

STORAGE MANAGEMENT APPARATUS AND CHIP

TECHNICAL FIELD

5 The disclosure relates to the technical field of chips, in particular, relates to a storage management apparatus and chip.

BACKGROUND

With the development of science and technology, human society is rapidly entering an era of intelligence. The important feature of the era of intelligence is that people may acquire more and more types of data, the amount of data that people may acquire is
10 becoming larger and larger, and the demand for the data processing speed is becoming higher and higher.

Chips are the cornerstone of data processing, which essentially determine the ability to process data. From the perspective of application fields, the chips are mainly developed in two routes. One is general-purpose chips, for example, central processing
15 units (CPU) and the like, which can provide great flexibility but have low effective computing power when processing algorithms in a specific field; the other is special-purpose chips, for example, tensor processing units (TPU) and the like, which have high effective computing power in certain specific fields but are poor in processing capacity or even have no processing capacity in flexible and universal fields.

20 Due to various data types and numerous data volumes in the era of intelligence, the chips are required to not only process changing algorithms in different fields with extremely high flexibility, but also rapidly process a great number of dramatically increasing data volumes with extremely strong processing capacities.

25 SUMMARY

1. Objective

The present disclosure aims to provide a storage management apparatus and a chip. The storage management apparatus is capable of parsing an instruction received from an network on chip, generating at least one data packet based on data released from an
30 external storage apparatus, and sending all the data packets to the network on chip; all data packets generated by the storage management apparatus can be sent to the

corresponding processing cores through the network on chip, so that each of the processing cores that require the same data does not need to fetch the data from the storage apparatus separately, thereby reducing the delay, saving power consumption of the chip and the storage apparatus, and saving the time taken for the processing cores that require the same data to read the data.

2. Technical Solution

To solve the above problems, according to a first aspect of the present disclosure, there is provided a storage management apparatus, disposed between a network on chip and a storage apparatus, including: an instruction parser, configured to parse an instruction received from the network on chip, and generate a control signal according to the instruction; and a data packet generator, configured to generate at least one data packet based on data released from the storage apparatus according to the control signal, and send all the data packets to the network on chip.

The storage management apparatus provided by the embodiment of the present disclosure is capable of parsing an instruction received from an network on chip, generating at least one data packet based on data released from an external storage apparatus, and sending all the data packets to the network on chip; by providing the storage management apparatus between the network on chip and the storage apparatus, data packets are sent to all the processing cores that require the same data by the storage management apparatus, so that each of the processing cores that require the same data does not need to fetch the data from the storage apparatus separately, thereby reducing the delay, saving power consumption of the chip and the storage apparatus, and saving the time taken for the processing cores that require the same data to read the data.

Further, the data packet generator is configured to generate, based on the released data and a processing core indicated by the instruction, the data packet corresponding to the processing core according to the control signal.

Further, a plurality of processing cores are indicated by the instruction; and the data packet generator is configured to generate the data packets corresponding to each processing core based on the released data and the processing cores indicated by the instruction, and send the generated plurality of data packets to the network on chip in parallel.

Further, the instruction includes fetch information and storage information.

Optionally, the apparatus further includes an access indicator; the access indicator is configured to receive the fetch information, and instruct the storage apparatus to release the data indicated by the fetch information according to the fetch information; and the access indicator is further configured to receive the storage information, and
5 generate a storage address of the data packet corresponding to the processing core indicated by each instruction according to the storage information.

Optionally, the data packet generator is configured to generate an address of the processing core indicated by the instruction according to the control signal, and generate the data packet based on the released data, the storage address and the address
10 of the processing core indicated by the instruction.

Optionally, the data packet generator includes switches, and the number of the switches is greater than or equal to the number of processing cores connected to the network on chip; wherein each processing core corresponds to one of the switches, and the control signal is used to control states of the switches.

Optionally, the data packet generator is configured to turn on the switch corresponding to the processing core indicated by the instruction according to the control signal, and send the data packets to the network on chip in parallel, so as to send the data packets to the respective corresponding processing cores through the network
15 on chip.

Optionally, the switch includes a control terminal, an input terminal and an output terminal; the input terminal of each of the switches is configured to receive the storage address sent by the access indicator and the data released from the storage apparatus; and the control terminal of each of the switches is configured to implement an on state or an off state according to the control signal received.

