

**ACCURATE GENERATION OF MULTIPLE DIMENSIONS
OF COMPUTER LOAD**

5

BACKGROUND

[0001] The subject matter discussed in the background section should not be assumed to be prior art merely as a result of its mention in the background section. Similarly, a problem mentioned in the background section or associated with the subject matter of the background section should not be assumed to have been previously recognized in the prior art. The subject matter in the background section merely represents different approaches, which in and of themselves may also correspond to implementations of the claimed inventions.

10

[0002] Load generators can create a maximum load to measure a maximum resource load capacity of a computer. However, existing load generators do not allow the creation of an arbitrary load. Also, existing load generators run a battery of tests and can assign a single metric to computer resource availability. However, existing load generators do not allow the fine-grained measurement of different types of computer resources, and in particular at arbitrary loads.

15

[0003] U.S. Patent Application Publication Number 2014/068335 discloses incrementally provisioning cloud resources 122 and vUsers 120A-120H in order to induce a load on a service provider system 112 that will test the ability of the service provider system 112 to satisfy a performance goal under load. The goal is typically specified by an agreed upon service level, commonly known as a service level agreement (SLA). U.S. 2014/068335 explains that the server 114 predicts cloud resources that will be required to generate a target load (e.g., 100,000 vUsers). U.S. 2014/068335 requires the provisioning of outside resources (i.e., the cloud resources 122) in order to simulate the vUsers and induce loads on the service provider system 112. The premise of needing to add servers and other cloud-based resources to simulate a large number of vUsers is contrary to generating a load on a machine using the machine's own hardware. U.S. 2014/068335 is not interested in generating a specified resource utilization. Instead, it measures performance under load, irrespective of resource utilization.

20

25

[0004] Accordingly, an opportunity arises to develop a load generator that helps determine resource load capacity of different types of computer resources at arbitrary levels.

30

SUMMARY

[0005] One aspect of the technology is a method of adjustably generating multiple dimensions of computer resource load to test capacity of the multiple dimensions of the computer resource load.

5 [0006] A utilization test generates the multiple dimensions of resource load on the first computer, the multiple dimensions including at least two of: a processor utilization dimension, a main memory utilization dimension, a mass storage utilization dimension, and a network utilization dimension. The utilization test performs iterations of attempts to generate said each dimension of the resource load on the first computer, in an amount equal to a changeable target utilization of said each dimension of the resource load on the first computer. The utilization test
10 adjusts the load generated to attempt to correct for inequality between (i) the actual utilization of a particular one of the multiple dimensions of the resource load on the first computer, and (ii) the changeable target utilization of the particular one of the multiple dimensions of the resource load on the first computer.

15 [0007] Other features and aspects of the technology disclosed appear in the drawings, specification and claims that follow.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] The included drawings are for illustrative purposes and serve only to provide examples of possible structures and process operations for one or more implementations of this disclosure. These drawings in no way limit any changes in form and detail that may be made by
20 one skilled in the art without departing from the spirit and scope of this disclosure. A more complete understanding of the subject matter may be derived by referring to the detailed description and claims when considered in conjunction with the following figures, wherein like reference numbers refer to similar elements throughout the figures.

25 [0009] FIG. 1 is a block diagram that illustrates example environments for running a resource utilization test.

[0010] FIGS. 2-5 are simplified block diagrams which show the resource utilization test application running on a general purpose operating system or on a unikernel.

[0011] FIG. 6 is an example block diagram of a utilization test application with multiple
30 closed loops to generate desired levels of resource load.

[0012] FIG. 7 is a simplified representation of different components that comprise an actual utilization of a particular computer resource.

[0013] FIG. 8 is a graph showing the superior performance of a closed loop load generator over an open loop load generator over time.

[0014] FIG. 9 is a graph showing the superior performance of a closed loop load generator over an open loop load generator over various changeable target utilizations of CPU.

5 [0015] FIG. 10 is a graph showing the performance of multiple dimensions of computer resource load at the same time.

[0016] FIG. 11 is a histogram of CPU load error in a virtual machine with sufficient underlying resources.

10 [0017] FIG. 12 is a histogram of CPU load error in a virtual machine with insufficient underlying resources.

[0018] FIG. 13 is a simplified block diagram of a datacenter cloud with resource utilization test applications running in multiple instances of virtual machines on different physical servers.

[0019] FIG. 14 is a block diagram of an example computer system.

DETAILED DESCRIPTION

15 [0020] The following detailed description is made with reference to the figures. Sample implementations are described to illustrate the technology disclosed, not to limit its scope, which is defined by the claims. Those of ordinary skill in the art will recognize a variety of equivalent variations on the description that follows.

20 [0021] FIG. 1 is a block diagram that illustrates example environments for running a resource utilization test.

[0022] A virtual machine operating system 20, or virtual machine instance, runs on virtual machine hardware 28a allotted to the virtual machine operating system 20. Virtual machine operating system 20 is managed by virtualization software 30a executing on physical server 10a. Virtualization software 30a can also manage additional virtual machines. Virtualization software
25 30a can be native or hosted, and manages one or more virtual machines, permitting multiple concurrent instances of operating systems on the same computer hardware resources. The computer hardware resources include physical servers 10a to 10z, interconnected by network 2 shown as a network cloud. Physical servers 10a to 10z include processors, memories, and nontransitory computer readable media with computer readable instructions executable by a
30 computer performing the technology described herein. Any number of servers 10a-10z may reside on network 2, and any number of virtual machines 20 may reside on each physical server 10a-10z. For example, virtualization software 30y is executing on physical server 10y, and physical server 10z does not have executing virtualization software.

