 202 are zero, the memory address offset is zero, as indicated by the box 230. The processor can then access the element $V[0][0]$ by adding the base memory address (i.e., P) to the memory address offset (i.e., 0) to determine a memory address, and accessing the stored data in the storage medium based on the determined memory address.

[0034] FIG. 2F illustrates accessing the element $V[0][1]$ according to the program 212. For example, after the program has iterated through first inner loop, the nested loop index variables are updated (i.e., $i=0$ and $j=0$), the processor may execute an instruction “*IterateTensor*” that updates the tensor index elements 202 as the program enters the second inner loop (i.e., $i=0$ and $j=1$). In some implementations, the tensor traversal unit updates the tensor index elements 202 by incrementing the tensor index element having the lowest index by a predetermined value. For example, referring to FIG. 2F, the tensor traversal unit may increment the tensor index element 228 (i.e., $X_{1,1}$) by 1. In some other implementations, the tensor traversal unit may update the tensor index elements 202 by decrementing the tensor index element having the lowest index by a predetermined value.

[0035] In some implementations, the tensor traversal unit compares the updated values in the tensor index elements 202 with the values in the tensor bound elements 204, and determines whether a threshold has been satisfied. For example, the tensor traversal unit may compare the updated value (i.e., 1) in the tensor index element 228 with the value (i.e., 2) in the tensor bound element 216, and determines whether the updated value in the tensor index element 228 is smaller than the value in the tensor bound element 216. Here, the updated value in the tensor index element 228 is smaller than the value in the tensor bound element 216, and the updated value in the tensor index element 228 remains 1.

[0036] The processor can then access the element $V[0][1]$ by executing the instruction “*LocateTensor*” to locate the memory address that corresponds to $V[0][1]$. Referring to FIG. 2F, all values in the tensor index elements 202 except the tensor index element 228 are zero. The tensor index element 228 has a value of 1, and the arithmetic logic unit may determine that the memory address offset is 1, as indicated by the box 232. The processor can then access the element $V[0][1]$ by adding the base memory address (i.e., P) to the memory address offset (i.e., 1) to determine a memory address, and accessing the stored data in the storage medium based on the determined memory address.

[0037] FIG. 2G illustrates accessing the element $V[1][0]$ according to the program 212. For example, after the program has iterated through inner loop for the second time, the nested loop index variables are updated (i.e., $i=1$ and $j=0$), and the processor may execute an instruction “*IterateTensor*” that updates the tensor index elements 202. For example, the tensor traversal unit may increment the tensor index element 228 (i.e., $X_{1,1}$) by 1. The tensor traversal unit may compare the updated value (i.e., 2) in the tensor index element 228 with the value (i.e., 2) in the tensor bound element 216, and determines that the updated value in the tensor index element 228 is not smaller than the value in the tensor bound element 216. In some implementations, in response to determining that the threshold has not been satisfied, the tensor traversal unit may carry over the updated value to the next tensor index element and reset the current tensor index element. For example, since the updated value in the tensor index element 228 is not smaller than the value in the tensor bound element 216, the tensor traversal unit

may increment the tensor index element 226 by 1, and reset the tensor index element 228 back to 0.

[0038] The processor can then access the element $V[1][0]$ by executing the instruction “*LocateTensor*” to locate the memory address that corresponds to $V[1][0]$. Referring to FIG. 2G, all values in the tensor index elements 202 except the tensor index element 226 are zero. The tensor index element 226 has a value of 1, and the arithmetic logic unit may determine that the memory address offset is 2, as indicated by the box 234. The processor can then access the element $V[1][0]$ by adding the base memory address (i.e., P) to the memory address offset (i.e., 2) to determine a memory address, and accessing the stored data in the storage medium based on the determined memory address.

[0039] FIG. 2H illustrates accessing the element $V[3][1]$ according to the program 212. For example, the program may continue to iterate through the inner loop and the outer loop until the nested loop index variables are at their upper bound (i.e., $i=3$ and $j=1$). The processor can access the element $V[3][1]$ by executing the instruction “*LocateTensor*” to locate the memory address that corresponds to $V[3][1]$. Referring to FIG. 2H, all values in the tensor index elements 202 except the tensor index elements 226 and 228 are zero. The tensor index element 226 has a value of 3, and the tensor index element 228 has a value of 1. The arithmetic logic unit may determine that the memory address offset is 7, as indicated by the box 236. The processor can then access the element $V[3][1]$ by adding the base memory address (i.e., P) to the memory address offset (i.e., 7) to determine a memory address, and accessing the stored data in the storage medium based on the determined memory address.

