CLAIMS

1. A system, comprising:

one or more computing devices configured to:

receive a work request directed to a work target;

in response to a determination that a token population of a normal-mode token bucket associated with the work target meets a first threshold criterion, consume one or more tokens from the normal-mode token bucket in accordance with a normal-mode token consumption policy, and accept the work request for execution; and

in response to a determination that the token population of the normal-mode token bucket does not meet the first threshold criterion,

determine whether a token population of at least one bucket of a burst-mode token bucket set meets a second threshold criterion;

in response to a determination that the token population of the at least one bucket of the burst-mode token bucket set meets the second threshold criterion, consume one or more tokens from the at least one bucket of the burst-mode token bucket set based at least in part on a burst-mode token consumption policy, and accept the work request for execution, and

in response to a determination that the token population of the at least one bucket of the burst-mode token bucket set does not meet the second threshold criterion, reject the work request.

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

assign a provisioned throughput capacity to the work target, indicative of a maximum rate of work operations to be performed at the work target during a normal mode of operation;

refill the normal-mode token bucket at a rate based at least in part on the provisioned throughput capacity; and

refill at least one bucket of the burst-mode token bucket set at another rate based at least in part on the provisioned throughput capacity.

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

in response to a determination that, at the end of a particular time period in which an average rate at which work requests were received was less than a provisioned rate designated for the work target, one or more work tokens remain in the normal-mode token bucket, increase the token population of at least one bucket of the burst-mode token bucket set by an amount based at least in part on the number of work tokens that remain in the normal-mode token bucket.

4. The system as recited in claim 1, wherein the burst-mode bucket set comprises a local-burst-limit token bucket whose maximum token population is indicative of a maximum burst rate supported for work operations at the work target, and a shared-resource token bucket whose token population is indicative of an available throughput capacity of a resource shared by a plurality of work objects, wherein, in response to said determination that the token population of the normal-mode token bucket does not meet the first threshold criterion, the one or more computing devices are further configured to:

accept the work request for execution in response to determining that the token population of the local-burst-limit token bucket exceeds zero and the token population of the shared-resource token bucket exceeds zero.

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

generate, prior to the acceptance of the work request for execution, an estimate of an amount of work to be performed in response to the work request;

determine, subsequent to the acceptance of the work request, the amount of work performed corresponding to the work request; and

in response to a determination of a difference between the amount of work performed and an estimate, modify, based at least in part on the difference, the token population of at least one bucket of one or more of: (a) the burst-mode token bucket set or (b) the burst-mode token bucket set.

6. A method, comprising:

performing, by one or more computing devices:

receiving a work request directed to a work target;

determining a token population of a normal-mode token bucket associated with the work target; and

in response to determining that the token population of the normal-mode token bucket meets a first threshold criterion,

determining that the token population of at least one bucket of a burst-mode token bucket set associated with the work target meets a second threshold criterion,

accepting the work request for execution; and

modifying the token population of at least one bucket of the burst-mode token bucket set based at least in part on a burst-mode token consumption policy.

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

assigning a provisioned throughput capacity to the work target, indicative of a maximum rate of work operations to be performed at the work target during a normal mode of operation;

refilling the normal-mode token bucket at a rate based at least in part on the provisioned throughput capacity; and

refilling at least one bucket of the burst-mode token bucket set at a different rate based at least in part on the provisioned throughput capacity.

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

in response to determining that, at the end of a particular time period in which the work target operated in normal mode, one or more work tokens remain in the normal-mode token bucket, increasing the token population of at least one bucket of the burst-mode token bucket set by an amount based at least in part on the number of work tokens that remain in the normal-mode token bucket.

9. The method as recited in claim 6, wherein the burst mode bucket set comprises a local-burst-limit token bucket whose maximum token population is indicative of a maximum burst rate supported for work operations at the work target, and a shared-resource token bucket whose token population is indicative of an available throughput capacity of a resource shared by

a plurality of work objects, wherein, in response to said determining that the token population of the normal-mode token bucket meets a first threshold criterion, the method further comprises performing, by the one or more computing devices:

accepting the work request for execution in response to determining that the token population of the local-burst-limit token bucket exceeds zero and the token population of the shared-resource token bucket exceeds zero.

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

generating, prior to accepting the work request for execution, an estimate of an amount of work to be performed in response to the work request;

determining, after said initiating, the amount of work performed corresponding to the work request; and

in response to determining a difference between the amount of work and the estimate, modifying, based at least in part on the difference, the token population of at least one bucket of one or more of: (a) the burst-mode token bucket set or (b) the burst-mode token bucket set.

11. The method as recited in claim 6, wherein the work target comprises at least a portion of a storage object, wherein the work request comprises at least one of (a) a read operation or (b) a write operation.

12. The method as recited in claim 6, wherein the burst-mode bucket set comprises a read-burst token bucket used for admission control for work requests comprising read operations, and a write-burst token bucket used for admission control for work requests comprising write operations.

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

monitoring a work request arrival rate over a time period; and

modifying a maximum token population limit of at least one bucket of the burst-mode token bucket set based at least in part on a result of monitoring the work request arrival rate.

