

CHIP AND CHIP-BASED DATA PROCESSING METHOD

TECHNICAL FIELD

[0001] Embodiments of the present specification relate to the field of data processing technologies, and more specifically, to chips and chip-based data processing methods.

BACKGROUND

[0002] With the development of artificial intelligence technologies, the need for computing power is also growing. In particular, neural network prediction as an online service has a very high requirement for a processing delay. In this case, software running on a server CPU generally cannot satisfy this need. In the existing technology, the processing delay is shortened by using a hardware acceleration method. In this case, hardware needs to have certain flexibilities, so as to adapt to an upgrade of a neural network structure and a change in a length of to-be-processed data. When a large computing power is needed, hardware sends to-be-processed data to a chip through a Peripheral Component Interconnect Express (PCIE) bus. After the chip performs computation, the chip transmits a result to a program memory through the PCIE bus. Operator units in a chip are generally designed to be relatively universal and specific operations are specified by using control parameters delivered by a server CPU. For example, a common convolution operator unit is used to process various convolution operations. However, if a program includes multiple calls of operator units in the chip, the overhead of data exchange between the server CPU and the chip will be very large. Currently, a common practice is to design each operator unit, such as convolution control, in the chip into formatted binary instructions, then map calls at each layer of a neural network to these instructions, integrate them into a sequence, and then send the sequence to the chip. Logic on the chip parses each instruction and starts a corresponding operator unit.

[0003] Therefore, a more effective chip-based data processing solution is needed.

SUMMARY

[0004] Embodiments of the present specification are intended to provide a more effective chip-based data processing solution to reduce deficiencies in the existing technology.

5 **[0005]** To achieve the previous objective, an aspect of the present specification provides a chip-based data processing method, where the chip includes an embedded CPU and multiple operator units, and is configured to execute a computational task, and the method is performed on the chip, and includes the following: obtaining to-be-processed data; and performing, by the embedded CPU, the following steps to
10 compute the computational task: configuring the multiple operator units, so that the multiple operator units respectively correspond to different types of computation included in the computational task, and computing the computational task by invoking the multiple operator units based on the to-be-processed data, to obtain a processing result.

15 **[0006]** Another aspect of the present specification provides a chip-based data processing method, where a chip includes an embedded CPU and multiple operator units, and is configured to perform predicted computation of a neural network, and the method is performed on the chip, and includes the following: obtaining to-be-processed data; and performing computation at each layer of the neural network
20 based on the to-be-processed data, to obtain a processing result, where the embedded CPU performs the following steps to perform computation at each layer of the neural network: configuring the multiple operator units, so that the multiple operator units respectively correspond to different types of computation included in the layer, and performing computation at each neuron in the layer by invoking the multiple operator
25 units based on a value of each neuron in an upper layer in the neural network, to obtain a value of each neuron in the layer.

[0007] In some embodiments, in the data processing method, the chip is connected to a server.

[0008] In some embodiments, in the data processing method, the to-be-processed
30 data is transmitted to the chip through the server.

[0009] In some embodiments, in the data processing method, the embedded CPU performs the steps by running a program-based assembly instruction, and the program is transmitted to the chip through the server.

[0010] In some embodiments, in the data processing method, the program is a C language program.

[0011] In some embodiments, in the data processing method, the invoking the multiple operator units includes the following: invoking the multiple operator units by using a function call in the C language program.

[0012] In some embodiments, in the data processing method, the chip further includes a DMA module, and the method further includes the following: after the processing result is obtained, controlling, by the embedded CPU, the DMA module to write the processing result into a server memory.

[0013] In some embodiments, in the data processing method, the invoking the multiple operator units includes the following: invoking the multiple operator units asynchronously.

[0014] In some embodiments, in the data processing method, the configuring the multiple operator units includes the following: determining a length of the to-be-processed data, and obtaining a parameter to be sent to each operator unit based on the length.

