automatic scaling, as demand for or capacity requirements of the application change. The computing resources provided by the computing resource provider may be made available in discrete units, which may be referred to as instances. An instance may be run on directly on a physical server hardware platform, within a virtual machine or some combination of the two. Various types and configurations of instances may be made available, including different sizes of resources executing different operating systems (OS) and/or hypervisors and with various installed software applications, runtimes and the like.

SUMMARY

[0004a] According to a first aspect of the present invention, there is provided a system for allocating computing resources in a provider network, the system comprising: one or more computing devices, the one or more computing devices comprising one or more resource slots, each resource slot configured to host a computing instance; at least one memory having stored therein computer instructions that, upon execution by one or more processors of the system, at least cause the system to: allocate a plurality of resource slots to ~~a plurality of entities according to a first pricing policy~~ a first user according to an allocation policy of the provider network; ~~receive an indication that at least a portion of the resource slots associated with a first entity of the plurality of entities is associated with a second pricing policy and accessible to one or more users identified by the first entity, the second pricing policy being designated by the first entity; and allocate the indicated resource slots to the one or more users according to the second pricing policy.~~ allow the first user to allocate at least a portion of the plurality of resource slots to a second user who is authorized to access and use the allocated portion according to an allocation policy established by the first user, wherein the allocation policy established by the first user is different from the allocation policy of the provider network, wherein the provider network is configured to deny requests to access the allocated portion of the plurality of resource slots not authorized by the allocation policy established by the first user; and maintain the allocation of the plurality of resource slots to the first user according to the allocation policy of the provider network while the allocated portion is accessed by the second user according to the allocation policy established by the first user; wherein the allocation policy established by the first user comprises: allowing the second user to submit a bid including a bid amount and a request for resource slots; comparing the bid amount with a settable minimum amount of the allocation policy established by the first user; and conditionally allocating one or more resource slots to the second user while the bid amount remains above the settable minimum amount.

[0004b] According to a second aspect of the present invention, there is provided a method for allocating computing resources in a provider network, the method comprising: in a computing environment comprising a plurality of computing devices and where capacity of ~~each~~ the computing devices are ~~is~~ logically segmented into one or more resource slots, associating a plurality of resource slots with a first user according to ~~a first set of pricing policies~~an allocation policy of the provider network; and allowing the user to identify one or more of the associated resource slots and ~~identify~~ indicate other users allowed to access the one or more of the associated resource slots according to ~~a second set of pricing policies, the second set of pricing policies being designated by the user.~~ an allocation policy determined by the user, wherein the allocation policy determined by the user is different from the allocation policy of the provider network, and wherein the provider network is configured to deny requests to access the identified resource slots not authorized by the allocation policy determined by the user; wherein the allocation policy determined by the user comprises: allowing the other users to submit bids including a bid amount and a request for resource slots; comparing a given bid amount with a settable minimum amount of the allocation policy determined by the first user; and conditionally allocating one or more resource slots to a given one of the other users while the given bid amount remains above the settable minimum amount.

[0004c] According to a third aspect of the present invention, there is provided a non-transitory computer-readable storage medium having stored thereon computer-readable instructions, the computer-readable instructions comprising instructions that upon execution on one or more computing devices, to at least: in a computing environment comprising a plurality of computing devices and where capacity of ~~each~~ the computing devices ~~is~~ are logically segmented into one or more resource slots, generate a user interface allowing a user to identify one or more resource slots associated with the user and ~~identify~~ select other users allowed to access the one or more resource slots associated with the user according to ~~a pricing policy, the pricing policy being designated by the user; and cause the computing environment to at least allow the other users to bid for the one or more resource slots associated with the user in accordance with the pricing policy.~~an allocation policy determined by the user, wherein a provider network is configured to deny requests to access the identified resource slots not authorized by the allocation policy determined by the user; and cause the computing environment to at least allow the other users to request the one or more resource slots associated with the user in accordance with the allocation policy determined by the user; wherein the allocation policy determined by the user comprises: allowing the other users to submit bids including a bid amount and a request for resource slots; comparing a given bid amount with a settable minimum amount of the allocation policy determined by the first user;

and conditionally allocating one or more resource slots to a given one of the other users while the given bid amount remains above the settable minimum amount.

BRIEF DESCRIPTION OF DRAWINGS

[0005] Throughout the drawings, reference numbers may be reused to indicate correspondence between referenced elements. The drawings are provided to illustrate example embodiments described herein and are not intended to limit the scope of the disclosure:

[0006] FIG. 1 is a diagram illustrating a mechanism for pooling instances in accordance with the present disclosure;

[0007] FIG. 2 is a diagram illustrating an example computing system that may be used in some embodiments;

[0008] FIG. 3 is a diagram illustrating an example computing system that may be used in some embodiments;

[0009] FIG. 4 is a diagram illustrating a system for pooling instances in accordance with the present disclosure;

[0010] FIG. 5 is a diagram illustrating a mechanism for pooling instances in accordance with the present disclosure;

[0011] FIG. 6 is a diagram illustrating a system for pooling instances in accordance with the present disclosure;