[0023] Virtual machine operating system 20 is an environment in which resource utilization test application 22a executes. Resource utilization test application 22a performs resource adjustment for resource load capacity test 23a. Examples of different types, or dimensions of resource load that are adjusted, include a CPU or process utilization dimension, a main memory utilization dimension, a mass storage utilization dimension, and a network utilization dimension.

[0024] Physical server 10z does not have executing virtualization software. Resource adjustment for resource load capacity test 23z functions in a manner similar to resource adjustment for resource load capacity test 23a, but in a non-virtual context on physical machine operating system 19.

[0025] FIGS. 2-5 are simplified block diagrams which show the resource utilization test application running on a general purpose operating system or on a unikernel.

[0026] FIG. 2 shows a virtual machine instance 20, supporting a unikernel operating system 52. Unikernel operating system 52 provides a single address space which allows inspection of any aspect of the operating system. Accordingly, unikernel operating system 52 helps guarantee accuracy of the load generation performed by resource utilization test application 22a.

[0027] FIG. 3 shows a virtual machine instance 20, supporting a general purpose operating system 50. General purpose operating system 50 provides user space 40 with multiple address spaces including an address space with the Resource Utilization Test Application 22a, which obscures inspection of aspects of the operating system. Accordingly, general purpose operating system 50 obstructs accuracy of the load generation performed by resource utilization test application 22a. Nevertheless, resource utilization test application 22a is an improved load generator compared to alternatives in general purpose operating system 50.

[0028] FIG. 4 is similar to FIG. 2 except that a physical machine 19 rather than a virtual machine supports unikernel operating system 52. FIG. 5 is similar to FIG. 3 except that a physical machine 19 rather than a virtual machine supports general purpose operating system 50.

[0029] FIG. 6 is an example block diagram of a utilization test application with multiple closed loops to generate desired levels of resource load.

[0030] Utilization test control 250 generates changeable target utilizations 208, 218, 228, and 238 for the different dimensions of resource load. The changeable target utilizations 208, 218, 228, and 238 can be manually specified by a user, or generated automatically by the utilization test application. Automatically generated changeable target utilizations 208, 218, 228, and 238 can cycle through a range of values to perform the utilization test application multiple times.

[0031] Network I/O generator 202 generates network utilization dimension resource load, such as one or more of: transmit packets per second and transactions per second. The amount of

network utilization dimension resource load generated by the network I/O generator 202 is referred to as load generated 206. The load generated 206 is initially equal to the changeable target utilization 208. Actual utilization 207 is the amount of resource load for the network utilization dimension that is actually generated on the computer. Actual utilization 207 is different from load generated 206, because actual utilization 207 for the network utilization dimension on the computer includes contributions in addition to the load generated 206 generated by network I/O generator 202. After actual utilization 207 is measured by actual utilization measurement and output 240, actual utilization 207 is provided to closed loop controller 204. Responsive to the actual utilization 207, closed loop controller 204 changes the load generated 206, in an attempt to make the actual utilization 207 equal to changeable target utilization 208.

[0032] The other 3 loops are similar, but are described for completeness.

[0033] Mass storage I/O generator 212 generates mass storage utilization dimension resource load, such as one or more of: read operations per second and write operations per second. The amount of mass storage dimension resource load generated by the mass storage I/O generator 212 is referred to as load generated 216. The load generated 216 is initially equal to the changeable target utilization 218. Actual utilization 217 is the amount of resource load for the mass storage utilization dimension that is actually generated on the computer. Actual utilization 217 is different from load generated 216, because actual utilization 217 for the mass storage utilization dimension on the computer includes contributions in addition to the load generated 216 generated by mass storage I/O generator 212. After actual utilization 217 is measured by actual utilization measurement and output 240, actual utilization 217 is provided to closed loop controller 214. Responsive to the actual utilization 217, closed loop controller 214 changes the load generated 216, in an attempt to make the actual utilization 217 equal to changeable target utilization 218.

[0034] Main memory I/O generator 222 generates main memory utilization dimension resource load, such as one or more of: read operations per second and write operations per second. The amount of main memory utilization dimension resource load generated by the main memory I/O generator 222 is referred to as load generated 226. The load generated 226 is initially equal to the changeable target utilization 228. Actual utilization 227 is the amount of resource load for the network utilization dimension that is actually generated on the computer. Actual utilization 227 is different from load generated 226, because actual utilization 227 for the main memory utilization dimension on the computer includes contributions in addition to load generated 226 generated by main memory I/O generator 222. After actual utilization 227 is

measured by actual utilization measurement and output 240, actual utilization 227 is provided to closed loop controller 224. Responsive to the actual utilization 227, closed loop controller 224 changes the load generated 226, in an attempt to make the actual utilization 227 equal to changeable target utilization 228.