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

providing, to a client, an indication of a pricing policy associated with burst-mode operations at the work target; and

modifying a maximum token population limit of at least one bucket of the burst-mode token bucket set based at least in part on a burst-mode capacity increase request received from a client in accordance with the pricing policy.

15. A system comprising:

one or more processors; and

one or more memories, the one or more memories storing program instructions that when executed on one or more processors:

receive a work request directed to a work target; and

in response to a determination, based at least in part on a first threshold criterion, that the work target is in a burst mode of operation,

determine that the token population of at least one bucket of a burst-mode token bucket set associated with the work target meets a second threshold criterion; and

accept the work request for execution.

Claims

1. A system, comprising:

one or more computing devices configured to:

receive a work request directed to a work target;

in response to a determination that a token population of a normal-mode token bucket associated with the work target meets a first threshold criterion, consume one or more tokens from the normal-mode token bucket in accordance with a normal-mode token consumption policy, and accept the work request for execution; and

in response to a determination that the token population of the normal-mode token bucket does not meet the first threshold criterion,

determine whether a token population of at least one bucket of a burst-mode token bucket set meets a second threshold criterion;

in response to a determination that the token population of the at least one bucket of the burst-mode token bucket set meets the second threshold criterion, consume one or more tokens from the at least one bucket of the burst-mode token bucket set based at least in part on a burst-mode token consumption policy, and accept the work request for execution, and

in response to a determination that the token population of the at least one bucket of the burst-mode token bucket set does not meet the second threshold criterion, reject the work request.

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

assign a provisioned throughput capacity to the work target, indicative of a maximum rate of work operations to be performed at the work target during a normal mode of operation;

refill the normal-mode token bucket at a rate based at least in part on the provisioned throughput capacity; and

refill at least one bucket of the burst-mode token bucket set at another rate based at least in part on the provisioned throughput capacity.

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

in response to a determination that, at the end of a particular time period in which an average rate at which work requests were received was less than a provisioned rate designated for the work target, one or more work tokens remain in the normal-mode token bucket, increase the token population of at least one bucket of the burst-mode token bucket set by an amount based at least in part on the number of work tokens that remain in the normal-mode token bucket.

4. The system as recited in claim 1, wherein the burst-mode bucket set comprises a local-burst-limit token bucket whose maximum token population is indicative of a maximum burst rate supported for work operations at the work target, and a shared-resource token bucket whose token population is indicative of an available throughput capacity of a resource shared by a plurality of work objects, wherein, in response to said determination that the token population of the normal-mode token bucket does not meet the first threshold criterion, the one or more computing devices are further configured to:

accept the work request for execution in response to determining that the token population of the local-burst-limit token bucket exceeds zero and the token population of the shared-resource token bucket exceeds zero.

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

generate, prior to the acceptance of the work request for execution, an estimate of an amount of work to be performed in response to the work request;

determine, subsequent to the acceptance of the work request, the amount of work performed corresponding to the work request; and

in response to a determination of a difference between the amount of work performed and an estimate, modify, based at least in part on the difference, the token population of at least one bucket of one or more of: (a) the burst-mode token bucket set or (b) the burst-mode token bucket set.

6. A method, comprising:

performing, by one or more computing devices:

receiving a work request directed to a work target;

determining a token population of a normal-mode token bucket associated with the work target; and

in response to determining that the token population of the normal-mode token bucket meets a first threshold criterion,

determining that the token population of at least one bucket of a burst-mode token bucket set associated with the work target meets a second threshold criterion,

accepting the work request for execution; and

modifying the token population of at least one bucket of the burst-mode token bucket set based at least in part on a burst-mode token consumption policy.

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

assigning a provisioned throughput capacity to the work target, indicative of a maximum rate of work operations to be performed at the work target during a normal mode of operation;

refilling the normal-mode token bucket at a rate based at least in part on the provisioned throughput capacity; and

refilling at least one bucket of the burst-mode token bucket set at a different rate based at least in part on the provisioned throughput capacity.

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

in response to determining that, at the end of a particular time period in which the work target operated in normal mode, one or more work tokens remain in the normal-mode token bucket, increasing the token population of at least one bucket of the burst-mode token bucket set by an amount based at least in part on the number of work tokens that remain in the normal-mode token bucket.

9. The method as recited in claim 6, wherein the burst mode bucket set comprises a local-burst-limit token bucket whose maximum token population is indicative of a maximum burst rate supported for work operations at the work target, and a shared-resource token bucket whose token population is indicative of an available throughput capacity of a resource shared by a plurality of work objects, wherein, in response to said determining that the token population of the normal-mode token bucket meets a first threshold criterion, the method further comprises performing, by the one or more computing devices:

accepting the work request for execution in response to determining that the token population of the local-burst-limit token bucket exceeds zero and the token population of the shared-resource token bucket exceeds zero.

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

generating, prior to accepting the work request for execution, an estimate of an amount of work to be performed in response to the work request;

determining, after said initiating, the amount of work performed corresponding to the work request; and

in response to determining a difference between the amount of work and the estimate, modifying, based at least in part on the difference, the token population of at least one bucket of one or more of: (a) the burst-mode token bucket set or (b) the burst-mode token bucket set.

