

PRE-CONFIGURE AND PRE-LAUNCH COMPUTE RESOURCES

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims the benefit of U.S. Patent Application No. 13/914,225, filed June 10, 2013, the disclosure of which is incorporated herein by reference in its entirety.

BACKGROUND

[0002] Service providers provide computing resources to various entities, such as corporations, universities, government agencies and other types of customers. The services offered by the service providers allow the entities flexibility in resource allocation, improved scalability, reduced operating costs and so on.

[0003] Service providers typically provide computing resources in the form of instances. The instances are sometimes implemented as virtual machines, virtual machine instances or the like. The ability to provide computing resources in the form of instances allows the service providers to dynamically scale the computing resources available to meet the needs and requirements of the various entities that use the services.

SUMMARY OF INVENTION

[0003a] According to a first aspect of the present invention, there is provided a system for provisioning and launching compute instances, the system comprising:

- one or more datacenters comprising a plurality of computing devices configured to communicate with each other over network connections; and

- one or more memories having stored therein computer-readable instructions that, upon execution on the system, cause the system at least to:

- determine an expected demand for a plurality of compute instances, the plurality of compute instances being associated with a plurality of machine images and a plurality of compute instance types;

- determine one or more computing devices from the plurality of computing devices configured to host the plurality of compute instances based at least in part on the expected demand and an expected availability of the plurality of computing devices to support a balanced distribution of the compute instances across the computing devices;

prior to receiving a request from a customer node to launch a compute instance, launch and idle the plurality of compute instances on the one or more computing devices based at least in part on the plurality of machine images and the plurality of compute instance types; and

in response to receiving the request from the customer node to launch the compute instance, activate the compute instance from one of the plurality of machine images that are idle on the one or more computing devices.

[0003b] According to a second aspect of the present invention, there is provided a method, comprising:

providing a plurality of pre-configured machine images on a plurality of compute nodes, wherein the plurality of pre-configured machine images is based at least in part on an expected demand for compute instances and an expected availability of the plurality of compute nodes to support a balanced distribution of the compute instances across the computing devices;

launching the compute instances prior to receiving the request from a user node;

idling the launched compute instances prior to receiving the request from the user node;

after receiving a request from the user node to launch a compute instance, determining that the compute instance is associated with at least a type of machine image of the plurality of pre-configured machine images; and

completing a launch of the compute instance using the associated type of machine image on a compute node of the plurality of compute nodes based at least in part on the determination by activating the idled compute instance in response to receiving the request from the user node.

[0003c] According to a third aspect of the present invention, there is provided a method comprising:

determining a demand for a plurality of compute instances associated with a plurality of machine images;

determining a plurality of compute nodes configured to launch the plurality of compute instances from the plurality of machine images, the determining being based at least in part on a constraint associated with launching the plurality of compute instances on the plurality of compute nodes, wherein the constraint is based at least in part on an available compute capacity at the plurality of compute nodes or availabilities of compute nodes across a zone or a region and is based at least in part on a number of compute instances configured to be concurrently launched on a compute node of the plurality of compute nodes to support a balanced distribution of the compute instances across the compute nodes;

configuring the plurality of machine images on the plurality of computing nodes prior to receiving a request from a user node to launch a compute instance of the plurality of compute instances; and

after receiving the request from the user node, completing a launch of a specific type of compute instance using one of the plurality of configured machine images on one of the plurality of compute nodes.

BRIEF DESCRIPTION OF DRAWINGS

[0004] Throughout the drawings, reference numbers may be re-used 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.

[0005] FIG. 1 illustrates an example operating environment of datacenters, according to embodiments.

[0006] FIG. 2 illustrates an example configuration of a datacenter, according to embodiments.

[0007] FIG. 3 is a flowchart depicting an example procedure for launching an instance.

[0008] FIG. 4 is a flowchart depicting another example procedure for launching an instance, according to embodiments.

[0009] FIG. 5 is a flowchart depicting an example procedure for determining an expected demand, according to embodiments.