[0040] The processor may then execute the instruction “*IterateTensor*” that updates the tensor index elements 202. For example, the tensor traversal unit may increment the tensor index element 228 (i.e., $X_{1,1}$) by 1. The tensor traversal unit may compare the updated value (i.e., 2) in the tensor index element 228 with the value (i.e., 2) in the tensor bound element 216, and determines that the updated value in the tensor index element 228 is not smaller than the value in the tensor bound element 216. Since the updated value in the tensor index element 228 is not smaller than the value in the

tensor bound element 216, the tensor traversal unit may increment the tensor index element 226 by 1, and reset the tensor index element 228 back to 0. The tensor traversal unit may compare the updated value (i.e., $3+1=4$) in the tensor index element 226 with the value (i.e., 4) in the tensor bound element 214, and determines that the updated value in the tensor index element 226 is not smaller than the value in the tensor bound element 214. Since the updated value in the tensor index element 228 is not smaller than the value in the tensor bound element 214, the tensor traversal unit may increment the next tensor index element (i.e., $X_{1,3}$) by 1, and reset the tensor index element 226 back to 0. However, since the corresponding tensor bound element (i.e., $Y_{1,3}$) is 0, the threshold is not satisfied. The tensor traversal unit will continue resetting the first row of the tensor index elements 202 until all of the tensor index elements at the first row are reset to 0.

[0041] In the example described above in references to FIGS. 2B-2H, the order of the tensor traversal for the element V is $V[0][0]$, $V[0][1]$, $V[1][0]$, $V[1][1]$, $V[2][0]$, $V[2][1]$, $V[3][0]$, and $V[3][1]$. In some implementations, a program may traverse the element V in the order of $V[0][0]$, $V[1][0]$, $V[2][0]$, $V[3][0]$, $V[0][1]$, $V[1][1]$, $V[2][1]$, and $V[3][1]$, by setting tensor bound information in the element 214 (i.e., $Y_{1,2}$) to 2 and the element 216 (i.e., $Y_{1,1}$) to 4, and by setting the dimension multiplier information in the element 222 (i.e., $Z_{1,2}$) is set to 1 and the element 224 (i.e., $Z_{1,1}$) is set to 2.

[0042] FIG. 3 shows an example set of tensor status elements 300 of a tensor traversal unit. The tensor status elements 300 may correspond to the tensor status elements 122 of the tensor traversal unit 106 or the tensor status elements 200. The tensor traversal unit 300 includes a group of tensor index elements 302, a group of tensor bound elements 304, a group of dimension multiplier elements 306, and a group of branch target elements 308. The branch target elements 308 may be physically or logically arranged as a 2-D array, where each element of the branch target elements 308 has a corresponding element in the tensor index elements 302. In some implementations, each row of the branch target elements 308 may represent branch target information for a tensor. In some implementations, each column of the branch target elements 308 may represent branch target information for nested loop index variable values that are associated with a tensor. In some implementations, for a N -

dimension tensor, where there are N columns for the tensor index elements 302, the tensor bound elements 304, and the dimension multiplier elements 306, there may be $N+1$ columns of corresponding branch targets 308 because there are $N+1$ cases for meeting the bounds threshold.

[0043] In some implementations, branch target information may include a memory address of an instruction that is to be executed by a processor upon the completion of a loop iteration in a nested loop. For example, when the value in a tensor index element is updated, the processor may execute an instruction according to the memory address stored in the corresponding branch target element. Referring to FIG. 3, a program 312 may be stored in the storage medium 104 or another storage medium that can be executed by the processing unit 102. The program 312 specifies a character array variable V having a first dimension of 4 and a second dimension of 2. The program 312 specifies a nested for-loop for traversing the variable V , where the for-loop traverses the first dimension of V in an outer loop tracked by a nested loop index variable i , and traverses the second dimension of V in an inner loop tracked by a nested loop index variable j . The program 312 also specifies a condition for exiting the inner loop, which requires the processor to execute a *minimum* function instruction to compare the current value of the nested loop index variable i and the number 2. In some implementations, the memory address of the *minimum* function instruction may be stored in the branch target element 324, such that in the event that there is a carry-over from the tensor index element 326 to the tensor index element 328, the *minimum* function instruction will be executed by the processor to determine the branch condition for the inner loop and to set the tensor bound information for the inner loop. Similarly, the memory address of a different instruction may be stored in the branch target element 322, such that in the event that there is an increment in the tensor index element 326 without having a carry-over to the tensor index element 328, the different instruction will be executed by the processor.

