WHAT IS CLAIMED IS:

1. A distributed storage service, comprising:

a plurality of computing devices that implement, using resources of a provider network that includes a plurality of availability containers with independent failure profiles:

a service access subsystem configured to receive client requests formatted according to one or more industry-standard file system interfaces from a plurality of compute instances of a virtual computing service implemented at the provider network;

a metadata subsystem configured to implement sequential consistency semantics on at least a subset of file store operations, and

a storage subsystem configured to store at least respective data portions of one or more file stores, wherein a particular data portion of a particular file store of the one or more file stores is organized as a replica group comprising a plurality of extent replicas including a first extent replica at a first availability container of the provider network and a second extent replica at a second availability container of the provider network;

wherein, in response to a particular client request received at the service access subsystem, the plurality of computing devices is configured to:

perform an atomic metadata operation comprising a group of file system metadata modifications, including a first metadata modification at a first node of the metadata subsystem and a second metadata modification at a second node of the metadata subsystem; and

apply at least one modification at a plurality of extent replicas at the storage subsystem prior to a transmission of a response to the particular client request.

2. The system as recited in claim 1, wherein the plurality of computing devices is configured to:

utilize a replicated state machine to generate a response to a particular read request for which respective physical read operations are performed at a plurality of storage devices.

3. The system as recited in claim 1, wherein the service access subsystem, the metadata subsystem and the storage subsystem are each implemented using respective sets of resources of the provider network, wherein the plurality of computing devices is further configured to:

detect one or more of: (a) a potential performance bottleneck at a particular subsystem of a set of subsystems comprising the service access subsystem, the metadata subsystem and the storage subsystem or (b) a node health state change requiring additional resources to be deployed at the particular subsystem; and

initiate a deployment of additional resources of the provider network to the particular subsystem, without modifying the number of resources used for remaining subsystems of the set.

4. The system as recited in claim 1, wherein the plurality of computing devices are further configured to:

utilize a consensus-based protocol to replicate log records of changes to a state of the particular file store; and

store a representation of the state of the particular file store as a plurality of erasure-coded replicas.

5. The system as recited in claim 1, wherein the plurality of computing devices are further configured to:

store, at a particular node of the storage subsystem, a particular extent replica belonging to a second replica group that includes at least a subset of data content of one or more file stores including the particular file store; and

store, at the particular node of the storage subsystem, a particular extent replica of a different replica group that includes at least a subset of metadata of one or more file stores including the particular file store.

6. The system as recited in claim 1, wherein the plurality of computing devices are further configured to:

distribute metadata and data of the particular file store among a plurality of physical storage devices including at least one solid-state disk (SSD device) and one rotating disk device.

7. A method, comprising:

performing, by one or more computing devices:

receiving a particular client request, formatted in accordance with an industry-standard file system interface, at an access subsystem of a multi-tenant storage service;

determining, at the access subsystem, that the client request meets authentication and authorization requirements;

initiating, in response to the particular client request, an atomic metadata operation comprising a group of file system metadata modifications, including a first metadata modification at a first node of a metadata subsystem of the storage service and a second metadata modification at a second node of the metadata subsystem;

verifying, in response to the particular client request, that a plurality of replicas of at least one data modification at a storage subsystem of the storage service have been saved; and

storing a record of completion of the particular client request, wherein the record is to be used, asynchronously with respect to the particular client request, to generate a billing amount to a customer of the storage service in accordance with a usage-based pricing policy.

8. The method as recited in claim 7, wherein the access subsystem, the metadata subsystem and the storage subsystem are each implemented using respective sets of resources of a provider network, further comprising performing, by one or more computing devices of the plurality of computing devices:

initiating, in response to a detection of a triggering condition, a deployment of additional resources of the provider network to a particular subsystem of a set of subsystems comprising the access subsystem, the metadata subsystem and the storage subsystem, without modifying the number of resources used for remaining subsystems of the set.

9. The method as recited in claim 7, further comprising performing, by the plurality of computing devices:

utilizing a consensus-based protocol to replicate log records of changes to a state of the particular file store; and

storing a representation of the state of the particular file store as a plurality of erasure-coded replicas.