11. The method as recited in claim 6, wherein the work target comprises at least a portion of a storage object, wherein the work request comprises at least one of (a) a read operation or (b) a write operation.

12. The method as recited in claim 6, wherein the burst-mode bucket set comprises a read-burst token bucket used for admission control for work requests comprising read operations, and a write-burst token bucket used for admission control for work requests comprising write operations.

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

monitoring a work request arrival rate over a time period; and

modifying a maximum token population limit of at least one bucket of the burst-mode token bucket set based at least in part on a result of monitoring the work request arrival rate.

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

providing, to a client, an indication of a pricing policy associated with burst-mode operations at the work target; and

modifying a maximum token population limit of at least one bucket of the burst-mode token bucket set based at least in part on a burst-mode capacity increase request received from a client in accordance with the pricing policy.

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

receive a work request directed to a work target; and

in response to a determination, based at least in part on a first threshold criterion, that the work target is in a burst mode of operation,

determine that the token population of at least one bucket of a burst-mode token bucket set associated with the work target meets a second threshold criterion; and

accept the work request for execution.

16. The non-transitory computer-accessible storage medium as recited in claim 15, wherein the program instructions when executed on one or more processors

assign a provisioned throughput capacity to the work target, indicative of a maximum rate of work operations to be performed at the work target during a normal mode of operation;

refill a normal-mode token bucket associated with the work target at a rate based at least in part on the provisioned throughput capacity, wherein the first threshold criterion is based at least in part on a token population of the normal-mode token bucket; and

refill at least one bucket of the burst-mode token bucket set at another rate based at least on the provisioned throughput capacity.

17. The non-transitory computer-accessible storage medium as recited in claim 15, wherein the instructions when executed on the one or more processors:

in response to a determination that, at the end of a particular time period, one or more work tokens remain in a normal-mode token bucket associated with the work target, increase the token population of at least one bucket of the burst-mode token bucket set by an amount based at least in part on the number of work tokens that remain in the normal-mode token bucket.

18. The non-transitory computer-accessible storage medium as recited in claim 15, wherein the burst mode bucket set comprises a local-burst-limit token bucket whose maximum token population is indicative of a maximum burst rate supported for work operations at the work target, and a shared-resource token bucket whose token population is indicative of an available throughput capacity of a resource shared by a plurality of work objects, wherein, in response to a determination that the first threshold criterion has been met, the instructions when executed on the one or more processors:

accept the work request for execution in response to determining that the token population of the local-burst-limit token bucket exceeds zero and the token population of the shared-resource token bucket exceeds zero.

19. The non-transitory computer-accessible storage medium as recited in claim 15, wherein the instructions when executed on the one or more processors:

in response to the determination that the token population of the at least one bucket of the burst-mode token bucket set meets the second threshold criterion, consume at least one token from the at least one bucket in accordance with a token consumption policy.

20. The non-transitory computer-accessible storage medium as recited in claim 15, wherein the work target comprises at least a portion of a partition of a database table, wherein the work request comprises at least one of (a) a read operation or (b) a write operation.

21. A system, comprising:

one or more computing devices configured to:

instantiate one or more token buckets associated with workload management at a work target, wherein the work target is configured to operate in a plurality of modes of operation with respective sets of admission control parameters for accepting work requests for execution, and wherein a decision to accept a work request is made based at least in part on a token population of at least one bucket of the one or more token buckets;

determine a token pricing policy to be applied to a particular token bucket of the one or more token buckets, wherein the token pricing policy indicates (a) one or more applicability criteria for the token pricing policy, including an identification of at least a burst mode of operation of the plurality of modes of operation during which the token pricing policy is to be applied, and (b) a pricing amount, to be charged to a client, for a change to a token population of the particular token bucket;

record, over a time period, one or more changes to the token population of the particular token bucket, wherein the one or more changes occur during a burst mode of operation of the work target in which the one or more applicability criteria are met; and

generate a billing amount to be charged to the client based at least in part on the one or more changes recorded over the time period.

22. The system as recited in claim 21, wherein the one or more token buckets comprise a normal-mode bucket whose token population is used to determine a mode of operation of the work target, and one or more burst-mode buckets whose token population is used to determine whether to accept a work request during the burst mode of operation, wherein particular token bucket comprises a burst-mode bucket of the one or more burst-mode buckets.

23. The system as recited in claim 22, wherein the one or more burst-mode buckets comprise a local-burst-limit bucket whose token population is indicative of the available throughput capacity at the work target, and a shared-resource capacity bucket whose token population is indicative of available throughput capacity at a resource shared by a plurality of work targets, wherein the particular bucket comprises the shared-capacity bucket, and wherein a particular applicability criterion of the one or more applicability criteria indicates that the pricing policy is to apply in response to a determination that the token population of the local-burst-limit bucket is no greater than a specified threshold.

24. The system as recited in claim 21, wherein the change to the token population of the particular token bucket comprises a reduction in the token population as a result of one or more tokens being consumed to accept a received work request for execution.

25. The system as recited in claim 21, wherein the change to the token population of the particular token bucket comprises a reduction in the token population as a result of one or more tokens being transferred to a different bucket in response to a token sale operation implemented using a token marketplace.

