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 the respective 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.