[0044] FIG. 4 is a flow diagram that illustrates an example of a process 400 for determining an address of a multi-dimensional tensor variable. The process 400 may be performed by a system of one or more computers, e.g., the computing system 100. The system includes a tensor traversal unit having multiple tensor index elements and

multiple dimension multiplier elements, where each of the dimension multiplier elements has a corresponding tensor index element. In some implementations, the system may include multiple tensor bound elements, where each of the multiple tensor bound elements may have a corresponding tensor index element in the multiple tensor index elements. In some implementations, the system includes one or more arithmetic logic units.

[0045] The system obtains 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 (402). For example, the processing unit 102 may execute an instruction for accessing a particular element of a tensor.

[0046] In some implementations, the instruction may represent 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, where the first loop is iterated using a first index variable, and where the second loop is iterated using a second index variable. For example, a program 212 may be stored in the storage medium 104 or another storage medium that can be executed by the processing unit 102. The program 212 may specify a character array variable V having a first dimension of 4 and a second dimension of 2. The program 212 may specify a nested for-loop for traversing the variable V , where the for-loop traverses the first dimension of V in an outer loop tracked by a nested loop index variable i , and traverses the second dimension of V in an inner loop tracked by a nested loop index variable j .

[0047] The system determines, using one or more tensor index elements of multiple tensor index elements and one or more dimension multiplier elements of multiple dimension multiplier elements, an address of the particular element (404). In some implementations, the address of the particular element may be an address offset from another element of the N -dimensional tensor. For each tensor index element of the one or more tensor index elements, the system may multiply a value stored in the respective tensor index element by a value stored in the corresponding tensor index element of the one or more multiple tensor index elements to generate a respective multiplied value.

The system may determine the address offset by determining a sum of the multiplied values. For example, an arithmetic logic unit (e.g., the arithmetic logic unit 106) may determine a memory address offset by calculating a sum of the multiplicative product between values stored in each of the tensor index elements 202 and values stored in the corresponding dimension multiplier elements 206.

[0048] In some implementations, the system may set values stored in the multiple tensor index elements, the multiple dimension multiplier elements, and the multiple tensor bound elements to an initialization value. For example, a processor may execute an instruction “*InitializeElements*” that initialize the tensor status elements 200. After initialization, each element of the tensor status elements 200 may be set to zero.

[0049] In some implementations, the system may set a value stored in 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. The system may set a value stored in 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. For example, a processor may execute an instruction “*InitializeElements*” that initialize the tensor status elements 200. After initialization, each element of the tensor status elements 200 may be set to zero. The system may set a value stored in a first dimension multiplier element of the multiple dimension multiplier elements to a predetermined value. The system may set a value stored in a second dimension multiplier element of the multiple dimension multiplier elements to the upper bound value of the first index variable. For example, a processor may execute an instruction “*SetTensorBound*” that sets tensor bound information in the tensor bound elements 204. After setting tensor bound information in the tensor bound elements 204, the element 214 (i.e., $Y_{1,2}$) is set to 4, and the element 216 (i.e., $Y_{1,1}$) is set to 2.

[0050] In some implementations, the system may set a value stored in a first tensor index element of the multiple tensor index elements to a current value of the first index variable. The system may set a value stored in a second tensor index element of the multiple tensor index elements to a current value of the second index variable. The system may multiply the value stored in the first dimension multiplier element by the

value stored in the first tensor index element to obtain a first multiplied value. The system may multiply the value stored in the second dimension multiplier element by the value stored in the second tensor index element to obtain a second multiplied value. The system may determine the address by at least summing the first multiplied value and the second multiplied value. For example, an arithmetic logic unit multiplies the value stored in element $X_{1,N}$ by the value stored in element $Z_{1,N}$, the value stored in element $X_{1,N-1}$ by the value stored in element $Z_{1,N-1}$, . . . , the value stored in element $X_{1,2}$ by the value stored in element $Z_{1,2}$, and the value stored in element $X_{1,1}$ by the value stored in element $Z_{1,1}$. The arithmetic logic unit then sums all the multiplied products together to determine the memory address that corresponds to the element being accessed.

