

CLAIMS

1. A system, comprising:
one or more computing devices configured to:
 - select a particular component of a multi-tenant file storage service as a coordinator of a distributed transaction;
 - generate, at the coordinator, a transaction preparation message comprising at least: (a) an indication of a respective proposed page modification to be performed at each page of a set of physical data pages managed by the service, and (b) an order in which respective page-level commit decisions associated with the proposed page modifications are to be made;
 - propagate the transaction preparation message sequentially among nodes of a storage node chain collectively storing the set of physical data pages, wherein the storage node chain ends at a decider node designated by the coordinator as responsible for a decision to commit the distributed transaction;
 - in response to receiving the transaction preparation message at a particular storage node of the chain other than the decider node,
 - store, at a persistent repository, subsequent to a commit analysis of a proposed modification to a physical data page of the particular storage node, a record of an intent to perform the proposed modification, wherein the commit analysis includes an examination of a logical timestamp corresponding to a read of the physical data page;
 - lock the physical data page of the particular storage node; and
 - transmit the transaction preparation message to a next storage node of the storage node chain;
 - in response to receiving the transaction preparation message at the decider node,
 - determine, subsequent to a commit analysis of a proposed modification to a physical data page of the decider node, that the distributed transaction is to be committed;

initiate the proposed modification to the physical data page of decider node; and
initiate a propagation of a transaction commit message to other nodes of the chain; and
in response to receiving the transaction commit message at a particular storage node of the chain,
initiate a proposed modification corresponding to the record of the intent to commit; and
release the lock on the physical data page.

2. The system as recited in claim 1, wherein the logical timestamp corresponds to a read request issued by the coordinator.

3. The system as recited in claim 1, wherein the coordinator is implemented at a selected storage node of the service, wherein the selected storage node is a member of the chain.

4. The system as recited in claim 1, wherein the one or more computing devices are configured to propagate the transaction commit message to other nodes of the chain in accordance with one of: (a) a sequential propagation policy in which the transaction commit message is transmitted to the other nodes in reverse order, relative to the order in which the transaction preparation message was propagated among the nodes of the storage node chain, or (b) a parallel propagation policy in which the decider node transmits the transaction commit message to a plurality of the other nodes.

5. The system as recited in claim 1, wherein the one or more computing devices are further configured to:

store, at the particular storage node, prior to transmitting the transaction preparation message to the next storage node, a transaction state record corresponding to the distributed transaction;

modify, at the particular storage node subsequent to initiating the proposed modification corresponding to the record of the intent to commit, the transaction state record to indicate that the distributed transaction has been committed; and retain the transaction state record at the particular storage node for a configurable time period.

6. A method, comprising:

performing, by one or more computing devices:

generating, at a particular component of a multi-tenant storage service, wherein the particular component is designated as a coordinator of a distributed transaction, a transaction preparation message comprising at least: (a) an indication of a respective proposed page modification to be performed at each page of a set of physical data pages managed by the service, and (b) an order in which respective commit decisions associated with the proposed page modifications are to be made;

propagating the transaction preparation message sequentially among nodes of a storage node chain collectively storing the set of physical data pages;

in response to receiving the transaction preparation message at a particular storage node of the chain other than a terminal node of the chain,

storing, at a persistent repository, subsequent to a commit analysis of a proposed modification to a first physical data page of the particular storage node, a record of an intent to commit the proposed modification; and

transmitting the transaction preparation message to a next storage node of the storage node chain;

in response to receiving the transaction preparation message at the terminal node, determining, subsequent to a commit analysis of a proposed modification to a second physical data page of the terminal node, that the distributed transaction is to be committed;

initiating the proposed modification to the second physical data page of the decider node; and

propagating a transaction commit message to other nodes of the chain; and
in response to receiving the transaction commit message at the particular storage
node of the chain, initiating a proposed modification corresponding to the
record of the intent to commit.

7. The method as recited in claim 6, wherein the transaction preparation message comprises an indication of a logical timestamp associated with a read request that was directed to a physical data page of at least the particular node of the chain prior to the propagation of the transaction preparation message, and wherein said commit analysis at the particular storage node is based at least in part on a comparison of the logical timestamp with an update timestamp associated with the physical data page.

