

CLAIMS

1. A method, comprising:

performing, by one or more computing devices:

receiving, at a storage node, a write request for a data volume persisted in at least one block-based storage device maintained at the storage node;

updating at least one corresponding entry of a plurality of entries in a page cache maintained in system memory at the storage node for the write request, wherein the plurality of entries in the page cache are periodically or aperiodically flushed to the at least one block-based storage device in order to persistently update the data volume;

storing a log record describing the update to the at least one corresponding page cache entry in a page cache write log maintained in a persistent storage device along with a plurality of other log records maintained in the persistent storage device, wherein the log record and the plurality of other log records describe a state of the page cache including the updated page cache entry to be restored to the page cache in the event of a system failure resulting in a loss of data in the page cache; and

in response to updating the page cache entry in the page cache maintained in the system memory and storing the log record in the page cache write log maintained in the persistent storage device, acknowledging the received write request for the data volume as complete.

2. The method of claim 1, further comprising:

upon recovery from a system failure:

obtaining from the persistent storage device maintaining the page cache write log, the log record and the plurality of other log records; and

applying the log record and the plurality of other log records maintained in the page cache write log to write to the page cache a state of the page cache prior to the system failure.

3. The method of claim 1, wherein the storage node is one of two or more storage nodes that maintain replicas of the data volume, wherein at least one of the two or more storage nodes different than the storage node does not maintain a respective page cache write log for write requests with respect to the data volume.

4. The method of claim 1, further comprising:
upon recovery from a system failure of the storage nodes, identifying the storage node that maintains the respective page cache write log as maintaining a current replica of the data volume prior to the system failure in order to make the current replica of the data volume available for processing subsequent input/output requests for the data volume.

5. The method of claim 1, further comprising:
detecting a trim event for the page cache write log;
in response to detecting the trim event:
identifying a trim location in the page cache write log;
performing one or more flush operations to write at least some of the plurality of page cache entries from the system memory to the at least one block-based storage device; and
upon completion of the one or more flush operations, reclaiming one or more portions of the persistent storage device maintaining the page cache write log for storing additional log records describing updates to the page cache.

6. The method of claim 5, further comprising:
in response to detecting the trim event, determining a flush operation size for the one or more flush operations.

7. The method of claim 1, wherein the at least one block-based storage device is at least one of a plurality of block-based storage devices maintained at the storage node, and wherein the method further comprises:

receiving another write request for another data volume persisted in at least one other block-based storage device of the plurality of block-based storage devices maintained at the storage node; and
performing said updating, said storing, and said acknowledging for the other write request.

8. The method of claim 1, wherein the one or more computing devices implement a storage node of a plurality of storage nodes that together implement a block-based storage service, wherein the storage node is one of one or more storage nodes in the block-based storage service that implement page cache write logging, wherein other ones of the plurality storage nodes do not implement page cache write logging, and wherein the method further comprises:

performing, by one or more computing devices implementing a block-based storage service control plane:

evaluating the other ones of the plurality of storage nodes to determine a durability score for each other storage node that does not implement page cache write logging; and

based, at least in part on the durability scores, selecting one or more storage nodes of the other storage nodes to implement page cache write logging.

9. The method of claim 1, wherein the storage node is one of a plurality of storage nodes implementing a network-based block-based storage service, wherein the write request is received from a virtual compute instance implemented by a network-based virtual compute service, wherein the network-based block-based storage service and the network-based virtual compute service are implemented together as part of a same network-based services provider network.

10. A system configured to implement the method of one of claims 1-9.

11. A non-transitory, computer-readable storage medium, storing program instructions that when executed by the one or more computing devices cause the one or more computing devices to implement the method of one of claims 1-9.