Further optionally, the instruction parser is configured to send the control signal to the control terminal to control the switch to be turned on; and the data packet generator generates the data packet from the address of the processing core, the storage address obtained through the input terminal and the data.

Optionally, the instruction includes control bits configured to control the switches;
30 the control bit corresponding to the processing core indicated by each instruction is a preset value; and the control signal is generated when the instruction parser determines that the control bit is a preset value.

Preferably, the instruction includes the number of fetches, and the fetch information includes the starting address of the fetch addresses; the access indicator is configured to generate the fetch addresses according to the starting address of the fetch addresses and the number of fetches, and instruct the storage apparatus to release the data according to the fetch addresses.

Further preferably, the storage information includes the starting address of the storage addresses of the data in the processing core indicated by the instruction, and the access indicator is configured to generate the storage addresses according to the starting address of the storage addresses and the number of fetches.

Optionally, the instruction includes: control bits, information bits, starting address for data fetching, number of fetches, and starting address of the storage addresses of the data in each processing core that waits to receive the data.

The control bits are used to control states of the switches, and the information bits are used to indicate the processing cores for receiving the data.

Optionally, the storage management apparatus further includes an instruction cache, configured to receive the instruction sent by the network on chip, cache the instruction, and send the cached instruction to the instruction parser.

According to a second aspect of the present disclosure, there is provided a chip, including a network on chip, a plurality of processing cores connected to the network on chip, a storage apparatus, and the storage management apparatus provided in the first aspect.

According to a third aspect of the present disclosure, there is provided a circuit board, including one or more chips provided in the second aspect.

According to a fourth aspect of the present disclosure, there is provided an electronic device, including one or more circuit boards provided in the third aspect.

According to a fifth aspect of the present disclosure, there is provided a storage management method, including: parsing an instruction received from a network on chip, and generating a control signal according to the instruction; and generating at least one data packet based on data released from a storage apparatus according to the control signal, and send all the data packets to the network on chip.

According to a sixth aspect of the present disclosure, there is provided a computer storage medium storing a computer program, which, upon execution by a processor, implements the storage management method of the fifth aspect.

According to a seventh aspect of the present disclosure, there is provided an electronic device, including a memory, a processor, and a computer program stored in the memory and executable by the processor, wherein the processor executing the program implements the storage management method of the fifth aspect.

5 According to an eighth aspect of the present disclosure, there is provided a computer program product, including computer instructions, wherein, when the computer instructions are executed by a computing device, the computing device performs the storage management method of the fifth aspect.

3. Beneficial Effects

10 The storage management apparatus provided by the embodiment of the present disclosure is capable of parsing an instruction received from an network on chip, generating at least one data packet based on data released from an external storage apparatus, and sending all the data packets to the network on chip; by providing the storage management apparatus between the network on chip and the storage apparatus,
15 data packets are sent to all the processing cores that require the same data by the storage management apparatus, so that each of the processing cores that require the same data does not need to fetch the data from the storage apparatus separately, thereby reducing the delay, saving power consumption of the chip and the storage apparatus, and saving the time taken for the processing cores that require the same data to read the data.

20 **BRIEF DESCRIPTION OF THE DRAWINGS**

FIG. 1 is a schematic flowchart of multiple cores obtaining same data in the prior art.

FIG. 2 is a schematic flowchart of multiple cores obtaining same data in the prior art.

25 FIG. 3 is a schematic structural diagram of a storage management apparatus according to an embodiment of the present disclosure.

FIG. 4 is a schematic structural diagram of a storage management apparatus according to another embodiment of the present disclosure.

30 FIG. 5 is a schematic structural diagram of a chip according to an embodiment of the present disclosure.

FIG. 6 is a schematic flowchart of a storage management method according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

In order to make the objective, technical solutions and advantages of the present disclosure clearer, the present disclosure is further described in detail below in conjunction with specific embodiments and with reference to the accompanying drawings. It should be understood that these descriptions are exemplary only and are not intended to limit the scope of the present disclosure. In addition, in the following description, descriptions of well-known structures and techniques are omitted to avoid unnecessary confusion about the concepts of the present disclosure.

In neural network computing, multi-core or many-core chip architectures are often used. How to harness as much computing power out of the many cores as possible becomes the key to the performance of the chip. This depends on how the cores can efficiently read the data into the cores for computation.