[0015] In some embodiments, in the data processing method, the chip is configured to perform predicted computation of multiple neural networks, the obtaining to-be-processed data includes the following: obtaining the to-be-processed data and a neural network type corresponding to the to-be-processed data, and the performing computation at each layer of the neural network based on the to-be-processed data, to obtain a processing result includes the following: performing computation at each layer of a neural network of the type corresponding to the to-be-processed data based on the to-be-processed data and the neural network type, to obtain a processing result.

[0016] Another aspect of the present specification provides a chip, configured to execute a computational task, where the chip includes the following: multiple operator units; and an embedded CPU, configured to perform the following steps to compute the computational task: configuring the multiple operator units, so that the multiple operator units respectively correspond to different types of computation included in the computational task, and computing the computational task by invoking the multiple operator units based on to-be-processed data input from the outside of the chip, to obtain a processing result.

[0017] Another aspect of the present specification provides a chip, configured to perform predicted computation of a neural network, where the chip includes the

following: multiple operator units; and an embedded CPU, configured to perform computation at each layer of the neural network based on to-be-processed data input from the outside of the chip, to obtain a processing result, where the embedded CPU performs the following steps to perform computation at each layer of the neural network: configuring the multiple operator units, so that the multiple operator units respectively correspond to different types of computation included in the layer, and performing computation at each neuron in the layer by invoking the multiple operator units based on a value of each neuron in an upper layer in the neural network, to obtain a value of each neuron in the layer.

10 **[0018]** In some embodiments, the chip is connected to a server.

[0019] In some embodiments, in the chip, the to-be-processed data is transmitted to the chip through the server.

[0020] In some embodiments, in the chip, the embedded CPU performs the steps by running a program-based assembly instruction, and the program is transmitted to the chip through the server.

[0021] In some embodiments, in the chip, the program is a C language program.

[0022] In some embodiments, in the chip, the embedded CPU is further configured to invoke the multiple operator units by using a function call in a C language.

20 **[0023]** In some embodiments, in the chip, the chip further includes a DMA module, and the embedded CPU is further configured to control the DMA module to write the processing result into a server memory after the processing result is obtained.

[0024] In some embodiments, in the chip, the embedded CPU is further configured to invoke the multiple operator units asynchronously.

25 **[0025]** In some embodiments, in the chip, the embedded CPU is further configured to determine a length of the to-be-processed data, and obtain a parameter to be sent to each operator unit based on the length.

[0026] In some embodiments, in the chip, the chip is an ASIC chip or an FPGA chip.

30 **[0027]** In some embodiments, the chip is configured to perform predicted computation of multiple neural networks, and the embedded CPU is further configured to perform computation at each layer of a neural network of a type corresponding to the to-be-processed data input from the outside of the chip based on the to-be-processed data and the neural network type, to obtain a processing result.

[0028] Another aspect of the present specification provides a computing device, including a memory and a processor, where the memory stores executable code, and the processor executes the executable code to perform the previously described data processing method.

5 [0029] According to the chip-based data processing solution in the embodiments of the present specification, because the embedded CPU is introduced, control of a relatively small operator can be completed inside the chip. Calling a C language function greatly reduces space of a program instruction. In addition, a C program on the embedded CPU can perfectly perform an operation of a control type, for example,
10 nesting various if/for/while statements. The C program on the embedded CPU can be used to determine a length of input data, and flexibly generate a parameter to be sent to each operator unit during computation. In addition, when a structure of the neural network is upgraded, a new program can be transmitted to the chip through the server, to perform predicted computation of the updated neural network.

15 BRIEF DESCRIPTION OF DRAWINGS

[0030] By describing the embodiments of the present specification with reference to the accompanying drawings, the embodiments of the present specification can be made clearer:

[0031] FIG. 1 is a schematic diagram illustrating a data processing system 100,
20 according to an embodiment of the present specification;

[0032] FIG. 2 shows a chip-based data processing method, according to an embodiment of the present specification;

[0033] FIG. 3 is a schematic diagram illustrating a neural network;

[0034] FIG. 4 shows a chip 400, according to an embodiment of the present
25 specification.

DESCRIPTION OF EMBODIMENTS

[0035] The following describes the embodiments of the present specification with reference to the accompanying drawings.

