

WHAT IS CLAIMED IS:

1. A data processing method, comprising:

obtaining to-be-processed data inputted to a first calculating unit in a plurality of
5 calculating units, wherein the to-be-processed data comprises data of a first bit width;

obtaining a processing parameter of the first calculating unit, wherein the
processing parameter comprises a parameter of a second bit width; and

obtaining an output result of the first calculating unit based on the to-be-processed
data and the processing parameter,

10 wherein a bit width of to-be-processed data inputted to a second calculating unit
in the plurality of calculating units is different from a bit width of the to-be-processed
data inputted to the first calculating unit, and/or

a bit width of a processing parameter inputted to the second calculating unit is
different from a bit width of the processing parameter inputted to the first calculating
15 unit.

2. The method according to claim 1, wherein obtaining the to-be-processed data
inputted to the first calculating unit in the plurality of calculating units comprises:

obtaining first configuration information of the first calculating unit, wherein
20 the first configuration information comprises the first bit width to indicate
that the to-be-processed data inputted to the first calculating unit is of the first bit width,
and

at least two calculating units in the plurality of calculating units use different
first bit widths; and

obtaining, based on the first bit width, to-be-processed data whose bit width is the first bit width.

3. The method according to claim 1, wherein obtaining the processing parameter
5 of the first calculating unit comprises:

obtaining second configuration information of the first calculating unit, wherein
the second configuration information comprises the second bit width to
indicate that the processing parameter inputted to the first calculating unit is of the
second bit width, and

10 at least two calculating units in the plurality of calculating units use different
second bit widths; and

obtaining, based on the second bit width, a processing parameter whose bit width
is the second bit width.

15 4. The method according to any of claims 1-3, wherein the to-be-processed data
comprises input data from a plurality of input channels, the input data comprises at least
one input data block, and obtaining the output result of the first calculating unit based
on the to-be-processed data and the processing parameter comprises:

for each input channel of the plurality of input channels,
20 obtaining a target input data block from the at least one input data block for
the input channel;

obtaining a processing parameter block associated with the target input data
block from the processing parameter, wherein the processing parameter block has a
same size as the target input data block;

transforming the target input data block and the associated processing parameter block respectively according to a first transforming relationship, so to obtain a first matrix corresponding to the target input data block and a second matrix corresponding to the processing parameter;

5 performing a multiplication operation on the first matrix and the second matrix to obtain a multiplication operation result of the input channel;

accumulating the multiplication operation result of each of the plurality of input channels to obtain a third matrix of a target size; and

transforming the third matrix according to a second transforming relationship to
10 obtain the output result of the first calculating unit.

5. The method according to claim 4, wherein the output result of the first calculating unit comprises output results of a plurality of output channels, and after transforming the third matrix according to the second matrix transforming relationship
15 to obtain the output result of the first calculating unit, the method further comprises:

outputting the output results of the plurality of output channels in parallel.

6. The method according to claim 4, wherein obtaining the to-be-processed data inputted to the first calculating unit in the plurality of calculating units comprises:

20 inputting the input data from the plurality of input channels in parallel into a plurality of first storage areas, wherein

a number of the first storage areas is the same as a number of the input channels,
and

input data from different input channels is inputted into different first storage areas.

7. The method according to claim 6, wherein each of the plurality of first storage areas comprises a plurality of input line buffers, a number of lines of the input data is the same as a number of columns of the input data, a number of lines of the target input data block is the same as a number of input line buffers in a corresponding first storage area, and obtaining the target input data block from the at least one input data block for the input channel comprises:

reading data in parallel from a plurality of input line buffers of the input channel to obtain the target input data block.

10

8. The method according to claim 6 or 7, wherein two adjacent input data blocks in the input data have overlapping data therebetween.

9. The method according to claim 5, wherein outputting the output results of the plurality of output channels in parallel comprises:

in a case of outputting operation results of the plurality of output channels at a time, adding biases respectively to the output results of the plurality of output channels and outputting the output results added with the biases.

20

10. The method according to claim 5 or 9, further comprising:

inputting the output results of the plurality of output channels in parallel into a plurality of second storage areas, wherein

a number of the second storage areas is the same as a number of the output channels, and

output results of different output channels are input into different second storage areas.

11. The method according to claim 10, wherein each of the second storage areas
5 comprises a plurality of output line buffers;

the output results comprise a plurality of line of output data and a plurality of columns of output data;

a target output data block is obtained by reading data in parallel from the plurality of output line buffers in a bus-aligned manner and is written into a memory, wherein a
10 number of lines of the target output data block is the same as a number of columns of the target output data block.