For example, in the field of image recognition, a chip has N cores, and all the cores process images independently and in parallel; and the algorithms and weights used by each core are the same. Then, the method to maximize the computing power of the chip with the least delay and power consumption is that all cores share the same set of weights. That is, the multiple cores fetch the same data.

FIG. 1 is a schematic flowchart of multiple cores obtaining same data in the prior art.

As shown in FIG. 1, a CPU includes N processing cores and a Shared Memory (SM). The N processing cores are the processing core $C1$ to the processing core Cn , and each processing core independently reads the required data from the SM. When each processing core needs the same data, such as the same weights or other same parameters, the process of data fetching by the multiple processing cores is as follows:

The processing core $C1$ sends a fetch instruction to the SM, and fetches the data released by the SM. When the processing core $C1$ completes the fetching, the other processing cores issue fetch instructions to the SM in turn, and fetch the data stored in the same location from the SM again.

The single arrows in FIG. 1 indicate the directions of data transmissions in which multiple cores acquire the same data, and the dashed lines indicate that the cores are capable of performing bidirectional transmission of data with the shared memory module SM.

Usually, this method has the following disadvantages:

When multiple processing cores need to use the same data, each processing core needs to read it independently. On the one hand, the same data is read from the SM multiple times, resulting in high power consumption of the SM; on the other hand, when only one copy of this data is stored in the SM, multiple processing cores are required to read it one by one, and the processing cores in the back of the queue need to wait for the preceding processing cores to complete reading before they can read, causing a delay in reading the data by the processing cores in the back of the queue. Sometimes, computation can only be performed after the processing core reads the data, which causes the processing cores in the back of the queue to wait for a long time, wasting computing power and reducing the performance of the chip.

FIG. 2 is a schematic flowchart of multiple cores obtaining same data in the prior art.

As shown in FIG. 2, a GPU includes N processing cores and an SM. The N processing cores are processing cores C1 to Cn. The SM stores multiple copies of the same data consecutively. A processing core reads the required data from the SM, and the SM releases all copies of the same data that have been consecutively stored for the multiple processing cores to read. The process of data fetching by the multiple processing cores is as follows:

The first processing core C1 sends a fetch instruction to the SM, and the SM releases the same data stored consecutively. The processing cores C1- Cn that need the same data receive the same data released by the SM, respectively. This approach requires the SM to store multiple copies of the same data consecutively, wasting the memory of the SM.

The single arrows in FIG. 2 indicate the directions of data transmissions in which multiple cores acquire the same data, and the dashed lines indicate that the cores are capable of performing bidirectional transmission of data with the shared memory module SM.

In order to solve the above problems, the technical solution of the present disclosure is provided.

The chip provided by an embodiment of the present disclosure will be described in detail below. In the description of the present disclosure, it should be noted that the terms "first", "second", "third" and "fourth" are only used for descriptive purposes and cannot be construed as indicating or implying relative importance. In addition, the

technical features involved in the different embodiments of the present disclosure described below can be combined with each other as long as they do not conflict with each other.

FIG. 3 is a schematic structural diagram of a storage management apparatus
5 according to an embodiment of the present disclosure.

In this embodiment, the storage management apparatus is disposed between a network on chip and a storage apparatus, in which the network on chip NoC connects N processing cores, and N is an integer greater than or equal to 1.

As shown in FIG. 3, the storage management apparatus includes: an instruction
10 parser ID and a data packet generator PG-SW.

The instruction parser ID is configured to parse an instruction received from the network on chip, and generate a control signal according to the instruction.

The data packet generator PG-SW is configured to generate at least one data packet based on data released from the storage apparatus according to the control signal, and
15 send all the data packets to the network on chip.

The storage management apparatus is capable of parsing an instruction received from a network on chip, generating at least one data packet based on data released from an external storage apparatus, and sending all the data packets to the network on chip. By providing the storage management apparatus between the network on chip and the
20 storage apparatus, data packets are sent to all the processing cores that require the same data by the storage management apparatus, so that the processing cores that require the same data do not need to fetch the data from the storage apparatus separately, thereby reducing the delay, saving power consumption of the chip and the storage apparatus, and saving the time taken for the processing cores that require the same data to read the
25 data.