[0036] FIG. 1 is a schematic diagram illustrating a data processing system 100,
30 according to an embodiment of the present specification. As shown in FIG. 1, the

system 100 includes a server 11 and a chip 12, which are connected through a PCIE bus, for example. The server 11 includes a server CPU 111 and a memory 112. The chip 12 includes an embedded CPU 121, multiple operator units 122, and a DMA module 123. The chip 12 is configured to execute a computational task, for example, predicted computation of a neural network. When the server 11 needs to invoke the chip 12 to perform predicted computation of a neural network, etc., the server 11 transmits, by using the server CPU 111, for example, to-be-processed data used for prediction of the neural network from the memory to the chip through the PCIE bus. The embedded CPU 121 in the chip 12 schedules each operator unit 122 by running a program to complete computation at each layer of the neural network based on the to-be-processed data, to obtain a final processing result, namely, a prediction result of the neural network. After the processing result is obtained, the embedded CPU 121 controls the DMA module to transmit the processing result to the server memory 112 through the PCIE bus.

[0037] FIG. 2 shows a chip-based data processing method, according to an embodiment of the present specification. A chip includes an embedded CPU and multiple operator units, and is configured to execute a computational task, and the method is performed on the chip, and includes the following steps:

[0038] Step 202: Obtain to-be-processed data.

[0039] Step 204: The embedded CPU performs the following steps to compute the computational task: configuring the multiple operator units, so that the multiple operator units respectively correspond to different types of computation included in the computational task, and computing the computational task by invoking the multiple operator units based on the to-be-processed data, to obtain a processing result.

[0040] First, in step 202, the to-be-processed data is obtained. In some embodiments, the chip is connected to a server. When invoking the chip for computation, the server transmits the to-be-processed data to the chip. It can be understood that the chip is not limited to be connected to the server. For example, the chip can be standalone hardware used to compute a task, and the chip obtains the to-be-processed data by using an input module of the chip.

[0041] In some embodiments, the chip is located on a PCIE card, and is an ASIC chip or an FPGA chip. The server transmits the to-be-processed data to the chip through a PCIE bus by using a server CPU. In some embodiments, the server

transmits the to-be-processed data to the chip by using an application program running on the server. In some embodiments, the chip can simultaneously receive to-be-processed data transmitted by multiple application programs, and the chip collects the multiple pieces of to-be-processed data by using the embedded CPU.

5 **[0042]** In step 204, the embedded CPU performs the following steps to compute the computational task: configuring the multiple operator units, so that the multiple operator units respectively correspond to different types of computation included in the computational task, and computing the computational task by invoking the multiple operator units based on the to-be-processed data, to obtain a processing
10 result.

[0043] The embedded CPU performs the computation by running a program-based assembly instruction. The program is, for example, a C language program. The chip translates the program into an assembly language program by using a system-on-a-chip (SOC) of the chip, for example. The program is transmitted by the
15 server to the chip.

[0044] During computation, the embedded CPU first configures the multiple operator units by running a program, so that the multiple operator units respectively correspond to different types of computation included in the computational task. The different types of computation include, for example, matrix multiplication, vector
20 addition, and non-linear operation, etc. Then the embedded CPU continues to run the program to perform a specific computation process. The computation is performed based on the to-be-processed data. When the computation includes various types of computation, each corresponding operator unit is invoked to complete the computation. In a complex computational task, multiple rounds of computation may
25 be needed to obtain a final processing result. In each round of computation, an operator unit is reconfigured by using a predetermined program, so as to satisfy a computation need of the round.

[0045] In some embodiments, for example, the computational task is predicted computation of a neural network. In this case, the embedded CPU performs
30 computation at each layer of the neural network based on the to-be-processed data, to obtain a processing result, namely, a prediction result of the neural network. The embedded CPU performs the following steps to perform computation at each layer of the neural network: configuring the multiple operator units, so that the multiple operator units respectively correspond to different types of computation included in

the layer, and performing computation at each neuron in the layer by invoking the multiple operator units based on a value of each neuron in an upper layer in the neural network, to obtain a value of each neuron in the layer.

