



- (51) International Patent Classification: US 16/017,954 (CON)
G06F 9/48 (2006.01) Filed on 25 June 2018 (25.06.2018)
- (21) International Application Number: PCT/US2019/038520 US 16/017,970 (CON)
Filed on 25 June 2018 (25.06.2018)
- (22) International Filing Date: 21 June 2019 (21.06.2019)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
16/017,954 25 June 2018 (25.06.2018) US
16/017,970 25 June 2018 (25.06.2018) US
- (63) Related by continuation (CON) or continuation-in-part (CIP) to earlier applications:
- (71) Applicant: **AMAZON TECHNOLOGIES, INC.** [US/US]; PO Box 81226, Seattle, Washington 98108-1226 (US).
- (72) Inventors: **MULLEN, Niall**; c/o Amazon Technologies, Inc., 410 Terry Avenue North, Seattle, Washington 98109-5210 (US). **PIWONKA, Philip Daniel**; c/o Amazon Technologies, Inc., 410 Terry Avenue North, Seattle, Washington 98109-5210 (US). **WAGNER, Timothy Allen**; c/o Amazon Technologies, Inc., 410 Terry Avenue North, Seattle, Washington 98109-5210 (US). **BROOKER, Marc John**; c/o Amazon Technologies, Inc., 410 Terry Avenue North, Seattle, Washington 98109-5210 (US).

(54) Title: EXECUTION OF AUXILIARY FUNCTIONS IN AN ON-DEMAND NETWORK CODE EXECUTION SYSTEM

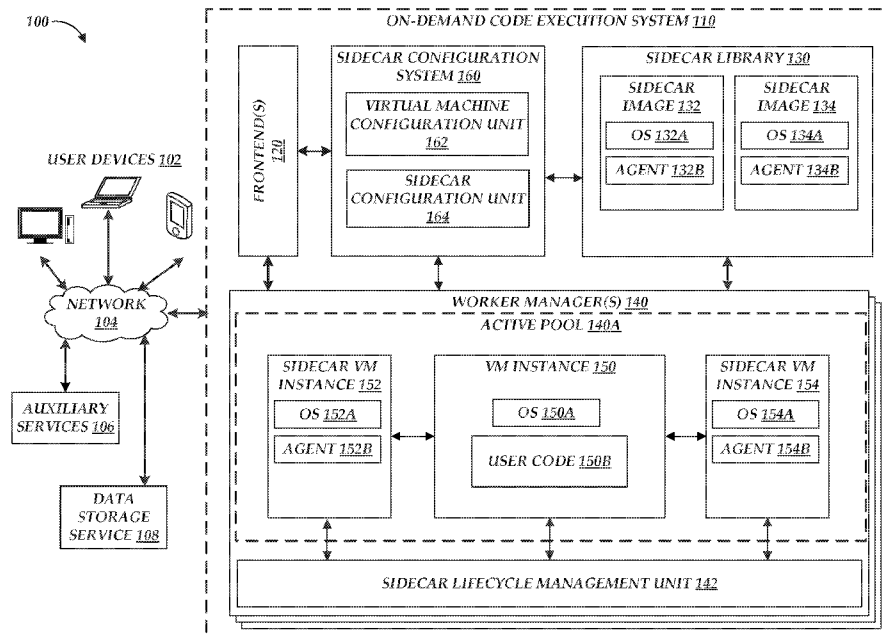


Fig. 1

(57) Abstract: Systems and methods are described for providing auxiliary functions in an on-demand code execution system in a manner that enables efficient execution of code. A user may generate a task on the system by submitting code. The system may determine the auxiliary functions that the submitted code may require when executed on the system, and may provide these auxiliary functions by provisioning sidecar virtual machine instances that work in conjunction with the virtual machine instance executing the submitted code. The sidecars may provide auxiliary functions on a per-task, per-user, or per-request basis, and the lifecycles of the sidecars may be determined based on the lifecycles of the virtual machine instances that execute submitted code. Auxiliary functions may thus be provided only when needed, and may be provided securely by preventing a user from accessing the sidecars of other users.



(74) **Agent: URIBE, Mauricio A.**; Knobbe Martens Olson & Bear LLP, 2040 Main Street, Fourteenth Floor, Irvine, California 92614 (US).

(81) **Designated States** (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

- *with international search report (Art. 21(3))*

EXECUTION OF AUXILIARY FUNCTIONS IN AN ON-DEMAND NETWORK CODE EXECUTION SYSTEM

BACKGROUND

[0001] Computing devices can utilize communication networks to exchange data. Companies and organizations operate computer networks that interconnect a number of computing devices to support operations or to provide services to third parties. The computing systems can be located in a single geographic location or located in multiple, distinct geographic locations (e.g., interconnected via private or public communication networks). Specifically, data centers or data processing centers, herein generally referred to as a “data center,” may include a number of interconnected computing systems to provide computing resources to users of the data center. The data centers may be private data centers operated on behalf of an organization or public data centers operated on behalf, or for the benefit of, the general public.

[0002] To facilitate increased utilization of data center resources, virtualization technologies allow a single physical computing device to host one or more instances of virtual machines that appear and operate as independent computing devices to users of a data center. With virtualization, the single physical computing device can create, maintain, delete, or otherwise manage virtual machines in a dynamic manner. In turn, users can request computer resources from a data center, including single computing devices or a configuration of networked computing devices, and be provided with varying numbers of virtual machine resources.

[0003] In some scenarios, virtual machine instances may be configured according to a number of virtual machine instance types to provide specific functionality. For example, various computing devices may be associated with different combinations of operating systems or operating system configurations, virtualized hardware resources and software applications to enable a computing device to provide different desired functionalities, or to provide similar functionalities more efficiently. These virtual machine instance type configurations are often contained within a device image, which includes static data containing the software (e.g., the OS and applications together with their configuration and

data files, etc.) that the virtual machine will run once started. The device image is typically stored on the disk used to create or initialize the instance. Thus, a computing device may process the device image in order to implement the desired software configuration.

BRIEF DESCRIPTION OF THE DRAWINGS

[0004] FIG. 1 is a block diagram depicting an illustrative environment in which an on-demand code execution system can operate to execute tasks corresponding to code, which may be submitted by users of the on-demand code execution system, and to provision and execute sidecar virtual machine instances to facilitate execution of the submitted code;

[0005] FIG. 2 depicts a general architecture of a computing device providing a sidecar configuration system that is configured to facilitate configuration of sidecar virtual machine instances used to facilitate execution of tasks on the on-demand code execution system of FIG. 1;

[0006] FIG. 3 is a flow diagram depicting illustrative interactions for submitting code corresponding to a task to the on-demand code execution system of FIG. 1, and for the on-demand code execution system to determine sidecar configurations to facilitate execution of the task on the on-demand code execution system;

[0007] FIG. 4 is a flow diagram depicting illustrative interactions for executing a task utilizing the determined sidecar configuration provided on the on-demand code execution system of FIG. 1;

[0008] FIG. 5 is a flow diagram depicting illustrative interactions for managing the lifecycles of sidecar virtual machine instances on the on-demand code execution system of FIG. 1;

[0009] FIG. 6 is a flow diagram depicting illustrative interactions between sidecar virtual machine instances and virtual machine instances that execute code corresponding to tasks on the on-demand code execution system of FIG. 1;

[0010] FIG. 7 is a flow chart depicting an illustrative routine for determining a sidecar configuration based on code submitted to the on-demand code execution system of FIG. 1; and

[0011] FIG. 8 is a flow chart depicting an illustrative routine for managing lifecycles of sidecars during execution of tasks on the on-demand code execution system of FIG. 1.

DETAILED DESCRIPTION

[0012] Generally described, aspects of the present disclosure relate to an on-demand code execution system. The on-demand code execution system enables rapid execution of code, which may be supplied by users of the on-demand code execution system. More specifically, embodiments of the present disclosure relate to implementing auxiliary functions for an on-demand code execution system through the use of “sidecar” virtual machine instances. As described in detail herein, the on-demand code execution system may provide a network-accessible service enabling users to submit or designate computer-executable code to be executed by virtual machine instances on the on-demand code execution system. Each set of code on the on-demand code execution system may define a “task,” and implement specific functionality corresponding to that task when executed on a virtual machine instance of the on-demand code execution system. Individual implementations of the task on the on-demand code execution system may be referred to as an “execution” of the task (or a “task execution”). The on-demand code execution system can further enable users to trigger execution of a task based on a variety of potential events, such as detecting new data at a network-based storage system, transmission of an application programming interface (“API”) call to the on-demand code execution system, or transmission of a specially formatted hypertext transport protocol (“HTTP”) packet to the on-demand code execution system. Thus, users may utilize the on-demand code execution system to execute any specified executable code “on-demand,” without requiring configuration or maintenance of the underlying hardware or infrastructure on which the code is executed. Further, the on-demand code execution system may be configured to execute tasks in a rapid manner (e.g., in under 100 milliseconds [ms]), thus enabling execution of tasks in “real-time” (e.g., with little or no perceptible delay to an end user).

[0013] The on-demand code execution system may instantiate virtual machine instances to execute the specified tasks on demand. The on-demand code execution system may further instantiate “sidecar” virtual machine instances, which enable users to control or

monitor the execution of a task and the virtual machine instance upon which it executes. Illustratively, a sidecar virtual machine instance (which may be referred to herein as a “sidecar”) may implement one or more functions for controlling, securing, filtering, monitoring, or managing the virtual machine instance that executes the task code. By implementing these functions in a sidecar or sidecars, the on-demand code execution system can effectively separate these functions from the virtual machine instances executing task code. The sidecar implementation thus improves efficiency with regard to resource utilization, since (as described in more detail below) the sidecars can be made available only when needed. The sidecar implementation further improves security for individual users, since an attacker who compromises one sidecar does not gain access to the sidecars or virtual machine instances of other users.

[0014] As used herein, “auxiliary functions” may refer generally to functions that facilitate the execution of user-submitted task code. For example, auxiliary functions may include encapsulation, logging, tracing, debugging, scanning, profiling, validating input, validating output, or other functions that relate to task code execution. It will be understood by one skilled in the art, however, that these examples are not limiting and that sidecars performing other functions are within the scope of the present disclosure. In some embodiments, auxiliary functions may include control plane functions that execute with administrator-level privileges. Sidecars may be instantiated to perform these functions on a per-user, per-task, or per-call basis, and may thus provide individual users with access to individualized control planes for their virtual machine instances. For example, a sidecar that encapsulates network traffic may be made available to individual users, and may translate packets that are transported on a physical substrate network to a virtual network that the user can access from the user’s virtual machine instance. The on-demand code execution system can thus provide network encapsulation via a sidecar, and can do so without allowing a virtual machine instance that runs user code to access the substrate network and potentially de-encapsulate the traffic of other users.

[0015] As will be appreciated by one of skill in the art in light of the present disclosure, the embodiments disclosed herein improves the ability of computing systems, such as on-demand code execution systems, to execute code in an efficient manner. Moreover, the presently disclosed embodiments address technical problems inherent within

computing systems; specifically, the limited nature of computing resources with which to execute code, the resource overhead associated with providing “always-on” auxiliary functionality, the inefficiencies caused by provisioning functionality that is not utilized, and the security issues caused by providing a common control plane to multiple users. These technical problems are addressed by the various technical solutions described herein, including the provisioning of an execution environment with sidecar virtual machine instances that provide user-specific or task-specific functionality. Thus, the present disclosure represents an improvement on existing data processing systems and computing systems in general.

[0016] The on-demand code execution system may include a virtual machine instance manager configured to receive user code (threads, programs, etc., composed in any of a variety of programming languages) and execute the code in a highly scalable, low latency manner, without requiring user configuration of a virtual machine instance. Specifically, the virtual machine instance manager can, prior to receiving the user code and prior to receiving any information from a user regarding any particular virtual machine instance configuration, create and configure virtual machine instances according to a predetermined set of configurations, each corresponding to any one or more of a variety of run-time environments. Thereafter, the virtual machine instance manager receives user-initiated requests to execute code, and identifies a pre-configured virtual machine instance to execute the code based on configuration information associated with the request. The virtual machine instance manager can further allocate the identified virtual machine instance to execute the user’s code at least partly by creating and configuring containers inside the allocated virtual machine instance, and provisioning the containers with code of the task as well as an dependency code objects. Various embodiments for implementing a virtual machine instance manager and executing user code on virtual machine instances is described in more detail in U.S. Patent No. 9,323,556, entitled “PROGRAMMATIC EVENT DETECTION AND MESSAGE GENERATION FOR REQUESTS TO EXECUTE PROGRAM CODE,” and filed September 30, 2014 (the “556 Patent”), the entirety of which is hereby incorporated by reference.

[0017] As used herein, the term “virtual machine instance” is intended to refer to an execution of software or other executable code that emulates hardware to provide an

environment or platform on which software may execute (an “execution environment”). Virtual machine instances are generally executed by hardware devices, which may differ from the physical hardware emulated by the virtual machine instance. For example, a virtual machine may emulate a first type of processor and memory while being executed on a second type of processor and memory. Thus, virtual machines can be utilized to execute software intended for a first execution environment (e.g., a first operating system) on a physical device that is executing a second execution environment (e.g., a second operating system). In some instances, hardware emulated by a virtual machine instance may be the same or similar to hardware of an underlying device. For example, a device with a first type of processor may implement a plurality of virtual machine instances, each emulating an instance of that first type of processor. Thus, virtual machine instances can be used to divide a device into a number of logical sub-devices (each referred to as a “virtual machine instance”). While virtual machine instances can generally provide a level of abstraction away from the hardware of an underlying physical device, this abstraction is not required. For example, assume a device implements a plurality of virtual machine instances, each of which emulate hardware identical to that provided by the device. Under such a scenario, each virtual machine instance may allow a software application to execute code on the underlying hardware without translation, while maintaining a logical separation between software applications running on other virtual machine instances. This process, which is generally referred to as “native execution,” may be utilized to increase the speed or performance of virtual machine instances. Other techniques that allow direct utilization of underlying hardware, such as hardware pass-through techniques, may be used, as well.

[0018] While a virtual machine executing an operating system is described herein as one example of an execution environment, other execution environments are also possible. For example, tasks or other processes may be executed within a software “container,” which provides a runtime environment without itself providing virtualization of hardware. Containers may be implemented within virtual machines to provide additional security, or may be run outside of a virtual machine instance.