- 5 [0035] Processor or CPU load generator 232 generates processor utilization dimension resource load, such as one or more of: CPU seconds used and CPU seconds total. The amount of processor utilization dimension resource load generated by the processor generator 232 is referred to as load generated 236. The load generated 236 is initially equal to the changeable target utilization 238. Actual utilization 237 is the amount of resource load for the processor
- 10 load utilization dimension that is actually generated on the computer. Actual utilization 237 is different from load generated 236, because actual utilization 237 for the processor utilization dimension on the computer includes contributions in addition to load generated 236 generated by processor load generator 232. After actual utilization 237 is measured by actual utilization measurement and output 240, actual utilization 237 is provided to closed loop controller 234.
- 15 Responsive to the actual utilization 237, closed loop controller 234 changes the load generated 236, in an attempt to make the actual utilization 237 equal to changeable target utilization 238.

- [0036] Closed loop controllers 204, 214, 224, and 234 can be implemented in hardware or software or a combination. More generally, closed loop controller 204, 214, 224, and 234 can perform cascade compensation or feedback compensation. Examples of cascade compensation
- 20 are proportional plus integral (PI) control, proportional plus derivative (PD) control, and proportional plus integral plus derivative (PID) control. Examples of tuning the closed loop are manual tuning, Ziegler-Nichols tuning, Tyreus Luyben tuning, software tool tuning, and Cohen-Coon tuning.

- [0037] Although the example of the utilization test application shows 4 dimensions of
- 25 resource load, other examples can include 3, 2, 1, or more than 4. Other example can divide a particular dimension of resource load into different characteristics of the particular dimension of resource load.

[0038] FIG. 7 is a simplified representation of different components that comprise an actual utilization of a particular computer resource.

- 30 [0039] Actual utilization for a particular computer resource 301 can be, for example, actual utilization 207, 217, 227, or 237 on FIG. 6. Actual utilization includes not just the load generated for the particular computer resource 310, but additional utilization contributions. The load generated for a particular computer resource 301 can be, for example, load generated 206, 216, 226, or 236 on FIG. 6.

[0040] Several shown contributions occur from load generator closed loops for other dimensions of computer load. For example, in an example where the actual utilization for a particular computer resource 301 is actual utilization 237 for the processor load dimension, additional contributions to the processor load dimension of computer load come from load generator closed loops for the network utilization dimension, the mass storage utilization dimension, and/or the main memory utilization dimension. Such additional contributions are shown as utilization contributions for the particular computer resource (e.g., processor load dimension) 320, 330, and 340, which respectively occur from actual utilization for different computer resource #1 321 (e.g., network utilization dimension), actual utilization for different computer resource #2 331 (e.g., mass storage dimension), and actual utilization for different computer resource #3 341 (e.g., main memory utilization dimension).

[0041] Other shown contributions occur from outside of the utilization test or even from outside the computer altogether. Utilizations for the particular computer resource 350, 360, and 370 respectively occur from a process on the virtual machine outside of the utilization test 351, a process on the physical machine outside of the utilization test 361, and a process on the physical server outside of the virtual machine 371. Strictly speaking, the process on the virtual machine outside of the utilization test 351, the process on the physical machine outside of the utilization test 361, and the process on the physical server outside of the virtual machine 371 can represent ceilings on resource availability. However, such ceilings can be accounted for by negative contributions.

[0042] FIG. 8 is a graph showing the superior performance of a closed loop load generator over an open loop load generator over time.

[0043] This graph shows an initial comparison of a naive CPU load generator (open loop 420) vs. the PID controlled CPU load generator (closed loop 410) over a 120 second duration at 10% CPU utilization. The x-axis represents time and the y-axis represents the observed error. In this case positive error represents insufficient CPU load and negative error represents excessive CPU load. Both controllers bounce around an average error value due to normal system variation. However, the average PID controlled error is clearly much closer to 0 than in the open loop case.

[0044] FIG. 9 is a graph showing the superior performance of a closed loop load generator over an open loop load generator over various changeable target utilizations of CPU.

[0045] The same comparison as in FIG. 8 is made over a range of target CPU loads. This time, the x-axis is the target load and the y-axis is the average error value observed over the test duration. The error bars indicate the maximum and minimum error values observed during the

test run. The average error value in nearly all PID controlled cases (closed loop 510) is 0, whereas the naive CPU generator (open loop 520) has a clear bias dependent on the target CPU load. Also, both generators performed nearly identically at 0% and 100% load.

[0046] FIG. 10 is a graph showing the performance of multiple dimensions of computer resource load at the same time.

[0047] The graph shows the measured error (variance percentage in the graph) when running three load generators using the unikernel based load generator inside a virtual machine on a lightly loaded hypervisor.

[0048] The generators were configured as follows:

[0049] CPU utilization (trace 650): 50%

[0050] Memory Read (trace 610): 1000 MB/s

[0051] Memory Write (trace 620): 100 MB/s

[0052] Disk Read (trace 630): 20 MB/s

[0053] Disk Write (trace 640): 2 MB/s

[0054] There is some variance in every result category, however the average error is 0. Additionally, this shows that multiple load generators can work together at the same time, to create an accurate load over time.