[0051] In some implementations, the system may receive an instruction to update the value stored in the first tensor index element. For example, the processor may execute an instruction “*IterateTensor*” that updates the tensor index elements 202. After receiving the instruction to update the value stored in the first tensor index element, the system may determine that a difference between the value stored in the first tensor index element and the value stored in the first tensor bound element satisfies a threshold. In response to determining that the difference between the value stored in the first tensor index element and the value stored in the first tensor bound element satisfies the threshold, the system may update the value stored in the first tensor index element. For example, referring to FIG. 2F, the tensor traversal unit may increment the tensor index element 228 (i.e., $X_{1,1}$) by 1. The updated value in the tensor index element 228 is smaller than the value in the tensor bound element 216, and the updated value in the tensor index element 228 remains 1

[0052] In some implementations, the first index variable may be incremented by a first incremental value each time the first loop completes. The system may increment the value stored in the first tensor index element by the first incremental value. For example, after the program enters the second inner loop (i.e., $i=0$ and $j=1$), the tensor traversal unit may increment the tensor index element 228 (i.e., $X_{1,1}$) by 1.

[0053] In some implementations, the system may receive an instruction to update the value stored in the first tensor index element. For example, the processor may execute an instruction “*IterateTensor*” that updates the tensor index elements 202. After receiving the instruction to update the value stored in the first tensor index element, the system may determine that a difference between the value stored in the first tensor index element and the value stored in the first tensor bound element does not satisfy a threshold. In response to determining that the difference between the value stored in the first tensor index element and the value stored in the first tensor bound element does not satisfy the threshold, the system may reset the value stored in the first tensor index element, and update the value stored in the second tensor index element. For example, the tensor traversal unit may increment the tensor index element 228 (i.e., $X_{1,1}$) by 1. The tensor traversal unit may compare the updated value (i.e., 2) in the tensor index element 228 with the value (i.e., 2) in the tensor bound element 216, and determines that the updated value in the tensor index element 228 is not smaller than the value in the tensor bound element 216. Since the updated value in the tensor index element 228 is not smaller than the value in the tensor bound element 216, the tensor traversal unit may increment the tensor index element 226 by 1, and reset the tensor index element 228 back to 0.

[0054] In some implementations, the second index variable may be incremented by a second incremental value each time the second loop completes. The system may increment the value stored in the second tensor index element by the second incremental value. For example, after the program enters the next loop of the outer loop (i.e., $i=1$ and $j=0$), the tensor traversal unit may increment the tensor index element 226 (i.e., $X_{1,2}$) by 1.

[0055] The system outputs data indicating the determined address for accessing the particular element of the N -dimensional tensor (406). For example, the processing unit 102 may access, using a memory address offset value, a particular element of a N -dimensional array variable in the storage medium.

[0056] In some implementations, the system includes multiple branch target elements, where each of the multiple branch target elements has a corresponding tensor index

element in the multiple tensor index elements, and where 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 respective branch target element. For example, the tensor traversal unit 300 includes a group of branch target elements 308 that store branch target information. Branch target information may include a memory address of an instruction that is to be executed by a processor upon the completion of a loop iteration in a nested loop.

[0057] Embodiments of the subject matter and the functional operations described in this specification can be implemented in digital electronic circuitry, in tangibly-embodied computer software or firmware, in computer hardware, including the structures disclosed in this specification and their structural equivalents, or in combinations of one or more of them. Embodiments of the subject matter described in this specification can be implemented as one or more computer programs, i.e., one or more modules of computer program instructions encoded on a tangible non transitory program carrier for execution by, or to control the operation of, data processing apparatus. Alternatively or in addition, the program instructions can be encoded on an artificially generated propagated signal, e.g., a machine-generated electrical, optical, or electromagnetic signal, that is generated to encode information for transmission to suitable receiver apparatus for execution by a data processing apparatus. The computer storage medium can be a machine-readable storage device, a machine-readable storage substrate, a random or serial access memory device, or a combination of one or more of them.

[0058] The term “data processing apparatus” encompasses all kinds of apparatus, devices, and machines for processing data, including by way of example a programmable processor, a computer, or multiple processors or computers. The apparatus can include special purpose logic circuitry, e.g., an FPGA (field programmable gate array) or an ASIC (application specific integrated circuit). The apparatus can also include, in addition to hardware, code that creates an execution environment for the computer program in question, e.g., code that constitutes processor firmware, a protocol stack, a database management system, an operating system, or a combination of one or more of them.