26. A method, comprising:

performing, by one or more computing devices,

instantiating one or more token buckets associated with workload management at a work target, wherein a decision to accept a work request directed to the work target is made based at least in part on a token population of at least one bucket of the one or more token buckets;

determining a token pricing policy to be applied to a particular token bucket of the one or more token buckets during a burst mode of operation of the work target, wherein the token pricing policy indicates a pricing amount to be charged to a client for a change to a token population of the particular token bucket;

recording, over a time period, one or more changes to the token population of the particular token bucket, wherein the one or more changes are implemented during the burst mode of operation; and

generating a billing amount to be charged to the client based at least in part on the one or more changes recorded over the time period.

27. The method as recited in claim 26, wherein the one or more token buckets comprise a normal-mode bucket whose token population is used to determine a mode of operation of the work target, and one or more burst-mode buckets whose token population is used to determine whether to accept a work request during the burst mode of operation, wherein particular token bucket comprises a burst-mode bucket of the one or more burst-mode buckets.

28. The method as recited in claim 27, wherein the one or more burst-mode buckets comprise a local-burst-limit bucket whose token population is indicative of the throughput capacity at the work target, and a shared-resource capacity bucket whose token population is indicative of available throughput capacity at a resource shared by a plurality of work targets, wherein the particular bucket comprises the shared-capacity bucket, and wherein, in accordance with the pricing policy, the method further comprises performing, by the one or more computing devices:

in response to determining that the token population of the local-burst-limit bucket meets a threshold criterion, applying the pricing policy to a particular change of token population of the shared-resource capacity bucket.

29. The method as recited in claim 26, wherein the change to the token population of the particular token bucket comprises a reduction in the token population as a result of one or more tokens being consumed to accept a received work request for execution.

30. The method as recited in claim 26, wherein the change to the token population of the particular token bucket comprises a reduction in the token population as a result of one or more tokens being transferred to a different bucket in response to a token sale operation implemented using a token marketplace.

31. The method as recited in claim 30, further comprising performing, by the one or more computing devices:

implementing one or more programmatic interfaces enabling a particular client to indicate, via the token marketplace, that tokens of the particular token bucket are available for transfer to another token bucket.

32. The method as recited in claim 31, wherein the token pricing policy indicates that an amount to be charged for a transfer of a token to a different bucket from the particular bucket is to be determined based at least in part on one or more auction bids received for the transfer.

33. The method as recited in claim 26, further comprising performing, by the one or more computing devices:

implementing one or more programmatic interfaces enabling a client to select a token pricing policy from among a plurality of token pricing policies.

34. The method as recited in claim 26, wherein the pricing policy indicates a client-specified budget limit, further comprising performing, by the one or more computing devices:

determining to implement a change to the population of the particular token bucket based at least in part on the client-specified budget limit.

35. The method as recited in claim 26, wherein the work target comprises a first partition of a storage object, wherein the storage object includes at least the first partition and a second partition, wherein a particular change of the one or more changes to the token population of the particular bucket is a result of a transfer of a token from the particular bucket to a different bucket used for workload management of the second partition.

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

determine a token pricing policy to be applied to a particular token bucket of one or more token buckets used for admission control decisions at a work target, wherein the token pricing policy is to be applied during a burst mode of operation of the work target, and wherein the token pricing policy indicates a pricing amount to be charged to a client for a change to a token population of the particular token bucket;

record, over a time period, one or more changes to the token population of the particular token bucket during the burst mode of operation; and

generate a billing amount to be charged to the client based at least in part on the one or more changes recorded over the time period.

37. The non-transitory computer-accessible storage medium as recited in claim 36, wherein the one or more token buckets comprise a normal-mode bucket whose token population is used to determine whether to accept a work request during a normal mode of operation of the work target, and one or more burst-mode buckets whose token population is used to determine whether to accept a work request during the burst mode of operation, wherein the particular token bucket comprises a burst-mode bucket of the one or more burst-mode buckets, and wherein the instructions when executed on the one or more processors:

determine a different token pricing policy to be applied to token population changes of the normal-mode bucket.

38. The non-transitory computer-accessible storage medium as recited in claim 37, wherein the one or more burst-mode buckets comprise a local-burst-limit bucket whose token population is indicative of available throughput capacity at the work target, and a shared-resource capacity bucket whose token population is indicative of available throughput capacity at a resource shared by a plurality of work targets, wherein the particular bucket comprises the shared-capacity bucket, and wherein, in accordance with the pricing policy, the instructions when executed on the one or more processors:

in response to a determination that the token population of the local-burst-limit bucket meets a threshold criterion, apply the pricing policy to a particular change of token population of the shared-resource capacity bucket.

39. The non-transitory computer-accessible storage medium as recited in claim 36, wherein the change to the token population of the particular token bucket comprises a reduction in the token population as a result of one or more tokens being transferred to a different bucket in response to a token sale operation implemented using a token marketplace.

40. The non-transitory computer-accessible storage medium as recited in claim 36, wherein the particular bucket is configured with a token refill rate and a maximum token population limit, wherein the pricing policy comprises an indication of a different amount to be charged to the client for a change to at least one of: (a) the token refill rate, or (b) the maximum token population limit.