12. The method according to any of claims 4-11, wherein before performing the multiplication operation on the first matrix and the second matrix, the method further
15 comprises:

obtaining third configuration information;

in a case where the third configuration information indicates that the first calculating unit supports a floating-point operation, processing floating-point data in the to-be-processed data.

20

13. A data processing apparatus, comprising:

a first obtaining module configured to obtain to-be-processed data inputted to a first calculating unit in a plurality of calculating units, wherein the to-be-processed data comprises data of a first bit width;

a second obtaining module configured to obtain a processing parameter of the first calculating unit, wherein the processing parameter comprises a parameter of a second bit width; and

5 a processing module configured to obtain an output result of the first calculating unit based on the to-be-processed data and the processing parameter,

wherein a bit width of to-be-processed data inputted to a second calculating unit in the plurality of calculating units is different from a bit width of the to-be-processed data inputted to the first calculating unit, and/or

10 a bit width of a processing parameter inputted to the second calculating unit is different from a bit width of the processing parameter inputted to the first calculating unit.

14. The apparatus according to claim 13, wherein the first obtaining module is further configured to:

15 obtain first configuration information of the first calculating unit, wherein

the first configuration information comprises the first bit width to indicate that the to-be-processed data inputted to the first calculating unit is of the first bit width, and

20 at least two calculating units in the plurality of calculating units use different first bit widths; and

obtain, based on the first bit width, to-be-processed data whose bit width is the first bit width,

the second obtaining module is further configured to:

obtain second configuration information of the first calculating unit, wherein

the second configuration information comprises the second bit width to indicate that the processing parameter inputted to the first calculating unit is of the second bit width, and

at least two calculating units in the plurality of calculating units use different
5 second bit widths; and

obtain, based on the second bit width, a processing parameter whose bit width is the second bit width.

15. The apparatus according to claim 13 or 14, wherein the to-be-processed data
10 comprises input data from a plurality of input channels, the input data comprises at least one input data block, and the processing module is further configured to:

for each input channel of the plurality of input channels,

obtain a target input data block from the at least one input data block for the
input channel;

15 obtain a processing parameter block associated with the target input data block from the processing parameter, wherein the processing parameter block has a same size as the target input data block;

transform the target input data block and the associated processing parameter block respectively according to a first transforming relationship, so to obtain a first
20 matrix corresponding to the target input data block and a second matrix corresponding to the processing parameter;

perform a multiplication operation on the first matrix and the second matrix to obtain a multiplication operation result of the input channel;

accumulate the multiplication operation result of each of the plurality of input

channels to obtain a third matrix of a target size; and

transform the third matrix according to a second transforming relationship to obtain the output result of the first calculating unit.

5 16. The apparatus according to claim 15, wherein the output result of the first calculating unit comprises output results of a plurality of output channels, and the apparatus further comprises:

an outputting module configured to output the output results of the plurality of output channels in parallel, wherein outputting the output results of the plurality of
10 output channels in parallel comprises:

in a case of outputting operation results of the plurality of output channels at a time, adding biases respectively to the output results of the plurality of output channels and outputting the output results added with the biases;

the outputting module is further configured to:

15 input the output results of the plurality of output channels in parallel into a plurality of second storage areas, wherein a number of the second storage areas is the same as a number of the output channels, and output results of different output channels are input into different second storage areas.

20 17. The apparatus according to claim 15, wherein the first obtaining module is further configured to:

input the input data from the plurality of input channels in parallel into a plurality of first storage areas, wherein a number of the first storage areas is the same as a number of the input channels, and input data from different input channels is inputted into

different first storage areas;

each of the plurality of first storage areas comprises a plurality of input line buffers,
a number of lines of the input data is the same as a number of columns of the input data,
a number of lines of the target input data block is the same as a number of input line

5 buffers in a corresponding first storage area;

the processing module is further configured to read data in parallel from a plurality
of input line buffers of the input channel to obtain the target input data block.

18. A data processing device, comprising:

10 a processor; and

a memory for storing a processor executable program,

wherein the program is executed by the processor to cause the processor to
implement a method according to any of claims 1-12.

15 19. A computer-readable storage medium having a computer program stored
thereon, wherein the computer program is executed by a processor to cause the
processor to implement a method according to any of claims 1-12.

20 20. A computer program product comprising machine executable instructions,
wherein the machine executable instructions are read and executed by a computer to
cause the computer to implement a method according to any of claims 1-12.