[0059] A computer program (which may also be referred to or described as a program, software, a software application, a module, a software module, a script, or code) can be written in any form of programming language, including compiled or interpreted languages, or declarative or procedural languages, and it can be deployed in any form, including as a stand alone program or as a module, component, subroutine, or other unit suitable for use in a computing environment. A computer program may, but need not, correspond to a file in a file system. A program can be stored in a portion of a file that holds other programs or data, e.g., one or more scripts stored in a markup language document, in a single file dedicated to the program in question, or in multiple coordinated files, e.g., files that store one or more modules, sub programs, or portions of code. A computer program can be deployed to be executed on one computer or on multiple computers that are located at one site or distributed across multiple sites and interconnected by a communication network.

[0060] The processes and logic flows described in this specification can be performed by one or more programmable computers executing one or more computer programs to perform functions by operating on input data and generating output. The processes and logic flows can also be performed by, and apparatus can also be implemented as, special purpose logic circuitry, e.g., an FPGA (field programmable gate array), an ASIC (application specific integrated circuit), or a GPGPU (General purpose graphics processing unit).

[0061] Computers suitable for the execution of a computer program include, by way of example, can be based on general or special purpose microprocessors or both, or any other kind of central processing unit. Generally, a central processing unit will receive instructions and data from a read only memory or a random access memory or both. The essential elements of a computer are a central processing unit for performing or executing instructions and one or more memory devices for storing instructions and data. Generally, a computer will also include, or be operatively coupled to receive data from or transfer data to, or both, one or more mass storage devices for storing data, e.g., magnetic, magneto optical disks, or optical disks. However, a computer need not have such devices. Moreover, a computer can be embedded in another device, e.g., a mobile telephone, a personal digital assistant (PDA), a mobile audio or video player, a

game console, a Global Positioning System (GPS) receiver, or a portable storage device, e.g., a universal serial bus (USB) flash drive, to name just a few.

[0062] Computer readable media suitable for storing computer program instructions and data include all forms of non volatile memory, media and memory devices, including by way of example semiconductor memory devices, e.g., EPROM, EEPROM, and flash memory devices; magnetic disks, e.g., internal hard disks or removable disks; magneto optical disks; and CD ROM and DVD-ROM disks. The processor and the memory can be supplemented by, or incorporated in, special purpose logic circuitry.

[0063] To provide for interaction with a user, embodiments of the subject matter described in this specification can be implemented on a computer having a display device, e.g., a CRT (cathode ray tube) or LCD (liquid crystal display) monitor, for displaying information to the user and a keyboard and a pointing device, e.g., a mouse or a trackball, by which the user can provide input to the computer. Other kinds of devices can be used to provide for interaction with a user as well; for example, feedback provided to the user can be any form of sensory feedback, e.g., visual feedback, auditory feedback, or tactile feedback; and input from the user can be received in any form, including acoustic, speech, or tactile input. In addition, a computer can interact with a user by sending documents to and receiving documents from a device that is used by the user; for example, by sending web pages to a web browser on a user's client device in response to requests received from the web browser.

[0064] Embodiments of the subject matter described in this specification can be implemented in a computing system that includes a back end component, e.g., as a data server, or that includes a middleware component, e.g., an application server, or that includes a front end component, e.g., a client computer having a graphical user interface or a Web browser through which a user can interact with an implementation of the subject matter described in this specification, or any combination of one or more such back end, middleware, or front end components. The components of the system can be interconnected by any form or medium of digital data communication, e.g., a communication network. Examples of communication networks include a local area network ("LAN") and a wide area network ("WAN"), e.g., the Internet.

[0065] The computing system can include clients and servers. A client and server are generally remote from each other and typically interact through a communication network. The relationship of client and server arises by virtue of computer programs running on the respective computers and having a client-server relationship to each other.

[0066] While this specification contains many specific implementation details, these should not be construed as limitations on the scope of any invention or of what may be claimed, but rather as descriptions of features that may be specific to particular embodiments of particular inventions. Certain features that are described in this specification in the context of separate embodiments can also be implemented in combination in a single embodiment. Conversely, various features that are described in the context of a single embodiment can also be implemented in multiple embodiments separately or in any suitable subcombination. Moreover, although features may be described above as acting in certain combinations and even initially claimed as such, one or more features from a claimed combination can in some cases be excised from the combination, and the claimed combination may be directed to a subcombination or variation of a subcombination.

