

CLAIMS

1. A data access method, wherein the method is applied to a distributed storage system (1100), the distributed storage system comprises m storage nodes (1110, 1120, 1200, 1300), the m storage nodes comprise a first storage node (1110), each storage node comprises at least one memory (1320, 1330), each memory comprises a non-volatile storage medium,
5 and m is an integer greater than 1; and
the method comprises:
receiving, by the first storage node (1110), a first data read request from a terminal, wherein the first data read request carries first location information, and the first location information is used to describe a location of first data in a data body to which the first data
10 belongs; and
if the first storage node (1110) determines, based on the first location information, that a memory (1320, 1330) in which the first data is located is a first memory, and the first memory belongs to the first storage node, when a state information set currently records that the first memory is in a trusted access state, reading, by the first storage node, the first
15 data directly from the first memory,
and when a state information set currently records that the first memory (1320, 1330) is in an untrusted access state, collecting other strips of a stripe to which the first data belongs, and obtaining the first data by using the other strips, which comprise one or more parity strips, and performing a correctness check,
20 and sending a first data read response used for responding to the first data read request to the terminal, wherein the first data read response carries the read or obtained first data.
2. The method according to claim 1, wherein the method further comprises:
if the first memory (1320, 1330) belongs to a second storage node (1120) in the m storage
25 nodes (1110, 1120, 1200, 1300), the first memory is in a trusted access state, and the first storage node (1110) and a management node that is configured to release a state information set are in a connected state, forwarding, by the first storage node, the first data read request to the second storage node; and
after receiving the first data read request from the first storage node (1110), reading, by
30 the second storage node (1120), the first data from the first memory (1320, 1330), and sending the first data read response used for responding to the first data read request to the first storage node or the terminal, wherein the first data read response carries the read first data.

3. The method according to claim 1, wherein the method further comprises:
if the first memory (1320, 1330) belongs to a second storage node (1120) in the m storage nodes (1110, 1120, 1200, 1300), and the first memory is in a trusted access state, or if the first memory belongs to a second storage node in the m storage nodes, the first memory is
5 in a trusted access state, and the first storage node (1110) and a management node that is configured to release a state information set are in a connected state, adding, by the first storage node into the first data read request, an identifier of the state information set currently cached by the first storage node, and sending, to the second storage node, the first data read request into which the identifier of the state information set is added; and
10 after the second storage node (1120) receives the first data read request from the first storage node (1110), if it is determined by comparison that the identifier of the state information set carried in the first data read request is the same as an identifier of a state information set currently cached by the second storage node, reading the first data from the first memory (1320, 1330), and sending the first data read response used for
15 responding to the first data read request to the first storage node or the terminal, wherein the first data read response carries the read first data.
4. The method according to claim 1, wherein the method further comprises:
if the first memory (1320, 1330) is in an untrusted access state, x memories in N memories
20 in which strips of a stripe to which the first data belongs are located belong to the first storage node (1110), and N-x memories in the N memories belong to y storage nodes (1110, 1120, 1200, 1300) different from the first storage node, sending, by the first storage node to the y storage nodes, a data read request that carries a stripe identifier of the stripe and an identifier of the state information set;
25 after receiving the data read request from the first storage node (1110), comparing, by each of the y storage nodes (1110, 1120, 1200, 1300), the identifier of the state information set carried in the data read request with an identifier of a state information set currently cached by the storage node, and if it is determined by comparison that the identifier of the state information set carried in the data read request is the same as the identifier of the
30 state information set currently cached by the storage node, reading a corresponding strip of the stripe from a corresponding memory (1320, 1330) comprised in the storage node, and sending the read corresponding strip of the stripe to the first storage node; and
performing, by the first storage node (1110) based on strips of the stripe that are collected from the y storage nodes (1110, 1120, 1200, 1300) and the first storage node, a check
35 operation to obtain the first data, and sending the first data read response used for responding to the first data read request to the terminal, wherein the first data read response carries the obtained first data.

5. The method according to any one of claims 1 to 4, wherein one of the m storage nodes (1110, 1120, 1200, 1300) is the management node; and

the method further comprises:

when a state of the first memory (1320, 1330) changes from a first state to a second state, sending, by the first storage node (1110), a memory state change report to the management node, wherein the memory state change report indicates that the first memory is in the second state, the first state is different from the second state, and the first state and the second state each comprise any one of the following states:

an offline state, a data reconstruction state, or a trusted access state;

after receiving the memory (1320, 1330) state change report from the first storage node (1110), updating, by the management node, a state that is of the first memory and recorded in a state information set cached by the management node into the second state, and updating a version number of the state information set cached by the management node;

and

sending, by the management node, an updated state information set to a storage node other than the management node in the m storage nodes (1110, 1120, 1200, 1300); and

using, by the first storage node (1110), the state information set from the management node to update the state information set currently cached by the first storage node.

