Further, the format conversion unit 405 converts the first matrix in the compressed format into the N first column vectors, converts the second matrix in the compressed format into the N first row vectors, and writes the N first column vectors and the N first row vectors into the ~~first~~ second cache 406. The vector outer product processing engine 401 obtains the first row vectors and the second row vectors from the second cache 406.

[00115] In a second case, one of the fifth matrix and the sixth matrix is a matrix in the uncompressed format, and the other matrix is a matrix in the compressed format. An example in which the fifth matrix is a matrix in the uncompressed format, and the sixth matrix is a matrix in the compressed format is used for description. If the sixth matrix is a compressed matrix in a CSC or CSR format, the format conversion unit converts both the fifth matrix and the sixth matrix into compressed matrices in a COO format.

[00116] Optionally, the format conversion unit 405 is further configured to convert a matrix in the compressed format into a matrix in a target compressed format. For example, when both the first matrix and the second matrix are in the CSC format or the CSR format, the format conversion unit converts both the first matrix and the second matrix into the COO format.

[00117] In this example, the matrix calculation apparatus may convert the matrix in the uncompressed format into the matrix in the compressed format by using the format conversion unit, so that the matrix calculation apparatus can support both calculation of the matrix in the compressed format and calculation of the matrix in the uncompressed format. Optionally, the format conversion unit may convert a matrix in a non-target compressed format into a matrix in the target compressed format (for example, the COO format). In this example, the matrix calculation apparatus may convert a matrix in another compressed format into a matrix in the target compressed format, and perform matrix calculation on the matrix in the target compressed format. The matrix calculation apparatus provided in this application may support calculation of matrices in various formats.

[00118] Optionally, to improve applicability of the matrix calculation unit, high-precision matrix calculation may be implemented based on a low-precision matrix calculation apparatus. Precision of element values included in the first column vector and the first row vector is first precision. The format conversion unit 405 splits the first column vector into X second column vectors, and splits the first row vector into X second row vectors, where the second column vector and the second row vector include element values of second precision, and the first precision is

calculation circuit is configured to perform one or more steps in step 1301 to step 1303 or one or more steps in step 1401 to step 1408 in the foregoing method embodiments. In actual application, the matrix calculation circuit may be an ASIC, an FPGA, a logic circuit, or the like.

[00170] Another embodiment of this application provides a matrix calculation system or a chip. A structure of the system or the chip may be shown in FIG. 3, and includes a processor (using a central processing unit as an example) 1 and a matrix calculation apparatus 1. The processor 1 is configured to send a calculation instruction to the matrix calculation apparatus 1. The matrix calculation apparatus 12 is configured to perform one or more steps in step 1301 to step 1303 or one or more steps in step 1401 to step 1408 in the foregoing method embodiments.

[00171] Still another embodiment of this application provides a matrix calculation device. A structure of the device may be shown in FIG. 2. The device may be specifically a PCIe card, an SoC, a processor, a server including the foregoing hardware, and the like. Refer to FIG. 2. The device includes a memory 201, a processor 202, a communication interface 203, and a bus 204. The communication interface 203 may include an input interface and an output interface.

[00172] The processor 202 may be configured to perform one or more steps in step 1301 to step 1303 or one or more steps in step 1401 to step 1408 in the foregoing method embodiments. In some feasible embodiments, the processor 202 may include a matrix calculation unit, and the matrix calculation unit may be configured to support the processor in performing one or more steps in the foregoing method embodiments. In actual application, the matrix calculation unit may be an ASIC, an FPGA, a logic circuit, or the like. Certainly, the matrix calculation unit may alternatively be implemented by using software. This is not specifically limited in this embodiment of this application.

[00173] It should be noted that the components of the matrix calculation circuit, the matrix calculation system, the matrix calculation device, and the like provided in embodiments of this application are separately configured to implement functions of corresponding steps in the foregoing method embodiments. Because the steps are described in detail in the foregoing method embodiments, details are not described herein again.