[0019] The foregoing aspects and many of the attendant advantages of this disclosure will become more readily appreciated as the same become better understood by

reference to the following description, when taken in conjunction with the accompanying drawings.

[0020] FIG. 1 is a block diagram of an illustrative operating environment 100 in which an on-demand code execution system 110 may operate based on communication with user computing devices 102, auxiliary services 106, and network-based data storage services 108. By way of illustration, various example user computing devices 102 are shown in communication with the on-demand code execution system 110, including a desktop computer, laptop, and a mobile phone. In general, the user computing devices 102 can be any computing device such as a desktop, laptop or tablet computer, personal computer, wearable computer, server, personal digital assistant (PDA), hybrid PDA/mobile phone, mobile phone, electronic book reader, set-top box, voice command device, camera, digital media player, and the like. The on-demand code execution system 110 may provide the user computing devices 102 with one or more user interfaces, command-line interfaces (CLIs), application programming interfaces (APIs), and/or other programmatic interfaces for generating and uploading user-executable code (e.g., including metadata identifying dependency code objects for the uploaded code), invoking the user-provided code (e.g., submitting a request to execute the user codes on the on-demand code execution system 110), scheduling event-based jobs or timed jobs, tracking the user-provided code, and/or viewing other logging or monitoring information related to their requests and/or user codes. Although one or more embodiments may be described herein as using a user interface, it should be appreciated that such embodiments may, additionally or alternatively, use any CLIs, APIs, or other programmatic interfaces.

[0021] The illustrative environment 100 further includes one or more auxiliary services 106, which can interact with the one-demand code execution environment 110 to implement desired functionality on behalf of a user. Auxiliary services 106 can correspond to network-connected computing devices, such as servers, which generate data accessible to the one-demand code execution environment 110 or otherwise communicate to the one-demand code execution environment 110. For example, the auxiliary services 106 can include web services (e.g., associated with the user computing devices 102, with the on-demand code execution system 110, or with third parties), databases, really simple syndication (“RSS”) readers, social networking sites, or any other source of network-

accessible service or data source. In some instances, auxiliary services 106 may be associated with the on-demand code execution system 110, e.g., to provide billing or logging services to the on-demand code execution system 110. In some instances, auxiliary services 106 actively transmit information, such as API calls or other task-triggering information, to the on-demand code execution system 110. In other instances, auxiliary services 106 may be passive, such that data is made available for access by the on-demand code execution system 110. For example, components of the on-demand code execution system 110 may periodically poll such passive data sources, and trigger execution of tasks within the on-demand code execution system 110 based on the data provided. While depicted in FIG. 1 as distinct from the user computing devices 102 and the on-demand code execution system 110, in some embodiments, various auxiliary services 106 may be implemented by either the user computing devices 102 or the on-demand code execution system 110.

[0022] The illustrative environment 100 further includes one or more network-based data storage services 108, configured to enable the on-demand code execution system 110 to store and retrieve data from one or more persistent or substantially persistent data sources. Illustratively, the network-based data storage services 108 may enable the on-demand code execution system 110 to store information corresponding to a task, such as code or metadata, to store additional code objects representing dependencies of tasks, to retrieve data to be processed during execution of a task, and to store information (e.g., results) regarding that execution. The network-based data storage services 108 may represent, for example, a relational or non-relational database. In another example, the network-based data storage services 108 may represent a network-attached storage (NAS), configured to provide access to data arranged as a file system. The network-based data storage services 108 may further enable the on-demand code execution system 110 to query for and retrieve information regarding data stored within the on-demand code execution system 110, such as by querying for a number of relevant files or records, sizes of those files or records, file or record names, file or record creation times, etc. In some instances, the network-based data storage services 108 may provide additional functionality, such as the ability to separate data into logical groups (e.g., groups associated with individual accounts, etc.). While shown as distinct from the auxiliary services 106, the network-based data storage services 108 may in some instances also represent a type of auxiliary service 106.

[0023] The user computing devices 102, auxiliary services 106, and network-based data storage services 108 may communicate with the on-demand code execution system 110 via a network 104, which may include any wired network, wireless network, or combination thereof. For example, the network 104 may be a personal area network, local area network, wide area network, over-the-air broadcast network (e.g., for radio or television), cable network, satellite network, cellular telephone network, or combination thereof. As a further example, the network 104 may be a publicly accessible network of linked networks, possibly operated by various distinct parties, such as the Internet. In some embodiments, the network 104 may be a private or semi-private network, such as a corporate or university intranet. The network 104 may include one or more wireless networks, such as a Global System for Mobile Communications (GSM) network, a Code Division Multiple Access (CDMA) network, a Long Term Evolution (LTE) network, or any other type of wireless network. The network 104 can use protocols and components for communicating via the Internet or any of the other aforementioned types of networks. For example, the protocols used by the network 104 may include Hypertext Transfer Protocol (HTTP), HTTP Secure (HTTPS), Message Queue Telemetry Transport (MQTT), Constrained Application Protocol (CoAP), and the like. Protocols and components for communicating via the Internet or any of the other aforementioned types of communication networks are well known to those skilled in the art and, thus, are not described in more detail herein.

[0024] The on-demand code execution system 110 is depicted in FIG. 1 as operating in a distributed computing environment including several computer systems that are interconnected using one or more computer networks (not shown in FIG. 1). The on-demand code execution system 110 could also operate within a computing environment having a fewer or greater number of devices than are illustrated in FIG. 1. Thus, the depiction of the on-demand code execution system 110 in FIG. 1 should be taken as illustrative and not limiting to the present disclosure. For example, the on-demand code execution system 110 or various constituents thereof could implement various Web services components, hosted or “cloud” computing environments, and/or peer to peer network configurations to implement at least a portion of the processes described herein.

[0025] Further, the on-demand code execution system 110 may be implemented directly in hardware or software executed by hardware devices and may, for instance, include

one or more physical or virtual servers implemented on physical computer hardware configured to execute computer executable instructions for performing various features that will be described herein. The one or more servers may be geographically dispersed or geographically co-located, for instance, in one or more data centers. In some instances, the one or more servers may operate as part of a system of rapidly provisioned and released computing resources, often referred to as a “cloud computing environment.”

[0026] In the example of FIG. 1, the on-demand code execution system 110 is illustrated as connected to the network 104. In some embodiments, any of the components within the on-demand code execution system 110 can communicate with other components of the on-demand code execution system 110 via the network 104. In other embodiments, not all components of the on-demand code execution system 110 are capable of communicating with other components of the virtual environment 100. In one example, only the frontend 120 (which may in some instances represent multiple frontends 120) may be connected to the network 104, and other components of the on-demand code execution system 110 may communicate with other components of the environment 100 via the frontends 120.

[0027] In FIG. 1, users, by way of user computing devices 102, may interact with the on-demand code execution system 110 to provide executable code, and establish rules or logic defining when and how such code should be executed on the on-demand code execution system 110, thus establishing a “task.” For example, a user may wish to run a piece of code in connection with a web or mobile application that the user has developed. One way of running the code would be to acquire virtual machine instances from service providers who provide infrastructure as a service, configure the virtual machine instances to suit the user’s needs, and use the configured virtual machine instances to run the code. In order to avoid the complexity of this process, the user may alternatively provide the code to the on-demand code execution system 110, and request that the on-demand code execution system 110 execute the code. The on-demand code execution system 110 can handle the acquisition and configuration of compute capacity (e.g., containers, instances, etc., which are described in greater detail below) based on the code execution request, and execute the code using the compute capacity. The on-demand code execution system 110 may automatically scale up and down based on the volume, thereby relieving the user from the burden of having

to worry about over-utilization (e.g., acquiring too little computing resources and suffering performance issues) or under-utilization (e.g., acquiring more computing resources than necessary to run the codes, and thus overpaying). In accordance with embodiments of the present disclosure, and as described in more detail below, the on-demand code execution system 110 may configure the virtual machine instances with customized operating systems to execute the user's code more efficiency and reduce utilization of computing resources.

[0028] To enable interaction with the on-demand code execution system 110, the system 110 includes one or more frontends 120, which enable interaction with the on-demand code execution system 110. In an illustrative embodiment, the frontends 120 serve as a "front door" to the other services provided by the on-demand code execution system 110, enabling users (via user computing devices 102) to provide, request execution of, and view results of computer executable code. The frontends 120 include a variety of components to enable interaction between the on-demand code execution system 110 and other computing devices. For example, each frontend 120 may include a request interface providing user computing devices 102 with the ability to upload or otherwise communicate user-specified code to the on-demand code execution system 110 and to thereafter request execution of that code. In one embodiment, the request interface communicates with external computing devices (e.g., user computing devices 102, auxiliary services 106, etc.) via a graphical user interface (GUI), CLI, or API. The frontends 120 process the requests and makes sure that the requests are properly authorized. For example, the frontends 120 may determine whether the user associated with the request is authorized to access the user code specified in the request.

[0029] References to user code as used herein may refer to any program code (e.g., a program, routine, subroutine, thread, etc.) written in a specific program language. In the present disclosure, the terms "code," "user code," and "program code," may be used interchangeably. Such user code may be executed to achieve a specific function, for example, in connection with a particular web application or mobile application developed by the user. As noted above, individual collections of user code (e.g., to achieve a specific function) are referred to herein as "tasks," while specific executions of that code (including, e.g., compiling code, interpreting code, or otherwise making the code executable) are referred to as "task executions" or simply "executions." Tasks may be written, by way of

non-limiting example, in JavaScript (e.g., node.js), Java, Python, and/or Ruby (and/or another programming language). Tasks may be “triggered” for execution on the on-demand code execution system 110 in a variety of manners. In one embodiment, a user or other computing device may transmit a request to execute a task may, which can generally be referred to as “call” to execute of the task. Such calls may include the user code (or the location thereof) to be executed and one or more arguments to be used for executing the user code. For example, a call may provide the user code of a task along with the request to execute the task. In another example, a call may identify a previously uploaded task by its name or an identifier. In yet another example, code corresponding to a task may be included in a call for the task, as well as being uploaded in a separate location (e.g., storage of an auxiliary service 106 or a storage system internal to the on-demand code execution system 110) prior to the request being received by the on-demand code execution system 110. As noted above, the code for a task may reference additional code objects maintained at the on-demand code execution system 110 by use of identifiers of those code objects, such that the code objects are combined with the code of a task in an execution environment prior to execution of the task. The on-demand code execution system 110 may vary its execution strategy for a task based on where the code of the task is available at the time a call for the task is processed. A request interface of the frontend 120 may receive calls to execute tasks as Hypertext Transfer Protocol Secure (HTTPS) requests from a user. Also, any information (e.g., headers and parameters) included in the HTTPS request may also be processed and utilized when executing a task. As discussed above, any other protocols, including, for example, HTTP, MQTT, and CoAP, may be used to transfer the message containing a task call to the request interface 122.

[0030] A call to execute a task may specify one or more third-party libraries (including native libraries) to be used along with the user code corresponding to the task. In one embodiment, the call may provide to the on-demand code execution system 110 a file containing the user code and any libraries (and/or identifications of storage locations thereof) corresponding to the task requested for execution. In some embodiments, the call includes metadata that indicates the program code of the task to be executed, the language in which the program code is written, the user associated with the call, and/or the computing resources (e.g., memory, etc.) to be reserved for executing the program code. For example, the

program code of a task may be provided with the call, previously uploaded by the user, provided by the on-demand code execution system 110 (e.g., standard routines), and/or provided by third parties. Illustratively, code not included within a call or previously uploaded by the user may be referenced within metadata of the task by use of a URI associated with the code. In some embodiments, such resource-level constraints (e.g., how much memory is to be allocated for executing a particular user code) are specified for the particular task, and may not vary over each execution of the task. In such cases, the on-demand code execution system 110 may have access to such resource-level constraints before each individual call is received, and the individual call may not specify such resource-level constraints. In some embodiments, the call may specify other constraints such as permission data that indicates what kind of permissions or authorities that the call invokes to execute the task. Such permission data may be used by the on-demand code execution system 110 to access private resources (e.g., on a private network). In some embodiments, individual code objects may also be associated with permissions or authorizations. For example, a third party may submit a code object and designate the object as readable by only a subset of users. The on-demand code execution system 110 may include functionality to enforce these permissions or authorizations with respect to code objects.

[0031] In some embodiments, a call may specify the behavior that should be adopted for handling the call. In such embodiments, the call may include an indicator for enabling one or more execution modes in which to execute the task referenced in the call. For example, the call may include a flag or a header for indicating whether the task should be executed in a debug mode in which the debugging and/or logging output that may be generated in connection with the execution of the task is provided back to the user (e.g., via a console user interface). In such an example, the on-demand code execution system 110 may inspect the call and look for the flag or the header, and if it is present, the on-demand code execution system 110 may modify the behavior (e.g., logging facilities) of the container in which the task is executed, and cause the output data to be provided back to the user. In some embodiments, the behavior/mode indicators are added to the call by the user interface provided to the user by the on-demand code execution system 110. Other features such as source code profiling, remote debugging, etc. may also be enabled or disabled based on the indication provided in a call.

[0032] To manage requests for code execution, the frontend 120 can include an execution queue (not shown in FIG. 1), which can maintain a record of requested task executions. Illustratively, the number of simultaneous task executions by the on-demand code execution system 110 is limited, and as such, new task executions initiated at the on-demand code execution system 110 (e.g., via an API call, via a call from an executed or executing task, etc.) may be placed on the execution queue 124 and processed, e.g., in a first-in-first-out order. In some embodiments, the on-demand code execution system 110 may include multiple execution queues, such as individual execution queues for each user account. For example, users of the on-demand code execution system 110 may desire to limit the rate of task executions on the on-demand code execution system 110 (e.g., for cost reasons). Thus, the on-demand code execution system 110 may utilize an account-specific execution queue to throttle the rate of simultaneous task executions by a specific user account. In some instances, the on-demand code execution system 110 may prioritize task executions, such that task executions of specific accounts or of specified priorities bypass or are prioritized within the execution queue. In other instances, the on-demand code execution system 110 may execute tasks immediately or substantially immediately after receiving a call for that task, and thus, the execution queue may be omitted.