41. A system, comprising:

one or more computing devices configured to:

configure a first work target and a second work target to utilize a shared resource in response to work requests accepted for execution, wherein the first work target has a first provisioned throughput rate, the second work target has a second provisioned throughput rate, and the shared resource has a throughput limit;

configure the first work target and the second work target with respective token bucket sets for admission control of work requests, wherein each token bucket set comprises one or more buckets whose token population is used to determine whether to accept a work request for execution;

determine (a) an arrival rate ratio indicative of relative rates at which work requests are received at the first and second work targets during a first time interval, (b) a provisioned throughput ratio based at least in part on the first and second provisioned throughput rates, and (c) a combined number of tokens to be distributed among the bucket sets of the first and second work targets for admission control during a second time interval, wherein the combined number is based at least in part on the throughput limit of the shared resource;

add a particular number of tokens, no greater than the combined number, to a particular bucket of the first work target based at least in part on the arrival rate ratio and the provisioned throughput ratio; and

accept a particular work request directed to the first work target for execution during the second time interval based at least in part on the token population of the particular bucket of the first work target.

42. The system as recited in claim 41, wherein the combined number of tokens is determined at least in part by subtracting, from the throughput limit of the shared resource, the sum of the first and second provisioned throughput rates.

43. The system as recited in claim 41, wherein the bucket set of the first work target comprises a normal-mode bucket whose token population is examined for admission control during a normal mode of operation in which the work request arrival rate is less than a threshold, and a burst-mode bucket whose token population is examined for admission control during a burst mode of operation in which the work request arrival rate is not less than the threshold, and wherein the particular bucket comprises the burst-mode bucket.

44. The system as recited in claim 41, wherein the first work target comprises at least a portion of a first storage object managed by a network-accessible service, wherein the second work target comprises at least a portion of a second storage object managed by the network-accessible service, and wherein the shared resource comprises a storage device at which the first work target and the second work target are stored.

45. The system as recited in claim 41, wherein the particular work request comprises one or more of: (a) a read operation, or (b) a write operation.

46. A method, comprising:

performing, by one or more computing devices:

configuring a first work target and a second work target to utilize a shared resource in response to work requests accepted for execution, wherein the first work target has a first provisioned throughput rate, and the second work target has a second provisioned throughput rate;

configuring the first work target and the second work target with respective token bucket sets for admission control of work requests, wherein each token bucket set comprises one or more buckets whose token population is used to determine whether to accept a work request for execution;

determining (a) a first metric indicative of work request arrival rates at the first and second work targets during a first time interval, and (b) a second metric associated with the first and second provisioned throughput rates;

adding a particular number of tokens to a particular bucket of the first work target based at least in part on the first metric, the second metric, and a throughput limit of the shared resource; and

accepting a particular work request directed to the first work target for execution during a second time interval based at least in part on the token population of the particular bucket of the first work target.

47. The method as recited in claim 46, further comprising performing, by the one or more computing devices:

determining a combined number of tokens to be distributed among the bucket sets of the first and second work targets based at least in part on the throughput limit of the shared resource, wherein the particular number of tokens is no greater than the combined number.

48. The method as recited in claim 47, wherein the combined number of tokens is determined at least in part by subtracting, from the throughput limit of the shared resource, the sum of the first and second provisioned throughput rates.

49. The method as recited in claim 46, wherein the bucket set of the first work target comprises a normal-mode bucket whose token population is examined for admission control during a normal mode of operation in which the work request arrival rate is less than a threshold, and a burst-mode bucket whose token population is examined for admission control during a burst mode of operation in which the work request arrival rate is not less than the threshold, and wherein the particular bucket comprises the burst-mode bucket.

50. The method as recited in claim 46, wherein the first work target comprises at least a portion of a first storage object managed by a network-accessible service, wherein the second work target comprises at least a portion of a second storage object managed by the network-accessible

service, and wherein the shared resource comprises a storage device at which the first work target and the second work target are stored.

51. The method as recited in claim 46, wherein the first work target comprises at least a portion of a first table partition managed by a network-accessible multi-tenant database service, and wherein the second work target comprises at least a portion of a second table managed by the network-accessible multi-tenant database service.

52. The method as recited in claim 46, wherein the shared resource comprises a data structure of a software module.

53. The method as recited in claim 46, wherein the particular work request comprises one or more of: (a) a read operation, or (b) a write operation.

54. The method as recited in claim 46, further comprising performing, by the one or more computing devices:

determining the particular number of tokens based at least in part on a particular function of the first metric, the second metric, and the throughput limit;

monitoring one or more additional metrics associated with the first and second work targets; and

modifying the particular function based at least in part on the one or more additional metrics.

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

configure a first work target and a second work target to utilize a shared resource in response to work requests accepted for execution, wherein the first work target has a first provisioned throughput rate, and the second work target has a second provisioned throughput rate;

configure the first work target and the second work target with respective token bucket sets for admission control of work requests, wherein each token bucket set comprises one or more buckets whose token population is used to determine whether to accept a work request for execution;

determine (a) a first metric indicative of work request arrival rates at the first and second work targets during a first time interval, and (b) a a second metric associated with the first and second provisioned throughput rates;

add a particular number of tokens to a particular bucket of the first work target based at least in part on the first metric, the second metric, and a throughput limit of the shared resource; and

accept a particular work request directed to the first work target for execution during a second time interval based at least in part on the token population of the particular bucket of the first work target.