6. The method according to any one of claims 1 to 5, wherein the method further comprises: receiving, by the first storage node (1110), a first data write request from the terminal, wherein the first data write request carries second data and second location information, and the second location information is used to describe a location of the second data in a data body to which the second data belongs;

determining, by the first storage node (1110) based on the second location information, W memories related to writing the second data, segmenting the second data into W-T data strips, and obtaining T parity strips by means of calculation by using the W-T data strips, wherein the T parity strips and the W-T data strips form a stripe comprising W strips, the W memories are in a one-to-one correspondence with the W strips, T and W are positive integers, and T is less than W;

when the state information set cached by the first storage node (1110) currently records that W1 memories in the W memories are in a non-offline state, sending, to y2 storage nodes (1110, 1120, 1200, 1300) to which the W1 memories belong, a data write request that carries a strip of the stripe and the identifier of the state information set; and

after receiving the data write request from the first storage node (1110), comparing, by each of the y2 storage nodes (1110, 1120, 1200, 1300), the identifier of the state information

set carried in the data write request with an identifier of a state information set currently cached by the storage node, and if it is determined by comparison that the identifier of the state information set carried in the data write request is the same as the identifier of the state information set currently cached by the storage node, writing a corresponding strip of the stripe into a corresponding memory (1320, 1330) comprised in the storage node, or if it is determined by comparison that the identifier of the state information set carried in the data write request is different from the identifier of the state information set currently cached by the storage node, refusing to write a corresponding strip of the stripe into a corresponding memory comprised in the storage node.

7. The method according to any one of claims 1 to 5, wherein the method further comprises: receiving, by the first storage node (1110), a second data write request from the terminal, wherein the second data write request carries third data and third location information, and the third location information is used to describe a location of the third data in a data body to which the third data belongs;

after determining, based on the third location information, W memories related to writing the third data, segmenting, by the first storage node (1110), the third data into W-T data strips, and obtaining T parity strips by means of calculation by using the W-T data strips, wherein the T parity strips and the W-T data strips form a stripe comprising W strips, the W memories are in a one-to-one correspondence with the W strips, T and W are positive integers, and T is less than W;

if the state information set cached by the first storage node (1110) currently records that W1 memories in the W memories are in a non-offline state, and the first storage node and the management node that is configured to release a state information set are in a connected state, sending, to y2 storage nodes (1110, 1120, 1200, 1300) to which the W1 memories belong, a data write request that carries a strip of the stripe; and

after receiving the data write request from the first storage node (1110), writing, by each of the y2 storage nodes (1110, 1120, 1200, 1300), a corresponding strip of the stripe into a corresponding memory (1320, 1330) comprised in the storage node.

8. The method according to claim 6 or 7, wherein W1 is less than W, and the state information set cached by the first storage node (1110) currently records that W2 memories in the W memories are in an offline state; and

the method further comprises:

generating, by the first storage node (1110), a first reconstruction log, wherein the first reconstruction log records a memory (1320, 1330) identifier of a second memory in the W2 memories, the first reconstruction log further records a strip identifier of a first strip

that corresponds to the second memory in the W strips, the first reconstruction log further records the stripe identifier of the stripe, and the second memory is any memory in the W2 memories; and

after the second memory (1320, 1330) in the W2 memories goes online again, and the second storage node (1120) to which the second memory belongs collects the first reconstruction log generated during an offline period of the second memory, obtaining, by the second storage node, the strip identifier, recorded in the first reconstruction log, of the first strip that needs to be written into the second memory, determining W-T memories in which W-T strips other than the first strip that are comprised in the stripe are located, reading the W-T strips from the W-T memories, performing a check operation by using the W-T strips to reconstruct the first strip that needs to be written into the second memory, and writing the reconstructed first strip into the second memory.

