

CLAIMS

WHAT IS CLAIMED IS:

1. A system for managing shared computing resources, the system comprising:
one or more processors; and
a computer-readable memory storing executable instructions that, when executed by the one or more processors, configure the system to:

receive a request to initialize a first instance of a first virtual machine, the first instance associated with a first expected usage amount of a computing resource and a first reserved amount of the computing resource, wherein the first expected usage amount is based at least in part on operating metrics determined from previously running at least an instance of a similar virtual machine, and wherein the first reserved amount corresponds to an amount of the computing resource reserved for the first instance by a customer associated with the first instance;

select a computing device of a plurality of computing devices configured to provide the computing resource, wherein a second instance of a second virtual machine is executing on the computing device, wherein the second instance is associated with a second expected usage amount and a second reserved amount of the computing resource, and wherein the selected computing device is selected based at least partly on:

calculating a total reserved amount, the total reserved amount comprising a sum of the first reserved amount and the second reserved amount;

determining that the total reserved amount is greater than an amount of the computing resource available on the computing device;

in response to the determination that the total reserved amount is greater than the amount of the computing resource available on the computing device, calculating a total expected usage amount, the total expected usage amount comprising a sum of the first expected usage amount and second expected usage amount; and

determining that the total expected usage amount is less than the amount of the computing resource available on the computing device; and cause, at least in part, the first instance to be initialized on the selected computing device, wherein the first instance and the second instance share the available amount of the computing resource.

2. The system of Claim 1, wherein the computing resource comprises one of central processing unit (CPU) capacity, memory space, network bandwidth, hard disk space, or electrical power.

3. The system of Claim 1, wherein the first instance and the second instance are associated with different customers.

4. The system of Claim 1, wherein the executable instructions further configure the system to:

cause, at least in part, a third instance of a third virtual machine to be initialized on the selected computing device,

wherein the third instance is associated with a third expected usage amount and a third reserved amount of the computing resource,

wherein the amount of the computing resource available on the computing device is further greater than a sum of the first expected usage amount, the second expected usage amount, and the third expected usage amount,

and wherein the first instance, the second instance, and the third instance share the available amount of the computing resource.

5. The system of Claim 1, wherein the executable instructions further configure the system to:

obtain a first measurement of usage of the computing resource by the first instance and a second measurement of usage of the computing resource by the second instance; and

in response to determining that the first measurement or the second measurement exceeds a threshold, transfer the first instance or the second instance to a second computing device.

6. A computer-implemented method for managing shared computing resources, the computer-implemented method comprising:

receiving, by a data center management system comprising one or more computing devices, a request to initialize a first software workload associated with a first operating constraint and a first reserved resource amount, wherein the first operating constraint is based at least in part on a first plurality of historical operating metrics regarding operation of the first software workload on a host computing device, and wherein the first reserved resource amount comprises an amount of a computing resource reserved for use by the first software workload;

in response to the request, identifying a first host computing device of a plurality of host computing devices based at least on the first operating constraint and one or more operating characteristics associated with the first host computing device,

wherein the first host computing device comprises a second software workload associated with a second operating constraint and a second reserved resource amount, and

wherein the identifying comprises:

calculating a total reserved resource amount, the total reserved resource amount comprising a sum of the first reserved resource amount and the second reserved resource amount;

determining that an available amount of the computing resource is less than the total reserved resource amount; and

in response to the determination that the available amount of the computing resource is less than the total reserved resource amount, determining that the one or more operating characteristics of the first host computing device are sufficient to satisfy both the first operating constraint and the second operating constraint during operation of the first software workload and the second software workload; and

causing, at least in part, the first software workload to be placed on the first host computing device.

7. The computer-implemented method of Claim 6, wherein the first software workload comprises a storage node, and wherein the identifying further comprises determining that placing the storage node on the first host computing device satisfies an operating constraint associated with a virtual machine that will use the storage node.

8. The computer-implemented method of Claim 6, wherein the first operating constraint relates to a first expected usage amount of the computing resource, wherein the second operating constraint relates to a second expected usage amount of the computing resource, and wherein the first software workload and the second software workload are enabled to concurrently consume the first expected usage amount and the second expected usage amount, respectively, of the computing resource.

9. The computer-implemented method of Claim 8, wherein an operating characteristic of the one or more operating characteristics comprises the available amount of the computing resource, the computer-implemented method further comprising:

preventing, at least in part, the first software workload and the second software workload from aggregately consuming more than the available amount of the computing resource.

10. The computer-implemented method of Claim 6 further comprising:

causing, at least in part, a third software workload to be initialized on the first host computing device, wherein the third software workload is associated with a third reserved resource amount and a third operating constraint;

wherein the method further comprises calculating an updated total reserved resource amount, the updated total reserved resource corresponding to a sum of the first reserved resource amount, the second reserved resource amount, and the third reserved resource amount; and

wherein the identifying further comprises:

determining that the available amount of the computing resource is less than the updated total reserved resource amount; and

in response to the determination that the available amount of the computing resource is less than the updated total reserved resource amount, determining that the one or more operating characteristics of the first host

computing device are sufficient to satisfy each of the first operating constraint, the second operating constraint, and the third operating constraint during operation of the first software workload, the second software workload, and the third software workload.