[0046] In other words, for each layer of the neural network, the embedded CPU reconfigures an operator unit, and sequentially performs computation at each neuron in the layer. In terms of the C language program, the program can be cycled in two loops. In the outer loop, the program is cycled n times to perform computation at each layer of the neural network except for an input layer, where n is a quantity of layers of the neural network except for the input layer. In the inner loop, for computation at each layer, the program is cycled m times to perform computation at each neuron in the layer, where m is a quantity of neurons.

[0047] FIG. 3 is a schematic diagram illustrating a neural network. As shown in FIG. 3, a left layer is an input layer, which includes two input neurons (1) and (2), an intermediate layer is a hidden layer, which includes three neurons (3), (4), and (5), and a right layer is an output layer, which includes output neuron (6). In the intermediate layer, parameters of neuron (3) for neurons in the upper layer are W_{13} and W_{23} , parameters of neuron (4) include W_{14} and W_{24} , and parameters of neuron (5) include W_{15} and W_{25} . Parameters of neuron (6) include W_{36} , W_{46} , and W_{56} .

[0048] During computation of the neural network shown in FIG. 3, first, computation is performed at the intermediate layer. For example, the computation at this layer includes matrix multiplication and vector addition. Therefore, the embedded CPU can configure two operator units respectively corresponding to the matrix multiplication and the vector addition, and respectively perform computation at the neurons (3), (4), and (5) by cycling a program. For example, during computation of neuron (3), to-be-processed data obtained from the input layer and the parameters (W_{13} and W_{23}) of neuron (3) can be input into the previously described two operator units to obtain an intermediate result of neuron (3), which can be considered as a value of neuron (3).

[0049] After computation at the intermediate layer is completed (i.e., values of the neurons (3), (4), and (5) are obtained), the embedded CPU enters a next cycle in the outer loop, in other words, the embedded CPU performs computation at the output layer. For example, the computation at the output layer includes sigmoid function computation. Therefore, the embedded CPU configures an operator unit for performing sigmoid function computation based on a running program, and performs

computation at neuron (6). During computation of neuron (6), the embedded CPU invokes the operator unit based on the running program, and inputs the values of the neuron (3), (4), and (5) and the parameters of neuron (6) into the operator unit, to obtain a value of neuron (6), where the value is a predicted value of the neural network.

[0050] It can be seen from the previous computation examples of the neural network shown in FIG. 3 that, during computation of each neuron, a corresponding operator unit is invoked to perform specific computation, to obtain a value of the neuron. When the neuron is located at the hidden layer of the neural network, the obtained value of the neuron is an intermediate result, and the intermediate result is stored in the chip or in a DDR memory in a PCIE card where the chip is located, and used for subsequent computation. When the neuron is located at the output layer of the neural network, the obtained value of the neuron is the predicted value of the neural network, and the embedded CPU configures a DMA descriptor, to control a DMA module to write the predicted value into a server memory.

[0051] The data processing method in the embodiments of the present specification is particularly applicable to predicted computation of a neural network. For computation of a neural network in a chip, currently, a common practice is to design each operator unit, such as convolution control, in the chip into formatted binary instructions, then map calls at each layer of the neural network to these instructions, integrate them into a sequence, and then send the sequence to the chip. Logic on the chip parses each instruction and starts a corresponding operator unit. However, the method is applicable to a relatively large operator. For control of a relatively small operator, the instruction sequence method has low efficiency, and a quantity of instruction sequences is very large. For an operation of a control type, the instruction sequence method can hardly be completed (e.g., it is difficult to nest multiple loops by using an instruction sequence). If an input data length adaptation problem exists, for example, if sizes of input images of a CNN network are different or input audio lengths of an LSTM network are different, an instruction sequence is unsuitable. If an instruction sequence is downloaded again each time input data lengths are different, the delay caused by this process is very long for the service.

