CLAIMS

What is claimed is:

~~1. A computer implemented method comprising:~~

~~under the control of one or more computer systems configured with executable instructions,~~

~~causing, by a host computing device running in a service provider environment, a virtual machine to be provisioned on the host computing device;~~

~~receiving, from a scaling service running in the service provider environment, a request to adjust resources allocated to the virtual machine; and~~

~~in response to receiving the request, adjusting allocation of one or more computing resources to the virtual machine.~~

~~2. The computer implemented method of claim 1, wherein the scaling service further performs:~~

~~monitoring one or more metrics associated with running the virtual machine;~~

~~detecting that the one or more metrics have passed at least one specified threshold; and~~

~~transmitting the request to adjust resources allocated to the virtual machine to the host computer device in response to detecting that the one or more metrics have passed the specified threshold.~~

~~3. The computer implemented method of claim 1, wherein a monitoring service executing on the host computing device is configured to~~

~~monitor one or more metrics associated with the virtual machine;~~

~~detect that the one or more metrics have passed a threshold; and~~

~~adjust the allocation of the one or more computing resources in response to detecting that the one or more metrics have passed the threshold.~~

~~4. The computer implemented method of claim 1, wherein the scaling service further performs:~~

~~receiving a request to adjust resources allocated to the virtual machine from a customer via an application programming interface (API); and~~

transmitting the request to adjust resources allocated to the virtual machine in response to receiving the request.

5. The computer implemented method of claim 1, further comprising:

continuing to allocate the one or more additional resources to the virtual machine until a predetermined limit is reached; and

provisioning one or more additional virtual machines to distribute a workload of the virtual machine across one or more additional virtual machines.

6. The computer implemented method of claim 1, wherein the virtual machine further includes a guest agent that reports one or more metrics associated with executing the workload.

7. The computer implemented method of claim 1, further comprising:

billing the user based at least in part on the one or more resources of the host computing device allocated to the virtual machine in response to the request from the scaling service.

8. The computer implemented method of claim 1, further comprising:

providing an electronic marketplace that enables a customer to obtain resources of the host computing device to be allocated to the virtual machine for a fee, the fee being based at least in part on demand and supply of the one or more resources.

9. The computer implemented method of claim 1, further comprising:

receiving, by a placement service, an API request requesting that the virtual machine be scalable;

determining, by the placement service, that the host computing device includes capacity to add additional resources to the virtual machine; and

provisioning, by the placement service, the virtual machine onto the host computing device.

10. A computing system, comprising:

at least one processor; and

memory including instructions that, when executed by the processor, cause the computing system to:

receive a web service request related to adjusting resources allocated to a virtual machine running in a service provider environment provision a virtual machine for a user; and

in response to receiving the request, cause a server hosting the virtual machine to allocate one or more computing resources to the virtual machine.

11. The computing system of claim 10, wherein the web service request contains one or more input parameters that specify a set of conditions in response to which the server is instructed to adjust the computing resources allocated to the virtual machine.

12. The computing device of claim 10, wherein the memory further comprises instructions that, when executed by the processor, cause the computing device to:

continue to allocate the one or more additional resources to the virtual machine until a predetermined limit is reached; and

provision one or more additional virtual machines to distribute a workload of the virtual machine across one or more additional virtual machines.

13. The computing device of claim 10, wherein the memory further comprises instructions that, when executed by the processor, cause the computing device to:

determine that the user has selected to the virtual machine instance of a type that is capable of being scaled by allocating the computing resource in response to the web service request; and

bill the user based at least in part on the type of the virtual machine selected by the user.

14. The computing device of claim 10, wherein the memory further comprises instructions that, when executed by the processor, cause the computing device to:

~~provide an electronic marketplace that enables a customer to purchase one or more additional resources of the computing device based at least in part on demand and supply of the one or more resources on the computing device.~~

~~15. The computing device of claim 10, wherein provisioning the virtual machine for the user further includes:~~

~~receive, by a placement service, an Application Programming Interface (API) request requesting that the virtual machine be scalable; and~~

~~provision, by the placement service, the virtual machine onto the host computing device.~~1. A computer implemented method for scaling a virtual machine, said method comprising:

associating, by one or more computer systems, an account with a customer;

providing, by the one or more computer systems, an application programming interface (API) to customers;

receiving, by the one or more computer systems, from at least one customer via the application programming interface (API), a request to launch an instance of a virtual machine image;

launching, by the one or more computer systems, the virtual machine image;

provisioning, by the one or more computer systems, a virtual machine instance for the at least one customer based on the launched virtual machine image, wherein the virtual machine instance is provisioned on a host computing device;

receiving, at the one or more computer systems, from the customer via the API, a customer-defined threshold associated with the virtual machine instance, wherein the customer-defined threshold is specified at the time of the provisioning of the virtual machine instance;

associating, by the one or more computer systems, one or more metrics with the virtual machine instance;

monitoring, by the one or more computer systems, the one or more metrics associated with the virtual machine instance during execution of the virtual machine instance; and

adjusting, by the one or more computer systems, allocation of one or more computing resources to the virtual machine instance based at least in part on the one or more metrics and the received customer-defined threshold.

2. The method of claim 1, wherein the method further comprises:

determining, by the one or more computer systems, that additional computing resources cannot be allocated to the virtual machine instance; and

provisioning, by the one or more computer systems, a second virtual machine instance for the customer based on a determining that the additional computing resources cannot be allocated to the virtual machine instance.

3. The method of claim 1, further comprising:

causing, at the one or more computer systems, an account associated with the customer to be charged a fee for running the virtual machine instance, wherein the fee is based at least in part on the adjusted computing resources allocated to the virtual machine instance.