CLAIMS

What is claimed is:

1. A distributed storage and replication system, wherein the system comprises at least one metadata control (MDC) module, multiple IO routing modules, and multiple object-based storage device (OSD) nodes, wherein the MDC module is adapted to: configure, for each OSD node, at least one logical partition (partition) corresponding to physical storage resources managed by each OSD node, wherein the at least one partition is a primary partition, a secondary partition, or any combination of primary and secondary partitions, a primary partition and a secondary partition corresponding to the primary partition constitute a partition group, a primary partition and a secondary partition in a same partition group are located on different OSD nodes, an OSD node on which a primary partition is located is a primary OSD node of a partition group that comprises the primary partition, and an OSD node on which a secondary partition is located is a secondary OSD node of a partition group that comprises the secondary partition; and generate a partition view according to the partition, wherein the partition view comprises information about an OSD on which a partition in a partition group is located; the IO routing module is adapted to route a received IO request to an OSD node; and the OSD node is adapted to execute, according to the IO request, storage of data corresponding to the IO request, wherein

the MDC determines that an OSD node in the system is a faulty OSD node, determines a partition on the faulty OSD node, updates a partition view of a partition group that comprises the partition on the faulty OSD node, and sends an updating notification to a primary OSD node in the updated partition view; and

the primary OSD node is adapted to process, according to the updated partition view after receiving the updating notification sent by the MDC module, replication of the data corresponding to the IO request;

wherein the partition view comprises a primary/secondary identity and a corresponding partition status that are of an OSD on which a partition in a partition group is located; and the primary OSD node is further adapted to update, according to the updated partition view, a partition view locally stored on the primary OSD node; and

the processing, according to the updated partition view, replication of the data

corresponding to the IO request comprises:

replicating, according to the updated locally-stored partition view, the data corresponding to the IO request from the IO routing module onto a secondary OSD node whose partition status is consistent in the updated locally-stored partition view, or onto a secondary OSD node whose partition status is consistent in the updated locally-stored partition view and a secondary OSD node whose partition status is inconsistent in the updated locally-stored partition view but that is recovering data.

2. The system according to claim 1, wherein the MDC module is further adapted to: generate an IO view, wherein the IO view comprises an identifier of a primary OSD node of a partition group; and send the IO view to the IO routing module and the OSD node on which the partition in the partition view is located; and

the primary OSD node is further adapted to: update, according to the updated partition view, an IO view locally stored on the primary OSD node, and process, according to the updated locally-stored IO view, replication of the data corresponding to the IO request.

3. The system according to claim 2, wherein

the MDC module is further adapted to: when determining that the partition on the faulty OSD node comprises a primary partition, update an IO view of a partition group that comprises the primary partition, and notify a secondary OSD node in the updated partition view of the updated IO view; and

the secondary OSD node in the updated partition view is adapted to: update a locally stored IO view according to the updated IO view, and process, according to the updated locally-stored IO view, replication of the data corresponding to the IO request.

4. The system according to claim 3, wherein

the MDC module is further adapted to notify the IO routing module of the updated IO view; and

the IO routing module updates, according to the updated IO view, an IO view locally stored on the IO routing module, and processes forwarding of the IO request according to the updated locally-stored IO view.

5. The system according to claim 3, wherein the updating a partition view of a partition group that comprises the partition on the faulty OSD node specifically comprises:

when the partition on the faulty OSD node comprises a secondary partition,

marking a partition status of the faulty OSD node in a partition view of a partition group that comprises the secondary partition as inconsistent; and

when the partition on the faulty OSD node comprises the primary partition, setting the faulty OSD node that serves as a primary OSD node in a partition view of the partition group that comprises the primary partition as a new secondary OSD node, marking a partition status corresponding to the new secondary OSD node as inconsistent, selecting a secondary OSD node whose partition status is consistent from an original secondary OSD node in the partition view of the partition group that comprises the primary partition, and setting the selected secondary OSD node as a new primary OSD node.

6. The system according to claim 5, wherein

after the faulty OSD node completes failback and data recovery, the MDC module is adapted to: further update the updated partition view and the updated IO view, send an updating notification to a primary OSD node in the further updated partition view, and send an updating notification to a secondary OSD node in the further updated partition view;

the primary OSD node in the further updated partition view is adapted to process, according to the further updated partition view, replication of the data corresponding to the IO request; and

the secondary OSD node in the further updated partition view is adapted to process, according to the further updated IO view, replication of the data corresponding to the IO request.

7. The system according to any one of claims 1 to 6, wherein the system comprises multiple hosts, the MDC module, the IO routing module, and the OSD node are separately deployed on at least one of the multiple hosts, and the OSD node is adapted to manage physical storage resources on the host.