8. The method as recited in claim 7, further comprising:
submitting, by the coordinator, the read request to the physical data page of the particular
node.

9. The method as recited in claim 6, wherein the coordinator is implemented at a
selected storage node of the service, wherein the selected storage node is a member of the chain.

10. The method as recited in claim 6, further comprising performing, by the one or
more computing devices:
determining the order in which the commit decisions are to be made in accordance with a
deadlock avoidance policy of the service.

11. The method as recited in claim 6, further comprising performing, by the one or
more computing devices:
storing, at the particular storage node, prior to transmitting the transaction preparation
message to the next storage node, a transaction state record corresponding to the
distributed transaction;

modifying, at the particular storage node subsequent to initiating the proposed modification corresponding to the record of the intent to commit, the transaction state record to indicate that the distributed transaction has been committed; and retaining the transaction state record at the particular storage node for a configurable time period.

12. The method as recited in claim 6, further comprising performing, by the one or more computing devices:

configuring a plurality of storage nodes of the service as a replication group responsible for storing a plurality of replicas of at least a portion of a storage object, wherein the at least a portion includes a particular physical data page which is to be accessed in response to the customer work request;

designating a particular node of the plurality of storage nodes as a master replica node of the replication group, responsible for coordinating write operations directed to the at least a portion of the storage object;

determining, by the coordinator, an identity of the master replica node; and

including the master replica node in the chain of storage nodes.

13. The method as recited in claim 6, further comprising performing, by the one or more computing devices:

determining, subsequent to storing a different record of an intent to commit a proposed modification by the master node, wherein the different record of the intent to commit corresponds to a different distributed transaction, that the master node has failed;

designating a new master node for the replication group; and

obtaining, at the new master node, the different record of an intent to commit from the persistent repository; and

in response to receiving the transaction commit message at the new master node, initiating a proposed modification corresponding to the record of the intent to commit.

14. The method as recited in claim 6, further comprising performing, by the one or more computing devices:

determining, by a different storage node, included within a different storage node chain associated with a different prepare transaction message for a different distributed transaction, that a proposed modification to a particular physical data page cannot be committed at the different storage node; and

propagating an abort transaction message associated with the different distributed transaction to one or more other members of the different storage node chain.

15. The method as recited in claim 6, wherein the storage service implements one or more file system programmatic interfaces, wherein the customer work request is formatted in accordance with a particular file system programmatic interface.

16. A non-transitory computer-accessible storage medium storing program instructions that when executed on one or more processors implement a first node of a distributed storage service configured to:

receive an indication that the first node has been selected as a coordinator of a distributed transaction to be implemented in response to a work request directed to the service;

identify one or more other nodes of the distributed storage service as members of a storage node chain collectively storing a group of physical data pages corresponding to proposed modifications to be performed as part of the distributed transaction, including a decider node designated as responsible for a decision to commit the distributed transaction;

generate a transaction preparation message comprising at least a representation of an order in which respective commit decisions associated with the proposed modifications are to be made; and

transmit the transaction preparation message to a selected node of the chain, for a sequential propagation among the nodes of the storage node chain, wherein the sequential propagation ends at the decider node.

17. The non-transitory computer-accessible storage medium as recited in claim 16, wherein the first node is a member of the storage node chain.

18. The non-transitory computer-accessible storage medium as recited in claim 16, wherein the instructions when executed at the one or more processors:

identify, at the first node, a particular physical data page, stored at a particular storage node of the chain, at which a proposed modification is to be applied in response to the work request;

issue a read request directed to the particular physical data page, prior to generation of the transaction preparation message; and

include, within the transaction preparation message, a read logical timestamp associated with the read request, wherein the read logical timestamp is used for a commit decision by the particular storage node.

19. The non-transitory computer-accessible storage medium as recited in claim 16, wherein the instructions when executed at the one or more processors:

determine the order in which the commit decisions are to be made in accordance with a deadlock avoidance policy.

20. The non-transitory computer-accessible storage medium as recited in claim 16, wherein the distributed storage service implements one or more file system programmatic interfaces.