[0052] According to the chip-based data processing solution in the embodiments of the present specification, because the embedded CPU is introduced, control of a relatively small operator can be completed inside the chip. Calling a C language

function greatly reduces space of a program instruction. In addition, the C program on the embedded CPU can perfectly perform an operation of a control type, for example, nesting various if/for/while statements. A C program on the embedded CPU can be used to determine a length of input data, and flexibly generate a parameter to be sent to each operator unit during computation. In addition, when a structure of the neural network is upgraded, a new program can be transmitted to the chip through the server, to perform predicted computation of the updated neural network.

[0053] In some embodiments, the embedded CPU is further configured to invoke the multiple operator units by using a function call in a C language.

[0054] In some embodiments, the chip further includes a DMA module, and the embedded CPU is further configured to control the DMA module to write the processing result into a server memory after the processing result is obtained.

[0055] In some embodiments, the embedded CPU is further configured to invoke the multiple operator units asynchronously. The multiple operator units are invoked asynchronously, so that the multiple operator units can work concurrently to the greatest extent possible, thereby improving hardware utilization.

[0056] In some embodiments, the embedded CPU is further configured to determine a length of the to-be-processed data or input data of each layer of the neural network, and obtain a parameter to be sent to each operator unit based on the length.

[0057] In some embodiments, the chip is configured to perform predicted computation of multiple neural networks, the obtaining to-be-processed data includes the following: obtaining the to-be-processed data and a neural network type corresponding to the to-be-processed data, and the performing computation at each layer of the neural network based on the to-be-processed data, to obtain a processing result includes the following: performing computation at each layer of a neural network of the type corresponding to the to-be-processed data based on the to-be-processed data and the neural network type, to obtain a processing result. The multiple neural networks include, for example, a convolutional neural network (CNN), a deep neural network (DNN), and a recurrent neural network (RNN). For example, the chip obtains control information (e.g., a neural network identifier) indicating a neural network type, to obtain the neural network type. After obtaining the neural network type, the chip selects a program corresponding to the type from the embedded CPU based on the neural network type, and runs the program, to perform computation of the neural network of the type.

[0058] FIG. 4 shows a chip 400, according to an embodiment of the present specification. The chip is configured to execute a computational task, and the chip includes the following: multiple operator units 41; and an embedded CPU 42, configured to perform the following steps to compute the computational task:

5 configuring the multiple operator units, so that the multiple operator units respectively correspond to different types of computation included in the computational task, and computing the computational task by invoking the multiple operator units based on to-be-processed data input from the outside of the chip, to obtain a processing result.

[0059] In some embodiments, the computational task is predicted computation of a neural network, and the embedded CPU 42 is configured to perform computation at each layer of the neural network based on to-be-processed data input from the outside of the chip, to obtain a processing result, where the embedded CPU performs the following steps to perform computation at each layer of the neural network:

10 configuring the multiple operator units, so that the multiple operator units respectively correspond to different types of computation included in the layer, and performing computation at each neuron in the layer by invoking the multiple operator units based on a value of each neuron in an upper layer in the neural network, to obtain a value of each neuron in the layer.

[0060] In some embodiments, the chip is connected to a server.

20 **[0061]** In some embodiments, in the chip, the to-be-processed data is transmitted to the chip through the server.

[0062] In some embodiments, in the chip, the embedded CPU performs the steps by running a program-based assembly instruction, and the program is transmitted to the chip through the server.

25 **[0063]** In some embodiments, in the chip, the program is a C language program.

[0064] In some embodiments, in the chip, the embedded CPU is further configured to invoke the multiple operator units by using a function call in a C language.

[0065] In some embodiments, in the chip, the chip further includes a DMA module 43, and the embedded CPU is further configured to control the DMA module to write the processing result into a server memory after the processing result is obtained.

30

[0066] In some embodiments, in the chip, the embedded CPU is further configured to invoke the multiple operator units asynchronously.

[0067] In some embodiments, in the chip, the embedded CPU is further configured to determine a length of the to-be-processed data, and obtain a parameter to be sent to each operator unit based on the length.

5 **[0068]** In some embodiments, in the chip, the chip is an ASIC chip or an FPGA chip.

[0069] In some embodiments, the chip is configured to perform predicted computation of multiple neural networks, and the embedded CPU is further configured to perform computation at each layer of a neural network of a type corresponding to the to-be-processed data input from the outside of the chip based on
10 the to-be-processed data and the neural network type, to obtain a processing result.