In an embodiment, the data packet generator PG-SW is configured to generate the data packet based on the data released from the storage apparatus according to the control signal, and it includes that, the data packet generator PG-SW is configured to generate, based on the released data and the processing core indicated by the instruction,
30 the data packet corresponding to the processing core according to the control signal.

In a preferred embodiment, the data packet generator is configured to generate, based on the released data and the processing core indicated by the instruction, the data packet corresponding to the processing core according to the control signal, and it

includes that, where a plurality of processing cores are indicated by the instruction, the data packet generator is configured to generate the data packets corresponding to each processing core based on the released data and the processing cores indicated by the instruction, and send the generated plurality of data packets to the network on chip in parallel, such that each data packet is sent to the processing core indicated by the instruction through the network on chip.

It should be noted that "corresponding to each processing core" may refer to corresponding to the address of each processing core. Preferably, "corresponding to each processing core" may also refer to corresponding to the address of each processing core, or corresponding to the storage address of the data in each processing core.

In an embodiment, the instruction includes fetch information and storage information.

In an embodiment, the storage management apparatus further includes an access indicator Addr-G; the access indicator Addr-G is configured to receive the fetch information, and instruct the storage apparatus to release the data indicated by the fetch information according to the fetch information, where the data can be parameters such as weights.

In a preferred embodiment, the instruction includes the number of fetches, and the fetch information includes the starting address of the fetch addresses; the access indicator is configured to generate the fetch addresses according to the starting address of the fetch addresses and the number of fetches, and instruct the storage apparatus to release the data according to the fetch addresses.

In this embodiment, the starting address of the fetch addresses is provided as the fetch information. On the one hand, it may save the number of bits in the instruction, so that the instruction can carry more other information; on the other hand, it is more convenient to fetch data according to the starting address of the fetch addresses and the number of fetches.

In an embodiment, the access indicator Addr-G is further configured to receive the storage information, and generate a storage address of the data packet corresponding to the processing core indicated by the instruction according to the storage information.

In an optional embodiment, the storage management apparatus further includes an instruction cache, configured to receive the instruction sent by the network on chip, cache the instruction, and send the cached instruction to the instruction parser ID.

When a plurality of instructions are sent by the network on chip, the instruction cache IS caches the instructions in the time order in which they are received, and sends the instructions to the instruction parser ID in the order in which they are cached.

Optionally, the instruction cache IS may be an instruction stack, which is used to temporarily store the received data access instructions, and the storage management apparatus MME may execute the instructions in the stack one by one.

In an embodiment, the storage information includes the starting address of the storage addresses of the data in the processing core indicated by the instruction, and the access indicator is configured to generate the storage addresses according to the starting address of the storage addresses and the number of fetches.

In this embodiment, the starting address of the storage addresses of the data in the processing core is provided as the storage information. The number of bits of the instruction may be reduced, compared to the storage information in the instruction being the storage addresses of the data in the processing core.

In an embodiment, the data packet generator is configured to generate the data packet based on the data released from the storage apparatus according to the control signal, and it includes that:

the data packet generator PG-SW is configured to generate an address of the processing core indicated by the instruction according to the control signal, and generate the data packet based on the released data, the storage addresses and the address of the processing core indicated by the instruction.

In an embodiment, the instruction includes control bits; the control bit corresponding to the processing core that waits to receive the data as indicated by each instruction is a preset value; and the control signal corresponding to the processing core is generated when the instruction parser ID determines that the control bit is a preset value.

In an embodiment, the instruction includes information bits for indicating the processing cores for receiving the data, the starting address for data fetching, the number of fetches, the starting address of the storage addresses in each processing core for receiving the data, and control bits for controlling switches.

In a specific embodiment, the format of the instruction is as follows:

Information bits	Starting address of fetch addresses	Starting address of storage addresses	Number of fetches	Reserved	Instruction code	Control bits
ID_C	Addr_S	Addr_	M	Reserved	Fetch instruction	C_C
147:144(4b)	143:128(16b)	127:112(16b)	111:96(16b)	95:4(84b)	11:4(8b)	3:0(4b)

It should be noted that the plurality of processing cores may be numbered first. Since the storage management apparatus needs to receive the instruction through the NoC, and also needs to send the generated data packets to one or more processing cores indicated by the instruction through the NoC, the storage management apparatus needs to be numbered together with the processing cores. For example, there is a total of N processing cores in the chip, and the storage management apparatus may be numbered 1, the first processing core C1 may be numbered 2, and the Nth processing core Cn may be numbered N+1.

