WHAT IS CLAIMED IS:

1. A system, comprising:

one or more computing devices configured to:

establish, to implement respective rates of provisioned workloads, respective sets of admission control parameters for each of a plurality of block-level storage devices implemented at a multi-tenant storage service;

generate an estimate, by a client-side component of the multi-tenant storage service, of a particular rate of work requests expected to be directed during a particular time period to at least a portion of a first block-level storage device implemented at a first storage server, wherein the particular rate exceeds a first provisioned rate;

identify, by the client-side component, one or more other storage servers, including a second storage server, at which respective rates of work requests are anticipated to be less than respective provisioned rates;

verify that the first storage server has a sufficient workload capacity during the particular time period to complete work requests at a rate higher than the first provisioned rate;

modify at least one admission control parameter of the first block-level storage device to enable the first storage server to accept work requests at up to a rate higher than the first provisioned rate during the particular time period; and

modify at least one admission control parameter of at least a particular block-level storage device at the second storage server to enable the second storage server to accept work requests at a rate no greater than a provisioned rate of the second storage server during the particular time period.

2. The system as recited in claim 1, wherein at least a portion of the client-side component is implemented within a virtualization management software stack at an instance host of a multi-tenant computing service.

3. The system as recited in claim 1, wherein to modify the at least one admission control parameter of the first block-level storage device, the one or more computing devices are further configured to increase a token refill rate of a work token bucket associated with the first block-level storage device.

4. The system as recited in claim 1, wherein said portion of the first block-level storage device comprises a first partition of a multi-partition block-level volume established for a particular client, and wherein at least a portion of the particular block-level storage device at the second storage server comprises a second partition of the multi-partition block-level volume.

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

re-set, after the particular time period, a particular admission control parameter of the first block-level storage device to enable the first storage server to accept work requests at no greater than the first provisioned rate.

6. The system as recited in claim 5, wherein, to re-set the particular admission parameter, a value of the particular admission control parameter is changed from a first setting to a second setting in accordance with a decay function over a re-set time period.

7. The system as recited in claim 5, wherein, to re-set the particular admission parameter, a value of the particular admission control parameter is changed from a first setting to a second setting in accordance with a step function.

8. A method, comprising:

performing, by one or more computing devices:

establishing, to implement respective workload limits, respective sets of one or more admission control parameters for a plurality of block-level storage devices implemented at a storage service;

generating, by a client-side component of the storage service, an estimate of a particular rate of work requests expected to be directed during a particular time period to at least a portion of a first block-level storage device implemented at a first storage server, wherein said particular rate of work

requests exceeds a first rate associated with the first block-level storage device;

identifying, by the client-side component, at least a portion of a particular block-level storage device to which a second rate of work requests directed during the particular time period is anticipated to be less than a second rate associated with the particular block-level storage device; and

modifying at least one admission control parameter of the first block-level storage device to enable the first multi-tenant storage server to accept work requests directed to the first block-level storage device at a rate higher than the first rate.

9. The method as recited in claim 8, wherein the first storage server is configured to implement block-level storage devices of a plurality of clients of the service.

10. The method as recited in claim 8, further comprising performing, by the one or more computing devices:

verifying, prior to said modifying, that the first storage server has a sufficient workload capacity during the particular time period to complete work requests at the rate higher than the first rate.

11. The method as recited in claim 8, wherein at least a portion of the client-side component is implemented within a virtualization management software stack at an instance host of a multi-tenant computing service.

12. The method as recited in claim 8, wherein said modifying at least one admission control parameter of the first block-level storage device comprises increasing a token refill rate in a work token bucket associated with the first block-level storage device.

13. The method as recited in claim 8, wherein said portion of the first block-level storage device comprises a first partition of a multi-partition block-level volume established for a particular client, and wherein the portion of the particular block-level storage device comprises a second partition of the multi-partition block-level volume.

14. The method as recited in claim 8, further comprising performing, by the one or more computing devices:

re-setting, after the particular time period, a particular admission control parameter of the first block-level storage device to enable the first storage server to accept work requests at no greater than the first rate.

15. The method as recited in claim 8, further comprising performing, by the one or more computing devices:

attaching the first block-level storage device to a plurality of compute instances including a first compute instance at a first instance host and a second compute instance at a second instance host, wherein the client-side component is instantiated at the first instance host;

obtaining, by the client-side component at the first instance host, an indication of a workload level of a second client-side component at the second instance host, to determine a change to be made to the at least one admission control parameter; and

wherein said indication of the workload level of the second client-side component is provided from the first storage server to the client-side component at the first instance host.

Claims

1. A system, comprising:

one or more computing devices configured to:

establish, to implement respective rates of provisioned workloads, respective sets of admission control parameters for each of a plurality of block-level storage devices implemented at a multi-tenant storage service;

generate an estimate, by a client-side component of the multi-tenant storage service, of a particular rate of work requests expected to be directed during a particular time period to at least a portion of a first block-level storage device implemented at a first storage server, wherein the particular rate exceeds a first provisioned rate, and wherein the particular time period is after the first provisioned rate has been established;

identify, by the client-side component, one or more other storage servers, including a second storage server, at which respective rates of work requests are anticipated to be less than respective provisioned rates;

verify that the first storage server has a sufficient workload capacity during the particular time period to complete work requests at a rate higher than the first provisioned rate;

modify at least one admission control parameter of the first block-level storage device to enable the first storage server to accept work requests at up to a rate higher than the first provisioned rate during the particular time period; and

modify at least one admission control parameter of at least a particular block-level storage device at the second storage server to enable the second storage server to accept work requests at a rate no greater than a provisioned rate of the second storage server during the particular time period.