10. The method as recited in claim 7, further comprising performing, by the plurality of computing devices:

storing, at a particular node of the storage subsystem, a particular replica belonging to a replica group storing data content of one or more file stores; and

storing, at the particular node of the storage subsystem, a particular replica of a different replica group storing metadata associated with one or more file stores.

11. The method as recited in claim 7, further comprising performing, by the plurality of computing devices:

allocating, in response to one or more write requests directed to a particular file store object, a first set of blocks of storage for write contents indicated in the write requests, and a second set of blocks of storage for metadata associated with the file store object, wherein sizes of blocks of the first set are selected according to a data block sizing policy, wherein sizes of blocks of the second set are selected according to a metadata block sizing policy, wherein at least one block of the first set differs in size from at least one block of the second set.

12. The method as recited in claim 11, further comprising performing, by the plurality of computing devices:

issuing, from the access subsystem, in response to a client request directed to the particular file store object subsequent to said allocating,

a page I/O (input/output) request to the storage subsystem for a particular metadata page to which a metadata block of the second set is mapped for physical storage, wherein a size of the metadata page differs from a size of the metadata block; and

a second page I/O request to the storage subsystem for a particular data page to which a data block of the first set is mapped for physical storage, wherein a size of the data page differs from a size of the data block, and wherein the size of the data page differs from the size of the metadata page.

13. The method as recited in claim 11, wherein the write request is received from a first client of the multi-tenant storage service, further comprising performing, by the plurality of computing devices:

determining, corresponding to a particular block of the second set, a reallocation ineligibility timeout, indicative of a minimum time period for which the particular block is not to be allocated to store metadata for another file store object; and

determining, corresponding to the particular block, a caching timeout indicative of a maximum period for which the particular block is to be retained at a node of the access subsystem before being re-validated with the metadata subsystem, wherein the caching timeout is set smaller than the reallocation ineligibility timeout.

14. The method as recited in claim 12, further comprising performing, by the plurality of computing devices:

retrieving, at the access subsystem from the metadata subsystem, the particular block;

caching, at the access subsystem, the particular block in accordance with the caching timeout; and

directing, from the access subsystem to the storage subsystem, one or more I/O requests directed to the particular file store object generated in response to additional client requests, without retrieving additional metadata from the metadata subsystem.

15. The method as recited in claim 7, further comprising performing, by the plurality of computing devices:

configuring, at the request of a particular client of a provider network, an isolated virtual network comprising a plurality of resources of a provider network, wherein respective private network addresses assigned to the plurality of resources are not accessible from the public Internet; and

configuring, to receive service requests at one or more nodes of the access subsystem, a particular private network address accessible from other resources of the isolated virtual network.

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

store, at a first data center, a plurality of extent replicas associated with a set of file stores, including a first extent replica comprising data of a first file store and a second extent replica comprising metadata associated with the first file store, wherein one or more additional extent replicas corresponding to the first extent replica and the second extent replica are stored at other data centers, and wherein the first extent replica is designated as a master data replica of a particular replica group corresponding to a particular file store object;

receive a write request directed to the master data replica from an access subsystem of the distributed storage service, wherein the access subsystem implements an industry-standard file system interface and is configured to identify the master replica using metadata managed by a metadata subsystem of the distributed storage service; and

coordinate, in response to the write request, respective updates to a plurality of members of the particular replica group using a consensus based state management protocol.

17. A non-transitory computer-accessible storage medium storing program instructions that when executed on one or more processors implement a node of a metadata subsystem of a distributed storage service, wherein the node is configured to:

coordinate storage of metadata associated with one or more file stores implemented at the distributed storage service;

receive, from an access subsystem of the distributed storage service, an internal operation request generated at the access subsystem in response to a client request formatted in accordance with an industry-standard file system interface;

determine that, to respond to the internal operation request, a plurality of metadata objects associated with the file store including a first metadata object and a second metadata object are to be modified in accordance with sequential consistency semantics, wherein at least a portion of the first metadata object is stored at a first extent replica group and at least a portion of the second metadata object is stored at a second extent replica group;

initiate an update protocol to implement updates to the first extent replica group and the second extent replica group in accordance with the sequential consistency semantics.