[0055] FIG. 11 is a histogram of CPU load error in a virtual machine with sufficient underlying resources.

[0056] This load generation technique compares the ability of hypervisor platforms to provide a desired level of performance. By configuring a specific load on our test tool, the resource utilization test application can determine whether or not the hypervisor can provide that amount of compute resource to the test tool.

[0057] This first histogram shows the range of measured CPU load error for the unikernel based load generator on a lightly loaded hypervisor. The average error for this test run is 0% (as in the initial tests discussed above). Additionally, the histogram shows a tight clustering of error values around 0% error.

[0058] FIG. 12 is a histogram of CPU load error in a virtual machine with insufficient underlying resources.

[0059] This second histogram shows the same test configuration, but run on an overloaded hypervisor. The measured error has a much larger variation on this test system. Additionally, the average error for this test run is ~ -5%, which indicates that the PID controlled load generator could not generated the desired load due to a lack of CPU resources.

[0060] FIG. 13 is a simplified block diagram of a datacenter cloud with resource utilization test applications running in multiple instances of virtual machines on different physical servers.

[0061] Network 700 includes multiple physical servers such as physical server 710 and physical server 740. Each physical server supports multiple virtual machine instances, such as virtual machine OS 720 and 721 on physical server 710, and virtual machine OS 750 and 751 on physical server 740. Each of the virtual machines supports a resource utilization test application, such as resource utilization test application 730, 731, 760, and 761 respectively on virtual machine OS 720, 721, 750, and 751. Although 2 virtual machines are shown in the example, each physical server can support an arbitrary number of virtual machines.

[0062] In one embodiment, control software remotely installs the resource utilization test applications (along with the virtual machines, unless the virtual machine was already installed). Then the control software runs all (or some subset) of the resource utilization test applications on all (or some subset) of the virtual machines on all (or some subset) of the physical servers and aggregates the results. The final output can indicate an aggregate capacity of the datacenter cloud supporting many virtual machines for different dimensions of computer load.

[0063] FIG. 14 is a block diagram of an example computer system, according to one implementation. Computer system 910 typically includes at least one processor 914 that communicates with a number of peripheral devices via bus subsystem 912. These peripheral devices may include a storage subsystem 924 including, for example, memory devices and a file storage subsystem, user interface input devices 922, user interface output devices 920, and a network interface subsystem 916. The input and output devices allow user interaction with computer system 910. Network interface subsystem 916 provides an interface to outside networks, including an interface to corresponding interface devices in other computer systems.

[0064] User interface input devices 922 may include a keyboard; pointing devices such as a mouse, trackball, touchpad, or graphics tablet; a scanner; a touch screen incorporated into the display; audio input devices such as voice recognition systems and microphones; and other types of input devices. In general, use of the term "input device" is intended to include all possible types of devices and ways to input information into computer system 910.

[0065] User interface output devices 920 may include a display subsystem, a printer, a fax machine, or non-visual displays such as audio output devices. The display subsystem may include a cathode ray tube (CRT), a flat-panel device such as a liquid crystal display (LCD), a projection device, or some other mechanism for creating a visible image. The display subsystem may also provide a non-visual display such as audio output devices. In general, use of the term

"output device" is intended to include all possible types of devices and ways to output information from computer system 910 to the user or to another machine or computer system.

[0066] Storage subsystem 924 stores programming and data constructs that provide the functionality of some or all of the modules and methods described herein. These software modules are generally executed by processor 914 alone or in combination with other processors.

[0067] Memory 926 used in the storage subsystem can include a number of memories including a main random access memory (RAM) 930 for storage of instructions and data during program execution and a read only memory (ROM) 932 in which fixed instructions are stored. A file storage subsystem 928 can provide persistent storage for program and data files, and may include a hard disk drive, a floppy disk drive, along with associated removable media, a CD-ROM drive, an optical drive, or removable media cartridges. The modules implementing the functionality of certain implementations may be stored by file storage subsystem 928 in the storage subsystem 924, or in other machines accessible by the processor.

[0068] Bus subsystem 912 provides a mechanism for letting the various components and subsystems of computer system 910 communicate with each other as intended. Although bus subsystem 912 is shown schematically as a single bus, alternative implementations of the bus subsystem may use multiple busses.

[0069] Computer system 910 can be of varying types including a workstation, server, computing cluster, blade server, server farm, or any other data processing system or computing device. Due to the ever-changing nature of computers and networks, the description of computer system 910 depicted in FIG. 9 is intended only as one example. Many other configurations of computer system 910 are possible having more or fewer components than the computer system depicted in FIG. 9.

Particular Implementations

[0070] One aspect of the technology is a method of adjustably generating multiple dimensions of computer resource load to test capacity of the multiple dimensions of the computer resource load.

[0071] The method comprises, executing, on a first computer, a utilization test that generates the multiple dimensions of resource load on the first computer.

[0072] The multiple dimensions include at least two of: a processor utilization dimension, a main memory utilization dimension, a mass storage utilization dimension, and a network utilization dimension.

[0073] For each dimension of two or more of the multiple dimensions, the utilization test performs iterations of generating a load in said each dimension of the resource load on the first computer, in an amount predicted to cause an actual utilization equal to a changeable target utilization of said each dimension on the first computer.