[0033] As noted above, tasks may be triggered for execution at the on-demand code execution system 110 based on explicit calls from user computing devices 102 (e.g., as received at the request interface). Alternatively or additionally, tasks may be triggered for execution at the on-demand code execution system 110 based on data retrieved from one or more auxiliary services 106 or network-based data storage services 108. To facilitate interaction with auxiliary services 106, the frontend 120 can include a polling interface (not shown in FIG. 1), which operates to poll auxiliary services 106 or data storage services 108 for data. Illustratively, the polling interface may periodically transmit a request to one or more user-specified auxiliary services 106 or data storage services 108 to retrieve any newly available data (e.g., social network “posts,” news articles, files, records, etc.), and to determine whether that data corresponds to a user-established criteria triggering execution a task on the on-demand code execution system 110. Illustratively, criteria for execution of a task may include, but is not limited to, whether new data is available at the auxiliary services 106 or data storage services 108, the type or content of the data, or timing

information corresponding to the data. In some instances, the auxiliary services 106 or data storage services 108 may function to notify the frontend 120 of the availability of new data, and thus the polling service may be unnecessary with respect to such services.

[0034] In addition to tasks executed based on explicit user calls and data from auxiliary services 106, the on-demand code execution system 110 may in some instances operate to trigger execution of tasks independently. For example, the on-demand code execution system 110 may operate (based on instructions from a user) to trigger execution of a task at each of a number of specified time intervals (e.g., every 10 minutes).

[0035] The frontend 120 can further include an output interface (not shown in FIG. 1) configured to output information regarding the execution of tasks on the on-demand code execution system 110. Illustratively, the output interface may transmit data regarding task executions (e.g., results of a task, errors related to the task execution, or details of the task execution, such as total time required to complete the execution, total data processed via the execution, etc.) to the user computing devices 102 or to auxiliary services 106, which may include, for example, billing or logging services. The output interface may further enable transmission of data, such as service calls, to auxiliary services 106. For example, the output interface may be utilized during execution of a task to transmit an API request to an external service 106 (e.g., to store data generated during execution of the task).

[0036] In some embodiments, the on-demand code execution system 110 may include multiple frontends 120. In such embodiments, a load balancer (not shown in FIG. 1) may be provided to distribute the incoming calls to the multiple frontends 120, for example, in a round-robin fashion. In some embodiments, the manner in which the load balancer distributes incoming calls to the multiple frontends 120 may be based on the location or state of other components of the on-demand code execution system 110. For example, a load balancer may distribute calls to a geographically nearby frontend 120, or to a frontend with capacity to service the call. In instances where each frontend 120 corresponds to an individual instance of another component of the on-demand code execution system, such as the active pool 140A described below, the load balancer may distribute calls according to the capacities or loads on those other components. As will be described in more detail below, calls may in some instances be distributed between frontends 120 deterministically, such that a given call to execute a task will always (or almost always) be routed to the same

frontend 120. This may, for example, assist in maintaining an accurate execution record for a task, to ensure that the task executes only a desired number of times. While distribution of calls via a load balancer is illustratively described, other distribution techniques, such as anycast routing, will be apparent to those of skill in the art.

[0037] To facilitate execution of tasks, the on-demand code execution system 110 includes one or more sidecar libraries 130, which in turn include one or more sidecar images. In the example illustrated in FIG. 1, the sidecar library 130 includes a sidecar image 132, which comprises an operating system 132A and an agent 132B, and a sidecar image 134, which comprises an operating system 134A and an agent 134B. In some embodiments, the operating system 132A and the operating system 134A may be the same operating system. As described in more detail below, the agents 132B and 134B may perform one or more auxiliary functions when configured to communicate with a virtual machine instance or instances. In some embodiments, the sidecar library 130 contains only agents that perform auxiliary functions (e.g., agents 132B and 134B), and a sidecar is created by provisioning a virtual machine instance with one or more of the agents from the sidecar library 130.

[0038] The on-demand code execution system 110 further includes a sidecar configuration system 160, which implements aspects of the present disclosure including, for example, the determination and configuration of virtual machine instances and sidecar instances for a particular task. In some embodiments, the sidecar configuration system 160 includes a virtual machine configuration unit 162, which may be invoked when the user submits code via the frontend 120 to determine a virtual machine configuration to use with the submitted code. As described in more detail below, the virtual machine configuration unit 162 may analyze the user's code and identify, for example, operating system "hooks," input and/or output redirections, or other modifications to facilitate interactions between a virtual machine instance and various sidecars before, during, and/or after execution of the user-submitted code. In various embodiments, the virtual machine configuration unit 162 may analyze the user's code or process user requests to determine a virtual machine configuration. In further embodiments, the on-demand code execution system 110 includes a sidecar configuration unit 164, which may identify the sidecars to be provisioned along with any configuration of the sidecars to facilitate interaction with the virtual machine instance(s). The sidecar configuration unit 164 may identify, for example, input validations that a sidecar

should perform, and may configure a sidecar to perform them. For example, the user-submitted code may process input data in a particular format, and thus a sidecar configuration may be determined to validate that the input data is in the format.

[0039] The on-demand code execution system further includes one or more worker managers 140 that manage the instances used for servicing incoming calls to execute tasks, and that manage the sidecars used to provide auxiliary functions for these instances. In the example illustrated in FIG. 1, each worker manager 140 manages an active pool 140A, which is a group (sometimes referred to as a pool) of virtual machine instances, implemented by one or more physical host computing devices, that are currently assigned to one or more users. Although the virtual machine instances are described here as being assigned to a particular user, in some embodiments, the instances may be assigned to a group of users, such that the instance is tied to the group of users and any member of the group can utilize resources on the instance. For example, the users in the same group may belong to the same security group (e.g., based on their security credentials) such that executing one member's task in a container on a particular instance after another member's task has been executed in another container on the same instance does not pose security risks. Similarly, the worker managers 140 may assign the instances and the containers according to one or more policies that dictate which requests can be executed in which containers and which instances can be assigned to which users. An example policy may specify that instances are assigned to collections of users who share the same account (e.g., account for accessing the services provided by the on-demand code execution system 110). In some embodiments, the requests associated with the same user group may share the same containers (e.g., if the user codes associated therewith are identical). In some embodiments, a task does not differentiate between the different users of the group and simply indicates the group to which the users associated with the task belong.

[0040] Once a triggering event to execute a task has been successfully processed by a frontend 120, the frontend 120 passes a request to a worker manager 140 to execute the task. In one embodiment, each frontend 120 may be associated with a corresponding worker manager 140 (e.g., a worker manager 140 co-located or geographically nearby to the frontend 120) and thus, the frontend 120 may pass most or all requests to that worker manager 140. In another embodiment, a frontend 120 may include a location

selector configured to determine a worker manager 140 to which to pass the execution request. In one embodiment, the location selector may determine the worker manager 140 to receive a call based on hashing the call, and distributing the call to a worker manager 140 selected based on the hashed value (e.g., via a hash ring). Various other mechanisms for distributing calls between worker managers 140 will be apparent to one of skill in the art. In accordance with embodiments of the present disclosure, the worker manager 140 can obtain a virtual machine configuration and/or sidecar configurations when provisioning a virtual machine instance.

[0041] As shown in FIG. 1, instances may have operating systems (OS), language runtimes, and containers. The containers may have individual copies of the OS, the runtimes, and user codes corresponding to various tasks loaded thereon. In the example of FIG. 1, the active pools 140A managed by a worker manager 140 includes the virtual machine instance 150. The instance 150 includes an operating system 150A and user code 150B. . In some embodiments, the worker managers 140 may maintain a list of instances in an active pool 140A. The list of instances may further specify the configuration (e.g., OS, runtime, container, etc.) of the instances. In some embodiments, the worker managers 140 may have access to a list of instances in a warming pool (e.g., including the number and type of instances). In other embodiments, the worker managers 140 requests compute capacity from a warming pool manager without having knowledge of the virtual machine instances in a warming pool.

[0042] The active pool 140A may further include one or more sidecar virtual machine instances, such as sidecar 152 and sidecar 154. As depicted in FIG. 1, the sidecar 152 includes an OS 152A and an agent 152B, and the sidecar 154 includes an OS 154A and an agent 154B. In some embodiments, one or both of the sidecars 152 and 154 may correspond to a provisioned instance of a sidecar image 132 or 134 from the sidecar library 130. The sidecars 152 and 154 may, as described in more detail below, provide one or more auxiliary functions in conjunction with the virtual machine instance 150 that executes user code 150B.

[0043] The worker manager 140 may further include a sidecar lifecycle management unit 142. As described in more detail below, the sidecar lifecycle management unit 142 may monitor the lifecycles of virtual machine instances, such as virtual machine

instance 150, and may ensure that the lifecycles of corresponding sidecar instances (e.g., sidecars 152 and 154) are synchronized with the virtual machine instance(s) to which they are attached. As described below, the sidecar lifecycle management unit 142 may determine whether a particular sidecar should precede, follow, or change its execution state in parallel when a virtual machine instance undergoes a change in execution state, and may cause sidecars to implement changes in execution state accordingly. In some embodiments, the sidecar lifecycle management unit 142 may be a component of the active pool 140A. In other embodiments, the sidecar lifecycle management unit 142 may sit outside the active pool 140A and facilitate the addition, removal, and/or the timing of the addition or removal of sidecars from the active pool 140A.

[0044] While some functionalities are generally described herein with reference to an individual component of the on-demand code execution system 110, other components or a combination of components may additionally or alternatively implement such functionalities. For example, a worker manager 140 may operate to configure virtual machine instances in a manner similar or identical to as described herein with reference to an OS configuration system 160. One skilled in the art will also understand that the present disclosure is not limited to the embodiment depicted in FIG. 1, in which one virtual machine instance 150 is in communication with two sidecars 152 and 154. In various embodiments, any number of sidecars may be in communication with any number of virtual machine instances, including one-to-many and many-to-many relationships between virtual machine instances and sidecars.

[0045] FIG. 2 depicts a general architecture of a computing system (referenced as sidecar configuration system 160) that operates to determine sidecar configurations within the on-demand code execution system 110. The general architecture of the sidecar configuration system 160 depicted in FIG. 2 includes an arrangement of computer hardware and software modules that may be used to implement aspects of the present disclosure. The hardware modules may be implemented with physical electronic devices, as discussed in greater detail below. The sidecar configuration system 160 may include many more (or fewer) elements than those shown in FIG. 2. It is not necessary, however, that all of these generally conventional elements be shown in order to provide an enabling disclosure. Additionally, the general architecture illustrated in FIG. 2 may be used to implement one or

more of the other components illustrated in FIG. 1. As illustrated, the sidecar configuration system 160 includes a processing unit 290, a network interface 292, a computer readable medium drive 294, and an input/output device interface 296, all of which may communicate with one another by way of a communication bus. The network interface 292 may provide connectivity to one or more networks or computing systems. The processing unit 290 may thus receive information and instructions from other computing systems or services via the network 104. The processing unit 290 may also communicate to and from memory 280 and further provide output information for an optional display (not shown) via the input/output device interface 296. The input/output device interface 296 may also accept input from an optional input device (not shown).

[0046] The memory 280 may contain computer program instructions (grouped as modules in some embodiments) that the processing unit 290 executes in order to implement one or more aspects of the present disclosure. The memory 280 generally includes random access memory (RAM), read only memory (ROM) and/or other persistent, auxiliary or non-transitory computer readable media. The memory 280 may store an operating system 284 that provides computer program instructions for use by the processing unit 290 in the general administration and operation of the sidecar configuration system 160. The memory 280 may further include computer program instructions and other information for implementing aspects of the present disclosure. For example, in one embodiment, the memory 280 includes a user interface unit 282 that generates user interfaces (and/or instructions therefor) for display upon a computing device, e.g., via a navigation and/or browsing interface such as a browser or application installed on the computing device. In addition, the memory 280 may include and/or communicate with one or more data repositories (not shown), for example, to access user program codes and/or libraries.

[0047] In addition to and/or in combination with the user interface unit 282, the memory 280 may include a virtual machine configuration unit 162 and a sidecar configuration unit 164 that may be executed by the processing unit 290. In one embodiment, the virtual machine configuration unit 162 and the sidecar configuration unit 164 individually or collectively implement various aspects of the present disclosure, e.g., generating or selecting sidecar configurations within the on-demand code execution system 110, determining virtual machine configurations, etc., as described further below.

[0048] While the virtual machine configuration unit 162 and the sidecar configuration unit 164 are shown in FIG. 2 as part of the sidecar configuration system 160, in other embodiments, all or a portion of the virtual machine configuration unit 162 and the sidecar configuration unit 164 may be implemented by other components of the on-demand code execution system 110 and/or another computing device. For example, in certain embodiments of the present disclosure, another computing device in communication with the on-demand code execution system 110 may include several modules or components that operate similarly to the modules and components illustrated as part of the sidecar configuration system 160.

[0049] In some embodiments, the sidecar configuration system 160 may further include components other than those illustrated in FIG. 2. For example, the memory 280 may further include an instance allocation unit for allocating execution environments to tasks, user code execution unit to facilitate execution of tasks within the execution environments, or a container manager for managing creation, preparation, and configuration of containers within virtual machine instances.

[0050] With reference to FIG. 3, illustrative interactions are depicted for determining and configuring the sidecars for an execution of user-submitted code. The interactions of FIG. 3 begin at (1), where a user device 102 may generate a request to execute task code on an on-demand code execution system 110. Illustratively, the user may generate code whose execution requires various auxiliary functions, and thus requires a sidecar or sidecars that provide these functions. In some embodiments, the user may generate or identify a list of auxiliary functions that the user expects to require during execution of the task code. At (2), the user device 102 submits the request to the frontend 120, such as by using a API or other interface of the frontend 120. The request may include, for example, the task code and a list of sidecars or auxiliary functions. At (3), the frontend 120 validates the submitted task code. Validation can include, for example, verifying that the task code can be executed by the on-demand code execution system 110.