[0070] Another aspect of the present specification provides a computing device, including a memory and a processor, where the memory stores executable code, and the processor executes the executable code to perform the data processing method shown in FIG. 2

15 **[0071]** According to the chip-based data processing solution in the embodiments of the present specification, because the embedded CPU is introduced, control of a relatively small operator can be completed inside the chip. Calling a C language function greatly reduces space of a program instruction. In addition, a C program on the embedded CPU can perfectly perform an operation of a control type, for example,
20 nesting various if/for/while statements. The C program on the embedded CPU can be used to determine a length of input data, and flexibly generate a parameter to be sent to each operator unit during computation. In addition, when a structure of the neural network is upgraded, a new program can be transmitted to the chip through the server, to perform predicted computation on the updated neural network.

25 **[0072]** The embodiments of the present specification are described in a progressive way. For same or similar parts of the embodiments, references can be made to the embodiments. Each embodiment focuses on a difference from other embodiments. Particularly, a system embodiment is basically similar to a method embodiment, and therefore, is described briefly. For related parts, references can be made to related
30 descriptions in the method embodiment.

[0073] Specific embodiments of the present specification are described above. Other embodiments fall within the scope of the appended claims. In some situations, the actions or steps described in the claims can be performed in an order different from the order in the embodiment and the desired results can still be achieved. In

addition, the process depicted in the accompanying drawings does not necessarily require a particular execution order to achieve the desired results. In some embodiments, multi-tasking and parallel processing may be advantageous.

[0074] A person of ordinary skill in the art should be further aware that, in

5 combination with the examples described in the embodiments disclosed in the present specification, units and algorithm steps can be implemented by electronic hardware, computer software, or a combination thereof. To clearly describe interchangeability between the hardware and the software, compositions and steps of each example are described above based on functions. Whether the functions are performed by
10 hardware or software depends on particular applications and design constraint conditions of the technical solutions. A person of ordinary skill in the art can use different methods to implement the described functions for each particular application, but it should not be considered that the embodiment goes beyond the scope of the present application.

15 **[0075]** Steps of methods or algorithms described in the embodiments disclosed in the present specification can be implemented by hardware, a software module executed by a processor, or a combination thereof. The software module can reside in a random access memory (RAM), a memory, a read-only memory (ROM), an electrically programmable ROM, an electrically erasable programmable ROM, a
20 register, a hard disk, a removable disk, a CD-ROM, or any other form of storage medium known in the art.

[0076] In the described specific embodiments, the objective, technical solutions, and benefits of the present disclosure are further described in detail. It should be understood that the descriptions are merely specific embodiments of the present
25 disclosure, but are not intended to limit the protection scope of the present disclosure. Any modification, equivalent replacement, or improvement made without departing from the spirit and principle of the present disclosure should fall within the protection scope of the present disclosure.

CLAIMS

What is claimed is:

1. A chip-based data processing method, wherein a chip comprises an embedded CPU and multiple operator units, and is configured to execute a computational task,
5 and the method is performed on the chip, and comprises:
 obtaining to-be-processed data; and
 performing, by the embedded CPU, the following steps to compute the computational task: configuring the multiple operator units, so that the multiple operator units respectively correspond to different types of computation comprised in
10 the computational task, and computing the computational task by invoking the multiple operator units based on the to-be-processed data, to obtain a processing result.
2. A chip-based data processing method, wherein a chip comprises an embedded CPU and multiple operator units, and is configured to perform predicted computation
15 of a neural network, and the method is performed on the chip, and comprises:
 obtaining to-be-processed data; and
 performing computation at each layer of the neural network based on the to-be-processed data, to obtain a processing result, wherein the embedded CPU performs the following steps to perform computation at each layer of the neural
20 network: configuring the multiple operator units, so that the multiple operator units respectively correspond to different types of computation comprised in the layer, and performing computation at each neuron in the layer by invoking the multiple operator units based on a value of each neuron in an upper layer in the neural network, to obtain a value of each neuron in the layer.
- 25 3. The data processing method according to claim 1 or 2, wherein the chip is connected to a server.
4. The data processing method according to claim 3, wherein the to-be-processed data is transmitted to the chip through the server.
5. The data processing method according to claim 3, wherein the embedded CPU
30 performs the steps by running a program-based assembly instruction, and the program is transmitted to the chip through the server.
6. The data processing method according to claim 5, wherein the program is a C language program.