56. The non-transitory computer-accessible storage medium as recited in claim 55, wherein the instructions, when executed on the one or more processors:

determine a combined number of tokens to be distributed among the bucket sets of the first and second work targets based at least in part on the throughput limit of the shared resource, wherein the particular number of tokens is no greater than the combined number.

57. The non-transitory computer-accessible storage medium as recited in claim 55, wherein the bucket set of the first work target comprises a normal-mode bucket whose token population is examined for admission control during a normal mode of operation in which the work request arrival rate is less than a threshold, and a burst-mode bucket whose token population is examined for admission control during a burst mode of operation in which the work request arrival rate is not less than the threshold, and wherein the particular bucket comprises the burst-mode bucket.

58. The non-transitory computer-accessible storage medium as recited in claim 55, wherein the first work target comprises at least a portion of a first storage object managed by a network-accessible service, wherein the second work target comprises at least a portion of a second storage object managed by the network-accessible service, and wherein the shared resource comprises a storage device at which the first work target and the second work target are stored.

59. The non-transitory computer-accessible storage medium as recited in claim 55, wherein the shared resource comprises a logical resource implemented in an operating system.

60. The non-transitory computer-accessible storage medium as recited in claim 55, wherein the particular work request comprises one or more of: (a) a read operation, or (b) a write operation.

61. A system, comprising:

one or more computing devices configured to:

configure a plurality of work targets with respective sets of token buckets for admission control of work requests, including a first work target with a first set of token buckets, wherein a decision to accept a work request for execution at the first work target is to be based at least in part on a token population of one or more token buckets of the first set;

in response to a determination that a token sharing evaluation criterion has been met at the first work target,

identify a second work target of the plurality of work targets with which to exchange token population information from the first work target, wherein the second work target is configured with a second set of token buckets;

determine, based at least in part on an analysis of respective token populations of a first bucket in the first set and a second bucket in the second set, whether to initiate a token transfer between the first and second buckets; and

in response to a determination to initiate a token transfer,

increase the token population of one bucket of the first and second buckets by a particular number of tokens, and decrease the token population of another bucket of the first and second buckets by the particular number of tokens, in accordance with a transfer direction identified for the token transfer; and

in response to receiving a work request directed to the first work target, accept the work request for execution based at least in part on the changed token population of the first bucket.

62. The system as recited in claim 61, wherein the plurality of work targets comprise a first collection of one or more storage objects created at a network-accessible storage service on behalf of a first client, and a second collection of one or more storage objects created at the network-accessible storage service on behalf of a second client, wherein the first work target comprises at least a portion of a storage object of the first collection, and wherein, to identify the second work target, the one or more computing devices are configured to select the second work target from the first collection.

63. The system as recited in claim 61, wherein the first work target comprises at least a portion of a base database table instantiated on behalf of a particular client of a network-accessible database service, and the second work target comprises at least a portion of a derived database table instantiated at the network-accessible database service to implement a secondary index on the base database table.

64. The system as recited in claim 61, wherein the first set of token buckets comprises a normal-mode bucket used for admission control during a normal mode of operation of the first work target, and one or more burst-mode token buckets used for admission control during a burst mode of operation at the first work target, wherein the one or more burst-mode token buckets include the first bucket.

65. The system as recited in claim 61, wherein the second work target is selected from among the plurality of work targets using a random selection policy in accordance with a gossip protocol.

66. A method, comprising:

performing, by one or more computing devices:

configuring a plurality of work targets with respective token bucket sets for admission control of work requests, including a first work target with a first

token bucket set, wherein a decision to accept a work request for execution at the first work target is to be based at least in part on a token population of one or more token buckets of the first token bucket set;

in response to determining that a token sharing evaluation criterion has been met at the first work target,

initiating a token transfer in a particular direction between a first bucket of the first token bucket set and a second bucket of a second token bucket set associated with a second work target of the plurality of work targets; and

accepting, in response to receiving a work request directed to the first work target, the work request for execution based at least in part on the token population of the first bucket subsequent to the token transfer.

67. The method as recited in claim 66, wherein the plurality of work targets comprise a first collection of one or more storage objects created at a network-accessible storage service on behalf of a first client, and a second collection of one or more storage objects created at the network-accessible storage service on behalf of a second client, wherein the first work target comprises at least a portion of a storage object of the first collection, further comprising performing, by the one or more computing devices:

selecting the second work target from the first collection.

68. The method as recited in claim 66, wherein the first work target comprises at least a portion of a base database table instantiated on behalf of a particular client of a network-accessible database service, and the second work target comprises at least a portion of a derived database table instantiated at the network-accessible database service to implement a secondary index on the base database table.

69. The method as recited in claim 66, wherein the first token bucket set comprises a normal-mode bucket used for admission control during a normal mode of operation of the first work target, and one or more burst-mode token buckets used for admission control during a burst mode of operation at the first work target, wherein the one or more burst-mode token buckets include the first bucket.

