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;

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;

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

completing a launch of the compute instance using the associated machine image on a compute node of the plurality of compute nodes based at least in part on the determination.

[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;

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 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.