11. The computer-implemented method of Claim 6 further comprising:
 - obtaining one or more operating metrics regarding operation of the first software workload on the first host computing device;
 - determining, based at least in part on the one or more operating metrics, that the one or more operating characteristics of the first host computing device are no longer sufficient to satisfy the first operating constraint; and
 - in response to the determination, transferring the first software workload or the second software workload to a second host computing device.
12. The computer-implemented method of Claim 6 further comprising:
 - obtaining one or more operating metrics regarding operation of the second software workload on the first host computing device;
 - determining, based at least in part on the one or more operating metrics, that the one or more operating characteristics of the first host computing device are no longer sufficient to satisfy the second operating constraint; and
 - in response to the determination, transferring the first software workload or the second software workload to a second host computing device.
13. The computer-implemented method of Claim 6, further comprising:
 - obtaining one or more operating metrics regarding the one or more operating characteristics of the first host computing device;
 - determining, based at least in part on the one or more operating metrics, that the one or more operating characteristics of the first host computing device are no longer sufficient to satisfy the first operating constraint or the second operating constraint; and
 - in response to the determination, transferring the first software workload or the second software workload to a second host computing device.
14. A system for managing shared computing resources, the system comprising:
 - one or more processors;

a computer-readable memory storing executable instructions that, when executed by the one or more processors, configure the system to:

- receive a request to initialize a first computer-executable program associated with a first operating constraint and a first reserved resource amount, wherein the first operating constraint is based at least in part on a first plurality of historical operating metrics regarding operation of the first computer-executable program on a host computing device, and wherein the first reserved resource amount comprises an amount of a computing resource reserved for use by the first computer-executable program; and

- identify a first host computing device of a plurality of host computing devices, the first host computing device associated with one or more operating characteristics,

- wherein the first host computing device comprises a second computer-executable program associated with a second operating constraint and a second reserved resource amount, and

- wherein the first host computing device is identified based at least in part on:

- calculating a total reserved resource amount, the total reserved resource amount comprising the sum of the first reserved resource amount and the second reserved resource amount;

- determining that an available amount of the computing resource is less than the total reserved resource amount; and

- in response to the determination that the available amount of the computing resource is less than the total reserved resource amount, determining that the one or more operating characteristics satisfy both the first operating constraint and the second operating constraint; and

- cause, at least in part, the first computer-executable program to execute on the first host computing device.

15. The system of Claim 14, wherein the first operating constraint relates to a first expected usage amount of the computing resource, wherein the second operating constraint relates to a second expected usage amount of the computing resource, and wherein the first computer-executable program and the second computer-executable program are enabled to concurrently consume the first expected usage amount and the second expected usage amount, respectively, of the computing resource.

16. The system of Claim 15, wherein the executable instructions further configure the system to prevent, at least in part, the first computer-executable program and the second computer-executable program from aggregately consuming more than the available amount of the computing resource.

17. The system of Claim 14, wherein the executable instructions further configure the system to:

- receive a request to initialize a third computer-executable program associated with a third operating constraint and a third reserved resource amount, wherein the third operating constraint is based at least in part on a second plurality of historical operating metrics regarding operation of the third computer-executable program on a host computing device, and wherein the third reserved resource amount comprises an amount of a third computing resource reserved for use by the third computer-executable program;

- cause, at least in part, the third computer-executable program to be initialized on the first host computing device; and

- calculate an updated total reserved resource amount, the updated total reserved resource amount corresponding to a sum of the first reserved resource amount, the second reserved resource amount, and the third reserved resource amount;

- wherein the identifying further comprises:

- determining that the available amount of the computing resource is less than the updated total reserved resource amount; and

- in response to the determination that the available amount of the computing resource is less than the updated total reserved resource amount, determining that the one or more operating characteristics of the first host computing device are sufficient to satisfy each of the first operating constraint, the

second operating constraint, and a third operating constraint during operation of the first computer-executable program, the second computer-executable program, and the third computer-executable program.

18. The system of Claim 14, wherein the executable instructions further configure the system to:

obtain one or more operating metrics regarding operation of the first computer-executable program on the first host computing device;

determine, based at least in part on the one or more operating metrics, that the one or more operating characteristics of the first host computing device are no longer sufficient to satisfy the first constraint; and

in response to the determination, transfer the first computer-executable program or the second computer-executable program to a second host computing device.

19. The system of Claim 14, wherein the executable instructions further configure the system to:

obtain one or more operating metrics regarding operation of the second computer-executable program on the first host computing device;

determine, based at least in part on the one or more operating metrics, that the one or more operating characteristics of the first host computing device are no longer sufficient to satisfy the second constraint; and

in response to the determination, transfer the first computer-executable program or the second computer-executable program to a second host computing device.

20. The system of Claim 14, wherein the executable instructions further configure the system to:

obtain one or more operating metrics regarding the one or more operating characteristics of the first host computing device;

determine, based at least in part on the one or more operating metrics, that the one or more operating characteristics of the first host computing device are no longer sufficient to satisfy the first constraint or the second constraint; and

in response to the determination, transfer the first computer-executable program or the second computer-executable program to a second host computing device.