[0012] FIG. 7 is a diagram illustrating a system for pooling instances in accordance with the present disclosure;

[0013] FIG. 8 is a diagram illustrating a system for bursting capacity in accordance with the present disclosure;

[0014] FIG. 9 is a diagram illustrating a system for pooling instances in accordance with the present disclosure;

[0015] FIG. 10 is a diagram illustrating an example user interface in accordance with the present disclosure;

CLAIMS

WHAT IS CLAIMED IS:

1. A system for allocating computing resources in a provider network, the system comprising:

one or more computing devices, the one or more computing devices comprising one or more resource slots, each resource slot configured to host a computing instance;

at least one memory having stored therein computer instructions that, upon execution by one or more processors of the system, at least cause the system to:

allocate a plurality of resource slots to a plurality of entities according to a first pricing policy;

receive an indication that at least a portion of the resource slots associated with a first entity of the plurality of entities is associated with a second pricing policy and accessible to one or more users identified by the first entity, the second pricing policy being designated by the first entity; and

allocate the indicated resource slots to the one or more users according to the second pricing policy.

2. The system according to claim 1, further comprising computer readable instructions that, upon execution by the one or more processors of the system, cause the system to at least allow the one or more users identified by the first entity to bid for the indicated resource slots in accordance with the second pricing policy.

3. The system according to claim 2, further comprising computer readable instructions that, upon execution by the one or more processors of the system, cause the system to at least provide a user interface for allowing the one or more users identified by the first entity to bid for the indicated resource slots.

4. The system according to claim 1, further comprising computer readable instructions that, upon execution by the one or more processors of the system, cause the system to at least provide a user interface for facilitating selection of the indicated resource slots, the one or more users identified by the first entity, and the second pricing policy.

5. A method for allocating computing resources in a provider network, the method comprising:

in a computing environment comprising a plurality of computing devices and where capacity of each computing device is logically segmented into one or more resource slots, associating a plurality of resource slots with a first user according to a first set of pricing policies; and

allowing the user to identify one or more of the associated resource slots and identify other users allowed to access the one or more of the associated resource slots according to a second set of pricing policies, the second set of pricing policies being designated by the user.

6. The method of claim 5, further comprising causing allocation of the one or more of the associated resource slots to the other users according to the second pricing policy.

7. The method of claim 5, further comprising allowing the other users to bid for the one or more of the associated resource slots in accordance with the second pricing policy.

8. The method of claim 5, further comprising providing a user interface operable to receive indication of the one or more of the associated resource slots, the other users allowed to access the one or more of the associated resource slots, and the second pricing policy.

9. The method of claim 5, wherein the second pricing policy indicates currency to use to acquire a resource slot, minimum/maximum currency levels, or a bidding algorithm.

10. The method of claim 9, wherein the currency relates to monetary units, CPU cycles, storage, memory, or network bandwidth.

11. The method of claim 7, wherein said bid comprises bidding for CPU cycles or network bandwidth.

12. The method of claim 5, further comprising allowing the user to disallow some of the other users from accessing the one or more of the associated resource slots.

13. The method of claim 5, wherein said allowing comprises receiving a request to allocate the one or more of the associated resource slots, associating the request with a request identifier, and associating the one or more of the associated resource slots with a pool identifier.

14. The method of claim 5, further comprising determining billing information for accessing the one or more of the associated resource slots.

15. A non-transitory computer-readable storage medium having stored thereon computer-readable instructions, the computer-readable instructions comprising instructions that upon execution on one or more computing devices, to at least:

in a computing environment comprising a plurality of computing devices and where capacity of each computing device is logically segmented into one or more resource slots, generate a user interface allowing a user to identify one or more resource slots associated with the user and identify other users allowed to access the one or more resource slots associated with the user according to a pricing policy, the pricing policy being designated by the user; and

cause the computing environment to at least allow the other users to bid for the one or more resource slots associated with the user in accordance with the pricing policy.

Claims

1. A system for allocating computing resources in a provider network, the system comprising:

one or more computing devices, the one or more computing devices comprising one or more resource slots, each resource slot configured to host a computing instance, and

at least one memory having stored therein computer-readable instructions that, upon execution by one or more processors of the system, at least cause the system to:

allocate a plurality of resource slots to a first user according to an allocation policy of the provider network;

allow the first user to allocate at least a portion of the plurality of resource slots to a second user who is authorized to access and use the allocated portion according to an allocation policy established by the first user, wherein the allocation policy established by the first user is different from the allocation policy of the provider network,

wherein the provider network is configured to deny requests to access the allocated portion of the plurality of resource slots not authorized by the allocation policy established by the first user; and

maintain the allocation of the plurality of resource slots to the first user according to the allocation policy of the provider network while the allocated portion is accessed by the second user according to the allocation policy established by the first user;

wherein the allocation policy established by the first user comprises:

allowing the second user to submit a bid including a bid amount and a request for resource slots;

comparing the bid amount with a settable minimum amount of the allocation policy established by the first user; and

conditionally allocating one or more resource slots to the second user while the bid amount remains above the settable minimum amount.