CLAIMS

WHAT IS CLAIMED IS:

1. A system for provisioning and launching compute instances, the system comprising:
one or more datacenters comprising a plurality of computing devices configured to communicate with each other over network connections; and
one or more memories having stored therein computer-readable instructions that, upon execution on the system, cause the system at least to:
determine an expected demand for a plurality of compute instances, the plurality of compute instances being associated with a plurality of machine images and a plurality of compute instance types;
determine one or more computing devices from the plurality of computing devices configured to host the plurality of compute instances based at least in part on the expected demand and an expected availability of the plurality of computing devices to support a balanced distribution of the compute instances across the computing devices;
prior to receiving a request from a customer node to launch a compute instance, launch and idle the plurality of compute instances on the one or more computing devices based at least in part on the plurality of machine images and the plurality of compute instance types; and
in response to receiving the request from the customer node to launch the compute instance, activate the compute instance from one of the plurality of machine images that are idle on the one or more computing devices.
2. The system of claim 1, wherein the expected demand is determined based at least in part on a history of requests received from customer nodes, and wherein the history of requests is sorted based at least in part on configuration parameters of compute instances associated with the requests.
3. The system of claim 1, wherein the instructions to determine an expected demand for a plurality of compute instances comprise instructions that, upon execution on the system, cause the system at least to:

analyze a history of requests for compute instances to determine a target for the plurality of compute instances and the one or more computing devices; and

revise the target based at least in part on statistical factors associated with the requests.

4. The system of claim 1, wherein the instructions to determine one or more computing devices from the plurality of computing devices comprise instructions that, upon execution on the system, cause the system at least to:

balance the expected demand with the expected availability to meet a constraint associated with launching and idling the plurality of compute instances; and

determine the one or more computing devices based at least in part on the balance.

5. A system of claim 1, wherein the instructions further cause the system at least to:

remove the compute instances that are idle on the one or more computing devices, if the compute instances are not launched in response to the request from the customer node within a predetermined timeframe.

6. A method, comprising:

providing a plurality of pre-configured machine images on a plurality of compute nodes, wherein the plurality of pre-configured machine images is based at least in part on an expected demand for compute instances and an expected availability of the plurality of compute nodes to support a balanced distribution of the compute instances across the computing devices;

launching the compute instances prior to receiving the request from a user node;

idling the launched compute instances prior to receiving the request from the user node;

after receiving a request from the user node to launch a compute instance, determining that the compute instance is associated with at least a type of machine image of the plurality of pre-configured machine images; and

completing a launch of the compute instance using the associated type of machine image on a compute node of the plurality of compute nodes based at least in part on the determination by activating the idled compute instance in response to receiving the request from the user node.

7. The method of claim 6, wherein the expected demand is based at least in part on previously received requests for compute instances.

8. The method of claim 6, wherein the expected demand comprises types of a plurality of compute instances associated with the plurality of pre-configured machine images and identifiers associated with the plurality of pre-configured machine images.

9. The method of claim 8, wherein the expected demand further comprises a timeframe within which requests for the plurality of compute instances from user nodes are expected to be received and regions expected to be indicated in the requests from the user nodes.

10. The method of claim 9, wherein the method further comprises the step of removing the compute instances that are idle on the one or more computing devices, if the compute instances are not launched in response to the request from the user nodes within the timeframe.

11. The method of claim 6, wherein the machine image is pre-configured for launch on the compute node based at least in part on:

- caching the machine image from a root storage device that stores the machine image;
- modifying a configuration file of the machine image based at least in part on a configuration of the compute instance; and

- allocating a storage volume to the compute instance based at least in part on the configuration of the compute instance.

12. The method of claim 11, wherein the completing a launch of the compute instance comprises booting an operating system of the machine image on the compute node based at least in part on the configuration file that is modified.

13. A method comprising:

- determining a demand for a plurality of compute instances associated with a plurality of machine images;

- determining a plurality of compute nodes configured to launch the plurality of compute instances from the plurality of machine images, the determining being based at least in part on a constraint associated with launching the plurality of compute instances on the plurality of compute nodes, wherein the constraint is based at least in part on an available compute capacity at the plurality of compute nodes or availabilities of compute nodes across a zone or a region and is based at least in part on a number of compute instances configured to be concurrently launched on a compute node of the plurality of compute nodes to support a balanced distribution of the compute instances across the compute nodes;

- configuring the plurality of machine images on the plurality of computing nodes prior to receiving a request from a user node to launch a compute instance of the plurality of compute instances; and

- after receiving the request from the user node, completing a launch of a specific type of compute instance using one of the plurality of configured machine images on one of the plurality of compute nodes.

14. The method of claim 13, wherein the constraint is based at least in part on a fee associated with executing an application or an operating system associated with one of the plurality of machine images or a fee associated with launching one of the plurality of compute instances from one of the plurality of machine images on one of the plurality of compute nodes.