2. The system as recited in claim 1, wherein at least a portion of the client-side component is implemented within a virtualization management software stack at an instance host of a multi-tenant computing service.

3. The system as recited in claim 1, wherein to modify the at least one admission control parameter of the first block-level storage device, the one or more computing devices are further configured to increase a token refill rate of a work token bucket associated with the first block-level storage device.

4. The system as recited in claim 1, wherein said portion of the first block-level storage device comprises a first partition of a multi-partition block-level volume established for a particular client, and wherein at least a portion of the particular block-level storage device at the second storage server comprises a second partition of the multi-partition block-level volume.

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

re-set, after the particular time period, a particular admission control parameter of the first block-level storage device to enable the first storage server to accept work requests at no greater than the first provisioned rate.

6. The system as recited in claim 5, wherein, to re-set the particular admission parameter, a value of the particular admission control parameter is changed from a first setting to a second setting in accordance with a decay function over a re-set time period.

7. The system as recited in claim 5, wherein, to re-set the particular admission parameter, a value of the particular admission control parameter is changed from a first setting to a second setting in accordance with a step function.

8. A method, comprising:

performing, by one or more computing devices:

establishing, to implement respective workload limits, respective sets of one or more admission control parameters for a plurality of block-level storage devices implemented at a storage service;

generating, by a client-side component of the storage service, an estimate of a particular rate of work requests expected to be directed during a particular time period to at least a portion of a first block-level storage device implemented at a first storage server, wherein said particular rate of work requests exceeds a first rate associated with the first block-level storage device, and wherein the particular time period is after the first rate has been established;

identifying, by the client-side component, at least a portion of a particular block-level storage device to which a second rate of work requests directed during the particular time period is anticipated to be less than a second rate associated with the particular block-level storage device;

verifying that the first storage server has a sufficient workload capacity during the particular time period to complete work requests at a rate higher than the first rate; and

modifying at least one admission control parameter of the first block-level storage device to enable the first multi-tenant storage server to accept work requests directed to the first block-level storage device at the rate higher than the first rate.

9. The method as recited in claim 8, wherein the first storage server is configured to implement block-level storage devices of a plurality of clients of the service.

10. The method as recited in claim 8, further comprising performing, by the one or more computing devices,

said verifying prior to said modifying.

11. The method as recited in claim 8, further comprising performing, by the one or more computing devices:

modifying at least one admission control parameter of the particular block-level storage device to enable the corresponding storage server to accept work requests at a rate no greater than the second rate.

12. The method as recited in claim 8, wherein at least a portion of the client-side component is implemented within a virtualization management software stack at an instance host of a multi-tenant computing service.

13. The method as recited in claim 8, wherein said modifying at least one admission control parameter of the first block-level storage device comprises increasing a token refill rate in a work token bucket associated with the first block-level storage device.

14. The method as recited in claim 8, wherein said portion of the first block-level storage device comprises a first partition of a multi-partition block-level volume established for a particular client, and wherein the portion of the particular block-level storage device comprises a second partition of the multi-partition block-level volume.

15. The method as recited in claim 8, further comprising performing, by the one or more computing devices:

re-setting, after the particular time period, a particular admission control parameter of the first block-level storage device to enable the first storage server to accept work requests at no greater than the first rate.

16. The method as recited in claim 8, further comprising performing, by the one or more computing devices:

attaching the first block-level storage device to a plurality of compute instances including a first compute instance at a first instance host and a second compute instance at a second instance host, wherein the client-side component is instantiated at the first instance host;

obtaining, by the client-side component at the first instance host, an indication of a workload level of a second client-side component at the second instance host, to determine a change to be made to the at least one admission control parameter.

17. The method as recited in claim 16, wherein said indication of the workload level of the second client-side component is provided from the first storage server to the client-side component at the first instance host.

18. A non-transitory computer-accessible storage medium storing program instructions that when executed on one or more processors:

generate an estimate of a particular rate of work requests expected to be directed during a particular time period to at least a portion of a first block storage device implemented at a first storage server of a storage service, wherein said particular rate of work requests exceeds a first rate associated with the first block storage device, and wherein the particular time period is after the first rate has been established;

identify, at a client-side component of the storage service, at least one other storage server at which a second rate of work requests directed to at least a portion of a particular block storage device during the particular time period is anticipated to be less than a second rate associated with the particular block storage device;

verify that the first storage server is expected to have a sufficient workload capacity during the particular time period to complete work requests at a rate higher than the first rate; and

modify at least one admission control parameter associated with the first block storage device to enable the first storage server to accept work requests at the rate higher than the first rate.

19. The non-transitory computer-accessible storage medium as recited in claim 18, wherein the first storage server is configured to implement storage devices of a plurality of clients of the storage service.

20. The non-transitory computer-accessible storage medium as recited in claim 18, wherein the instructions, when executed on the one or more computing devices:

verify that the first storage server is expected to have the sufficient workload capacity prior to modifying the at least one admission control parameter.

21. The non-transitory computer-accessible storage medium as recited in claim 18, wherein the instructions, when executed on the one or more computing devices:

modify at least one admission control parameter of the particular storage device to enable the other storage server to accept work requests at a rate no greater than the second rate.

22. The non-transitory computer-accessible storage medium as recited in claim 18, wherein at least a portion of the client-side component is implemented within a virtualization management software stack at an instance host of a multi-tenant computing service.