18. A non-transitory computer-accessible storage medium storing program instructions that when executed on one or more processors implement a node of an access subsystem of a distributed storage service, wherein the node is configured to:

expose one or more network addresses to enable a plurality of clients of the distributed storage service to submit service requests in accordance with one or more industry-standard file system interfaces;

receive, in accordance with a particular interface of the one or more file system industry-standard interfaces, an I/O request directed to a particular object of a file store comprising a plurality of blocks, wherein each block is mapped to one or more physical pages by the storage service;

obtain metadata pertaining to the particular object from a metadata subsystem of the distributed storage service;

determine, using the metadata, (a) a particular node of a storage subsystem of the distributed storage subsystem that stores at least one replica of contents of a particular logical block of the plurality of blocks and interacts with other nodes of the storage subsystem using a consensus-based protocol to implement update operations, and (b) an offset within the particular logical block that is to be accessed in response to the file I/O request; and

transmit an internal I/O request indicating the offset to the particular node of the storage subsystem.

WHAT IS CLAIMED IS:

1. A system, comprising:

a plurality of computing devices that implement, using resources of a provider network that includes a plurality of availability containers with independent failure profiles:

a service access subsystem configured to receive client requests formatted according to one or more industry-standard file system interfaces from a plurality of compute instances of a virtual computing service implemented at the provider network;

a metadata subsystem configured to implement sequential consistency semantics on at least a subset of file store operations; and

a storage subsystem configured to store at least respective data portions of one or more file stores, wherein a particular data portion of a particular file store of the one or more file stores is organized as a replica group comprising a plurality of extent replicas including a first extent replica at a first availability container of the provider network and a second extent replica at a second availability container of the provider network;

wherein, in response to a particular client request received at the service access subsystem, the plurality of computing devices is configured to:

perform an atomic metadata operation comprising a group of file system metadata modifications, including a first metadata modification at a first node of the metadata subsystem and a second metadata modification at a second node of the metadata subsystem; and

apply at least one modification at a plurality of extent replicas at the storage subsystem prior to a transmission of a response to the particular client request.

2. The system as recited in claim 1, wherein the plurality of computing devices is configured to:

utilize a replicated state machine to generate a response to a particular read request for which respective physical read operations are performed at a plurality of storage devices.

3. The system as recited in claim 1, wherein the service access subsystem, the metadata subsystem and the storage subsystem are each implemented using respective sets of resources of the provider network, wherein the plurality of computing devices is further configured to:

detect one or more of: (a) a potential performance bottleneck at a particular subsystem of a set of subsystems comprising the service access subsystem, the metadata subsystem and the storage subsystem or (b) a node health state change requiring additional resources to be deployed at the particular subsystem; and

initiate a deployment of additional resources of the provider network to the particular subsystem, without modifying the number of resources used for remaining subsystems of the set.

4. The system as recited in claim 1, wherein the plurality of computing devices are further configured to:

utilize a consensus-based protocol to replicate log records of changes to a state of the particular file store; and

store a representation of the state of the particular file store as a plurality of erasure-coded replicas.

5. The system as recited in claim 1, wherein the plurality of computing devices are further configured to:

store, at a particular node of the storage subsystem, a particular extent replica belonging to a second replica group that includes at least a subset of data content of one or more file stores including the particular file store; and

store, at the particular node of the storage subsystem, a particular extent replica of a different replica group that includes at least a subset of metadata of one or more file stores including the particular file store.

6. The system as recited in claim 1, wherein the plurality of computing devices are further configured to:

distribute metadata and data of the particular file store among a plurality of physical storage devices including at least one solid-state disk (SSD device) and one rotating disk device.