- 5 [0074] The utilization test measures the actual utilization of a particular one of the multiple dimensions during generation of the resource load on the first computer.

[0075] The actual utilization is a result of at least:

[0076] (i) load generated for the particular one of the multiple dimensions on the first computer. The load generated is generated by the utilization test.

- 10 [0077] (ii) a first utilization contribution to the particular one of the multiple dimensions on the first computer. The first utilization contribution is generated outside the utilization test and on the first computer.

[0078] Said particular one of the multiple dimensions of the resource load on the first computer is one of said two or more of the multiple dimensions, such that said particular one of the multiple dimensions of the resource load on the first computer has the changeable target utilization.

- [0079] The utilization test adjusts the load generated to attempt to realize and maintain the target utilization of the first computer for said particular one of the multiple dimensions by correcting for an inequality between (i) the actual utilization resulting from the load generated and the first utilization contribution, and (ii) the changeable target utilization of the particular one of the multiple dimensions on the first computer.

[0080] In one embodiment a subsequent one of the iterations is after a prior one of the iterations. In the prior one of the iterations the utilization test measures the actual utilization, and in response, in the subsequent one of the iterations the utilization test adjusts the load generated.

- 25 [0081] In one embodiment a subsequent one of the iterations is after a prior one of the iterations. In the prior one of the iterations the utilization test measures the actual utilization of the particular one of the multiple dimensions on the first computer, and in response, in the subsequent one of the iterations the utilization test adjusts the load generated for the particular one of the multiple dimensions on the first computer.

- 30 [0082] In one embodiment the actual utilization of the particular one of the multiple dimensions of the resource load on the first computer is the result of at least:

[0083] (i) the load generated,

[0084] (ii) the first utilization contribution, and

[0085] (iii) a second utilization contribution of another one of the multiple dimensions on the first computer. The second utilization contribution is generated by the utilization test for said another one of the multiple dimensions.

5 [0086] In one embodiment the particular one of the multiple dimensions on the first computer is the processor utilization dimension, and said another one of the multiple dimensions is any of the main memory utilization dimension, the mass storage utilization dimension, and the network utilization dimension.

[0087] In one embodiment the two or more of the multiple dimensions includes the particular one of the multiple dimensions and another one of the multiple dimensions on the first computer.
10 The actual utilization measured by the utilization test for said another one of the multiple dimensions is a result of at least:

[0088] (i) a load generated for said another one of the multiple dimensions on the first computer. The load generated for said another one of the multiple dimensions is generated by the utilization test for said another one of the multiple dimensions; and

15 [0089] (ii) another first utilization contribution on the first computer generated outside the utilization test on the first computer.

[0090] In one embodiment the utilization test has the ability to store different values of the changeable target utilization, the actual utilization, the load generated, and the first utilization contribution for different ones of the multiple dimensions of resource load on the first computer.

20 [0091] In one embodiment the multiple dimensions measure basic behavior of an OS.

[0092] In one embodiment the first computer is a physical computer.

[0093] In one embodiment the first computer is a virtual machine on a physical computer.

[0094] In one embodiment in the subsequent one of the iterations, the load generated is further adjusted to attempt to correct for the virtual machine undergoing changes in access to
25 resources of the physical computer.

[0095] In one embodiment the utilization test runs in a unikernel operating system on the first computer.

[0096] In one embodiment the utilization test runs in a general purpose operating system.

[0097] In one embodiment the changeable target utilization is changeable to intermediate
30 values in between full utilization and no utilization of the particular one of the multiple dimensions of the first computer.

[0098] In one embodiment the changeable target utilization for the different dimensions is varied over time, and the utilization test performs a new set of iterations responsive to the changeable target utilization being changed.

[0099] In one embodiment the changeable target utilization for the different dimensions is determined by a user.

[00100] In one embodiment the changeable target utilization for the different dimensions is determined by an algorithm automatically.

5 [00101] One embodiment further comprises:

[00102] reporting the actual utilization and a difference between the actual utilization and the changeable target utilization for the multiple dimensions.

[00103] In one embodiment the utilization test includes a PID controller loop with a PID controller output that causes an actual utilization for the multiple dimensions of resource load of the first computer, in an attempt to realize the changeable target utilization for the multiple dimensions of resource load of the first computer.

[00104] Another aspect of the technology is a method of adjustably generating multiple dimensions of computer resource load to test capacity of the multiple dimensions of the computer resource load.

15 [00105] The method comprises executing, on a first computer, a utilization test that generates the multiple dimensions of resource load on the first computer.

[00106] The multiple dimensions include at least two of: a processor utilization dimension, a main memory utilization dimension, a mass storage utilization dimension, and a network utilization dimension.

20 [00107] For each dimension of two or more of the multiple dimensions, the utilization test performs, with a PID controller, iterations of generating a load in said each dimension of the resource load on the first computer, in an amount predicted to cause an actual utilization equal to a changeable target utilization of said each dimension on the first computer.

[00108] Another aspect of the technology is an apparatus configured to adjustably generate multiple dimensions of computer resource load to adjustably test availability of multiple dimensions of computer resource load, comprising a first computer that includes a processor, memory, and instructions in the memory that implement a method as described herein.