2. The system according to claim 1, further comprising computer-readable instructions that, upon execution by the one or more processors of the system, cause the system to at least allow the second user to bid for indicated resource slots in accordance with the allocation policy established by the first user.

3. The system according to claim 2, further comprising computer-readable

instructions that, upon execution by the one or more processors of the system, cause the system to at least provide a user interface for allowing the second user to bid for the indicated resource slots.

4. The system according to claim 2, further comprising computer-readable instructions that, upon execution by the one or more processors of the system, cause the system to at least provide a user interface for facilitating selection of the indicated resource slots, the second user, and the allocation policy established by the first user.

5. A method for allocating computing resources in a provider network, the method comprising:

in a computing environment comprising a plurality of computing devices and where capacity of the computing devices are logically segmented into one or more resource slots, associating a plurality of resource slots with a first user according to an allocation policy of the provider network; and

allowing the user to identify one or more of the associated resource slots and indicate other users allowed to access the one or more of the associated resource slots according to an allocation policy determined by the user, wherein the allocation policy determined by the user is different from the allocation policy of the provider network, and wherein the provider network is configured to deny requests to access the identified resource slots not authorized by the allocation policy determined by the user;

wherein the allocation policy determined by the user comprises:

allowing the other users to submit bids including a bid amount and a request for resource slots;

comparing a given bid amount with a settable minimum amount of the allocation policy determined by the first user; and

conditionally allocating one or more resource slots to a given one of the other users while the given bid amount remains above the settable minimum amount.

6. The method of claim 5, further comprising providing a user interface operable to receive indication of the one or more of the associated resource slots, the other users allowed to access the one or more of the associated resource slots, and the allocation policy determined by the user.

7. The method of claim 5, wherein the allocation policy determined by the user indicates currency for acquiring a resource slot, minimum/maximum currency levels, or a bidding algorithm.

8. The method of claim 7, wherein the currency relates to monetary units.

9. The method of claim 7, wherein the currency relates to CPU cycles, storage, memory, or network bandwidth.

10. The method of claim 5, wherein said bid comprises bidding for CPU cycles or network bandwidth.

11. The method of claim 5, further comprising allowing the user to disallow some of the other users from accessing the one or more of the associated resource slots.

12. The method of claim 5, wherein said allowing comprises receiving a request to allocate the one or more of the associated resource slots, associating the request with a request identifier, and associating the one or more of the associated resource slots with a pool identifier.

13. The method of claim 5, further comprising determining billing information for accessing the one or more of the associated resource slots.

14. The method of claim 5, further comprising determining which of the other users have bid for at least one of the one or more of the associated resource slots in accordance with the allocation policy determined by the user and collecting billing information based on the determining.

15. The method of claim 14, further comprising associating the determined other users with user identifiers and including the user identifiers with records associated with the determined other users.

16. The method of claim 15, wherein the records are formed as key value pairs.

17. The method of claim 5, wherein the bid comprises a maximum price to be paid for one or more of the associated resource slots.

18. The method of claim 17, wherein one of the one or more of the associated resource slots is allocated to one of the other users when the maximum price exceeds a predetermined price.

19. The method of claim 18, wherein the predetermined price is based on supply and demand of the one or more of the associated resource slots.

20. The method of claim 19, wherein the predetermined price is selectable by the user.

21. A non-transitory computer-readable storage medium having stored thereon computer-readable instructions, the computer-readable instructions comprising instructions that upon execution on one or more computing devices, to at least:

in a computing environment comprising a plurality of computing devices and where capacity of the computing devices are logically segmented into one or more resource slots, generate a user interface allowing a user to identify one or more resource slots associated with the user and select other users allowed to access the one or more resource slots associated with the user according to an allocation policy determined by the user, wherein a provider network is configured to deny requests to access the identified resource slots not authorized by the allocation policy determined by the user; and

cause the computing environment to at least allow the other users to request the one or more resource slots associated with the user in accordance with the allocation policy determined by the user;

wherein the allocation policy determined by the user comprises:

allowing the other users to submit bids including a bid amount and a request for resource slots;

comparing a given bid amount with a settable minimum amount of the allocation policy determined by the first user; and

conditionally allocating one or more resource slots to a given one of the other users while the given bid amount remains above the settable minimum amount.

22. The non-transitory computer-readable storage medium of claim 21, wherein the

allocation policy determined by the user comprises currency to use to acquire a resource slot, price levels, or a bidding algorithm.

23. The non-transitory computer-readable storage medium of claim 22, wherein the bidding algorithm comprises one of an ascending auction and a descending auction.

24. The non-transitory computer-readable storage medium of claim 21, comprising instructions that upon execution on one or more computing devices, at least cause determining billing information for allowing the other users to bid for the one or more resource slots associated with the user.

25. The non-transitory computer-readable storage medium of claim 24, comprising instructions that upon execution on one or more computing devices, at least cause determining which of the other users have bid for one or more of the one or more resource slots associated with the user in accordance with the allocation policy determined by the user and determining billing information based on the determined other users.