7. A method, comprising:

performing, by one or more computing devices:

receiving a particular client request, formatted in accordance with an industry-standard file system interface, at an access subsystem of a multi-tenant storage service;

determining, at the access subsystem, that the client request meets authentication and authorization requirements;

initiating, in response to the particular client request, an atomic metadata operation comprising a group of file system metadata modifications, including a first metadata modification at a first node of a metadata subsystem of the storage service and a second metadata modification at a second node of the metadata subsystem;

verifying, in response to the particular client request, that a plurality of replicas of at least one data modification at a storage subsystem of the storage service have been saved; and

storing a record of completion of the particular client request, wherein the record is to be used, asynchronously with respect to the particular client request, to generate a billing amount to a customer of the storage service in accordance with a usage-based pricing policy.

8. The method as recited in claim 7, wherein the access subsystem, the metadata subsystem and the storage subsystem are each implemented using respective sets of resources of a provider network, further comprising performing, by one or more computing devices of the plurality of computing devices:

initiating, in response to a detection of a triggering condition, a deployment of additional resources of the provider network to a particular subsystem of a set of subsystems comprising the access subsystem, the metadata subsystem and the storage subsystem, without modifying the number of resources used for remaining subsystems of the set.

9. The method as recited in claim 7, further comprising performing, by the plurality of computing devices:

utilizing a consensus-based protocol to replicate log records of changes to a state of the particular file store; and

storing a representation of the state of the particular file store as a plurality of erasure-coded replicas.

10. The method as recited in claim 7, further comprising performing, by the plurality of computing devices:

storing, at a particular node of the storage subsystem, a particular replica belonging to a replica group storing data content of one or more file stores; and

storing, at the particular node of the storage subsystem, a particular replica of a different replica group storing metadata associated with one or more file stores.

11. The method as recited in claim 7, further comprising performing, by the plurality of computing devices:

allocating, in response to one or more write requests directed to a particular file store object, a first set of blocks of storage for write contents indicated in the write requests, and a second set of blocks of storage for metadata associated with the file store object, wherein sizes of blocks of the first set are selected according to a data block sizing policy, wherein sizes of blocks of the second set are selected according to a metadata block sizing policy, wherein at least one block of the first set differs in size from at least one block of the second set.

12. The method as recited in claim 11, further comprising performing, by the plurality of computing devices:

issuing, from the access subsystem, in response to a client request directed to the particular file store object subsequent to said allocating,

a page I/O (input/output) request to the storage subsystem for a particular metadata page to which a metadata block of the second set is mapped for physical storage, wherein a size of the metadata page differs from a size of the metadata block; and

a second page I/O request to the storage subsystem for a particular data page to which a data block of the first set is mapped for physical storage, wherein a size of the data page differs from a size of the data block, and wherein the size of the data page differs from the size of the metadata page.

13. The method as recited in claim 11, wherein the write request is received from a first client of the multi-tenant storage service, further comprising performing, by the plurality of computing devices:

determining, corresponding to a particular block of the second set, a reallocation ineligibility timeout, indicative of a minimum time period for which the particular block is not to be allocated to store metadata for another file store object; and

determining, corresponding to the particular block, a caching timeout indicative of a maximum period for which the particular block is to be retained at a node of the access subsystem before being re-validated with the metadata subsystem, wherein the caching timeout is set smaller than the reallocation ineligibility timeout.

14. The method as recited in claim 12, further comprising performing, by the plurality of computing devices:

retrieving, at the access subsystem from the metadata subsystem, the particular block;

caching, at the access subsystem, the particular block in accordance with the caching timeout; and

directing, from the access subsystem to the storage subsystem, one or more I/O requests directed to the particular file store object generated in response to additional client requests, without retrieving additional metadata from the metadata subsystem.

15. The method as recited in claim 7, further comprising performing, by the plurality of computing devices:

configuring, at the request of a particular client of a provider network, an isolated virtual network comprising a plurality of resources of a provider network, wherein respective private network addresses assigned to the plurality of resources are not accessible from the public Internet; and

configuring, to receive service requests at one or more nodes of the access subsystem, a particular private network address accessible from other resources of the isolated virtual network.

16. The method as recited in claim 7, further comprising performing, by the plurality of computing devices:

distributing, by the storage subsystem, metadata and data of the particular file store among a plurality of physical storage devices including at least one solid-state disk (SSD device) and one rotating disk device.