9. A distributed storage system (1100), comprising m storage nodes (1110, 1120, 1200, 1300), wherein each storage node comprises at least one memory (1320, 1330), each memory comprises a non-volatile storage medium, and m is an integer greater than 1; and a first storage node (1110) in the m storage nodes (1110, 1120, 1200, 1300) is configured to: receive a first data read request from a terminal, wherein the first data read request carries first location information, and the first location information is used to describe a location of first data in a data body to which the first data belongs; and determine, based on the first location information, a first memory (1320, 1330) in which the first data is located; and if the first memory (1320, 1330) belongs to the first storage node (1110), when a state information set currently records that the first memory is in a trusted access state, directly read the first data from the first memory, and when a state information set currently records that the first memory (1320, 1330) is in an untrusted access state, collect other strips of a stripe to which the first data belongs, and obtain the first data by using the other strips, which comprise one or more parity strips, and performing a correctness check, and send a first data read response used for responding to the first data read request to the terminal, wherein the first data read response carries the read or obtained first data.
10. The distributed storage system (1100) according to claim 9, wherein the distributed storage system further comprises a management node that is configured to release a state information set; the first storage node (1110) is further configured to: if the first memory (1320, 1330) belongs to a second storage node (1120) in the m storage

nodes (1110, 1120, 1200, 1300), the first memory is in a trusted access state, and the first storage node (1110) and the management node that is configured to release a state information set are in a connected state, forward the first data read request to the second storage node; and

the second storage node (1120) is configured to:

after receiving the first data read request from the first storage node (1110), read the first data from the first memory (1320, 1330), and send the first data read response used for responding to the first data read request to the first storage node or the terminal, wherein the first data read response carries the read first data.

11. The distributed storage system (1100) according to claim 9, wherein

the first storage node (1110) is further configured to:

if the first memory (1320, 1330) belongs to a second storage node (1120) in the m storage nodes (1110, 1120, 1200, 1300), and the first memory is in a trusted access state, add, into the first data read request, an identifier of the state information set currently cached by the first storage node (1110), and send, to the second storage node, the first data read request into which the identifier of the state information set is added; and

the second storage node (1120) is configured to:

after receiving the first data read request from the first storage node (1110), compare the identifier of the state information set carried in the first data read request with an identifier of a state information set currently cached by the second storage node (1120), and if it is determined by comparison that the identifier of the state information set carried in the first data read request is the same as the identifier of the state information set currently cached by the second storage node, read the first data from the first memory (1320, 1330), and send the first data read response used for responding to the first data read request to the first storage node or the terminal, wherein the first data read response carries the read first data.

12. The distributed storage system (1100) according to claim 9, wherein

the first storage node (1110) is further configured to:

if it is determined that the first memory (1320, 1330) is in an untrusted access state, x memories in N memories in which strips of a stripe to which the first data belongs are located belong to the first storage node (1110), and N-x memories in the N memories belong to y storage nodes (1110, 1120, 1200, 1300) different from the first storage node, send, to the y storage nodes, a data read request that carries a stripe identifier of the stripe and an identifier of the state information set;

each of the y storage nodes (1110, 1120, 1200, 1300) is configured to:

after receiving the data read request from the first storage node (1110), compare the identifier of the state information set carried in the data read request with an identifier of a state information set currently cached by the storage node, and if it is determined by comparison that the identifier of the state information set carried in the data read request is the same as the identifier of the state information set currently cached by the storage node, read a corresponding strip of the stripe from a corresponding memory (1320, 1330) comprised in the storage node, and send the read corresponding strip of the stripe to the first storage node; and

the first storage node (1110) is further configured to:

perform, based on strips of the stripe that are collected from the y storage nodes (1110, 1120, 1200, 1300) and the first storage node (1110), a check operation to obtain the first data, and send the first data read response used for responding to the first data read request to the terminal, wherein the first data read response carries the obtained first data.

13. The distributed storage system (1100) according to any one of claims 9 to 12, wherein one of the m storage nodes (1110, 1120, 1200, 1300) is the management node;

the first storage node (1110) is further configured to:

when a state of the first memory (1320, 1330) changes from a first state to a second state, send a memory state change report to the management node, wherein the memory state change report indicates that the first memory is in the second state, the first state is different from the second state, and the first state and the second state each comprise any one of the following states:

an offline state, a data reconstruction state, or a trusted access state;

the management node is configured to:

after receiving the memory (1320, 1330) state change report from the first storage node (1110), update a state that is of the first memory and recorded in a state information set cached by the management node into the second state, and update a version number of the state information set cached by the management node; and

the management node sends an updated state information set to a storage node other than the management node in the m storage nodes (1110, 1120, 1200, 1300); and

the first storage node (1110) is further configured to use the state information set from the management node to update the state information set currently cached by the first storage node.