It should be understood that the storage management apparatus may also be numbered 0, the first processing core C1 may be numbered 1, and the Nth processing core Cn may be numbered N. In some examples of the present disclosure, the storage management apparatus is numbered 1, but it is not limited to this.

The instruction may be 148 bits.

In the instruction, the information bits may be represented by ID_C, which may be 4 bits, and are used to indicate the address of the processing core for receiving the data. The information bits are set to be N+1 bits, and the first bit in the N+1 bits corresponds to the storage management apparatus, the second to N+1th bits correspond to the N processing cores, respectively. The network on chip NoC uses the information bits to determine the recipient of the instruction or data packet. The recipient here may be a processing core or the storage management apparatus. That is, when the first bit in ID_C is 1, it means that the instruction or data packet is sent to the storage management apparatus; when a certain bit among the second to the N+ 1th bits in ID_C is 1, it means that the data packet is sent to the processing core corresponding to the certain bit.

The instruction code may be 8 bits. The instruction code is set with various information. For example, the fetch instruction code is set to 00000001; and when the instruction parser ID finds that the instruction code is 00000001, it indicates that this is a fetch instruction.

The number of control bits is $N+1$; the $N+1$ bits of the control bits are in a one-to-one correspondence with the $N+1$ bits of the information bits, where the second to $N+1$ th bits correspond to the N processing cores, respectively; and the control bit corresponding to the processing core indicated by each instruction is a preset value. For example, the preset value is 1, i.e., when the instruction parser identifies that the control bit is 1, a control signal is generated. For example, a chip has 4 processing cores, i.e., there are 5 control bits; the second to the fifth bits correspond to the four processing cores, respectively. When C-C is 10110, that is, the first and fourth bits are 0, and the second bit, the third bit, and the fifth bit are 1, it indicates that the first, second and fourth processing cores all need to receive the data.

It should be noted that, in the embodiment of the present disclosure, the first number on the right side of the $N+1$ bits of C-C is set to correspond to the first bit (corresponding to the storage management apparatus), and the second number on the right side corresponds to the second bit (corresponding to the first processing core), the last number on the right side corresponds to the last bit (the N th processing core); of course, it may also be set that the first number on the left corresponds to the first bit (the storage management apparatus), and the second number on the left corresponds to the second bit (the first processing core), and the last number on the left corresponds to the last bit (the N th processing core), which is not limited in the present disclosure.

The fetch information, the storage information and the number of fetches may be 16 bits, respectively.

The reserved means bits that are not coded.

FIG. 4 is a schematic structural diagram of a storage management apparatus according to another embodiment of the present disclosure.

In an embodiment, the data packet generator includes a Tx transmitter and an Rx receiver, in which the Tx transmitter includes switches and a packet header generator, and the number of switches is greater than or equal to the number of processing cores connected to the network on chip NoC. Each processing core corresponds to one of the switches, and the control signals are used to control the states of the switches.

In an embodiment, the data packet generator PG-SW is configured to send the plurality of data packets to the network on chip in parallel, and it includes that, the data packet generator is configured to turn on the switches corresponding to the processing cores indicated by the instruction according to the control signals, and send the data

packets to the network on chip in parallel, so as to send all the data packets to the respective corresponding processing cores through the network on chip.

In a specific embodiment, the switch includes a control terminal, an input terminal and an output terminal; the input terminal of each of the switches is configured to
5 receive the storage address sent by the access indicator and the data released from the storage apparatus; and the control terminal of each of the switches is configured to implement an on state or an off state according to the control signal received.

The instruction includes control bits configured to control the states of the switches; the control bit corresponding to the processing core indicated by each instruction is a
10 preset value; and the control signal is generated when the instruction parser ID determines that the control bit is a preset value.

In an embodiment, the storage management apparatus MME stores the received data in the storage apparatus through the Rx receiver.

In the embodiment shown in FIG. 4, the first processing core C1 sends an
15 instruction to the storage management apparatus MME through the network on chip NoC, where the cache IS caches the instruction and sends the instruction to the instruction parser ID.