[00174] All or some of the foregoing embodiments may be implemented using software, hardware, firmware, or any combination thereof. When software is used to implement embodiments, the foregoing embodiments may be implemented completely or partially in a form of a computer program product. The computer program product includes one or more computer

CLAIMS

~~What is claimed is:~~

1. A matrix calculation apparatus, comprising a vector outer product processing engine and an accumulator, wherein

the vector outer product processing engine is configured to calculate vector outer products of N first column vectors and N first row vectors, to obtain N intermediate result matrices, wherein the first column vector comprises first element values and row coordinates of the first element values, the first row vector comprises second element values and column coordinates of the second element values, the intermediate result matrix comprises third element values and position coordinates of the third element values, the position coordinates comprise the row coordinates of the first element values and the column coordinates of the second element values, the N first column vectors are obtained by converting a first matrix in a compressed format, the N first row vectors are obtained by converting a second matrix in the compressed format, and N is an integer greater than or equal to 1; and

the accumulator is configured to accumulate, based on indexes of the position coordinates of the third element values, third element values with same position coordinates in the N intermediate result matrices, to obtain a result matrix.

2. The apparatus according to claim 1, wherein the N intermediate result matrices comprise at least a first intermediate result matrix and a second intermediate result matrix, position coordinates of third element values in the first intermediate result matrix are first position coordinates, and position coordinates of third element values in the second intermediate result matrix are second position coordinates; and

the accumulator is further specifically configured to:

write, in a generation sequence of the N intermediate result matrices, the third element values in the first intermediate result matrix into corresponding positions in a cache based on the first position coordinates; and

read, based on the second position coordinates of the third element values in the second intermediate result matrix, cached values that are at positions corresponding to the second position coordinates in the cache, and accumulate the third element values in the second intermediate result matrix and the cached values, to obtain a result matrix in an uncompressed format.

3. The apparatus according to claim 2, wherein the apparatus further comprises a matrix compression unit, wherein

the matrix compression unit is configured to compress the result matrix in the uncompressed format to obtain a result matrix in a compressed format.

4. The apparatus according to claim 1, wherein the accumulator is further specifically configured to:

sort third element values in the N intermediate result matrices based on position coordinates of the third element values; and

compare position coordinates in N intermediate result matrices obtained through sorting, add up third element values with same position coordinates, and delete position coordinates of a zero-element value, to obtain a result matrix in a compressed format.

5. The apparatus according to claim 1, wherein the apparatus further comprises a format conversion unit, wherein

the format conversion unit is configured to: obtain the first matrix and the second matrix, convert the first matrix into the N first column vectors, and convert the second matrix into the N first row vectors.

6. The apparatus according to claim 5, wherein the apparatus further comprises the format conversion unit, wherein

the format conversion unit is further configured to: obtain a fifth matrix and a sixth matrix, perform format conversion on the fifth matrix to obtain the first matrix, and perform format conversion on the sixth matrix to obtain the second matrix, wherein at least one of the fifth matrix and the sixth matrix is a matrix in an uncompressed format.

7. The apparatus according to any one of claims 1 to 6, wherein precision of element values comprised in the first column vector and the first row vector is first precision;

the vector outer product processing engine is further configured to calculate vector outer products of X second column vectors and X second row vectors to obtain X^2 fourth matrices, wherein the fourth matrix comprises fourth element values and position coordinates of the fourth element values, the position coordinates of the fourth element values comprise the row coordinates of the first element values and the column coordinates of the second element values, and precision of the fourth element values is the first precision, wherein the X second column vectors are obtained by splitting the first column vector, the X second row vectors are obtained by splitting

the first row vector, precision of element values comprised in the second column vector and the second row vector is second precision, the first precision is higher than the second precision, and X is an integer greater than or equal to 2; and

the accumulator is further configured to accumulate, based on indexes of the position coordinates of the fourth element values, fourth element values with same position coordinates in the X^2 fourth matrices, to obtain the intermediate result matrix, wherein precision of the third element values in the intermediate result matrix is the first precision.

8. A matrix calculation method, wherein the method is applied to a matrix calculation apparatus, and the method comprises:

obtaining a first calculation instruction, wherein the first calculation instruction comprises N first column vectors and N first row vectors;

calculating vector outer products of the N first column vectors and the N first row vectors, to obtain N intermediate result matrices, wherein the first column vector comprises first element values and row coordinates of the first element values, the first row vector comprises second element values and column coordinates of the second element values, the intermediate result matrix comprises third element values and position coordinates of the third element values, the position coordinates comprise the row coordinates of the first element values and the column coordinates of the second element values, the N first column vectors are obtained by converting a first matrix in a compressed format, the N first row vectors are obtained by converting a second matrix in the compressed format, and N is an integer greater than or equal to 1; and

accumulating, based on indexes of the position coordinates of the third element values, third element values with same position coordinates in the N intermediate result matrices, to obtain a result matrix.