14. The distributed storage system (1100) according to any one of claims 9 to 12, wherein the first storage node (1110) is further configured to receive a first data write request from the terminal, wherein the first data write request carries second data and second location

information, and the second location information is used to describe a location of the second data in a data body to which the second data belongs;

the first storage node (1110) is further configured to:

determine, based on the second location information, W memories related to writing the second data, segment the second data into W-T data strips, and obtain T parity strips by means of calculation by using the W-T data strips, wherein the T parity strips and the W-T data strips form a stripe comprising W strips, the W memories are in a one-to-one correspondence with the W strips, T and W are positive integers, and T is less than W;

the first storage node (1110) is further configured to:

when the state information set cached by the first storage node (1110) currently records that W1 memories in the W memories are in a non-offline state, send, to y2 storage nodes (1110, 1120, 1200, 1300) to which the W1 memories belong, a data write request that carries a strip of the stripe and the identifier of the state information set; and

each of the y2 storage nodes (1110, 1120, 1200, 1300) is configured to:

after receiving the data write request from the first storage node (1110), compare the identifier of the state information set carried in the data write request with an identifier of a state information set currently cached by the storage node, and if it is determined by comparison that the identifier of the state information set carried in the data write request is the same as the identifier of the state information set currently cached by the storage node, write a corresponding strip of the stripe into a corresponding memory (1320, 1330) comprised in the storage node, or if it is determined by comparison that the identifier of the state information set carried in the data write request is different from the identifier of the state information set currently cached by the storage node, refuse to write a corresponding strip of the stripe into a corresponding memory comprised in the storage node.

15. The distributed storage system (1100) according to any one of claims 9 to 12, wherein the first storage node (1110) is further configured to receive a second data write request from the terminal, wherein the second data write request carries third data and third location information, and the third location information is used to describe a location of the third data in a data body to which the third data belongs;

the first storage node (1110) is further configured to:

after determining, based on the third location information, W memories related to the third data, segment the third data into W-T data strips, and obtain T parity strips by means of calculation by using the W-T data strips, wherein the T parity strips and the W-T data strips form a stripe comprising W strips, the W memories are in a one-to-one correspondence with the W strips, T and W are positive integers, and T is less than W;

the first storage node (1110) is further configured to:

determine states that are of the W memories and currently recorded in the state information set cached by the first storage node (1110), and if it is determined that W1 memories in the W memories are in a non-offline state, and the first storage node and the management node that is configured to release a state information set are in a connected state, send, to y2 storage nodes (1110, 1120, 1200, 1300) to which the W1 memories belong, a data write request that carries a strip of the stripe; and

each of the y2 storage nodes (1110, 1120, 1200, 1300) is configured to:

after receiving the data write request from the first storage node (1110), write a corresponding strip of the stripe into a corresponding memory (1320, 1330) comprised in the storage node.

16. The distributed storage system (1100) according to claim 14 or 15, wherein W1 is less than W, and the state information set cached by the first storage node (1110) currently records that W2 memories in the W memories are in an offline state;

the first storage node (1110) is further configured to generate a first reconstruction log, wherein the first reconstruction log records a memory (1320, 1330) identifier of a second memory in the W2 memories, the first reconstruction log further records a strip identifier of a first strip that corresponds to the second memory in the W strips, the first reconstruction log further records the stripe identifier of the stripe, and the second memory is any memory in the W2 memories; and

after the second memory (1320, 1330) in the W2 memories goes online again, and the second storage node (1120) to which the second memory belongs collects the first reconstruction log generated during an offline period of the second memory, the second storage node obtains the identifier, recorded in the first reconstruction log, of the first strip that needs to be written into the second memory; and

determines W-T memories in which W-T strips other than the first strip that are comprised in the stripe are located, reads the W-T strips from the W-T memories, performs a check operation by using the W-T strips to reconstruct the first strip that needs to be written into the second memory (1320, 1330), and writes the reconstructed first strip into the second memory.