17. A non-transitory computer-accessible storage medium storing program instructions that when executed on one or more processors implement a node of a storage subsystem of a distributed storage service, wherein the node is configured to:

store, at a first data center, a plurality of extent replicas associated with a set of file stores, including a first extent replica comprising data of a first file store and a second extent replica comprising metadata associated with the first file store, wherein one or more additional extent replicas corresponding to the first extent replica and the second extent replica are stored at other data centers, and wherein the first extent replica is designated as a master data replica of a particular replica group corresponding to a particular file store object;

receive a write request directed to the master data replica from an access subsystem of the distributed storage service, wherein the access subsystem implements an industry-standard file system interface and is configured to identify the master replica using metadata managed by a metadata subsystem of the distributed storage service; and

coordinate, in response to the write request, respective updates to a plurality of members of the particular replica group using a consensus based state management protocol.

18. The non-transitory computer-accessible storage medium as recited in claim 17, wherein the first file store is established on behalf of a particular customer of the distributed storage service, and wherein the node is further configured to store at least one addition extent replica of a different file store established on behalf of a different customer of the distributed storage service.

19. A non-transitory computer-accessible storage medium storing program instructions that when executed on one or more processors implement a node of a metadata subsystem of a distributed storage service, wherein the node is configured to:

coordinate storage of metadata associated with one or more file stores implemented at the distributed storage service;

receive, from an access subsystem of the distributed storage service, an internal operation request generated at the access subsystem in response to a client request formatted in accordance with an industry-standard file system interface;

determine that, to respond to the internal operation request, a plurality of metadata objects associated with the file store including a first metadata object and a second metadata object are to be modified in accordance with sequential consistency semantics, wherein at least a portion of the first metadata object is stored at a first extent replica group and at least a portion of the second metadata object is stored at a second extent replica group;

initiate an update protocol to implement updates to the first extent replica group and the second extent replica group in accordance with the sequential consistency semantics.

20. The non-transitory computer-accessible storage medium as recited in claim 19, wherein the first replica group comprises a first extent replica at a first data center and a second extent replica at a second data center, and wherein, in accordance with a durability policy associated with metadata of the file store, the node is further configured to:

verify that a particular modification is complete at the first and second extent replicas prior to generating a response to the internal operation request.

21. A non-transitory computer-accessible storage medium storing program instructions that when executed on one or more processors implement a node of an access subsystem of a distributed storage service, wherein the node is configured to:

expose one or more network addresses to enable a plurality of clients of the distributed storage service to submit service requests in accordance with one or more industry-standard file system interfaces;

receive, in accordance with a particular interface of the one or more file system industry-standard interfaces, an I/O request directed to a particular object of a file store comprising a plurality of blocks, wherein each block is mapped to one or more physical pages by the storage service;

obtain metadata pertaining to the particular object from a metadata subsystem of the distributed storage service;

determine, using the metadata, (a) a particular node of a storage subsystem of the distributed storage subsystem that stores at least one replica of contents of a particular logical block of the plurality of blocks and interacts with other nodes of the storage subsystem using a consensus-based protocol to implement update operations, and (b) an offset within the particular logical block that is to be accessed in response to the file I/O request; and

transmit an internal I/O request indicating the offset to the particular node of the storage subsystem.

22. The non-transitory computer-accessible storage medium as recited in claim 21, wherein the node is further configured to:

cache a particular metadata block associated with the particular object, wherein the particular metadata block is retrieved from the metadata subsystem, for a time period in accordance with a caching timeout, wherein a value of the caching timeout is set smaller than a reallocation ineligibility timeout associated with the particular metadata block, wherein the reallocation ineligibility timeout is indicative of a minimum time period for which storage used for the particular metadata block is not to be re-allocated; and

utilize the particular metadata block to issue one or more subsequent internal I/O requests associated with the particular object.

23. The non-transitory computer-accessible storage medium as recited in claim 21, wherein the node is configured with a private network address accessible from devices of an isolated virtual network established at the request of a particular client of the distributed storage service, and wherein the I/O request is received from a compute instance instantiated in the isolated virtual network.