[0051] At (4), the frontend 120 transmits a request to the sidecar configuration system 160 to determine a set of sidecars for the task. Thereafter, at (5), the sidecar configuration system 160 determines a suitable set of sidecars. Illustratively, the virtual machine configuration unit 162 of the sidecar configuration system 160 may analyze the

request to identify a set of sidecar virtual machine instances that will facilitate executing the task code. The sidecar configuration unit 164 may then configure the virtual machine instance and the sidecars that will be needed during task code execution. For example, the sidecar configuration system 160 may determine that the task code will require input validation during execution. The virtual machine configuration unit 162 may thus configure a virtual machine instance to receive processed input from a sidecar, and the sidecar configuration unit 164 may identify and configure a sidecar to perform the necessary input validation. As a further example, the sidecar configuration system 160 may determine that the user wishes to profile execution of the task code to determine whether and how the code can be optimized. The virtual machine configuration unit 162 may thus configure a virtual machine instance to output profiling data during task execution, and the sidecar configuration unit 164 may configure a sidecar that aggregates and reports the profiling data.

[0052] At (6), the sidecar configuration system may store the determined sidecar configuration, and the validated task code, in a storage device such as the data storage device 108. The on-demand code execution system 110 may thus reduce the time spent analyzing code and determining sidecar configurations when receiving further requests to execute the task code, as discussed in more detail below. In some embodiments, the on-demand code execution system 110 may determine sidecars on a per-request basis. For example, the request may contain a debugging flag or other information that indicates whether to include a particular sidecar when executing the task code.

[0053] In some embodiments, the ordering and implementation of operations described above may be modified, or these interactions may be carried out by additional or alternative elements of the on-demand code execution system 110. For example, in one embodiment, the virtual machine configuration unit 162 and the sidecar configuration unit 164 may be combined, and the determinations made by these units may be collectively considered a “sidecar configuration” that includes configuration of the virtual machine that executes the task code. As a further example, in another embodiment, the user device 102 may request a particular sidecar configuration for the submitted task code, and the sidecar configuration system 160 may validate, process, and/or implement this request.

[0054] Illustrative interactions for utilizing a predetermined sidecar configuration in conjunction with executing tasks on the on-demand code execution system 110 will be

described with reference to FIG. 4. At (1), the user device 102 may request task execution. In some embodiments, as described above, the frontend 120 may initiate task execution without receiving a request, in which case the interaction at (1) may be omitted.

[0055] At (2), the frontend 120 distributes the task for execution to the worker manager 140. Prior to distributing the task execution, the frontend 120 may undertake any of a number of additional interactions, such as queuing the request, validating the request, etc., as described in more detail within the '556 Patent, incorporated by reference above.

[0056] At (3), the worker manager 140 requests a sidecar configuration for the task. In some embodiments, as described above, a sidecar configuration may be determined when the code of the task is submitted for validation (e.g., by carrying out the interactions illustrated in FIG. 3). In other embodiments, a sidecar configuration may be provided by the user when the code of the task is submitted for validation. In further embodiments, a sidecar configuration may be determined on a per-request basis. At (4), the worker manager 140 receives the previously determined (or previously specified) sidecar configuration from the data storage device 108.

[0057] Thereafter, at (5), the worker manager 140 configures and executes a virtual machine instance and sidecars in accordance with the received sidecar configuration. In some embodiments, as described above, the worker manager 140 may obtain sidecar images from a library, such as the sidecar library 130 of FIG. 1, and configure these images in accordance with the configuration. In other embodiments, the worker manager 140 may obtain fully or partially preconfigured sidecars from a warming pool, and may perform additional configurations as needed (e.g., to cause the sidecar to communicate with a particular virtual machine instance). In still further embodiments, the worker manager 140 may obtain multiple virtual machine instances from a warming pool, and may configure some of the instances to execute task code and configure other instances to be sidecars (e.g., by provisioning the sidecar instances with agents that perform auxiliary functions).

[0058] Illustrative interactions for managing the lifecycles of sidecars in conjunction with the lifecycle of a “main” virtual machine instance (e.g., one that executes task code) will be described with reference to FIG. 5. At (1), the sidecar lifecycle management unit 142 monitors the execution of the virtual machine instance 150. Illustratively, the sidecar lifecycle management unit 142 may monitor inputs, outputs,

resource utilization, metrics generated by the virtual machine instance 150, metrics generated by the worker manager 140, or other information relating to the status of the virtual machine instance 150.

[0059] At (2), the sidecar lifecycle management unit 142 detects that the virtual machine instance 150 is about to undergo a change in execution state. Changes in execution state may include, for example, provisioning the virtual machine instance 150, starting execution of the task code, suspending execution of the task code, resuming execution of the task code, terminating execution of the task code, suspending the virtual machine instance 150, resuming the virtual machine instance 150, deprovisioning the virtual machine instance 150, and the like. Terminating execution of the task code may include both normal termination (e.g., the task code executing to completion) and abnormal termination (e.g., the task code halting execution due to an error, because the user requested termination, because a signal was received from the operating system, etc.). In some embodiments, the sidecar lifecycle management unit 142 detects a pending change in execution state prior to its occurrence. For example, the sidecar lifecycle management unit 142 may receive a notification that the worker manager 140 is provisioning, or is about to provision, the virtual machine instance 150. In other embodiments, the sidecar lifecycle management unit 142 may detect the change while or after it occurs. For example, the sidecar lifecycle management unit 142 may detect that the virtual machine instance has suspended or terminated execution of the task code because the task code has processed all of the available input data, because a threshold amount of computing resources has been consumed, or because the task code has exited with an error message.

[0060] Thereafter, at (3), the sidecar lifecycle management unit 142 determines what changes should occur to the execution states of sidecars associated with the virtual machine instance 150, such as sidecars 152 and 154. The changes in execution states for the sidecars may be determined based on multiple factors, such as the number of virtual machine instances associated with the sidecar, the auxiliary function(s) performed by the sidecar, and so forth. For example, the sidecar 152 may provide output validation for the virtual machine instance 150. The sidecar lifecycle management unit 142 may therefore determine that the sidecar 152 can enter an inactive state when the virtual machine instance 150 enters an inactive state, since the output validation function will not be required. As another example,

the sidecar 154 may provide control plane functions to both the virtual machine instance 150 and another virtual machine instance, which may execute the same task code or different task code on behalf of the same user. The sidecar lifecycle management unit 142 may thus determine that the sidecar 154 should only be deprovisioned if both of the virtual machine instances it supports have entered a deprovisioned state.

[0061] In some embodiments, the sidecar lifecycle management unit 142 may determine an order in which sidecar(s) and the virtual machine instance 150 should change their execution states, and may cause the execution states to change in the determined order. For example, the sidecar 154 may provide logging or debugging functions for the virtual machine instance 150. The sidecar lifecycle management unit 142 may thus determine that the sidecar 154 should enter a suspended execution state only after the virtual machine instance 150 does so, and should resume execution before the virtual machine instance 150 does so, in order to permit the sidecar 154 to maintain a complete record of logging or debugging information. As another example, the sidecar 152 may process and validate input data before sending it to the virtual machine instance 150, and thus the virtual machine instance 150 must be active before the sidecar 152 begins sending data to it. In some embodiments, the sidecar lifecycle management unit 142 may not have control over the timing of changes in the execution state of the virtual machine instance 150, and may thus schedule and prioritize changes to sidecar execution states based on the timing of changes in the execution state of the virtual machine instance 150.

[0062] In the example illustrated in FIG. 5, the sidecar lifecycle management unit 142 determines that a change in the execution state of the sidecar 152 should precede the change in the execution state of the virtual machine instance 150, and that a change in the execution state of the sidecar 154 should follow the change to the virtual machine instance 150. The sidecar lifecycle management unit 142 therefore, at (4), causes the sidecar 152 to change its execution state. In various embodiments, the sidecar lifecycle management unit 142 may cause the change in execution state by sending instructions to the sidecar 152, the worker manager 140, or another component of the on-demand code execution system 110. At (5), the sidecar 152 (or another component) implements the change in execution state.

[0063] Thereafter, at (6), the virtual machine instance 150 implements its change in execution state. In some embodiments, the worker manager 140 implements the change in

the virtual machine instance 150 execution state, for example by provisioning or deprovisioning the virtual machine instance 150. In other embodiments, the virtual machine instance 150 may implement the change. For example, the virtual machine instance 150 may begin, end, suspend, or resume execution of the task code.

[0064] At (7), the sidecar lifecycle management unit 142 causes a change in the execution state of the sidecar 154, in similar fashion to the interaction at (4). Thereafter, at (8), the sidecar 154 (or another component of the on-demand code execution system 110) implements the change in execution state for the sidecar 154.

[0065] In some embodiments, one or more of the interactions at (4), (6), and (7) may be consolidated into a single interaction. For example, the sidecar lifecycle management unit 142 may provide the worker manager 140 with a schedule or an ordered list for changing the execution states of the sidecar 152, virtual machine instance 150, and sidecar 154, and the worker manager 140 may implement the execution state changes in the indicated order. As another example, the sidecar lifecycle management unit 142 may indicate to the worker manager 140 that the sidecar 152 should resume execution prior to the virtual machine instance 150, and the sidecar 154 should resume after the virtual machine instance 150 resumes execution. Additionally, as discussed above, in some embodiments the interaction at (2) may detect a change that is occurring or has already occurred, rather than detecting a change that is about to occur.

[0066] FIG. 6 depicts an illustrative example of interactions between a virtual machine instance 604 and sidecars 606, 608, and 610. In the illustrated example, at (1), external input is provided to an input validation sidecar 606. The external input may be received from the frontend 120, or in some embodiments from another input source. At (2), the input validation sidecar 606 may validate the external input. For example, the input validation sidecar 606 may sanitize the external input (e.g., by inserting escape characters or removing special characters) or verify that the external input is in a format expected by the task code.

[0067] Thereafter, at (3), the input validation sidecar 606 may provide validated input to the virtual machine instance 604. In some embodiments, the virtual machine instance 604 may be configured to receive input from the input validation sidecar 606. In other embodiments, the sidecar 606 may be configured to transmit processed input to the

virtual machine instance 604, and no special configuration is performed on the virtual machine instance 604: Instead, the task code executing on the virtual machine instance 604 simply processes any input it receives, and the configuration of only receiving input from sidecar 606 is transparent to the virtual machine instance 604. In further embodiments, neither the input validation sidecar 606 nor the virtual machine instance 604 may be configured to be aware of the other, and the communication of validated data from the sidecar 606 to the virtual machine instance 604 may be handled by the worker manager 140.

[0068] At (4), the virtual machine instance 604 may process the validated input. Illustratively, the virtual machine instance 604 may process the validated input by executing the task code with the validated input, and may generate corresponding output. The virtual machine instance 604 may also, at (5), provide access to information that allows a profiling sidecar 608 to profile the task code. For example, the virtual machine instance 604 may generate log files, metrics, stack traces, memory dumps, or other data relating to the execution of the task code. The virtual machine instance 604 may then transmit the data, provide APIs for accessing the data, allow the profiling sidecar 608 to read filesystems or memory locations, or otherwise grant access to the profiling information. At (6), the profiling sidecar may obtain and/or receive the profiling information and process it to profile the virtual machine instance 604 and/or the task code executing on the virtual machine instance 604.

[0069] At (7), the virtual machine instance 604 may provide the output generated at (4) to an output validation sidecar 610, which at (8) may validate the output. For example, the output validation sidecar 610 may post-process the output, convert it to a particular format, verify that it falls within a specified range, or perform other validations.

[0070] At (9), in some embodiments, the output validation sidecar 610 may provide validated output to the frontend 120 or another component. In some embodiments, the interactions at (1) – (9) may be repeated until no further input is available or the user issues a request to stop executing the task code. Additionally, the example interactions depicted in FIG. 6 may be carried out in different orders, or more or fewer interactions may be carried out. For example, the interactions at (5) and (6) may be omitted, or may be carried out prior to (or concurrently with) the interaction at (4). As a further example, in some embodiments, the validated input produced by the input validation sidecar 606 may be

provided as input to another input validation sidecar, which may perform further validations before sending the resulting data to the virtual machine instance 604. It will thus be understood by one skilled in the art that many other combinations and arrangements of sidecars are within the scope of the present disclosure.

[0071] With reference to FIG. 7, one illustrative routine 700 for determining a sidecar configuration for executing a task on the on-demand code execution system 110 will be described. The routine 700 may be carried out, for example, by the sidecar configuration system 160 of FIG. 1. In one embodiment, the routine 700 (e.g., blocks 702-706) is carried out in conjunction with the creation of a task on the on-demand code execution system 110. In another embodiment, all or part of the routine 700 (e.g., blocks 708-716) may be carried out in conjunction with executing the task on the on-demand code execution system 110. All or part of the routine 700 may additionally or alternatively be carried out at other times, such as periodically (e.g., once per day, week, month, etc.) or based on detection that execution metrics or resource utilization metrics fall outside a threshold.

[0072] The routine 700 begins at block 702, where the sidecar configuration system 160 obtains code for a task (e.g., as submitted by a user). In one embodiment, the code for the task is represented as a code object, such as a compressed file including source code for the task. The routine 700 then continues to block 704, where the sidecar configuration system 160 determines an sidecar configuration for the task. As described above, the sidecar configuration system 160 may, in some embodiments, determine a sidecar configuration based on the submitted task code. In other embodiments, the sidecar configuration system 160 may determine a sidecar configuration based on user input, such as a request to profile execution of the task code or validate input provided to the task code. In some embodiments, the frontend 120 or another component of the on-demand code execution system 110 may present user interfaces that enable a user to select desired sidecars.

[0073] At block 706, the sidecar configuration system 160 may store the task code and the sidecar configuration in a data store, such as the data store service 108. In some embodiments, as described above, the sidecar configuration may include a configuration of the virtual machine instance that executes the task code. In other embodiments, the sidecar configuration may include instructions utilized by the worker manager 140.

[0074] Thereafter, at block 708, a request to execute the task code may be received. The request may be received by, for example, the frontend 120, the worker manager 140, or another component of the on-demand code execution system 110. In some embodiments, as described above, the task code may be executed without receiving a request, in which case block 708 may be omitted.

[0075] At block 710, the worker manager 140 provisions a virtual machine instance to execute the task code. In some embodiments, the worker manager 140 configures the virtual machine instance in accordance with the sidecar configuration. For example, the worker manager 140 may configure the virtual machine instance to accept input from a sidecar, provide output to a sidecar, provide access to internals of the virtual machine (e.g., stack or memory contents), or otherwise configure the virtual machine instance to work with sidecars. In other embodiments, the virtual machine instance may not be configured and the interactions with sidecars may be transparent to the virtual machine instance.