7. The data processing method according to claim 6, wherein the invoking the multiple operator units comprises: invoking the multiple operator units by using a function call in the C language program.

8. The data processing method according to claim 1 or 2, wherein the chip further comprises a DMA module, and the method further comprises: after the processing result is obtained, controlling, by the embedded CPU, the DMA module to write the processing result into a server memory.

9. The data processing method according to claim 1 or 2, wherein the invoking the multiple operator units comprises: invoking the multiple operator units asynchronously.

10. The data processing method according to claim 1 or 2, wherein the configuring the multiple operator units comprises: determining a length of the to-be-processed data, and obtaining a parameter to be sent to each operator unit based on the length.

11. The data processing method according to claim 2, wherein the chip is configured to perform predicted computation of multiple neural networks, the obtaining to-be-processed data comprises: obtaining the to-be-processed data and a neural network type corresponding to the to-be-processed data, and the performing computation at each layer of the neural network based on the to-be-processed data, to obtain a processing result comprises: performing computation at each layer of a neural network of the type corresponding to the to-be-processed data based on the to-be-processed data and the neural network type, to obtain a processing result.

12. A chip, configured to execute a computational task, wherein the chip comprises:

multiple operator units; and

an embedded CPU, configured to perform the following steps to compute the computational task: configuring the multiple operator units, so that the multiple operator units respectively correspond to different types of computation comprised in the computational task, and computing the computational task by invoking the multiple operator units based on to-be-processed data input from the outside of the chip, to obtain a processing result.

13. A chip, configured to perform predicted computation of a neural network, wherein the chip comprises:

multiple operator units; and

an embedded CPU, configured to perform computation at each layer of the neural network based on to-be-processed data input from the outside of the chip, to obtain a processing result, wherein the embedded CPU performs the following steps to perform computation at each layer of the neural network: configuring the multiple
5 operator units, so that the multiple operator units respectively correspond to different types of computation comprised in the layer, and performing computation at each neuron in the layer by invoking the multiple operator units based on a value of each neuron in an upper layer in the neural network, to obtain a value of each neuron in the layer.

10 14. The chip according to claim 12 or 13, wherein the chip is connected to a server.

15 15. The chip according to claim 14, wherein the to-be-processed data is transmitted to the chip through the server.

16. The chip according to claim 14, wherein the embedded CPU performs the steps by running a program-based assembly instruction, and the program is
15 transmitted to the chip through the server.

17. The chip according to claim 16, wherein the program is a C language program.

18. The chip according to claim 17, wherein the embedded CPU is further
20 configured to invoke the multiple operator units by using a function call in a C language.

19. The chip according to claim 12 or 13, wherein the chip further comprises a DMA module, and the embedded CPU is further configured to control the DMA module to write the processing result into a server memory after the processing result
25 is obtained.

20. The chip according to claim 12 or 13, wherein the embedded CPU is further configured to invoke the multiple operator units asynchronously.

21. The chip according to claim 12 or 13, wherein the embedded CPU is further configured to determine a length of the to-be-processed data, and obtain a parameter
30 to be sent to each operator unit based on the length.

22. The chip according to claim 12 or 13, wherein the chip is an ASIC chip or an FPGA chip.

23. The chip according to claim 13, wherein the chip is configured to perform predicted computation of multiple neural networks, and the embedded CPU is further

configured to perform computation at each layer of a neural network of a type corresponding to the to-be-processed data input from the outside of the chip based on the to-be-processed data and the neural network type, to obtain a processing result.

24. A computing device, comprising a memory and a processor, wherein the
5 memory stores executable code, and the processor executes the executable code to perform the method according to any one of claims 1 to 11.