The instruction parser ID parses the instruction as 00000001, indicating that data needs to be fetched from the storage apparatus; and the instruction parser ID parses out
20 the starting address for data fetching Addr_S, and it parses that the number of fetches is M and the control information C_C is 0110, indicating that it is required to send the M pieces of data to the first processing core C1 and the second processing core C2; the instruction parser ID parses out the starting address Addr_D1 of the storage addresses for storing the M pieces of data to the first processing core C1, and the starting address
25 Addr_D2 of the storage addresses for storing the M pieces of data to the second processing core C2. The instruction parser ID generates the first control signal C_C1, i.e., 0010, and the second control signal C_C2, i.e., 0100, respectively, according to the control information 0110.

The instruction parser ID sends the number of fetches M, Addr_S and Addr_D to
30 the access indicator Addr_G, and sends the first control signal C_C1 and the second control signal C_C2 to the first switch S1 corresponding to the first processing core and the second switch S2 corresponding to the second processing core in the data packet generator PG_SW, respectively.

Addr_G generates specific fetch addresses according to M and Addr_S, and fetches data from the external storage apparatus SM; and it generates specific storage addresses Addr_D1_N and Addr_D2_N according to M, Addr_D1 and Addr_D2, and sends Addr_D1_N and Addr_D2_N to PG_SW.

5 PG_SW turns on the first switch and the second switch according to the received first control signal C_C1 and second control signal C_C2. The packet header generator ID-GEN generates first information bits ID_C1 and second information bits ID_C2, respectively, according to the first control signal C_C1 and the second control signal C_C2, where the first information bits are 0010, and the second information bits are
10 0100.

It should be noted that, when the number of processing cores is less than 3, the fourth bit of the information bits may be set to 0.

PG_SW generates the first packet header according to the storage addresses Addr_D1_N obtained from the output terminal of the first switch and the first
15 information bits ID_C1, and generates the second packet header according to the storage addresses Addr_D2_N obtained from the output terminal of the second switch and the second information bits ID_C2. PG_SW also generates a first data packet based on the first packet header and the data released from the storage apparatus obtained from the output terminal of the first switch, and generates a second data packet based
20 on the second packet header and the data released from the storage apparatus obtained from the output terminal of the second switch; PG_SW sends the first data packet and the second data packet to the network on chip NoC in parallel.

The NoC sends the first data packet to the first processing core, and sends the second data packet to the second processing core.

25 In an embodiment, the data packet received by the storage management apparatus MME or the data packet sent by the MME may be in the following format:

Information bits	Storage addresses of data in processing core	Data
(ID_C,4bit)	(Addr_D,16bit)	(128bit)

It should be noted that, in the embodiment shown in FIG. 4, there are two processing cores; and when there are three processing cores, the data packet generator is provided with three switches. When the first processing core sends an instruction that
30 both the first and the second processing cores need data, but the third processing core

does not need data; the instruction parser ID may generate a third control signal to control the third switch to be in an off state, or the instruction parser ID may not generate the third control signal.

5 The storage management apparatus provided by the present disclosure can send data packets to all the processing cores that require the same data, so that each of the processing cores that require the same data does not need to fetch the data from the storage apparatus separately, thereby reducing the delay, saving power consumption of the chip and the storage apparatus, and saving the time taken for the processing cores that require the same data to read the data.

10 FIG. 5 is a schematic structural diagram of a chip according to an embodiment of the present disclosure.

As shown in FIG. 5, the chip includes a network on chip NoC, a plurality of the processing cores connected to the network on chip, a storage apparatus, and the storage management apparatus MME provided by the above embodiments.

15 The plurality of processing cores are the N processing cores, that is, the first processing core C1 to the Nth processing core Cn.

The storage management apparatus parses the instruction obtained from the network on chip, generates, according to the instruction and the data released from the storage apparatus, data packets corresponding to the processing cores that wait to receive the data as indicated by the instruction, and sends all the generated data packets to the network on chip; and all the data packets are sent to the processing cores through the network on chip.

25 It should be noted that, in FIG. 5, the single arrows indicate the directions of data transmissions provided by the embodiment of the present disclosure, and the dashed lines indicate the capability of bidirectional interaction; for example, the double arrow in dashed lines between the processing core C1 and the storage management apparatus MME indicates that the processing core and the storage management apparatus MME can transmit and receive data bidirectionally. The storage management apparatus MME and the storage apparatus SM can also transmit and receive data bidirectionally.