[0076] At block 712, the worker manager 140 provisions one or more sidecars in accordance with the sidecar configuration. In some embodiments, as described above, the worker manager 140 obtains sidecar images or agents from a library, such as the sidecar library 130, and uses these images or agents to provision and configure the sidecars. In other embodiments, the sidecar configuration may include instructions for provisioning a sidecar to user specifications, and the worker manager 140 may carry out these instructions to provision the sidecar accordingly.

[0077] At block 714, the worker manager 140 executes the virtual machine instance and the sidecar instance(s). In some embodiments, the worker manager 140 proceeds to manage the lifecycles of the virtual machine instance and the sidecar instance(s) by carrying out a routine such as the routine 800 described below with reference to FIG. 8. In other embodiments, the routine 700 ends with the execution of the provisioned virtual machine instances and sidecar(s).

[0078] With reference to FIG. 8, an illustrative routine 800 for managing sidecar lifecycles will be described. The routine 800 may be carried out, for example, by the sidecar lifecycle management unit 142 of FIG. 1. The routine 800 begins at block 802, where the sidecar lifecycle management unit 142 (or, in some embodiments, the worker manager 140 or another component of the on-demand code execution system 110) monitors execution of a

virtual machine instance. At decision block 804, a determination is made as to whether there is a pending change in the virtual machine instance's execution state. In some embodiments, as discussed above, the determination at decision block 804 may also include whether a change to the virtual machine instance's execution state has already occurred or is in progress. If no change in the virtual machine instance's execution state is detected, then the routine 800 branches to block 802 and continues monitoring.

[0079] If the determination at decision block 804 is that the virtual machine instance will change its execution state, then the routine 800 branches to block 806, where an unprocessed sidecar may be selected. At block 808, the sidecar lifecycle management unit 142 may determine whether the selected sidecar should change its execution state, and if so when it should change its execution state relative to the virtual machine instance. For example, the sidecar lifecycle management unit 142 may determine that the sidecar should change its execution state before, with, or after the virtual machine instance. In some embodiments, the sidecar lifecycle management unit 142 may determine a timing for the change to the selected sidecar's execution state relative to other sidecars. For example, the sidecar lifecycle management unit 142 may generate an ordered list of sidecar and virtual machine instance state changes, and may place the selected sidecar on the ordered list. In other embodiments, the sidecar lifecycle management unit 142 may determine that the selected sidecar does not require a change to its execution state, or that the selected sidecar will change to a different execution state than the virtual machine instance. For example, the sidecar lifecycle management unit 142 may determine that a sidecar should be deprovisioned after the virtual machine instance suspends execution of the task code, or that a new sidecar should be provisioned before the virtual machine instance resumes execution of the task code. In one embodiment, the sidecar lifecycle management unit 142 may ensure that a new sidecar virtual machine be created for each execution of task code within the virtual machine instance. Illustratively, where a sidecar virtual machine provides a security function, regeneration of a new sidecar for each execution of code may ensure that the function of the sidecar is not compromised, for example, by malicious data being passed to the sidecar virtual machine during a past code execution.

[0080] At decision block 810, the sidecar lifecycle management unit 142 may determine whether all of the sidecars have been processed by the routine 800. If not, the

routine 800 branches to block 806, where one of the remaining unprocessed sidecars may be selected, and the routine 800 then iterates through blocks 806-810 until all sidecars have been processed. The routine 800 then branches to block 812, where the sidecar lifecycle management unit 142 causes the virtual machine instance and the sidecars to change execution states in the determined order. In some embodiments, the determined order may be relative to the virtual machine instance's change, and may not specify a particular order among the sidecars. For example, the sidecar lifecycle management unit 142 may determine that sidecars A and B should both change their execution states before the virtual machine instance does, but that it does not matter sidecar A's execution state change comes before or after sidecar B's. In other embodiments, the sidecar lifecycle management unit 142 may determine that one sidecar should change its execution state before another sidecar does so, and change the execution states accordingly.

[0081] The blocks of the routines described above may vary in embodiments of the present disclosure. For example, in some implementations of either routine, the sidecar configuration system 160 may be configured to implement additional functionality, such as generating and storing additional sidecars. Thus, the routines may include additional blocks, or the blocks of the routines may be rearranged, according to various embodiments.

[0082] All of the methods and processes described above may be embodied in, and fully automated via, software code modules executed by one or more computers or processors. The code modules may be stored in any type of non-transitory computer-readable medium or other computer storage device. Some or all of the methods may alternatively be embodied in specialized computer hardware.

[0083] Conditional language such as, among others, "can," "could," "might" or "may," unless specifically stated otherwise, are otherwise understood within the context as used in general to present that certain embodiments include, while other embodiments do not include, certain features, elements and/or steps. Thus, such conditional language is not generally intended to imply that features, elements and/or steps are in any way required for one or more embodiments or that one or more embodiments necessarily include logic for deciding, with or without user input or prompting, whether these features, elements and/or steps are included or are to be performed in any particular embodiment.

[0084] Disjunctive language such as the phrase “at least one of X, Y or Z,” unless specifically stated otherwise, is otherwise understood with the context as used in general to present that an item, term, etc., may be either X, Y or Z, or any combination thereof (e.g., X, Y and/or Z). Thus, such disjunctive language is not generally intended to, and should not, imply that certain embodiments require at least one of X, at least one of Y or at least one of Z to each be present.

[0085] Unless otherwise explicitly stated, articles such as ‘a’ or ‘an’ should generally be interpreted to include one or more described items. Accordingly, phrases such as “a device configured to” are intended to include one or more recited devices. Such one or more recited devices can also be collectively configured to carry out the stated recitations. For example, “a processor configured to carry out recitations A, B and C” can include a first processor configured to carry out recitation A working in conjunction with a second processor configured to carry out recitations B and C.

[0086] Any routine descriptions, elements or blocks in the flow diagrams described herein and/or depicted in the attached figures should be understood as potentially representing modules, segments, or portions of code which include one or more executable instructions for implementing specific logical functions or elements in the routine. Alternate implementations are included within the scope of the embodiments described herein in which elements or functions may be deleted, or executed out of order from that shown or discussed, including substantially synchronously or in reverse order, depending on the functionality involved as would be understood by those skilled in the art.

[0087] It should be emphasized that many variations and modifications may be made to the above-described embodiments, the elements of which are to be understood as being among other acceptable examples. All such modifications and variations are intended to be included herein within the scope of this disclosure and protected by the following claims.

[0088] Examples of the embodiments of the present disclosure can be described in view of the following clauses:

[0089] Clause 1. A system comprising: a physical data store storing configuration information; and a computing device configured with executable instructions to: receive a request to execute user-submitted code on an on-demand code execution system; determine,

based at least in part on the request, a plurality of virtual machine instances, the plurality including a first virtual machine instance to execute the user-submitted code and a second virtual machine instance to perform an auxiliary function; cause the first virtual machine instance to execute the user-submitted code in accordance with the configuration; cause the second virtual machine instance to perform the auxiliary function in accordance with the configuration; monitor execution of the user-submitted code on the first virtual machine instance to detect a pending change in execution state for the first virtual machine instance; determine, based at least in part on the pending change in execution state for the first virtual machine instance, a change in execution state for the second virtual machine instance; determine an order for execution state changes for the first virtual machine instance and the second virtual machine instance; and cause the execution states for the first virtual machine instance and the second virtual machine instance to change in the determined order.

[0090] Clause 2. The system of Clause 1, wherein the pending change in execution state for the first virtual machine instance comprises one or more of provisioning the first virtual machine instance, starting execution of the user-submitted code, suspending the first virtual machine instance, suspending execution of the user-submitted code, resuming the first virtual machine instance, resuming execution of the user-submitted code, terminating execution of the user-submitted code, or deprovisioning the first virtual machine instance.

[0091] Clause 3. The system of Clause 1, wherein the computing device is further configured to provision the first virtual machine instance and the second virtual machine instance.

[0092] Clause 4. The system of Clause 1, wherein the order for changing the execution states for the first virtual machine instance and the second virtual machine instance is determined based at least in part on the auxiliary function performed by the second virtual machine instance.

[0093] Clause 5. The system of Clause 1, wherein the auxiliary function comprises one or more of encapsulation, logging, tracing, debugging, scanning, profiling, validating input, or validating output.

[0094] Clause 6. A computer-implemented method comprising: receiving a request to execute user-submitted code on an on-demand code execution system; obtaining

configuration information associated with the user-submitted code, the configuration information specifying at least a first virtual machine instance configured to execute the user-submitted code and a second virtual machine instance configured to perform an auxiliary function; causing the first virtual machine instance to execute the user-submitted code in accordance with the configuration information; causing the second virtual machine instance to perform the auxiliary function in accordance with the configuration information; identifying, based at least in part on execution of the user-submitted code on the first virtual machine instance, a pending change in execution state for the first virtual machine instance; determining, based at least in part on the pending change in execution state for the first virtual machine instance, a change in execution state for the second virtual machine instance; determining an order for execution state changes for at least the first virtual machine instance and the second virtual machine instance; and causing the execution states for the first virtual machine instance and the second virtual machine instance to change in the determined order.

[0095] Clause 7. The computer-implemented method of Clause 6 further comprising determining a provisioning order for the first virtual machine instance and the second virtual machine instance, wherein causing the first virtual machine instance to execute the user-submitted code and causing the second virtual machine instance to perform the auxiliary function are in accordance with the provisioning order.

[0096] Clause 8. The computer-implemented method of Clause 6, wherein the order for changing the execution states specifies that the change in execution state for the second virtual machine instance must precede the pending change in execution state for the first virtual machine instance.

[0097] Clause 9. The computer-implemented method of Clause 6 further comprising causing the pending change in execution state for the first virtual machine instance.

[0098] Clause 10. The computer-implemented method of Clause 6, wherein the configuration information further specifies at least a third virtual machine instance configured to perform a second auxiliary function, and wherein the determined order for execution state changes further includes an execution state change for the third virtual machine instance.

[0099] Clause 11. The computer-implemented method of Clause 10 further comprising causing the execution state changes for the first virtual machine instance, the second virtual machine instance, and the third virtual machine instance to occur in the determined order.

[0100] Clause 12. The computer-implemented method of Clause 6 further comprising configuring the first virtual machine instance to communicate with the second virtual machine instance.

[0101] Clause 13. The computer-implemented method of Clause 12, wherein configuring the first virtual machine instance to communicate with the second virtual machine instance comprises one or more of redirecting input, redirecting output, or enabling access to performance information.

[0102] Clause 14. The computer-implemented method of Clause 6, wherein the second virtual machine instance is configured to perform the auxiliary function for a plurality of virtual machine instances associated with a specified user.

[0103] Clause 15. The computer-implemented method of Clause 14, wherein each of the plurality of virtual machine instances is associated with different user-submitted code.

[0104] Clause 16. Non-transitory computer-readable media including computer-executable instructions that, when executed on an on-demand code execution system, cause the on-demand code execution system to: obtain configuration information associated with user-submitted code that is executable on an on-demand code execution system, the configuration information including a first configuration associated with a first virtual machine instance and a second configuration associated with a second virtual machine instance; cause the first virtual machine instance to execute the user-submitted code in accordance with the first configuration; cause the second virtual machine instance to perform an auxiliary function in accordance with the second configuration; determine, based at least in part on a pending change in execution state for the first virtual machine instance, a change in execution state for the second virtual machine instance; and cause execution states for the first virtual machine instance and the second virtual machine instance to change in an order determined based at least in part on the configuration information.

[0105] Clause 17. The non-transitory computer-readable media of Clause 16 wherein the instructions further cause the on-demand code execution system to receive a request to execute the user-submitted code.

[0106] Clause 18. The non-transitory computer-readable media of Clause 16 wherein the instructions further cause the on-demand code execution system to identify the pending change in execution state.

[0107] Clause 19. The non-transitory computer-readable media of Clause 16 wherein the instructions further cause the on-demand code execution system to instantiate the second virtual machine instance.

[0108] Clause 20. The non-transitory computer-readable media of Clause 16 wherein the instructions further cause the on-demand code execution system to select the second virtual machine instance from a plurality of preconfigured virtual machine instances.

[0109] Clause 21. A system comprising: a physical data store storing configuration information; and a computing device configured with executable instructions to: receive a request to execute a task on an on-demand code execution system, the request including user-submitted code and identifying unprocessed data to be processed by the user-submitted code; determine, based at least in part on the request, that a first virtual machine instance will be required to execute the user-submitted code and a second virtual machine instance will be required to pre-process the unprocessed data; provision the first virtual machine; configure the first virtual machine instance to: receive processed data from the second virtual machine; and execute the user-submitted code with the processed data; provision the second virtual machine instance; configure the second virtual machine instance to: process the unprocessed data to form the processed data; and transmit the processed data to the first virtual machine instance; determine that the user-submitted code has completed execution on the first virtual machine instance; and in response to the determination that the user-submitted code has completed execution, deprovision the first virtual machine instance and the second virtual machine instance.

[0110] Clause 22. The system of Clause 2 **Error! Reference source not found.**, wherein the second virtual machine instance performs at least one of validating, sanitizing, pre-processing, or scanning the unprocessed data to form the processed data.

[0111] Clause 23. The system of Clause 21, wherein the computing device is further configured to cause the first virtual machine instance to transmit results of executing the user-submitted code with the processed data to a third virtual machine instance.

[0112] Clause 24. The system of Clause 23, wherein the computing device is further configured to cause the third virtual machine instance to perform at least one of validating, sanitizing, post-processing, or scanning the results.

[0113] Clause 25. The system of Clause 23, wherein the computing device is further configured to deprovision the third virtual machine instance in response to the determination that the user-submitted code has completed execution.

[0114] Clause 26. A computer-implemented method comprising: receiving a request to execute user-submitted code on an on-demand code execution system; obtaining, based at least in part on the request, information identifying a plurality of virtual machine instances; provisioning a first virtual machine instance of the plurality of virtual machine instances; configuring the first virtual machine instance to: execute the user-submitted code; and transmit results of executing the user-submitted code to the second virtual machine instance; provisioning a second virtual machine instance of the plurality of virtual machine instances; configuring the second virtual machine instance to process the results from the first virtual machine instance to form processed results; and in response to a determination that the first virtual machine instance has ended execution of the user-submitted code, deprovisioning the second virtual machine instance.