70. The method as recited in claim 66, further comprising performing, by the one or more computing devices:

selecting the second work target using a random selection policy.

71. The method as recited in claim 66, further comprising performing, by the one or more computing devices:

selecting the second work target in accordance with a gossip protocol.

72. The method as recited in claim 61, further comprising performing, by the one or more computing devices:

determining whether the token sharing evaluation criterion has been met based at least in part on one or more of (a) an amount of time since a previous evaluation was initiated at the first work target, (b) an indication that the token population of the first token bucket has fallen below a first threshold value, (c) an indication that the token population of the first token bucket has risen above a second threshold value, or (d) an indication that a rate of rejection of work requests at the first work target is above a particular threshold value.

73. The method as recited in claim 66, wherein the first work target comprises a first replica of a plurality of replicas of a first logical partition of a storage object, wherein each replica of the plurality of replicas is assigned a role comprising one of (a) a master role and (b) a slave role, wherein admission control for a work request that includes a write operation directed to the first logical partition is performed at the replica to which the master role is assigned, wherein the first replica is assigned a first role with respect to the first logical partition, further comprising, performing by the one or more computing devices:

selecting a particular replica of a second logical partition of the storage object as the second work target, wherein the particular replica is assigned the same role with respect to the second logical partition as the role of the first replica with respect to the first logical partition.

74. The method as recited in claim 66, wherein said initiating the token transfer in the particular direction between the first bucket and the second bucket is based at least in part on a determination that, within a particular time window, no other token transfer between the first work target and the second work target was implemented.

75. The method as recited in claim 66, wherein the particular direction is from the token bucket with the greater token population among the first and second buckets, to the token bucket with the smaller token population among the first and second buckets, wherein the method comprises performing, by the one or more computing devices:

determining a number of tokens to be transferred as a function of the difference between the greater and smaller token populations.

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

determine a number of tokens to be transferred between a first bucket in a first token bucket set and a second bucket in a second token bucket set, wherein the first token bucket set is used for admission control of work requests directed at a first work target of a plurality of work targets, and wherein the second token bucket set is used for admission control at a second work target of the plurality of work targets;

initiate a transfer of the number of tokens in a particular direction between the first bucket and the second bucket; and

accept, in response to receiving a work request directed to the first work target, the work request for execution based at least in part on the token population of the first bucket subsequent to the transfer.

77. The non-transitory computer-accessible storage medium as recited in claim 76, wherein the plurality of work targets comprise a first collection of one or more storage objects created at a network-accessible storage service on behalf of a first client, and a second collection of one or more storage objects created at the network-accessible storage service on behalf of a second client, wherein the first work target comprises at least a portion of a storage object of the first collection, and wherein the instructions when executed at the one or more processors:

select the second work target from the first collection.

78. The non-transitory computer-accessible storage medium as recited in claim 76, wherein the first work target comprises at least a portion of a base database table instantiated on behalf of a particular client of a network-accessible database service, and the second work target comprises at least a portion of a derived database table instantiated at the network-accessible database service to implement a secondary index on the base database table.

79. The non-transitory computer-accessible storage medium as recited in claim 76, wherein the first token bucket set comprises a normal-mode bucket used for admission control during a normal mode of operation of the first work target, and one or more burst-mode token buckets used for admission control during a burst mode of operation at the first work target, wherein the one or more burst-mode token buckets include the first bucket.

80. The non-transitory computer-accessible storage medium as recited in claim 76, wherein the second work target is selected from among the plurality of work targets using a random selection policy in accordance with a gossip protocol.

81. A system, comprising:

one or more computing devices configured to:

determine a peak burst rate at which work requests directed to a work target are to be accepted for execution during a burst mode of operation, and a peak burst window size indicative of a maximum duration for which work requests at the peak burst rate are to be accepted;

determine a sustained burst rate smaller than the peak burst rate, and a sustained burst window size greater than the peak burst window size, wherein the sustained burst window size is indicative of a maximum duration for which work requests directed to the work target at the sustained burst rate are to be accepted during the burst mode of operation;

configure a first maximum token population, a first initial token population, and a first token consumption policy for a peak-burst token bucket to be used for admission control of work requests directed at the work target, wherein the

first maximum token population is based at least in part on the peak burst rate;

configure a second maximum token population, a second initial token population, and a second token consumption policy for a sustained-burst token bucket to be used for admission control of work requests directed at the work target, wherein the second maximum token population is based at least in part on the sustained burst rate; and

in response to receiving a work request directed at the work target during a burst mode of operation, determine to accept the work request for execution based at least in part on the respective token populations of the peak-burst token bucket and the sustained-burst token bucket.

82. The system as recited in claim 81, wherein the one or more computing devices are further configured to:

in response to a determination to accept the work request for execution, consume at least one token from the peak-burst token bucket in accordance with the first token consumption policy, and consume at least one token from the sustained-burst token bucket in accordance with the second token consumption policy.

83. The system as recited in claim 81, wherein the one or more computing devices are further configured to:

in response to receiving a particular work request directed at the work target, determine whether the work target is in a burst mode of operation based at least in part on a token population of a provisioned-capacity token bucket.