9. The method according to claim 8, wherein the N intermediate result matrices comprise at least a first intermediate result matrix and a second intermediate result matrix, position coordinates of third element values in the first intermediate result matrix are first position coordinates, and position coordinates of third element values in the second intermediate result matrix are second position coordinates; and

the accumulating, based on indexes of the position coordinates of the third element values, third element values with same position coordinates in the N intermediate result matrices, to obtain a result matrix comprises:

writing, in a generation sequence of the N intermediate result matrices, the third element values in the first intermediate result matrix into corresponding positions in a cache based on the first position coordinates; and

reading, based on the second position coordinates of the third element values in the second intermediate result matrix, cached values that are at positions corresponding to the second position coordinates in the cache, and accumulating the third element values in the second intermediate result matrix and the cached values, to obtain a result matrix in an uncompressed format.

10. The method according to claim 9, wherein the method further comprises:

compressing the result matrix in the uncompressed format to obtain a result matrix in a compressed format.

11. The method according to claim 8, wherein the accumulating, based on indexes of the position coordinates of the third element values, third element values with same position coordinates in the N intermediate result matrices, to obtain a result matrix comprises:

sorting third element values in the N intermediate result matrices based on position coordinates of the third element values; and

comparing position coordinates in N intermediate result matrices obtained through sorting, adding up third element values with same position coordinates, and deleting position coordinates of a zero-element value, to obtain a result matrix in a compressed format.

12. The method according to claim 8, wherein before the obtaining a first calculation instruction, the method further comprises:

obtaining a second calculation instruction, wherein the second calculation instruction comprises the first matrix and the second matrix; and

converting the first matrix into the N first column vectors, and converting the second matrix into the N first row vectors.

13. The method according to claim 12, wherein before the obtaining a second calculation instruction, the method further comprises:

obtaining a third calculation instruction, wherein the third calculation instruction comprises a fifth matrix and a sixth matrix, and at least one of the fifth matrix and the sixth matrix is a matrix in an uncompressed format; and

performing format conversion on the fifth matrix to obtain the first matrix in the compressed format, and performing format conversion on the sixth matrix to obtain the second matrix.

14. The method according to any one of claims 8 to 13, wherein precision of element values comprised in the first column vector and the first row vector is first precision;

the calculating vector outer products of the N first column vectors and the N first row vectors, to obtain N intermediate result matrices comprises:

calculating vector outer products of X second column vectors and X second row vectors to obtain X^2 fourth matrices, wherein the fourth matrix comprises fourth element values and position coordinates of the fourth element values, the position coordinates of the fourth element values comprise the row coordinates of the first element values and the column coordinates of the second element values, and precision of the fourth element values is the first precision, wherein the X second column vectors are obtained by splitting the first column vector, the X second row vectors are obtained by splitting the first row vector, precision of element values comprised in the second column vector and the second row vector is second precision, the first precision is higher than the second precision, and X is an integer greater than or equal to 2; and

the method further comprises:

accumulating, based on indexes of the position coordinates of the fourth element values, fourth element values with same position coordinates in the X^2 fourth matrices, to obtain the intermediate result matrix, wherein precision of the third element values in the intermediate result matrix is the first precision.

15. A matrix calculation circuit, wherein the matrix calculation circuit is configured to perform the method according to any one of claims 8 to 14.

16. A matrix calculation system, wherein the system comprises a processor and a matrix calculation apparatus, the processor is configured to send a calculation instruction to the matrix calculation apparatus, and the matrix calculation apparatus is configured to perform the method according to any one of claims 8 to 14.

17. A chip, wherein the chip comprises a processor, a matrix calculation apparatus is integrated into the processor, and the matrix calculation apparatus is configured to perform the matrix calculation method according to any one of claims 8 to 14.

18. A matrix calculation device, wherein the device comprises the matrix calculation system according to claim 16 or the chip according to claim 17.