[0115] Clause 27. The computer-implemented method of Clause 26, wherein deprovisioning the second virtual machine instance is further in response to a determination that the second virtual machine instance has finished processing the results.

[0116] Clause 28. The computer-implemented method of Clause 26 further comprising deprovisioning the first virtual machine instance in response to the determination that the first virtual machine instance has ended execution of the user-submitted code.

[0117] Clause 29. The computer-implemented method of Clause 26 further comprising causing the first virtual machine instance to receive processed data from a third virtual machine instance of the plurality of virtual machine instances, wherein the user-submitted code is executed with the processed data.

[0118] Clause 30. The computer-implemented method of Clause 29 further comprising provisioning the third virtual machine instance.

[0119] Clause 31. The computer-implemented method of Clause 29, wherein the third virtual machine instance is a preexisting virtual machine instance, the method further comprising configuring the third virtual machine instance to transmit the processed data to the first virtual machine instance.

[0120] Clause 32. The computer-implemented method of Clause 31 further comprising deconfiguring the third virtual machine in response to the determination that the first virtual machine instance has ended execution of the user-submitted code.

[0121] Clause 33. The computer-implemented method of Clause 29 further comprising deprovisioning the third virtual machine instance in response to one or more of: the determination that the first virtual machine instance has ended execution of the user-submitted code; or a determination that the third virtual machine instance has transmitted all of the processed data.

[0122] Clause 34. The computer-implemented method of Clause 26 further comprising determining, based at least in part on the request, the information identifying the plurality of virtual machines.

[0123] Clause 35. Non-transitory computer-readable media including computer-executable instructions that, when executed on an on-demand code execution system, cause the on-demand code execution system to: receive a request to execute user-submitted code; obtain, based at least in part on the request, configuration information identifying a plurality of virtual machine instances; configure a first virtual machine instance of the plurality of virtual machine instances to execute the user-submitted code in accordance with the configuration information; configure a second virtual machine instance of the plurality of virtual machine instances to perform at least one of transmitting input to or receiving output from the first virtual machine instance in accordance with the configuration information; and in response to a determination that the first virtual machine instance has ended execution of the user-submitted code, deconfigure the second virtual machine instance.

[0124] Clause 36. The non-transitory computer-readable media of Clause 35 wherein the instructions further cause the on-demand code execution system to provision the first virtual machine instance.

[0125] Clause 37. The non-transitory computer-readable media of Clause 35 wherein the instructions further cause the on-demand code execution system to determine that the first virtual machine instance has ended execution of the user-submitted code.

[0126] Clause 38. The non-transitory computer-readable media of Clause 35, wherein the instructions cause the on-demand code execution system to configure the second virtual machine instance to transmit input to and receive output from the first virtual machine instance.

[0127] Clause 39. The non-transitory computer-readable media of Clause 35, wherein the second virtual machine instance is reset to a default configuration in response to the determination that the first virtual machine instance has ended execution of the user-submitted code.

[0128] Clause 40. The non-transitory computer-readable media of Clause 35, wherein the first virtual machine instance is reset to a default configuration in response to the determination that the first virtual machine instance has ended execution of the user-submitted code.

CLAIMS

WHAT IS CLAIMED IS:

1. A system comprising:
 - a physical data store storing configuration information; and
 - a computing device configured with executable instructions to:
 - receive a request to execute user-submitted code on an on-demand code execution system;
 - determine, based at least in part on the request, plurality of virtual machine instances, the plurality including a first virtual machine instance to execute the user-submitted code and a second virtual machine instance to perform an auxiliary function;
 - cause the first virtual machine instance to execute the user-submitted code in accordance with the configuration;
 - cause the second virtual machine instance to perform the auxiliary function in accordance with the configuration;
 - monitor execution of the user-submitted code on the first virtual machine instance to detect a pending change in execution state for the first virtual machine instance;
 - determine, based at least in part on the pending change in execution state for the first virtual machine instance, a change in execution state for the second virtual machine instance;
 - determine an order for execution state changes for the first virtual machine instance and the second virtual machine instance; and
 - cause the execution states for the first virtual machine instance and the second virtual machine instance to change in the determined order.
2. The system of Claim 1, wherein the pending change in execution state for the first virtual machine instance comprises one or more of provisioning the first virtual machine instance, starting execution of the user-submitted code, suspending the first virtual machine instance, suspending execution of the user-submitted code, resuming the first virtual machine instance, resuming execution of the user-submitted code, terminating execution of the user-submitted code, or deprovisioning the first virtual machine instance.

3. The system of Claim 1, wherein the order for changing the execution states for the first virtual machine instance and the second virtual machine instance is determined based at least in part on the auxiliary function performed by the second virtual machine instance.

4. The system of Claim 1, wherein the auxiliary function comprises one or more of encapsulation, logging, tracing, debugging, scanning, profiling, validating input, or validating output.

5. A computer-implemented method comprising:

receiving a request to execute user-submitted code on an on-demand code execution system;

obtaining configuration information associated with the user-submitted code, the configuration information specifying at least a first virtual machine instance configured to execute the user-submitted code and a second virtual machine instance configured to perform an auxiliary function;

causing the first virtual machine instance to execute the user-submitted code in accordance with the configuration information;

causing the second virtual machine instance to perform the auxiliary function in accordance with the configuration information;

identifying, based at least in part on execution of the user-submitted code on the first virtual machine instance, a pending change in execution state for the first virtual machine instance;

determining, based at least in part on the pending change in execution state for the first virtual machine instance, a change in execution state for the second virtual machine instance;

determining an order for execution state changes for at least the first virtual machine instance and the second virtual machine instance; and

causing the execution states for the first virtual machine instance and the second virtual machine instance to change in the determined order.

6. The computer-implemented method of Claim 5 further comprising determining a provisioning order for the first virtual machine instance and the second virtual machine instance, wherein causing the first virtual machine instance to execute the user-

submitted code and causing the second virtual machine instance to perform the auxiliary function are in accordance with the provisioning order.

7. The computer-implemented method of Claim 5, wherein the order for changing the execution states specifies that the change in execution state for the second virtual machine instance must precede the pending change in execution state for the first virtual machine instance.

8. The computer-implemented method of Claim 5 further comprising causing the pending change in execution state for the first virtual machine instance.

9. The computer-implemented method of Claim 5, wherein the configuration information further specifies at least a third virtual machine instance configured to perform a second auxiliary function, and wherein the determined order for execution state changes further includes an execution state change for the third virtual machine instance.

10. The computer-implemented method of Claim 5 further comprising configuring the first virtual machine instance to communicate with the second virtual machine instance, wherein configuring the first virtual machine instance to communicate with the second virtual machine instance comprises one or more of redirecting input, redirecting output, or enabling access to performance information.

11. The computer-implemented method of Claim 5, wherein the second virtual machine instance is configured to perform the auxiliary function for a plurality of virtual machine instances associated with a specified user, and wherein each of the plurality of virtual machine instances is associated with different user-submitted code.

12. Non-transitory computer-readable media including computer-executable instructions that, when executed on an on-demand code execution system, cause the on-demand code execution system to:

obtain configuration information associated with user-submitted code that is executable on an on-demand code execution system, the configuration information including a first configuration associated with a first virtual machine instance and a second configuration associated with a second virtual machine instance;

cause the first virtual machine instance to execute the user-submitted code in accordance with the first configuration;

cause the second virtual machine instance to perform an auxiliary function in accordance with the second configuration;

determine, based at least in part on a pending change in execution state for the first virtual machine instance, a change in execution state for the second virtual machine instance; and

cause execution states for the first virtual machine instance and the second virtual machine instance to change in an order determined based at least in part on the configuration information.

13. The non-transitory computer-readable media of Claim 12 wherein the instructions further cause the on-demand code execution system to receive a request to execute the user-submitted code.

14. The non-transitory computer-readable media of Claim 12 wherein the instructions further cause the on-demand code execution system to instantiate the second virtual machine instance.

15. The non-transitory computer-readable media of Claim 12 wherein the instructions further cause the on-demand code execution system to select the second virtual machine instance from a plurality of preconfigured virtual machine instances.