84. The system as recited in claim 81, wherein said peak burst rate is based at least in part on a provisioned throughput capacity of the work target, and wherein said sustained burst rate is based at least in part on the provisioned throughput capacity.

85. The system as recited in claim 81, wherein the one or more computing devices are further configured to:

determine a first refill rate for the peak-burst token bucket based at least in part on a provisioned throughput capacity of the work target; and

determine a second refill rate for the sustained-burst token bucket based at least in part on a provisioned throughput capacity of the work target.

86. A method, comprising:

performing, by one or more computing devices:

determining a peak burst rate at which work requests directed to a work target are to be accepted during a burst mode of operation;

determining a sustained burst rate at which work requests directed to the work target are to be accepted for up to a particular duration during the burst mode of operation;

configuring a first maximum token population for a peak-burst token bucket to be used for admission control of work requests directed at the work target, and a second maximum token population for a sustained-burst token bucket to be used for admission control of work requests directed at the work target, wherein the first maximum token population is based at least in part on the peak burst rate and the second maximum token population is based at least in part on the sustained burst rate; and

in response to receiving a work request directed at the work target during a burst mode of operation, determining to accept the work request for execution based at least in part on a token population of at least one of: the peak-burst token bucket and the sustained-burst token bucket.

87. The method as recited in claim 86, further comprising performing, by the one or more computing devices:

determining a peak burst time window indicative of a maximum duration for which work requests at the peak burst rate are to be accepted for execution,

wherein said first maximum token population is based at least in part on the peak burst time window.

88. The method as recited in claim 86, wherein said second maximum token population is based at least in part on a sustained burst time window indicative of the particular duration.

89. The method as recited in claim 86, further comprising performing, by the one or more computing devices:

in response to determining to accept the work request for execution, consuming at least one token from the peak-burst token bucket, and consuming at least one token from the sustained-burst token bucket.

90. The method as recited in claim 86, further comprising performing, by the one or more computing devices:

in response to receiving a particular work request directed at the work target, determining whether the work target is in a burst mode of operation based at least in part on a token population of a provisioned-capacity token bucket.

91. The method as recited in claim 86, wherein the peak burst rate is based at least in part on a provisioned throughput capacity of the work target, and wherein the sustained burst rate is based at least in part on the provisioned throughput capacity.

92. The method as recited in claim 86, further comprising performing, by the one or more computing devices:

determining a first refill rate for the peak-burst token bucket based at least in part on a provisioned throughput capacity of the work target; and

determining a different refill rate for the sustained-burst token bucket based at least in part on a provisioned throughput capacity of the work target.

93. The method as recited in claim 86, further comprising performing, by the one or more computing devices:

in response to receiving the work request directed at the work target during the burst mode of operation, determining to accept the work request for execution based at least in part on a token population of a shared-resource token bucket representing available throughput capacity at a resource shared by a plurality of work targets.

94. The method as recited in claim 86, wherein the work target comprises at least a portion of a storage object, wherein the work request comprises a request for at least one of: (a) a read operation on the storage object and (b) a write operation on the storage object.

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

determine a peak burst rate at which work requests directed to a work target are to be accepted;

determine a sustained burst rate at which work requests directed to the work target are to be accepted for up to a particular duration;

configure a first maximum token population for a peak-burst token bucket to be used for admission control of work requests directed at the work target and a second maximum token population for a sustained-burst token bucket to be used for admission control of work requests directed at the work target, wherein the first maximum token population is based at least in part on the peak burst rate and the second maximum token population is based at least in part on the sustained burst rate; and

in response to receiving a work request directed at the work target, determine to accept the work request for execution based at least in part on a token population of at least one of: the peak-burst token bucket and the sustained-burst token bucket.

96. The non-transitory computer-accessible storage medium as recited in claim 95, wherein the instructions when executed on the one or more processors:

determine a peak burst time window indicative of a maximum duration for which work requests at the peak burst rate are to be accepted for execution;

wherein said first maximum token population is based at least in part on the peak burst time window.

97. The non-transitory computer-accessible storage medium as recited in claim 95, wherein said second maximum token population is based at least in part on a sustained burst time window indicative of the particular duration.

98. The non-transitory computer-accessible storage medium as recited in claim 95, wherein the instructions when executed on the one or more processors:

in response to a determination to accept the work request for execution, consume at least one token from the peak-burst token bucket, and consume at least one token from the sustained-burst token bucket.

99. The non-transitory computer-accessible storage medium as recited in claim 95, wherein the instructions when executed on the one or more processors:

in response to receiving a particular work request directed at the work target, determine, based at least in part on a token population of a provisioned-capacity token bucket associated with the work target, whether an admission control decision for the particular work request is to be based on the token population of one or more of: (a) the peak-burst token bucket and (b) the sustained-burst token bucket.

100. The non-transitory computer-accessible storage medium as recited in claim 95, wherein the instructions when executed on the one or more processors:

determine a first refill rate for the peak-burst token bucket based at least in part on a provisioned throughput capacity of the work target; and

determine a second refill rate for the sustained-burst token bucket based at least in part on a provisioned throughput capacity of the work target.