[00109] Another aspect of the technology is a non-transitory computer-readable medium storing code for adjustably generating multiple dimensions of computer resource load to test capacity of the multiple dimensions of the computer resource load. The code comprises instructions for performing a method as described herein.

30 [00110] This method and other implementations of the technology disclosed can include one or more of the following features and/or features described in connection with additional

methods disclosed. In the interest of conciseness, the combinations of features disclosed in this application are not individually enumerated and are not repeated with each base set of features.

[00111] Other implementations may include tangible computer-readable memory including computer program instructions that cause a computer to implement any of the methods described above. Yet another implementation may include a system including memory and one or more processors operable to execute instructions, stored in the memory, to perform any of the methods described above.

[00112] The following clauses describe aspects of various examples of adjustably generating multiple dimensions of computer resource load to test capacity of the multiple dimensions of the computer resource load.

[00113] 1. A method of adjustably generating multiple dimensions of computer resource load to test capacity of the multiple dimensions of the computer resource load, the method comprising:

executing, on a first computer, a utilization test that generates the multiple dimensions of resource load on the first computer, the multiple dimensions including at least two of:

a processor utilization dimension, a main memory utilization dimension, a mass storage utilization dimension, and a network utilization dimension,

wherein, for each dimension of two or more of the multiple dimensions, the utilization test performs iterations of generating a load in said each dimension of the resource load on the first computer, in an amount predicted to cause an actual utilization equal to a changeable target utilization of said each dimension on the first computer,

wherein the utilization test measures the actual utilization of a particular one of the multiple dimensions during generation of the resource load on the first computer, the actual utilization being a result of at least:

a load generated for the particular one of the multiple dimensions on the first computer, as generated by the utilization test; and

a first utilization contribution to the particular one of the multiple dimensions on the first computer, wherein the first utilization contribution is generated outside the utilization test and on the first computer, and

wherein the utilization test adjusts the generated load to attempt to realize and maintain the target utilization on the first computer for said particular one of the multiple dimensions by correcting for an inequality between (i) the actual utilization resulting from the load generated and the first utilization contribution, and (ii) the changeable target utilization of the particular one of the multiple dimensions on the first computer.

[00114] 2. The method of clause 1, wherein a subsequent one of the iterations is after a prior one of the iterations, in the prior one of the iterations the utilization test measures the actual utilization, and in response, in the subsequent one of the iterations the utilization test adjusts the load generated.

5 [00115] 3. The method of any of clauses 1-2, wherein a subsequent one of the iterations is after a prior one of the iterations, in the prior one of the iterations the utilization test measures the actual utilization of the particular one of the multiple dimensions on the first computer, and in response, in the subsequent one of the iterations the utilization test adjusts the load generated for the particular one of the multiple dimensions on the first computer.

10 [00116] 4. The method of any of clauses 1-3, wherein the actual utilization of the particular one of the multiple dimensions on the first computer is the combination of at least: the load generated, the first utilization contribution, and

a second utilization contribution of another one of the multiple dimensions on the first computer, the second utilization contribution generated by the utilization test for said another one of the multiple dimensions.

[00117] 5. The method of clauses 4, wherein the particular one of the multiple dimensions on the first computer is the processor utilization dimension, and said another one of the multiple dimensions is any of the main memory utilization dimension, the mass storage utilization dimension, and the network utilization dimension.

20 [00118] 6. The method of any preceding clause, wherein the two or more of the multiple dimensions includes the particular one of the multiple dimensions and another one of the multiple dimensions on the first computer, and the actual utilization measured by the utilization test for said another one of the multiple dimensions is a combination of at least:

a load generated for said another one of the multiple dimensions on the first computer, the load generated for said another one of the multiple dimensions generated by the utilization test for said another one of the multiple dimensions; and

another first utilization contribution on the first computer generated outside the utilization test on the first computer.

[00119] 7. The method of any preceding clause, wherein the utilization test has the ability to store different values of the changeable target utilization, the actual utilization, the load generated, and the first utilization contribution for different ones of the multiple dimensions of resource load on the first computer.

[00120] 8. The method of any preceding clause, wherein the multiple dimensions measure basic behavior of an OS.

[00121] 9. The method of any of clauses 1-8, wherein the first computer is a physical computer.

[00122] 10. The method of any of clauses 1-8, wherein the first computer is a virtual machine on a physical computer.

5 [00123] 11. The method of clause 10, wherein in the subsequent one of the iterations, the load generated is further adjusted to attempt to correct for the virtual machine undergoing changes in access to resources of the physical computer.

[00124] 12. The method of any preceding clause, wherein the utilization test runs in a unikernel operating system on the first computer.

10 [00125] 13. The method of any preceding clause, wherein the utilization test runs in a general purpose operating system.

[00126] 14. The method of any preceding clause, wherein the changeable target utilization is changeable to intermediate values in between full utilization and no utilization of the particular one of the multiple dimensions of the first computer.

15 [00127] 15. The method of any preceding clause, wherein the changeable target utilization for the different dimensions is varied over time, and the utilization test performs a new set of iterations responsive to the changeable target utilization being changed.