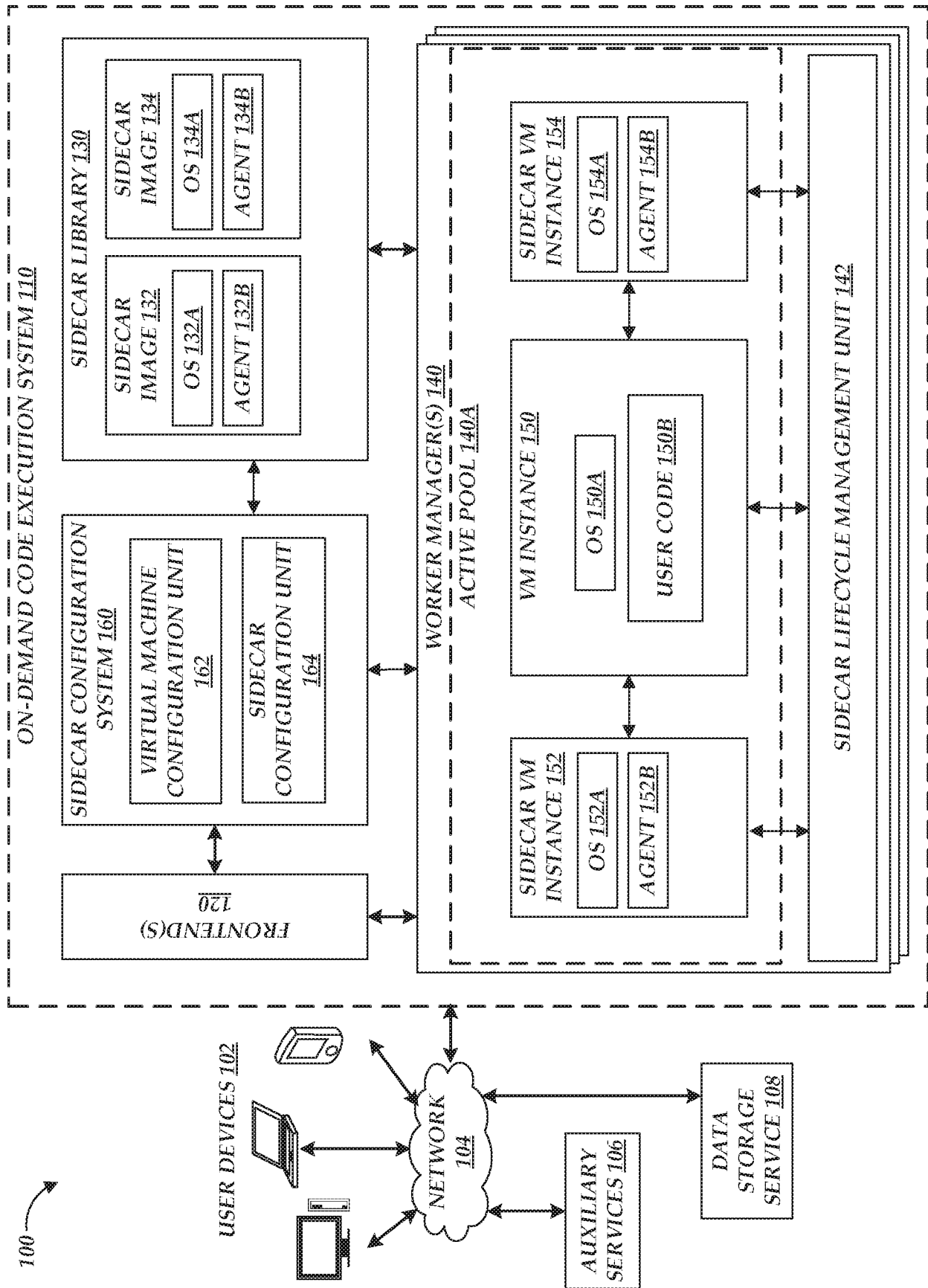


Fig. 1

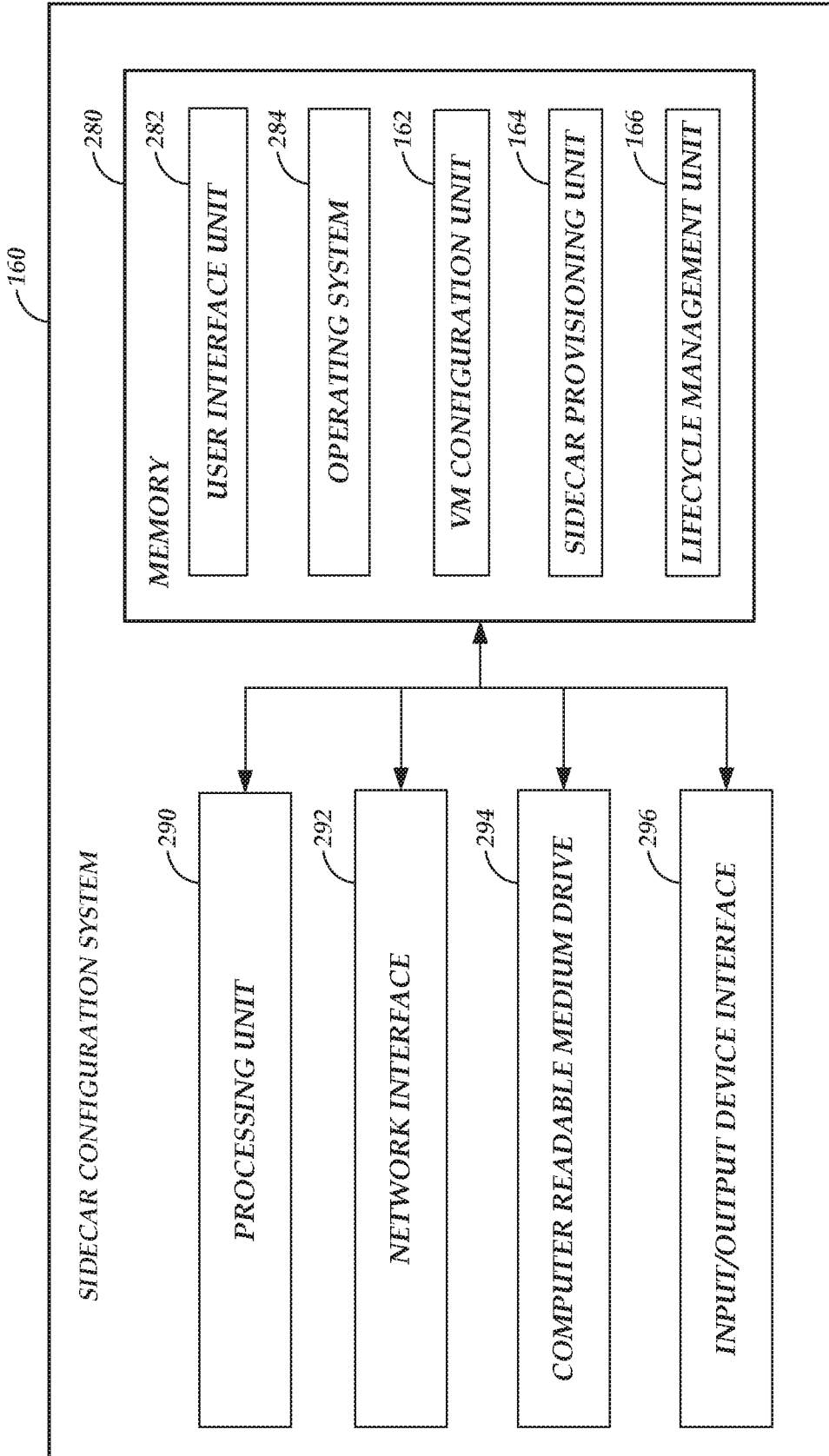


Fig. 2

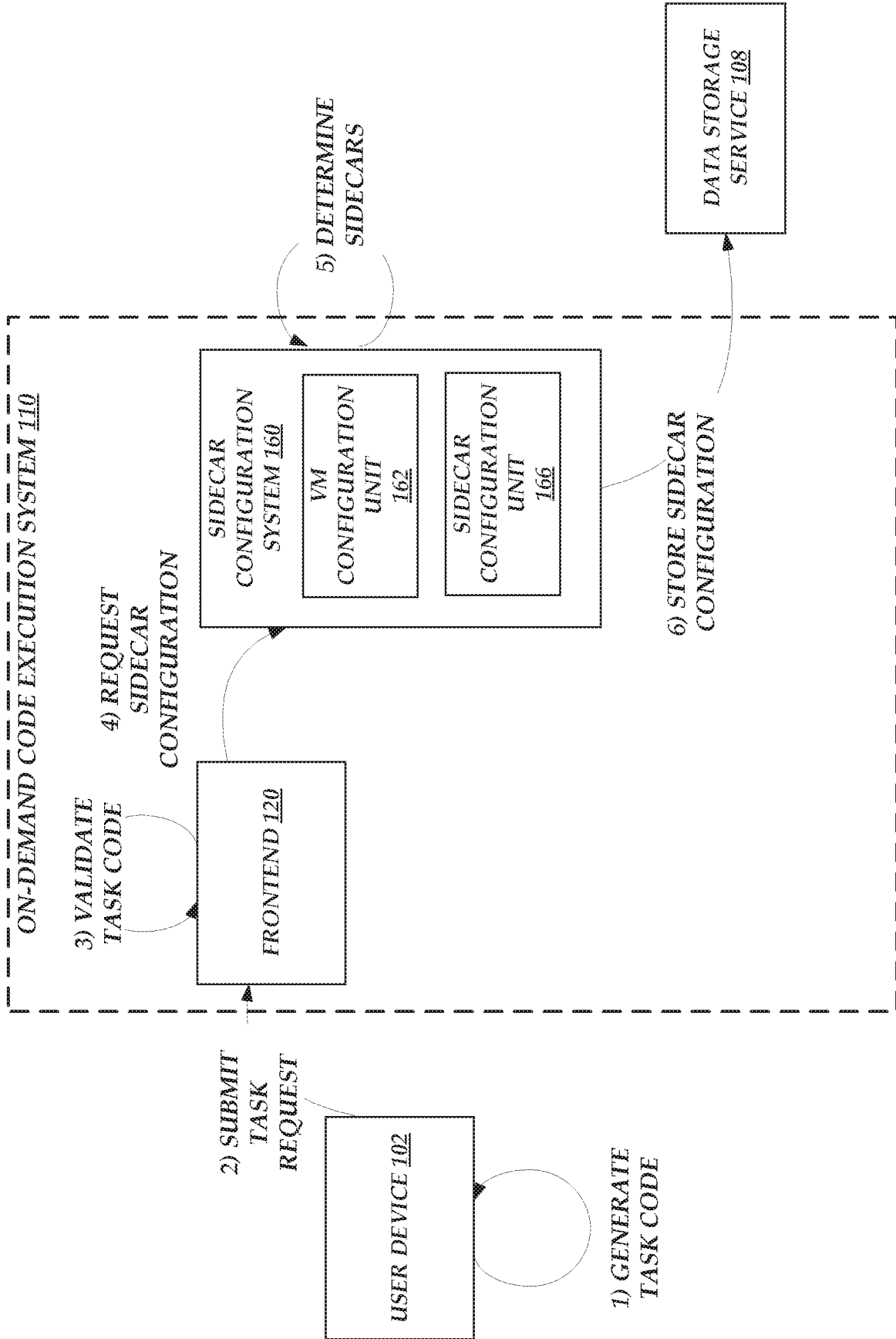


Fig. 3

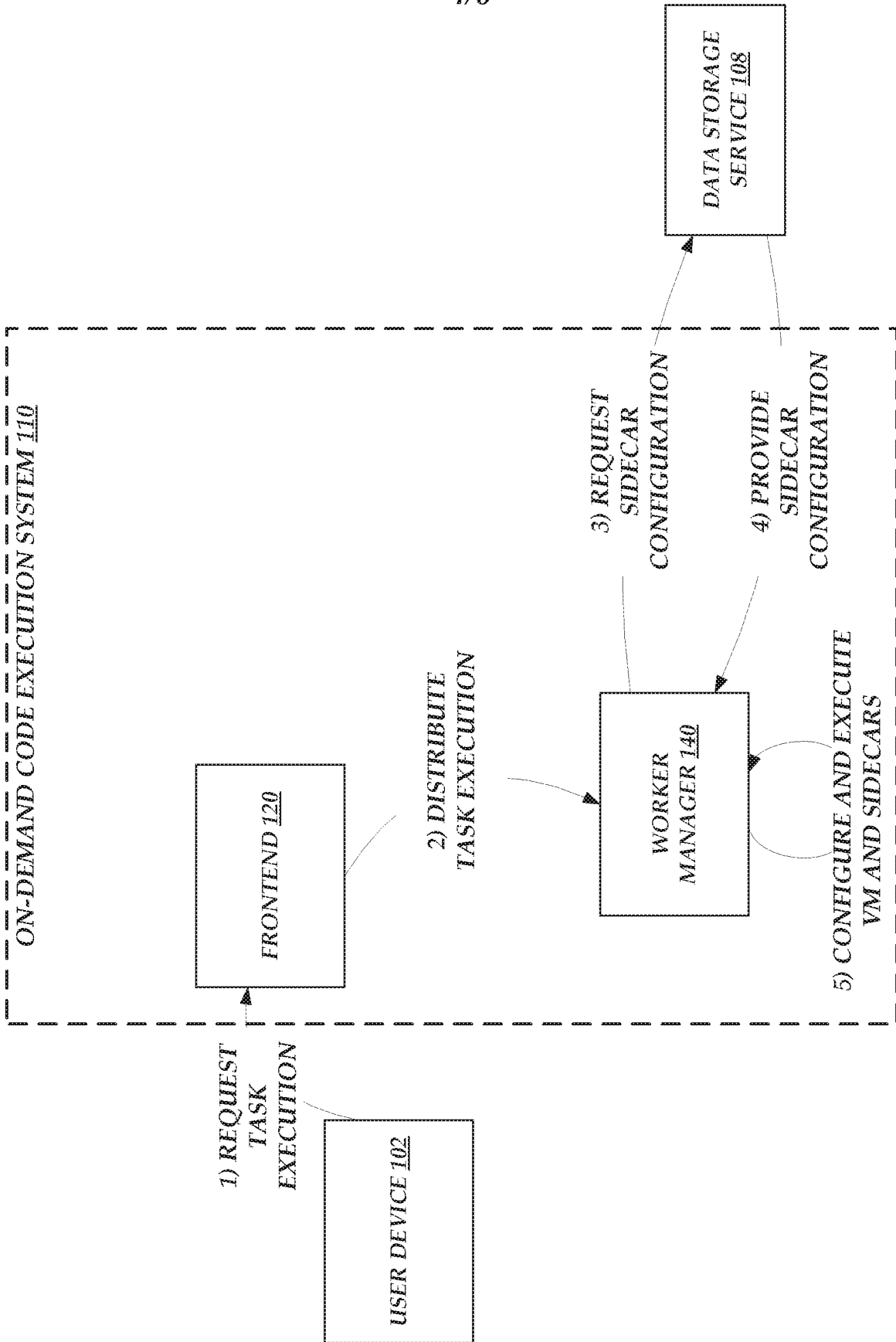


Fig. 4

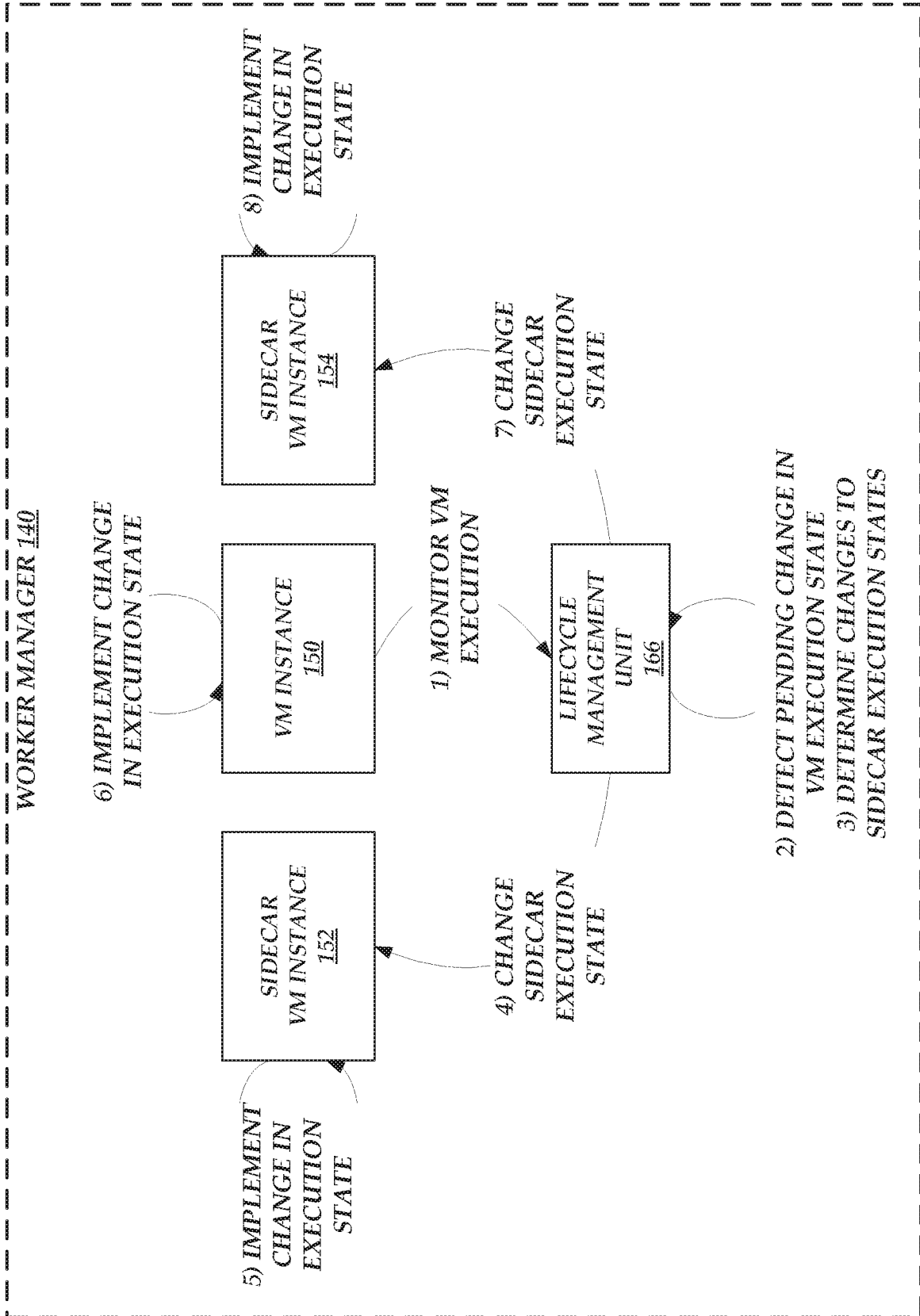


Fig. 5

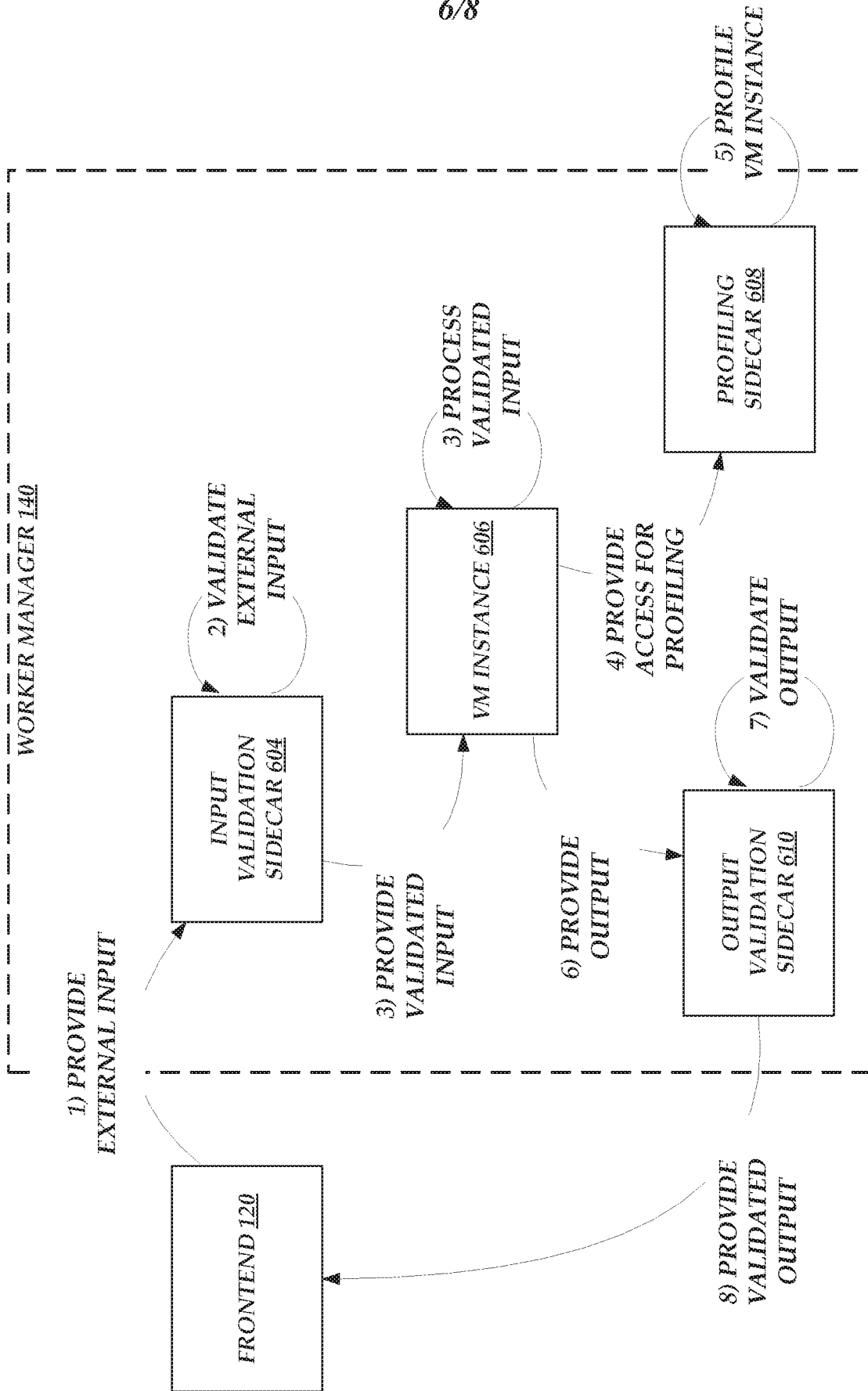


Fig. 6

7/8

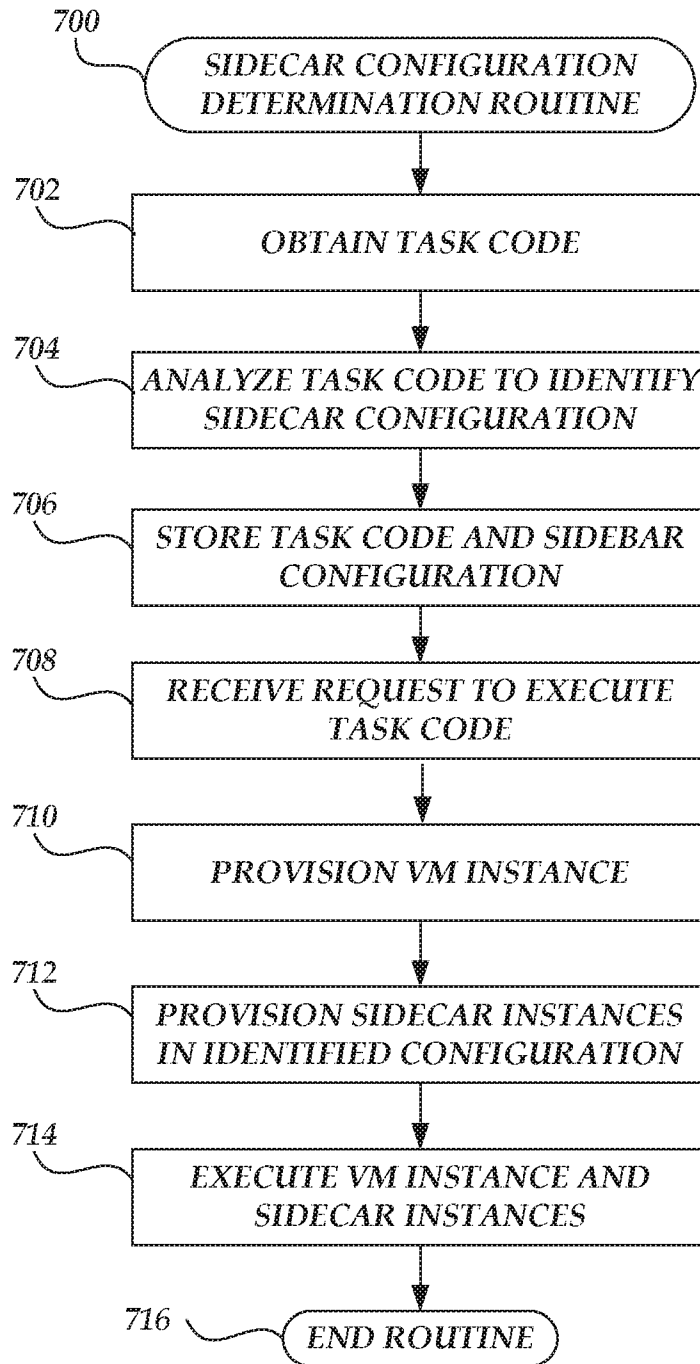


Fig. 7

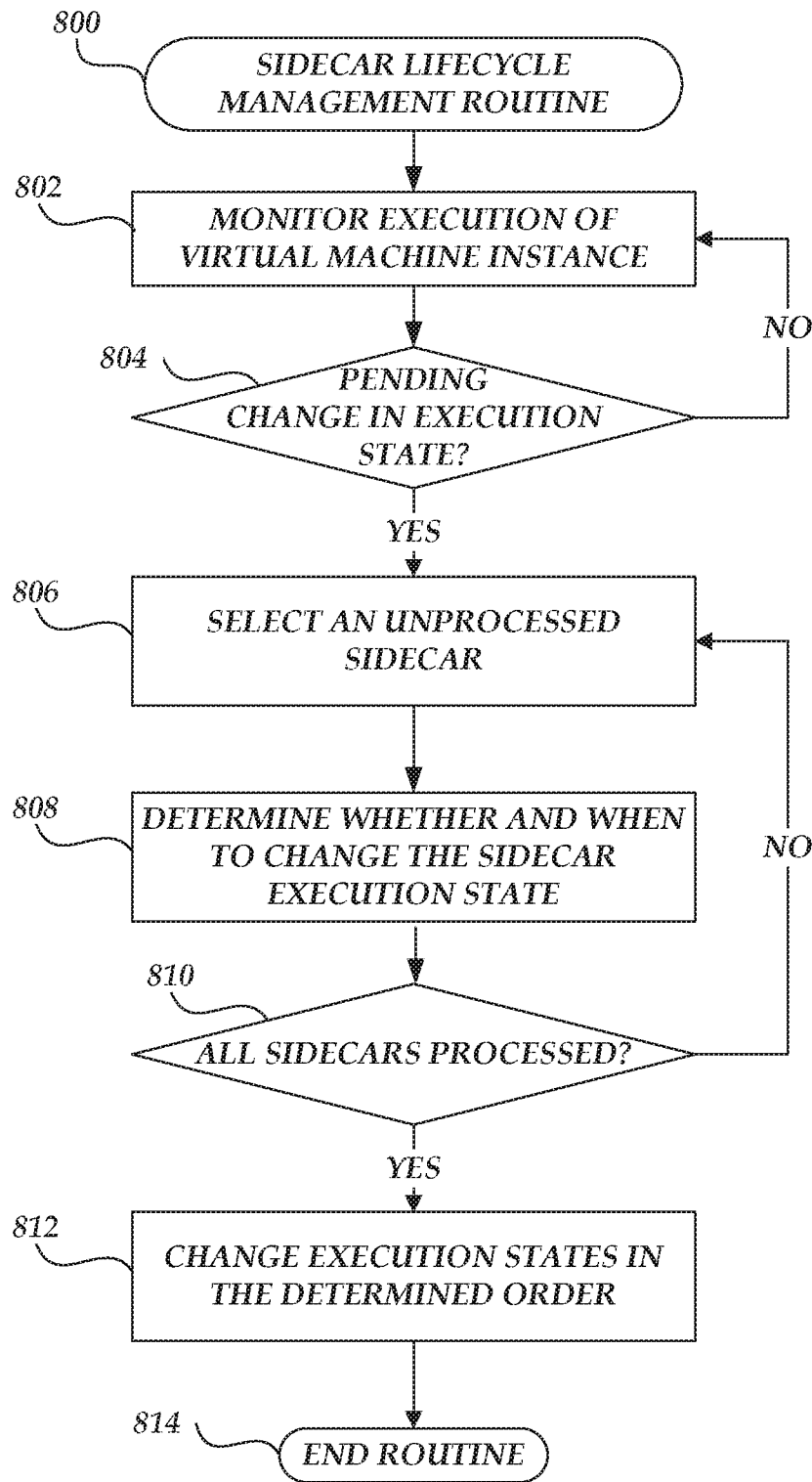


Fig. 8

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2019/038520

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F9/48
ADD.
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED
Minimum documentation searched (classification system followed by classification symbols)
G06F
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data, INSPEC, COMPENDEX

C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2017/371703 A1 (WAGNER TIMOTHY ALLEN [US] ET AL) 28 December 2017 (2017-12-28) abstract paragraph [0011] - paragraph [0023] paragraph [0067] - paragraph [0068] -----	1-15
A	US 10 002 026 B1 (WAGNER TIMOTHY ALLEN [US]) 19 June 2018 (2018-06-19) the whole document -----	1-15
A	WO 2018/098445 A1 (AMAZON TECH INC [US]) 31 May 2018 (2018-05-31) the whole document -----	1-15
A	US 2014/201735 A1 (KANNAN HARIHARAN [US] ET AL) 17 July 2014 (2014-07-17) the whole document -----	1-15
	-/--	

Further documents are listed in the continuation of Box C.

See patent family annex.

* Special categories of cited documents :

<p>"A" document defining the general state of the art which is not considered to be of particular relevance</p> <p>"E" earlier application or patent but published on or after the international filing date</p> <p>"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)</p> <p>"O" document referring to an oral disclosure, use, exhibition or other means</p> <p>"P" document published prior to the international filing date but later than the priority date claimed</p>	<p>"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention</p> <p>"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone</p> <p>"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art</p> <p>"&" document member of the same patent family</p>
---	---

Date of the actual completion of the international search 1 August 2019	Date of mailing of the international search report 14/08/2019
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Authorized officer Beltrán-Escavy, José

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2019/038520

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2018/150339 A1 (PAN CHENG [US] ET AL) 31 May 2018 (2018-05-31) the whole document	1-15