17. A storage node (1110, 1120, 1200, 1300) comprised in a distributed storage system comprising m storage nodes, wherein the storage node comprises at least one memory (1320, 1330), each memory comprises a non-volatile storage medium, and m is an integer greater than 1; and the storage node (1110, 1120, 1200, 1300) comprises:
a communications unit (1210), configured to receive a first data read request from a

terminal, wherein the first data read request carries first location information, and the first location information is used to describe a location of first data in a data body to which the first data belongs; and

a processing control unit (1220), configured to:

5 if it is determined, based on the first location information, that a memory (1320, 1330) in which the first data is located is a first memory, and the first memory belongs to the storage node (1110, 1120, 1200, 1300), when a state information set currently records that the first memory is in a trusted access state, read the first data directly from the first memory, and
10 when a state information set currently records that the first memory is in an untrusted access state, collect other strips of a stripe to which the first data belongs, and obtain the first data by using the other strips, which comprise one or more parity strips, and performing a correctness check, wherein the communications unit (1210) is further configured to send a first data read response used for responding to the first data read request to the terminal, wherein the first data read response carries the read or obtained
15 first data.

18. The storage node (1110, 1120, 1200, 1300) according to claim 17, wherein the communications unit (1210) is further configured to:

20 if the first memory (1320, 1330) belongs to a second storage node (1120) in the m storage nodes (1110, 1120, 1200, 1300), the first memory is in a trusted access state, and the storage node and a management node that is configured to release a state information set are in a connected state, forward the first data read request to the second storage node; and

the first data read request is used to trigger the second storage node (1120) to:

25 after receiving the first data read request from the storage node (1110, 1120, 1200, 1300), read the first data from the first memory (1320, 1330), and send the first data read response used for responding to the first data read request to the storage node or the terminal, wherein the first data read response carries the read first data.

30 19. The storage node (1110, 1120, 1200, 1300) according to claim 17, wherein the communications unit (1210) is further configured to:

35 if the first memory (1320, 1330) belongs to a second storage node (1120) in the m storage nodes (1110, 1120, 1200, 1300), and the first memory is in a trusted access state, or if the first memory belongs to a second storage node in the m storage nodes, the first memory is in a trusted access state, and the first storage node (1110) and a management node that is configured to release a state information set are in a connected state, add, into the first data read request, an identifier of the state information set currently cached by the first

storage node, and send, to the second storage node, the first data read request into which the identifier of the state information set is added; and

the first data read request is used to trigger the second storage node (1120) to:

after receiving the first data read request from the storage node, compare the identifier of the state information set carried in the first data read request with an identifier of a state information set currently cached by the second storage node (1120), and if it is determined by comparison that the identifier of the state information set carried in the first data read request is the same as the identifier of the state information set currently cached by the second storage node, read the first data from the first memory (1320, 1330), and send the first data read response used for responding to the first data read request to the storage node or the terminal, wherein the first data read response carries the read first data.

20. The storage node

according to any one of claims 17 to 19, wherein one of the m storage nodes

(1110, 1120, 1200, 1300) is the management node;

the communications unit (1210) is further configured to:

when a state of the first memory (1320, 1330) changes from a first state to a second state, send a memory state change report to the management node, wherein the memory state change report indicates that the first memory is in the second state, the first state is different from the second state, and the first state and the second state each comprise any one of the following states:

an offline state, a data reconstruction state, or a trusted access state; and

the memory (1320, 1330) state change report is used to trigger the management node to:

after receiving the memory (1320, 1330) state change report from the first storage node (1110), update a state that is of the first memory and recorded in a state information set cached by the management node into the second state, update a version number of the state information set cached by the management node, and send an updated state information set to a storage node other than the management node in the m storage nodes (1110, 1120, 1200, 1300); and

the processing control unit (1220) is configured to use the state information set from the management node to update the state information set currently cached by the storage node (1110, 1120, 1200, 1300).

21. A data access method, applied to a first storage node (1110), wherein the first storage node is located in a distributed storage system (1100) comprising m storage nodes (1110, 1120, 1200, 1300), each storage node comprises at least one memory (1320, 1330), each memory comprises a non-volatile storage medium, and m is an integer greater than 1; and the method comprises:

receiving, by the first storage node (1110), a first data read request, wherein the first data read request carries first location information, and the first location information is used to describe a location of first data in a data body to which the first data belongs; and
if the first storage node (1110) determines, based on the first location information, that a
5 memory (1320, 1330) in which the first data is located is a first memory, and the first memory belongs to the first storage node, when a state information set currently records that the first memory is in a trusted access state, directly reading, by the first storage node, the first data from the first memory, and when a state information set currently records that the first memory (1320, 1330) is in an untrusted access state, collecting other strips
10 of a stripe to which the first data belongs, and obtaining the first data by using the other strips, which comprise one or more parity strips, and performing a correctness check, and sending a first data read response used for responding to the first data read request, wherein the first data read response carries the read or obtained first data.