[00128] 16. The method of any preceding clause, wherein the changeable target utilization for the different dimensions is determined by a user.

20 [00129] 17. The method of any of clauses 1-15, wherein the changeable target utilization for the different dimensions is determined by an algorithm automatically.

[00130] 18. The method of any preceding clause, further comprising:

reporting the actual utilization and a difference between the actual utilization and the changeable target utilization for the multiple dimensions.

25 [00131] 19. The method of any preceding clause, wherein the utilization test includes a PID controller loop with a PID controller output that causes an actual utilization for the multiple dimensions of resource load of the first computer, in an attempt to realize the changeable target utilization for the multiple dimensions of resource load of the first computer.

[00132] 20. A method of adjustably generating multiple dimensions of computer resource
30 load to test capacity of the multiple dimensions of the computer resource load, the method comprising:

executing, on a first computer, a utilization test that generates the multiple dimensions of resource load on the first computer, the multiple dimensions including at least two of:

a processor utilization dimension, a main memory utilization dimension, a mass storage utilization dimension, and a network utilization dimension,

wherein, for each dimension of two or more of the multiple dimensions, the utilization test performs, with a PID controller, iterations of generating a load in said each dimension of the resource load on the first computer, in an amount predicted to cause an actual utilization equal to a changeable target utilization of said each dimension on the first computer.

[00133] 21. An apparatus configured to adjustably generate multiple dimensions of computer resource load to adjustably test availability of multiple dimensions of computer resource load, comprising:

a first computer that includes a processor, memory, and instructions in the memory that implement:

executing, on a first computer, a utilization test that generates the multiple dimensions of resource load on the first computer, the multiple dimensions including at least two of:

a processor utilization dimension, a main memory utilization dimension, a mass storage utilization dimension, and a network utilization dimension,

wherein, for each dimension of two or more of the multiple dimensions, the utilization test performs iterations of generating a load in said each dimension of the resource load on the first computer, in an amount predicted to cause an actual utilization equal to a changeable target utilization of said each dimension on the first computer,

wherein the utilization test measures the actual utilization of a particular one of the multiple dimensions during generation of the resource load on the first computer, the actual utilization being a result of at least:

a load generated for the particular one of the multiple dimensions on the first computer, as generated by the utilization test; and

a first utilization contribution to the particular one of the multiple dimensions on the first computer, wherein the first utilization contribution is generated outside the utilization test and on the first computer, and

wherein the utilization test adjusts the generated load to attempt to realize and maintain the target utilization on the first computer for said particular one of the multiple dimensions by correcting for an inequality between (i) the actual utilization resulting from the load generated and the first utilization contribution, and (ii) the changeable target utilization of the particular one of the multiple dimensions on the first computer.

[00134] 22. An apparatus configured to adjustably generate multiple dimensions of computer resource load to test capacity of the multiple dimensions of the computer resource load, comprising:

a first computer that includes a processor, memory, and instructions in the memory that
5 implement:

executing, on a first computer, a utilization test that generates the multiple dimensions of resource load on the first computer, the multiple dimensions including at least two of:

a processor utilization dimension, a main memory utilization dimension, a mass storage utilization dimension, and a network utilization dimension,

10 wherein, for each dimension of two or more of the multiple dimensions, the utilization test performs, with a PID controller, iterations of generating a load in said each dimension of the resource load on the first computer, in an amount predicted to cause an actual utilization equal to a changeable target utilization of said each dimension on the first computer.

[00135] While the technology disclosed is by reference to the preferred implementations and
15 examples detailed above, it is to be understood that these examples are intended in an illustrative rather than in a limiting sense. It is contemplated that modifications and combinations will readily occur to those skilled in the art, which modifications and combinations will be within the spirit of the technology disclosed and the scope of the following claims.

REPLACEMENT CLAIMS

1. (Amended) A method of adjustably generating multiple dimensions of computer resource load to test capacity of the multiple dimensions of the computer resource load, the method comprising:

executing, on a first computer (19, 20), a utilization test (22a, 22z) that generates the multiple dimensions of resource load (206, 216, 226, 236) on the first computer, the multiple dimensions including at least two of:

a processor utilization (236) dimension, a main memory utilization (226) dimension, a mass storage utilization (216) dimension, and a network utilization (206) dimension,

wherein, for each dimension of two or more of the multiple dimensions, the utilization test performs iterations of generating a load in said each dimension of the resource load on the first computer, in an amount predicted to cause an actual utilization equal to a changeable target utilization (208, 218, 228, 238) of said each dimension on the first computer,

wherein the utilization test measures the actual utilization (240) of a particular one of the multiple dimensions during generation of the resource load on the first computer, the actual utilization being a result (301) of at least:

a load generated for the particular one of the multiple dimensions on the first computer, as generated by the utilization test; and

a first utilization contribution to the particular one of the multiple dimensions on the first computer, wherein the first utilization contribution is generated outside the utilization test and on the first computer, and

wherein the utilization test adjusts the generated load (23a, 23z) to attempt to realize and maintain the target utilization (FIGS. 8-12) on the first computer for said particular one of the multiple dimensions by correcting for an inequality between (i) the actual utilization resulting from the load generated and the first utilization contribution, and (ii) the changeable target utilization of the particular one of the multiple dimensions on the first computer.