[0067] Similarly, while operations are depicted in the drawings in a particular order, this should not be understood as requiring that such operations be performed in the particular order shown or in sequential order, or that all illustrated operations be performed, to achieve desirable results. In certain circumstances, multitasking and parallel processing may be advantageous. Moreover, the separation of various system modules and components in the embodiments described above should not be understood as requiring such separation in all embodiments, and it should be understood that the described program components and systems can generally be integrated together in a single software product or packaged into multiple software products.

[0068] Particular embodiments of the subject matter have been described. Other embodiments are within the scope of the following claims. For example, the actions recited in the claims can be performed in a different order and still achieve desirable

results. As one example, the processes depicted in the accompanying figures do not necessarily require the particular order shown, or sequential order, to achieve desirable results. In certain implementations, multitasking and parallel processing may be advantageous.

[0069] What is claimed is:

CLAIMS

1. An apparatus for processing an instruction for accessing a N -dimensional tensor, the apparatus comprising:
 - multiple tensor index elements;
 - multiple dimension multiplier elements, wherein each of the dimension multiplier elements has a corresponding tensor index element; and
 - one or more processors configured to:
 - obtain 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;
 - 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.
2. The apparatus of claim 1,
 - wherein the address of the particular element is an address offset from another element of the N -dimensional tensor, and
 - wherein determining the address of the particular element comprises:
 - for each tensor index element of the one or more tensor index elements, multiplying a value stored in the respective tensor index element by a value stored in the corresponding tensor index element of the one or more multiple tensor index elements to generate a respective multiplied value; and
 - determining the address offset by determining a sum of the multiplied values.
3. 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.

4. 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.

5. The apparatus of claim 4, 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.

6. The apparatus of claim 5, wherein the one or more processors are further configured to:

set values stored in the multiple tensor index elements, the multiple dimension multiplier elements, and the multiple tensor bound elements to an initialization value.

7. The apparatus of claim 5, wherein the one or more processors are further configured to:

set a value stored in 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;

set a value stored in 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;

set a value stored in a first dimension multiplier element of the multiple dimension multiplier elements to a predetermined value; and

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

8. The apparatus of claim 7, wherein determining the address of the element of the N -dimensional tensor comprises:

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

multiplying the value stored in the second dimension multiplier element by the 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.

9. The apparatus of claim 8, wherein the one or more processors are further configured to:

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

10. The apparatus of claim 9,

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

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

11. The apparatus of claim 8, wherein the one or more processors are further configured to:

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

reset the value stored in the first tensor index element; and
update the value stored in the second tensor index element.

12. The apparatus of claim 11,
wherein the second index variable is incremented by a second incremental value each time the second loop completes, and

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

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

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

15. 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 respective branch target element.

16. A system comprising:
- one or more processors configured to perform linear algebra operations on 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;
 - multiple tensor index elements;
 - multiple dimension multiplier elements, wherein each of the dimension multiplier elements has a corresponding tensor index element; and
 - circuitry configured to:
 - obtain an instruction to access a particular element of the N -dimensional tensor;
 - 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, to the one or more processors, data indicating the determined address for accessing the particular element of the N -dimensional tensor.
17. The system of claim 16,
- wherein the address of the particular element is an address offset from another element of the N -dimensional tensor, and
 - wherein determining the address of the particular element comprises:
 - for each tensor index element of the one or more tensor index elements, multiplying a value stored in the respective tensor index element by a value stored in the corresponding tensor index element of the one or more multiple tensor index elements to generate a respective multiplied value; and
 - determining the address offset by determining a sum of the multiplied values.

18. A computer-implemented method, 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;
determining, using one or more tensor index elements of multiple tensor index elements and one or more dimension multiplier elements of multiple dimension multiplier elements, an address of the particular element; and
outputting data indicating the determined address for accessing the particular element of the N -dimensional tensor.
19. The method of claim 18,
wherein the address of the particular element is an address offset from another element of the N -dimensional tensor, and
wherein determining the address of the particular element comprises:
for each tensor index element of the one or more tensor index elements, multiplying a value stored in the respective tensor index element by a value stored in the corresponding tensor index element of the one or more multiple tensor index elements to generate a respective multiplied value; and
determining the address offset by determining a sum of the multiplied values.
20. The method of claim 18, 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.