

WHAT IS CLAIMED IS:

1. A system implementing a storage node, comprising:

one or more block-based storage devices maintaining one or more data volumes;

5 a system memory, comprising a page cache, wherein the page cache is periodically or aperiodically flushed to the one or more block-based storage devices in order to persistently update the one or more data volumes;

10 a persistent storage device maintaining a page cache write log, wherein the page cache write log maintains a plurality of log records describing updates to the page cache, wherein the plurality of log records are applicable to restore a prior state to the page cache in the event of a system failure resulting in a loss of data in the page cache;

the storage node, configured to:

15 receive a write request for a particular data volume of the one or more data volumes;

update at least one corresponding entry of a plurality of entries in the page cache for the write request;

store a log record describing the update to the at least one corresponding page cache entry in the page cache write log; and

20 in response to updating the page cache entry and storing the log record in the page cache write log, acknowledge the write request as complete.

2. The system of claim 1, wherein the storage node is further configured to:

upon recovery from a system failure:

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

30 apply the log record and the plurality of 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 system of claim 1, wherein storage node is further configured to:
detect a trim event for the page cache write log;
in response to detecting the trim event:

identify a trim location in the page cache write log;

perform one or more flush operations to write at least some of the plurality of
page cache entries from the system memory to the plurality of block-based
storage devices; and

upon completion of the one or more flush operations, reclaim 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.

4. The system of claim 1, wherein the storage node is implemented as part of a
network-based block-based storage service, wherein a least one other storage node implemented
as part of the block-based storage service maintains a replica of the particular data volume, and
wherein the at least one other storage node is configured to perform said receiving, said storing,
said updating, and said acknowledging.

5. 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 and storing the log record in the page cache write log, acknowledging the write request as complete.

6. The method of claim 5, 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.

7. The method of claim 5, 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.

8. The method of claim 7, further comprising:

upon recovery from a system failure of the two or more storage nodes, identifying one of the two or more storage nodes 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.

9. The method of claim 5, 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.

10. The method of claim 9, further comprising:
in response to detecting the trim event, determining a flush operation size for the one or
more flush operations.

5 11. The method of claim 5, 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
10 maintained at the storage node; and
performing said updating, said storing, and said acknowledging for the other write
request.

12. The method of claim 5, wherein the one or more computing devices implement a
15 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
20 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
25 of the other storage nodes to implement page cache write logging.

13. The method of claim 5, 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
30 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.

14. A system implementing a master storage node, comprising:
the master storage node, configured to:

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

5 update at least one corresponding entry of a plurality of entries in a page cache
maintained in system memory at the master 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;

10 store a log record describing the update to the at least one corresponding page cache entry
in a page cache write log maintained in persistent storage device along with a
plurality of other log records maintained in the persistent storage device that is
different than the at least one block-based storage device, wherein the log record
and the plurality of other log records describe a state of the page cache including
15 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 and storing the log record in the page cache
write log, acknowledging the write request as complete.

20 15. The system of claim 14, and further comprising:

one or more slave storage nodes, wherein the master storage node and the one or more
slave storage nodes each maintain a respective replica of the data volume, and wherein the
master storage node and the one or more slave storage nodes each perform the receiving, the
storing, the updating, and the acknowledging.

Claims

1. A system implementing a storage node, comprising:
 - one or more block-based storage devices maintaining one or more data volumes;
 - a system memory, comprising a page cache, wherein the page cache is periodically or aperiodically flushed to the one or more block-based storage devices in order to persistently update the one or more data volumes;
 - a persistent storage device maintaining a page cache write log, wherein the page cache write log maintains a plurality of log records describing updates to the page cache, wherein the plurality of log records are applicable to restore a prior state to the page cache in the event of a system failure resulting in a loss of data in the page cache;the storage node, configured to:
 - receive a write request for a particular data volume of the one or more data volumes;
 - update at least one corresponding entry of a plurality of entries in the page cache for the write request;
 - store a log record describing the update to the at least one corresponding page cache entry in the page cache write log; and
 - in response to updating the page cache entry and storing the log record in the page cache write log, acknowledge the write request as complete.
2. The system of claim 1, wherein the storage node is further configured to:
 - upon recovery from a system failure:
 - obtain from the persistent storage device maintaining the page cache write log, the log record and the plurality of log records; and
 - apply the log record and the plurality of 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 system of claim 1, wherein storage node is further configured to:

detect a trim event for the page cache write log;

in response to detecting the trim event:

identify a trim location in the page cache write log;

perform one or more flush operations to write at least some of the plurality of page cache entries from the system memory to the plurality of block-based storage devices; and

upon completion of the one or more flush operations, reclaim 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.

4. The system of claim 1, wherein the storage node is implemented as part of a network-based block-based storage service, wherein a least one other storage node implemented as part of the block-based storage service maintains a replica of the particular data volume, and wherein the at least one other storage node is configured to perform said receiving, said storing, said updating, and said acknowledging.

5. 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 and storing the log record in the page cache write log, acknowledging the write request as complete.

6. The method of claim 5, 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.

7. The method of claim 5, 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.

8. The method of claim 7, further comprising:

upon recovery from a system failure of the two or more storage nodes, identifying one of the two or more storage nodes 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.

9. The method of claim 5, 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.
10. The method of claim 9, further comprising:
in response to detecting the trim event, determining a flush operation size for the one or
more flush operations.
11. The method of claim 5, 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.
12. The method of claim 5, 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.

13. The method of claim 5, 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.

14. A non-transitory, computer-readable storage medium, storing program instructions that when executed by one or more computing devices cause the one or more computing devices to implement:

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 persistent storage device along with a plurality of other log records maintained in the persistent storage device that is different than the at least one block-based 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 and storing the log record in the page cache write log, acknowledging the write request as complete.

15. The non-transitory, computer-readable storage medium of claim 14, wherein the program instructions further cause the one or more computing devices to implement:

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.

16. The non-transitory, computer-readable storage medium of claim 15, wherein the one or more computing devices implement a master storage node and one or more slave storage nodes, wherein the master storage node and the one or more slave storage nodes each maintain a respective replica of the data volume, and wherein the master storage node and the one or more slave storage nodes each perform said receiving, said storing, said updating, and said acknowledging.

17. The non-transitory, computer-readable storage medium of claim 14, wherein the program instructions further cause the one or more computing devices to implement:

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.

18. The non-transitory, computer-readable storage medium of claim 17, wherein the program instructions further cause the one or more computing devices to implement:

in response to detecting the trim event, determining a flush operation size for the one or more flush operations.

19. The non-transitory, computer-readable storage medium of claim 14, 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 program instructions further cause the one or more computing devices to implement:

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.

20. The non-transitory, computer-readable storage medium of claim 14, wherein the storage node is one of a plurality of storage nodes implementing a network-based block-based storage service, wherein each of the plurality of storage nodes maintains a plurality of different data volumes for a plurality of different clients.