2. (Amended) The method of claim 1, wherein a subsequent one of the iterations is after a prior one of the iterations, in the prior one of the iterations the utilization test measures the actual utilization, and in response, in the subsequent one of the iterations the utilization test adjusts the load generated.

REPLACEMENT CLAIMS

3. (Amended) The method of any of claims 1-2, wherein a subsequent one of the iterations is after a prior one of the iterations, in the prior one of the iterations the utilization test measures the actual utilization of the particular one of the multiple dimensions on the first computer, and in response, in the subsequent one of the iterations the utilization test adjusts the load generated for the particular one of the multiple dimensions on the first computer.

4. (Amended) The method of any of claims 1-3, wherein the actual utilization of the particular one of the multiple dimensions on the first computer is the result of at least: the load generated, the first utilization contribution, and

a second utilization contribution of another one of the multiple dimensions on the first computer, the second utilization contribution generated by the utilization test for said another one of the multiple dimensions.

5. (Amended) The method of claim 4, wherein the particular one of the multiple dimensions on the first computer is the processor utilization dimension, and said another one of the multiple dimensions is any of the main memory utilization dimension, the mass storage utilization dimension, and the network utilization dimension.

6. (Amended) The method of any preceding claim, wherein the two or more of the multiple dimensions includes the particular one of the multiple dimensions and another one of the multiple dimensions on the first computer, and the actual utilization measured by the utilization test for said another one of the multiple dimensions is a result of at least:

a load generated for said another one of the multiple dimensions on the first computer, the load generated for said another one of the multiple dimensions generated by the utilization test for said another one of the multiple dimensions; and

another first utilization contribution on the first computer generated outside the utilization test on the first computer.

7. (Amended) The method of any preceding claim, wherein the utilization test has the ability to store different values of the changeable target utilization, the actual utilization, the load generated, and the first utilization contribution for different ones of the multiple dimensions of resource load on the first computer.

8. The method of any preceding claim, wherein the multiple dimensions measure basic behavior of an OS.

REPLACEMENT CLAIMS

9. The method of any of claims 1-8, wherein the first computer is a physical computer.
10. The method of any of claims 1-8, wherein the first computer is a virtual machine on a physical computer.
11. (Amended) The method of claim 10, wherein in the subsequent one of the iterations, the load generated is further adjusted to attempt to correct for the virtual machine undergoing changes in access to resources of the physical computer.
12. The method of any preceding claim, wherein the utilization test runs in a unikernel operating system on the first computer.
13. The method of any preceding claim, wherein the utilization test runs in a general purpose operating system.
14. The method of any preceding claim, wherein the changeable target utilization is changeable to intermediate values in between full utilization and no utilization of the particular one of the multiple dimensions of the first computer.
15. The method of any preceding claim, wherein the changeable target utilization for the different dimensions is varied over time, and the utilization test performs a new set of iterations responsive to the changeable target utilization being changed.
16. The method of any preceding claim, wherein the changeable target utilization for the different dimensions is determined by a user.
17. The method of any of claims 1-15, wherein the changeable target utilization for the different dimensions is determined by an algorithm automatically.
18. The method of any preceding claim, further comprising:

reporting the actual utilization and a difference between the actual utilization and the changeable target utilization for the multiple dimensions.
19. The method of any preceding claim, wherein the utilization test includes a PID controller loop with a PID controller output that causes an actual utilization for the multiple dimensions of resource load of the first computer, in an attempt to realize the changeable target utilization for the multiple dimensions of resource load of the first computer.

REPLACEMENT CLAIMS

20. (Amended) An apparatus configured to adjustably generate multiple dimensions of computer resource load to adjustably test availability of multiple dimensions of computer resource load, comprising:

a first computer that includes a processor, memory, and instructions in the memory that implement:

executing, on a first computer (19,20), a utilization test (22A, 22z) that generates the multiple dimensions of resource load (206, 216, 226, 236) on the first computer, the multiple dimensions including at least two of:

a processor utilization (236) dimension, a main memory utilization (226) dimension, a mass storage utilization (216) dimension, and a network utilization (206) dimension,

wherein, for each dimension of two or more of the multiple dimensions, the utilization test performs iterations of generating a load in said each dimension of the resource load on the first computer, in an amount predicted to cause an actual utilization equal to a changeable target utilization (208, 218, 228, 238) of said each dimension on the first computer,

wherein the utilization test measures the actual utilization (240) of a particular one of the multiple dimensions during generation of the resource load on the first computer, the actual utilization being a result (301) of at least:

a load generated for the particular one of the multiple dimensions on the first computer, as generated by the utilization test; and

a first utilization contribution to the particular one of the multiple dimensions on the first computer, wherein the first utilization contribution is generated outside the utilization test and on the first computer, and

wherein the utilization test adjusts the generated load (23a, 23z) to attempt to realize and maintain the target utilization (FIGS. 8-12) on the first computer for said particular one of the multiple dimensions by correcting for an inequality between (i) the actual utilization resulting from the load generated and the first utilization contribution, and (ii) the changeable target utilization of the particular one of the multiple dimensions on the first computer.