A	----- ZHANG ZHAONING ET AL: "VMThunder: Fast Provisioning of Large-Scale Virtual Machine Clusters", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE SERVICE CENTER, LOS ALAMITOS, CA, US, vol. 25, no. 12, 1 December 2014 (2014-12-01), pages 3328-3338, XP011564267, ISSN: 1045-9219, DOI: 10.1109/TPDS.2014.7 [retrieved on 2014-11-10] the whole document -----	1-15

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No PCT/US2019/038520

Patent document cited in search report	Publication date	Patent family member(s)	Publication date																												
US 2017371703	A1	28-12-2017	NONE																												

US 10002026	B1	19-06-2018	NONE																												

WO 2018098445	A1	31-05-2018	<table style="width: 100%; border: none;"> <tr> <td style="width: 15%;">AU</td> <td style="width: 15%;">2017363368</td> <td style="width: 15%;">A1</td> <td style="width: 15%;">04-07-2019</td> </tr> <tr> <td>KR</td> <td>20190082926</td> <td>A</td> <td>10-07-2019</td> </tr> <tr> <td>WO</td> <td>2018098445</td> <td>A1</td> <td>31-05-2018</td> </tr> </table>	AU	2017363368	A1	04-07-2019	KR	20190082926	A	10-07-2019	WO	2018098445	A1	31-05-2018																
AU	2017363368	A1	04-07-2019																												
KR	20190082926	A	10-07-2019																												
WO	2018098445	A1	31-05-2018																												

US 2014201735	A1	17-07-2014	<table style="width: 100%; border: none;"> <tr> <td style="width: 15%;">AU</td> <td style="width: 15%;">2014207721</td> <td style="width: 15%;">A1</td> <td style="width: 15%;">09-07-2015</td> </tr> <tr> <td>CA</td> <td>2896304</td> <td>A1</td> <td>24-07-2014</td> </tr> <tr> <td>CN</td> <td>104956332</td> <td>A</td> <td>30-09-2015</td> </tr> <tr> <td>EP</td> <td>2946291</td> <td>A1</td> <td>25-11-2015</td> </tr> <tr> <td>JP</td> <td>2016507100</td> <td>A</td> <td>07-03-2016</td> </tr> <tr> <td>US</td> <td>2014201735</td> <td>A1</td> <td>17-07-2014</td> </tr> <tr> <td>WO</td> <td>2014113330</td> <td>A1</td> <td>24-07-2014</td> </tr> </table>	AU	2014207721	A1	09-07-2015	CA	2896304	A1	24-07-2014	CN	104956332	A	30-09-2015	EP	2946291	A1	25-11-2015	JP	2016507100	A	07-03-2016	US	2014201735	A1	17-07-2014	WO	2014113330	A1	24-07-2014
AU	2014207721	A1	09-07-2015																												
CA	2896304	A1	24-07-2014																												
CN	104956332	A	30-09-2015																												
EP	2946291	A1	25-11-2015																												
JP	2016507100	A	07-03-2016																												
US	2014201735	A1	17-07-2014																												
WO	2014113330	A1	24-07-2014																												

US 2018150339	A1	31-05-2018	NONE																												