30 An embodiment of the present disclosure provides a circuit board, including one or more chips provided by the above embodiments.

An embodiment of the present disclosure provides an electronic device, including one or more circuit boards provided by the above embodiments.

FIG. 6 is a schematic structural diagram of a storage management method according to an embodiment of the present disclosure.

As shown in FIG. 6, the storage management method includes step S101-step S102:

Step S101, an instruction parser parses an instruction received from a network on
5 chip, and generates a control signal according to the instruction.

Step S102, a data packet generator generates at least one data packet based on data released from a storage apparatus according to the control signal, and sends all the data packets to the network on chip.

An embodiment of the present disclosure further provides a computer storage
10 medium storing a computer program, which, upon execution by a processor, implements the storage management method of the above embodiments.

Another embodiment of the present disclosure provides an electronic device, including a memory, a processor, and a computer program stored in the memory and executable by the processor, wherein the processor executing the program implements
15 the storage management method of the above embodiments.

A further embodiment of the present disclosure provides a computer program product, including computer instructions, wherein, when the computer instructions are executed by a computing device, the computing device performs the storage management method of the above embodiments.

20 It should be understood that the above specific embodiments of the present disclosure are used only for exemplary illustration or explanation of the principle of the present disclosure, but not to limit the present disclosure. Therefore, any modifications, equivalent replacements, improvements, etc. made without departing from the spirit and scope of the present disclosure should be included within the protection scope of
25 the present disclosure. Furthermore, the appended claims of this disclosure are intended to cover all variations and modifications that fall within the scope and boundaries of the appended claims, or the equivalents of such scope and boundaries.

CLAIMS

1. A storage management apparatus, disposed between a network on chip and a storage apparatus, comprising:

an instruction parser, configured to parse an instruction received from the network
5 on chip, and generate a control signal according to the instruction; and
a data packet generator, configured to generate at least one data packet based on data released from the storage apparatus according to the control signal, and send all the data packets to the network on chip.

2. The apparatus of claim 1, wherein
10 the data packet generator is further configured to generate, based on the released data and a processing core indicated by the instruction, the data packet corresponding to the processing core according to the control signal.

3. The apparatus according to claim 2, wherein
a plurality of processing cores are indicated by the instruction; and
15 the data packet generator is further configured to generate the data packets corresponding to each processing core based on the released data and the processing cores indicated by the instruction, and send the generated plurality of data packets to the network on chip in parallel.

3. The apparatus according to any of claim 1-3, wherein the instruction includes
20 fetch information and storage information.

4. The apparatus according to claim 3, wherein the apparatus further comprises an access indicator;

the access indicator is configured to receive the fetch information, and instruct the storage apparatus to release the data indicated by the fetch information according to the
25 fetch information; and

the access indicator is further configured to receive the storage information, and generate a storage address of the data packet corresponding to the processing core indicated by the instruction according to the storage information.

5. The apparatus of claim 4, wherein
30 the data packet generator is further configured to generate an address of the processing core indicated by the instruction according to the control signal, and generate the data packet based on the released data, the storage address and the address of the processing core indicated by the instruction.

6. The apparatus of claim 5, wherein the data packet generator comprises switches, and the number of the switches is greater than or equal to the number of processing cores connected to the network on chip;

5 wherein each processing core corresponds to one of the switches, and the control signal is used to control states of the switches.

7. The apparatus of claim 6, wherein

the data packet generator is further configured to turn on the switch corresponding to the processing core indicated by the instruction according to the control signal, and send the plurality of the data packets to the network on chip in parallel, so as to send
10 the plurality of data packets to the respective corresponding processing cores through the network on chip.

8. The apparatus according to claim 6 or 7, wherein the switch comprises a control terminal, an input terminal and an output terminal;

the input terminal of each of the switches is configured to receive the storage
15 address sent by the access indicator and the data released from the storage apparatus; and

the control terminal of each of the switches is configured to implement an on state or an off state according to the control signal received.

9. The apparatus according to any of claims 1-8, wherein

20 the instruction includes control bits; the control bit corresponding to the processing core indicated by each instruction is a preset value; and

the control signal is generated when the instruction parser determines that the control bit is a preset value.

10. A chip, comprising a network on chip, a plurality of processing cores connected
25 to the network on chip, a storage apparatus, and the storage management apparatus according to any one of claims 1-9.