8. A distributed storage and replication system, wherein the system comprises at least one metadata control (MDC) module, multiple IO routing modules, and multiple object-based storage device (OSD) nodes, wherein the MDC module is adapted to: configure, for each OSD node, at least one logical partition (partition) corresponding to physical storage resources managed by each OSD node, wherein the at least one partition is a primary partition, a secondary partition, or any combination of primary and secondary partitions, a primary partition and a secondary partition corresponding to the primary partition constitute a partition group, a primary partition and a

secondary partition in a same partition group are located on different OSD nodes, an OSD node on which a primary partition is located is a primary OSD node of a partition group that comprises the primary partition, and an OSD node on which a secondary partition is located is a secondary OSD node of a partition group that comprises the secondary partition; and generate a partition view according to the partition, wherein the partition view comprises information about an OSD on which a partition in a partition group is located and a primary/secondary identity and a corresponding partition status that are of an OSD on which a partition in a partition group is located; the IO routing module is adapted to route a received IO request to an OSD node; and the OSD node is adapted to execute, according to the IO request, storage of data corresponding to the IO request, wherein

the system comprises a memory and a processor, wherein

the memory is adapted to store a computer readable instruction, and the instruction is used to execute functions of the MDC module, the IO routing module, and the OSD node; and

the processor is adapted to: be connected to the memory, read the instruction in the memory, and enable, according to the instruction, the processor to execute the following operations:

determining that an OSD node in the system is a faulty OSD node, determining a partition on the faulty OSD node, updating a partition view of a partition group that comprises the partition on the faulty OSD node, and sending an updating notification to a primary OSD node on which the partition group in the updated partition view is located, so that the primary OSD node updates, according to the updated partition view, a partition view locally stored on the primary OSD node; replicates, according to the updated locally-stored partition view, the data corresponding to the IO request from the IO routing module onto a secondary OSD node whose partition status is consistent in the updated locally-stored partition view, or onto a secondary OSD node whose partition status is consistent in the updated locally-stored partition view and a secondary OSD node whose partition status is inconsistent in the updated locally-stored partition view but that is recovering data.

9. A method for managing data storage and replication in a distributed storage system, wherein the system comprises at least one metadata control (MDC) module, multiple IO routing modules, and multiple object-based storage device (OSD) nodes, wherein the MDC module is adapted to: configure, for each OSD node, at least one

logical partition (partition) corresponding to physical storage resources managed by each OSD node, wherein the at least one partition is a primary partition, a secondary partition, or any combination of primary and secondary partitions, a primary partition and a secondary partition corresponding to the primary partition constitute a partition group, a primary partition and a secondary partition in a same partition group are located on different OSD nodes, an OSD node on which a primary partition is located is a primary OSD node of a partition group that comprises the primary partition, and an OSD node on which a secondary partition is located is a secondary OSD node of a partition group that comprises the secondary partition; and generate a partition view according to the partition, wherein the partition view comprises information about an OSD on which a partition in a partition group is located; the IO routing module is adapted to route a received IO request to an OSD node; and the OSD node is adapted to execute, according to the IO request, storage of data corresponding to the IO request, wherein the method comprises:

determining that an OSD node in the system is a faulty OSD node, determining a partition on the faulty OSD node, updating a partition view of a partition group that comprises the partition on the faulty OSD node, and sending an updating notification to a primary OSD node in the updated partition view; and

the primary OSD node is adapted to process, according to the updated partition view after receiving the updating notification sent by the MDC module, replication of the data corresponding to the IO request; wherein the partition view comprises a primary/secondary identity and a corresponding partition status that are of an OSD on which a partition in a partition group is located; and the primary OSD node is further adapted to update, according to the updated partition view, a partition view locally stored on the primary OSD node; and

the processing, according to the updated partition view, replication of the data corresponding to the IO request comprises:

replicating, according to the updated locally-stored partition view, the data corresponding to the IO request from the IO routing module onto a secondary OSD node whose partition status is consistent in the updated locally-stored partition view, or onto a secondary OSD node whose partition status is consistent in the updated locally-stored partition view and a secondary OSD node whose partition status is inconsistent in the updated locally-stored partition view but that is recovering data.

10. The method according to claim 9, wherein the method further comprises:

when determining that the partition on the faulty OSD node comprises a primary partition, updating, by the MDC module, an IO view of a partition group that comprises the primary partition, and notifying a secondary OSD node in the updated partition view of the updated IO view; and

updating, by the secondary OSD node in the updated partition view, a locally stored IO view according to the updated IO view, and processing, according to the updated locally-stored IO view, replication of the data corresponding to the IO request.

11. The method according to claim 10, wherein the updating a partition view of a partition group that comprises the partition on the faulty OSD node specifically comprises:

when the partition on the faulty OSD node comprises a secondary partition, marking a partition status of the faulty OSD node in a partition view of a partition group that comprises the secondary partition as inconsistent; and

when the partition on the faulty OSD node comprises the primary partition, setting the faulty OSD node that serves as a primary OSD node in a partition view of the partition group that comprises the primary partition as a new secondary OSD node, marking a partition status corresponding to the new secondary OSD node as inconsistent, selecting a secondary OSD node whose partition status is consistent from an original secondary OSD node in the partition view of the partition group that comprises the primary partition, and setting the selected secondary OSD node as a new primary OSD node.

12. The method according to claim 11, wherein the method further comprises:

after the faulty OSD node completes failback and data recovery, further updating, by the MDC module, the updated partition view and the updated IO view, sending an updating notification to a primary OSD node in the further updated partition view, and sending an updating notification to a secondary OSD node in the further updated partition view;

processing, by the primary OSD node in the further updated partition view according to the further updated partition view, replication of the data corresponding to the IO request; and

processing, by the secondary OSD node in the further updated partition view according to the further updated IO view, replication of the data corresponding to the IO request.