CLAIMS

1. A chip-based data processing method, wherein a chip (12) comprises an embedded CPU (121) and multiple operator units (122), and is configured to perform prediction computation of a neural network, and the method is performed on the chip (12), and comprises:

obtaining (202) to-be-processed data; and

performing (204), by the embedded CPU (121), the following steps at each layer of the neural network based on the to be processed data to obtain a processing result: configuring the multiple operator units (122), so that the multiple operator units (122) one-to-one correspond to different types of computation comprised in the layer,

wherein the different types of computation comprise one of: a matrix multiplication, a vector addition, and a non-linear operation,

wherein the configuring the multiple operator units (122) comprises: determining a length of the to-be-processed data, and obtaining a parameter to be sent to each operator unit (122) based on the length,

and computing a computational task by invoking the multiple operator units (122) based on the to-be-processed data, to obtain a processing result.

2. The chip-based data processing method according to claim 1, wherein the method further comprises:

performing computation at each neuron in the layer by invoking the multiple operator units (122) 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.

3. The chip-based data processing method according to claim 1 or 2, wherein the chip (12) is connected to a server (11).

4. The chip-based data processing method according to claim 3, wherein the to-be-processed data is transmitted to the chip (12) through the server (11).

5. The chip-based data processing method according to claim 3, wherein the embedded CPU (121) performs the steps by running a program-based assembly instruction, and the program is transmitted to the chip (12) through the server (11).

6. The chip-based data processing method according to claim 5, wherein the program is a C language program.

7. The chip-based data processing method according to claim 6, wherein the invoking the multiple operator units (122) comprises: invoking the multiple operator units (122) by using a function call in the C language program.

8. The chip-based data processing method according to claim 1 or 2, wherein the chip (12) further comprises a DMA module (123), and the method further comprises: after the processing result is obtained, controlling, by the embedded CPU (121), the DMA module (123) to write the processing result into a server memory (112).

9. The chip-based data processing method according to claim 1 or 2, wherein the invoking the multiple operator units (122) comprises: invoking the multiple operator units asynchronously.

10. The chip-based data processing method according to claim 2, wherein the chip is configured to perform prediction 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 neural network 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.

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

multiple operator units (122); and

an embedded CPU (121), configured to perform a chip-based data processing method according to any one of the preceding claims.

12. The chip (12) according to claim 11, wherein the chip (12) is an ASIC chip or an FPGA chip.

13. A computing device, comprising a memory and a processor, wherein the memory stores executable code, and the processor executes the executable code to perform the chip-based data processing method according to any one of claims 1 to 10.